

IBM

KB

long

term

evaluation

study

010010

managing  
media

000010

migration  
in a deposit  
system







010010

managing  
media 000010  
migration  
in a deposit  
system

Dr. Raymond J. van Diessen  
and  
Ing. Ben J. van Rijnsoever

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	Koninklijke Bibliotheek
PO Box 9999	PO Box 90407
1000 CE Amsterdam	2509 LK The Hague
The Netherlands	The Netherlands

Title: **Managing Media Migration in a Deposit System**

ISBN: 90-6259-158-2  
Authors: Dr. Raymond J. van Diessen and Ing. Ben J. van Rijnsoever  
Date: December 2002  
Copyright: IBM / Koninklijke Bibliotheek

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

01 IBM / KB

000010

# 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.





# contents



1/	Summary	1
2/	Storage Performance Model	3
	2.1 Optimizing the Performance Model	6
3/	DNEP Component Architecture	7
	3.1 Details for Tape Library Speed	9
	3.2 Details for Optical Library Speed	10
	3.3 Details for RAID Disk System (ESS) Speed	11
4/	Migration Strategies within DIAS	13
	4.1 Functional Components Involved	13
	4.2 Possible Migration Strategies	16
	4.3 Performance Model	17
	4.4 Potential Optimizing Efforts	19
5/	Conclusions	23
	Appendix A: References	27
	Appendix B: Glossary	29



# //

# summary



Technology is changing rapidly. The life span of technology components can be as short as 5 years before new technological innovations make them obsolete. Storage technology used to store the digital objects in an electronic deposit system is no exception.

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, the most significant one being the amount of data that can be physically written to a specific storage device.

Most electronic deposit systems define their storage capacity needs in several TeraBytes ( $10^{12}$  Bytes). This requires media migration / refreshment to be consciously managed within the electronic deposit system to prevent a situation in which the time available for migration is insufficient for completing the process in time.

This report shows what type of migration strategies can be adopted: conversion, distribution, migration, and refreshment. It also shows how a performance model and resulting medium migration indicators are used to monitor the electronic deposit system's ability to execute a migration strategy. Finally, it shows the basic components involved in calculating these indicators.

The medium migration indicators alone will not guarantee that one can always prevent a situation in which the time available for migration is too short. Changes in the system architecture or load characteristics can drastically change the values of the indicators. This report also identifies the potential strategies that can be applied to reduce the time needed in these circumstances. First the critical components inside the complete chain of activities have to be identified. The weakest link can be in any of the basic component areas: processing of the digital object, I/O bus or storage medium.

This report also shows the actions to be taken to manage media migration / refreshment effectively inside an electronic deposit system, in particular the specific media migration issues involved within DIAS. Potential additional capacity required for media migration might be created using redundancy and parallelism.

Together, the indicators and the actions provide a framework for continuously monitoring an electronic deposit system's capabilities for performing potential media migrations triggered by decay or obsolescence of the storage medium.



# 2/

## 01 storage 0101000010 performance model

Technology is changing rapidly. The life span of technology components can be as short as 5 years before new technological innovations make them obsolete. Storage technology used to store the digital objects of an electronic deposit system is no exception.

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. One of the reasons for migration could be the better media decay characteristics offered by optical storage and the resulting reduction in read/write errors. 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 290 days does not even take into account the processing that has to be performed by the electronic deposit system. Before a digital object can be migrated some processing is required to retrieve the information needed to locate the object and to update this information after refreshment/migration.

The high volumes associated with media migration processes require an explicit management process that supports these activities. At any given moment one needs to

*The high volumes associated with media migration processes require an explicit management process that supports these activities. At any given moment one needs to be aware of the effort involved in migrating identified sets of digital objects to the same or to other media.*

be aware of the effort involved in migrating identified sets of digital objects to the same or to other media. The migration possibilities must be evaluated based on the life expectancy of the individual storage media. These types of storage performance models have to be specified in every electronic deposit system.

Individual storage performance models contain some basic components defining the characteristics of the storage media, their life expectancy and the operational window available for conducting the media migration activities.

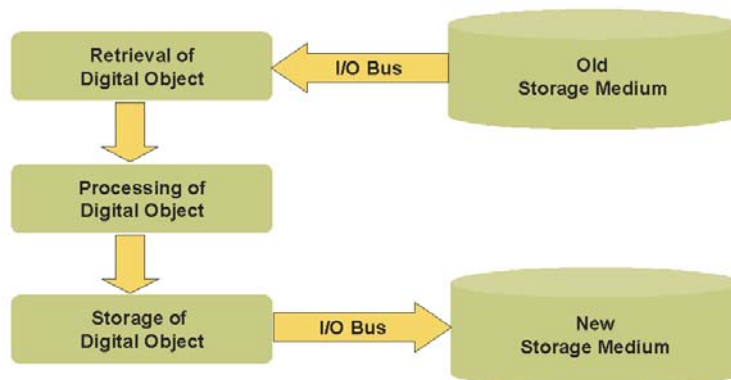


Figure 2.1 / Basic components of a storage performance model

Figure 2.1 graphically depicts the global components and the process around which specific storage performance models are built. Depending on the specific deposit system architecture, the basic model can be further refined to create a more detailed component architecture. For instance, this might include a staging area or DIAS to which the Dissemination Information Package (DIP) and Submission Information Package (SIP) are written during digital object retrieval and resubmitted after performing the migration activities.

The storage performance model also results in a several basic measurement units all defined in the same chosen time unit, e.g. seconds, hours, weeks, etc. The following list of items are encountered in every storage performance model:

- € **Medium Expected Lifetime (MEL)**  
The estimated amount of time the media will be supported and will be operational within the electronic deposit system.
- € **Medium Decay Time (MDT)**  
The estimated amount of time the medium should operate without substantial read and write errors.
- € **Medium Write Rate (MWR)**  
The maximum number of bytes potentially written to the medium in the chosen time unit.
- € **Medium Read Rate (MRR)**  
The maximum number of bytes potentially read from the medium in the chosen time unit.
- € **Digital Object Processing (DOP)**  
The number of time units needed to process a digital object.

€ **Operational Time Interval (OTI)**

The number of time units available within one operational cycle, e.g. day, week, month or year.

€ **Medium Migration Window (MMW)**

The percentage of the operational time units available to the electronic deposit system for media refreshment or migration activities.

The Medium Expected Lifetime, Operational Time Interval and Medium Decay Time all have to be specified in the same time interval, e.g. week, month, year. Otherwise these aspects can not directly be compared.

$$\text{MEL} > \frac{\left[ \frac{(\# \text{objects} \times \text{average}(\text{object size}))}{\text{MRR}} \right] + \left[ \frac{(\# \text{objects} \times \text{average}(\text{object size}))}{\text{MWR}} \right] + (\# \text{objects} \times \text{DOP})}{(\text{MMW} \times \text{OTI})}$$

Figure 2.2 / Medium migration indicator

Collectively these measurement units define the formula used to assess the deposit system's media migration ability. Key aspects are the available medium migration window, the number of digital objects to be migrated and their average size. The formula in Figure 2.2 specifies that the Medium Expected Lifetime (MEL) should always be greater than the amount of time needed to migrate the digital objects on the medium in question. Naturally the MEL always has to be less than the Medium Decay Time (MDT) to secure the correct availability of the digital objects stored on the medium.

The electronic deposit system must continuously monitor the results of the medium migration indicator in green, amber, or red status:

€ **Green**

The MEL is greater than the time needed to migrate including some contingency to be defined by the deposit operators, e.g. 10% of the computed media migration time.

€ **Amber**

The MEL is within the contingency window but, with optimal conditions, migration is still possible.

€ **Red**

The MEL is less than the time needed to migrate.

Normally the deposit should manage the medium migration monitors so that they are in the green range for each of the media used.

## 2.1 Optimizing the Performance Model

Sometimes certain media migration monitors will move into the amber or even the red area; this cannot always be avoided. In these circumstances the electronic deposit system processes or architecture must be modified to reduce the window needed to migrate from one medium to another.

The first activity that has to be performed is to establish the critical components inside the complete chain of activities. The weakest link could be in any of the basic components: DOP, I/O bus or storage medium. Duplicating specific components can enhance the performance. For instance, introducing multiple processing units to reduce the DOP time increases processing power. If the medium to which one is migrating has a bad MWR, one could try to find other media with shorter MWRs so that the time required falls within the defined MEL window. Sometimes the solution is a combination of performance improvements in multiple components. The exact analysis always depends on the individually defined storage performance models associated with each electronic deposit system.

# 3/ DNEP component architecture

01 101000010

Figure 3.1 shows the implementation of DIAS in the surrounding DNEP technical architecture with labels for each component that could influence the general performance.

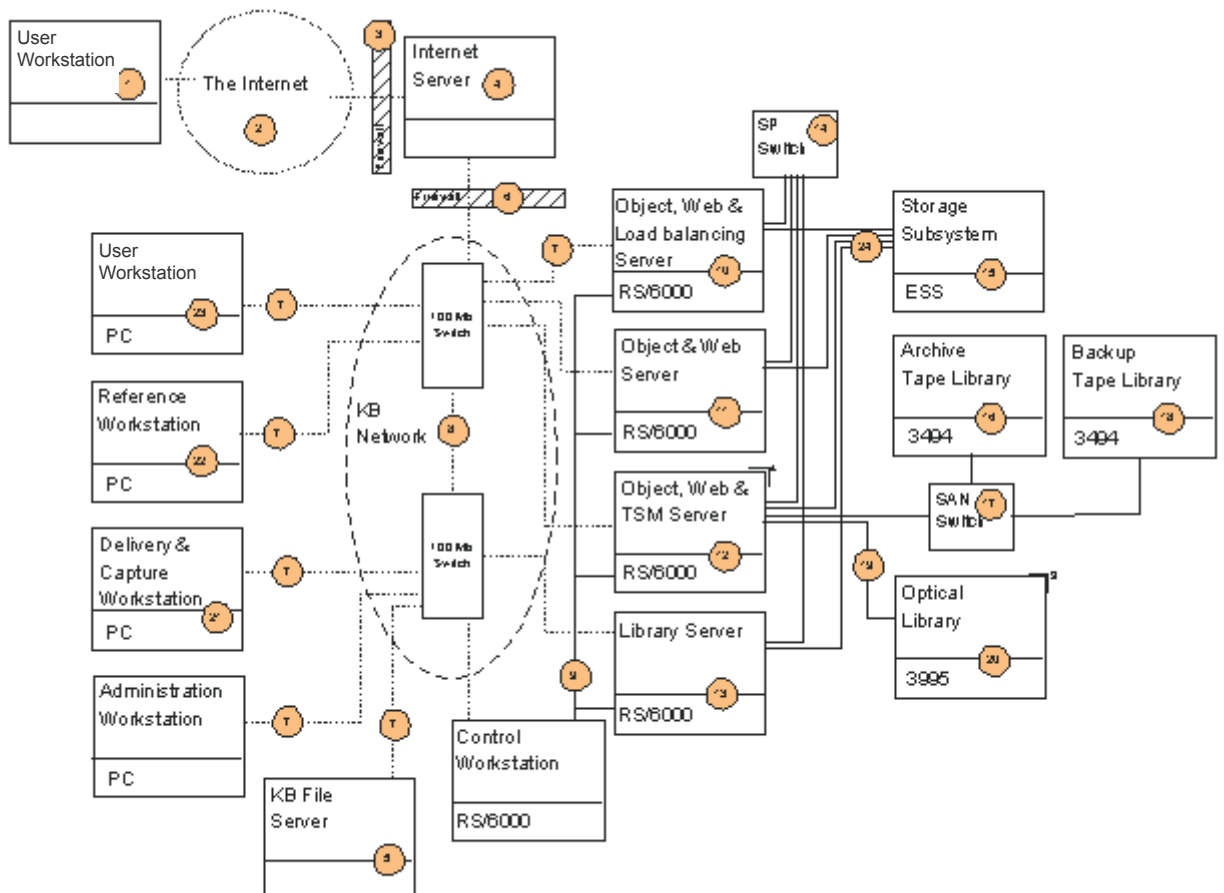


Figure 3.1 / DNEP component architecture

These labels will be used later in this report to reference individual components in the storage performance model. Media migration will be performed inside the DIAS core components: Library-, Object-, TSM servers. For this reason, components 1, 2, 3, 4, 5, 6, 21, 22 and 23 are not further analyzed within the storage performance model. These components are used during the normal ingest and delivery process and are thus key components in establishing operational performance characteristics for DIAS' processing of user requests.

Table 3.1 shows the parametric costs of the DIAS components relevant during media migration. The parametric costs identify the usage information regarding specific IT resources for specific units of work that will be executed within DIAS. The resources are divided into two categories:

€ **Connections**

All relevant I/O-related components with their associated maximum throughput.

€ **Functions**

The number of digital objects, i.e. bits, potentially processed in a certain amount of time by the business logic associated with distinct activities defined within DIAS.

	Label	Description	Speed (Kbyte/sec)	Latency (msec)	Utili- zation	Contin- gency	Quality
<b>Connections</b>	LAN(8)	LAN backbone	100.000	0	10%		
	IPC(10-12)	Object & Web Server	n/a	1	50%		Assumpt.
	SP (14)	SP-Switch	150.000	0	10%		Specs
	D-W(15)	ESS Disk Write	50.000	5	10%		Specs
	D-R(15)	ESS Disk Read	70.000	5	10%		Specs
	D-RC(15)	ESS Disk Read from Cache	100.000	3	10%		Specs
	D-NFS(15)	ESS Disk I/O Through NFS	45.455	6	10%		Measured
	SAN(17)	SAN-Switch	100.000	0			ROT
	T-SW(18)	Backup Tape Library (Sequential Write)	21.000	0			Measured
	T-RR(18)	Backup Tape Library (Random Read)	3.500	123000			Specs
	SCSI(19)	SCSI cable to Optical Lib	20.000	0			Specs
	O-SW(20)	Optical Library (Sequential Write)	4.000	10		25%	Specs
	O-RR(20)	Optical Library (Random Read)	2.800	12000			Specs
FC(24)	Fibre Channel to ESS	100.000	0			Specs	

Table 3.1 / Parametric costs

	Label	Description	Speed (Kbyte/sec)	Latency (msec)	Utili- zation	Contin- gency	Quality
Functions	Stager	Stager	1.000.000	5	75%		Assumpt.
	Destager	Destager	1.000.000	5	75%		Assumpt.
	TSMserver	TSM server	1.000.000	50	75%		Assumpt.
	OS-R	Object Server Read	26.600	670	75%		Measured
	OS-W	Object Server Write	1.000	20	75%		Assumpt.
	Unpack	Unpack	100.000	10	75%		Assumpt.
	CreateSIP	Pack to SIP	100.000	10	75%		Assumpt.
	Conversion	Conversion Prog	10.000	20	75%		Assumpt.

Table 3.1 / Parametric costs (continued)

Explanation of the table:

- ⊘ Label refers to the labels in Figure 3.1 or contains a brief code that identifies the item.
- ⊘ Speed refers to the maximum throughput of the component. For LAN connections, 25% is added to the data as transmission protocol overhead, which results in 10 bits per byte.
- ⊘ Latency refers to the total of the fixed delay times, such as propagation delay in a network or a context switch on a computer.
- ⊘ Utilization is the expected utilization of the component.
- ⊘ The Quality columns refer to the figures' reliability. This ranges from Assumption, Rule-of-thumb (ROT), to Specifications to be Measured.

### 3.1 Details for Tape Library Speed

The tape drive and robot have the following performance characteristics:

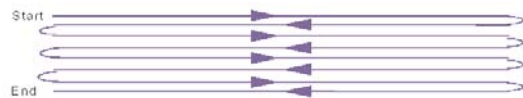
**Tape Drive 3590E**

- Tape capacity: 20 or 40 GigaBytes uncompressed.
- Data rate: 14 MegaBytes/sec uncompressed.
- Physical Tape Load: 25 seconds.
- Physical Tape Unload: 10 seconds (assumption).
- Scan to 50% of tape: 1.5 minutes.
- Rewind 50% of tape: 20 seconds (assumption).

**Tape Robot:**

- Average load time: 7 seconds.

The 3590 uses a serpentine recording technique: data is recorded to the end of the tape, the heads are moved, and the recording continues in the opposite direction, as shown in the figure.



The main benefit of this technique is that for a full tape, the recording ends at the start of the tape, making the rewind time negligible. Within the DIAS design, all new data will be

buffered in a disk pool and written in a single batch, so DIAS will benefit from this technique.

The tape robot and tape drives will be connected using a SAN. The SAN has a higher throughput than the tape devices and can thus be ignored in the performance model.

The DIAS software does not perform any data compression. Some files that will be stored in DIAS are already compressed (disk images, pictures); so we will use an average compression factor of 1.5.

Within the performance model, two situations are used: Sequential Writing and Random Reading. The values are calculated as follows:

#### Sequential Writing (MWR)

- € Latency first write: Robot + Tape Load + Scan = 7 + 25 + 90 = 122 seconds.
- € Latency sequential writes: 0 seconds.
- € Latency when tape is full (loading next tape): Unload + Robot + Robot + Load = 10 + 7 + 7 + 25 = 49 seconds.
- € Average Speed: 14 Mbytes/sec \* compression of 1.5 = 21MegaBytes/sec.
- € Time required to write one full 20G tape: 20G/14M = 1429 seconds (24 minutes).  
The delay caused by the load time of the first tape (122 sec) or subsequent tapes (49 sec) is far less than the uncertainty regarding the compression ratio and will consequently be ignored.
- € Resulting parameters:
  - Speed: 21 MegaBytes/sec.
  - Latency: 0 seconds (for sequential writing!).
  - Utilization: 0%.
  - Contingency: 0% (already factored into the lower compression ratio assumption).

#### Random Reading (MRR)

- € Latency when unit is idle: Robot + Load + Seek = 7 + 25 + 90 = 122 seconds.
- € Latency when unit is busy: Wait-For-Ready (WFR) + Rewind + Unload + Robot + Robot + Load + Seek = WFR + 20 + 10 + 7 + 7 + 25 + 90 = WFR + 159 seconds.  
The Latency when idle is used in the basic retrieval performance models. The latency when busy is used in the queue waiting time analysis. These latency values will have a huge impact when estimating data retrieval time. Hence it is important to validate this figure using a number of benchmarks.
- € Speed: the same as above: 21 MegaBytes/second.
- € Utilization: the effect of utilization will be estimated in the model separately.  
Consequently a value of 0% will be used in the parametric cost list.
- € Contingency: 0% (already factored into the assumptions).

## 3.2 Details for Optical Library Speed

The optical library has the following performance characteristics:

#### Optical Library 3995 C68

Disk Capacity:	8 x 5.2 GigaBytes .
Disk Unload (including spin-down):	3 seconds.
Cartridge Move:	3 seconds.
Disk Load (incl. spin-up to ready):	5.5 seconds.
Average Seek Time:	120 msec.

Latency:	10 msec.
I/O rate:	4.6 MegaBytes/second.
Number of Drives:	6.

Each optical library is connected to an RS/6000 SP node using a dedicated SCSI-2 interface. This interface has a maximum throughput of 20 MegaBytes/second. If all 6 drives were used simultaneously, this would limit the I/O rate to 3.3 MegaBytes/second. However, we assume that at least one drive is busy loading/unloading. This results in an average I/O rate of 4 MegaBytes/second.

Two situations are also used for the performance model: Sequential Writing and Random Reading. The values are calculated as follows:

#### Sequential Writing (MWR)

- ∄ Latency first write:  $\text{Disk Load} + \text{Seek} + \text{Latency} = 5.5 + 0.12 + 0.01 = 5.7$  seconds.
- ∄ Latency sequential writes: 10 msec.
- ∄ Latency when disk is full (loading the next disk) =  $\text{Disk Unload} + \text{Latency first write} = 3 + 5.7 = 8.7$  seconds.
- ∄ Speed: 4.0 MegaBytes/second.
- ∄ Time to write one full 5.2G disk:  $5.2 \text{ GigaBytes} / 4.0 \text{ MegaBytes per second} = 1330$  seconds (22 minutes). The delay caused by the latency is far less and will thus be ignored.
- ∄ Resulting parameters:
  - Speed: 4.0 MegaBytes/second.
  - Latency: 10 msec. (for sequential writing!).
  - Utilization: 0%.
  - Contingency: 25%.

#### Random Reading (MRR)

- ∄ Latency when unit is idle:  $\text{Load} + \text{Seek} + \text{Latency} = 5.5 + 0.12 + 0.01 = 5.7$  seconds
- ∄ Latency when unit is busy:  $\text{WFR} + \text{Unload} + \text{Latency-idle} = \text{WFR} + 3 + 5.7 = \text{WFR} + 8.7$  seconds.
- ∄ Speed: 4.0 MegaBytes/second.
- ∄ Utilization: the effect of utilization will be estimated in the model separately. Consequently a value of 0% will be used in the parametric costs list.
- ∄ Contingency: 25%.

## 3.3 Details for RAID Disk System (ESS) Speed

Within ESS the data in cache and disk will be automatically compressed. (The typical average value is 1:3). The exact transfer speed to and from the ESS is very difficult to predict. It depends on many factors such as:

- ∄ Load from other channels connected to the same ESS.
- ∄ Distribution of the data among different disk packs.
- ∄ Compression factor.
- ∄ Cache.

The ESS is connected with each RS/6000 SP unit by an individual fiber channel adapter. The transfer rate for these connections is 1 GigaByte/second.

Theoretically, peak transfer rate from the ESS could be higher than this value. The ESS uses a very large cache. Moreover, all data in both cache and disk is compressed, while data on the fiber channel is not.

However, in the performance model, we will use safe average values that are lower than 1 GigaByte/second. Hence the fiber channel component will be ignored in these performance models.

# 4/



# migration



# strategies within DIAS

The performance models are always based on the actual system architecture of the specific electronic deposit system. If the system architecture changes, the models should be updated as well. A general DIAS architecture can be found in the overview report also published as part of the Long-Term Preservation Report Series [Diessen and Steenbakkens 2002].

This chapter will explore the different basic strategies that can be applied within the media migration process and techniques that can be applied to reduce the migration window needed.

## 4.1 Functional Components Involved

During normal operations all digital objects are ingested and retrieved through the standard DIAS interfaces. The immutability of the digital object stored in the AIP is an important factor in guaranteeing the quality of the digital object once it is submitted to DIAS. Possible conversions required to render the object throughout various technology changes will never update the existing AIP, but instead create a new corresponding AIP. To preserve the immutability of digital objects, the technical meta-data, which changes because of technological innovations, is registered outside the AIP in a separate Preservation Subsystem [Diessen 2002].

The core of DIAS is built around standard IBM products. IBM's Content Manager is a standard solution for storage, management and distribution of all types of digital content, including text, images, audio and video. Content Manager itself can be further decomposed into two core components:

- 1) Library Server.
- 2) Object Server.

The Library Server performs the following functions:

- € It manages catalogue information.
- € It locates stored objects using a variety of search technologies.
- € It provides secured access to the objects held in the collection.
- € It communicates with the Content Manager Object Server.

The Library Server uses a database in which it stores information such as object types, indexes of all the objects stored, the authorized system users, and access control lists (ACL) for each object. The indexes of stored objects contain the key fields that represent the object's meta-data.

The Object Server is used to store the actual digital objects. There can be many Object Servers associated with one Library Server within any given electronic deposit system. An Object Server contains the following functions:

- € Destager.
- € Migrator.
- € Purger.
- € Stager.

*The performance models are always based on the actual system architecture of the specific electronic deposit system. If the system architecture changes, the models should be updated as well*

New objects will be stored in the staging area. Each Object Server has its own staging area. The Destager copies new objects from the staging area into the archive. Requested objects will be retrieved from the archive and placed in the staging area by the Stager. The staging area is also used as the caching area. New objects and retrieved objects will remain in the staging area under control of the Purger based on their access time. The Purger will ensure that the staging area contains neither too many nor too few objects, so as to provide acceptable object retrieval response times. The Migrator controls the migration of objects through a storage hierarchy. However, since objects will be stored in storage pools that are under control of the Storage Manager, object migration will be controlled from there.

The Object Server is designed to interface with another IBM software product called Tivoli Storage Manager (TSM) to provide a flexible and scalable storage solution.

The Storage Manager consists of a Storage Manager Server. The Server manages storage in so-called storage pools. Storage pools can be allocated on magnetic disk, optical disk and tape. Storage pools can be placed in a storage hierarchy that is under the control of the Storage Manager Server. Stored objects can migrate through the

storage hierarchy. This functionality is called Hierarchical Storage Management (HSM). Objects are migrated through the storage pools based on age in conformance with migration policies. Note that this is not space management. Space management is functionality that automatically migrates data (files) based on file size, the number of days since the file was last accessed, or a combination of the two and allows the user to automatically recall (a) data (file) and restore it to its original location in the file system once it is accessed. Space management is not part of the DIAS solution.

The Storage Manager Server is accessible via the Storage Manager Client software component. Through this interface a client can perform the following functions:

- ∉ Backup and Restore.
- ∉ Archive and Retrieve.

The Object Server uses the Storage Manager's Backup and Restore function to request storage (archiving) and retrieval of AIP objects. The primary storage pools must be disk storage pools. Caching must be enabled for these pools to operate. With caching enabled, the migration process leaves duplicate copies of the files on the disk after the server migrates these files to subordinate storage pools in the storage pool hierarchy. The copies remain in the disk storage pool, but in a cached state, so that subsequent retrieval requests can be handled quickly. However, if space is needed to store new data in the disk storage pool, the space occupied by cached files can be immediately reused for the new data. When space is needed, the server reclaims space by writing over the cached files. Files with the oldest retrieval date that occupy the largest amount of disk space are overwritten first.

The Backup and Restore function is also used to create backups of archived objects, system disks and database contents and is used to control restoration.

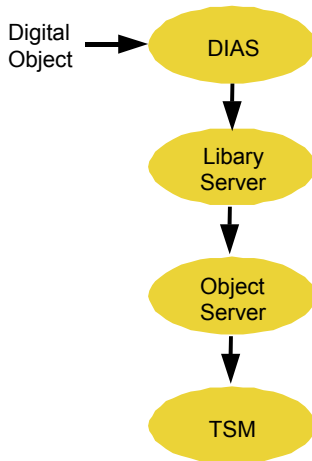


Figure 4.1 / DIAS functional component hierarchy for ingesting digital objects

Figure 4.1 shows the complete chain of major functional components involved in the ingestion process of a digital object within DIAS.

Frequently the reading speeds for specific storage devices are much faster than the writing speeds. Therefore, the buffers used between the different functional components have to actively control the buffers. When required, they have to reduce the speed at which specific functional components write their results to individual buffers in order to avoid a buffer overflow.

## 4.2 Possible Migration Strategies

Depending on the type of migration to be performed, it is possible to bypass part of the normal ingestion process. This will inevitably also reduce the processing time. Figure 4.2 shows the migration strategies involved.

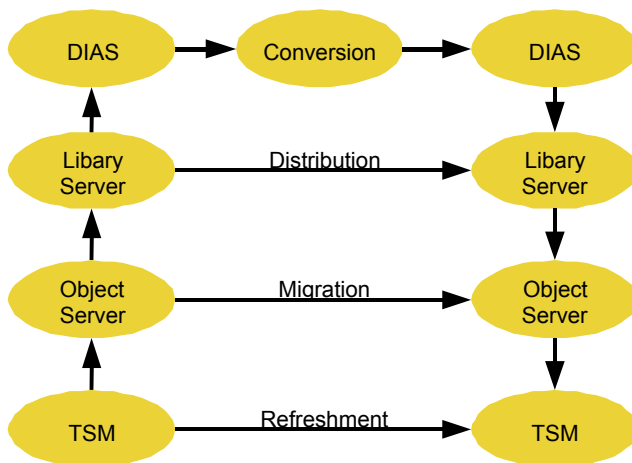


Figure 4.2 / Different migration strategies

The type of movement between storage devices and control components defines the migration strategy to be applied:

1. **Conversion** should be used when the digital object itself is converted. The original digital object is kept in the system and a new copy is created. This is a normal retrieval and ingest process, which involves the complete chain of components.
2. **Distribution** is used when the AIP is moved between Object Servers. The Library Server must manage the relationship between digital objects and Object Servers, which means that this migration must be controlled by that component. The desire to better distribute the total collection over the available Object Servers could be a trigger for this type of migration.
3. **Migration** is used to migrate AIPs to other types of storage devices, for example from disk to tape. The AIP remains unchanged but is moved. The Object Server must manage the storage device type and therefore this migration must be controlled by

this component. The AIP is still part of the same Object Server.

4. **Refreshment** is used to re-write the AIP bit stream on the same device and can be controlled by TSM. For example data migration from an old to a new tape.

Every migration strategy requires not only knowledge regarding conversion, but also a knowledge of the components involved: Library Server, Object Server, TSM. Because these migration strategies shortcut the normal process, using them incorrectly could result in inconsistent data structures. Potentially these inconsistencies could make the digital objects inaccessible.

*Every migration strategy requires not only knowledge regarding conversion, but also a knowledge of the components involved: Library Server, Object Server, TSM. Because these migration strategies shortcut the normal process, using them incorrectly could result in inconsistent data structures.*

## 4.3 Performance Model

In order to create the performance model, the processing flows must be analyzed in detail: The performance model defined in Figure 4.3, on the next page, identifies the high impact components that influence the performance of any given migration strategy.

Control processes, such as Library Server, TSM Library Manager, and Conversion Flow Control are not modelled since they will only have a marginal impact on migration performance. The model will be slightly different for each migration. For example: a disk that is directly accessed by the Stager or Destager could replace the TSM Server and Tape.

Special care has to be taken in managing the staging areas used to store the intermediate results of the different processing steps. Potential deadlock situations can occur if different processes use the same staging areas. The process will grind to a halt if the staging area fills up and a processing step then has to write additional information to the staging area before it can release space in the staging area. To prevent this from happening, the actual flows to and from the staging area have to be regulated.

The next step is to translate this figure into a calculation table, using the parametric cost values listed in Chapter 3. The actual calculation is quite complicated and will not be presented in this report. This calculation provides an estimate of the sequential processing time for a specific migration strategy. Table 4.2 gives an impression of the time required to migrate digital objects (AIPs) from tape to optical storage within DIAS using migration as the selected migration strategy.

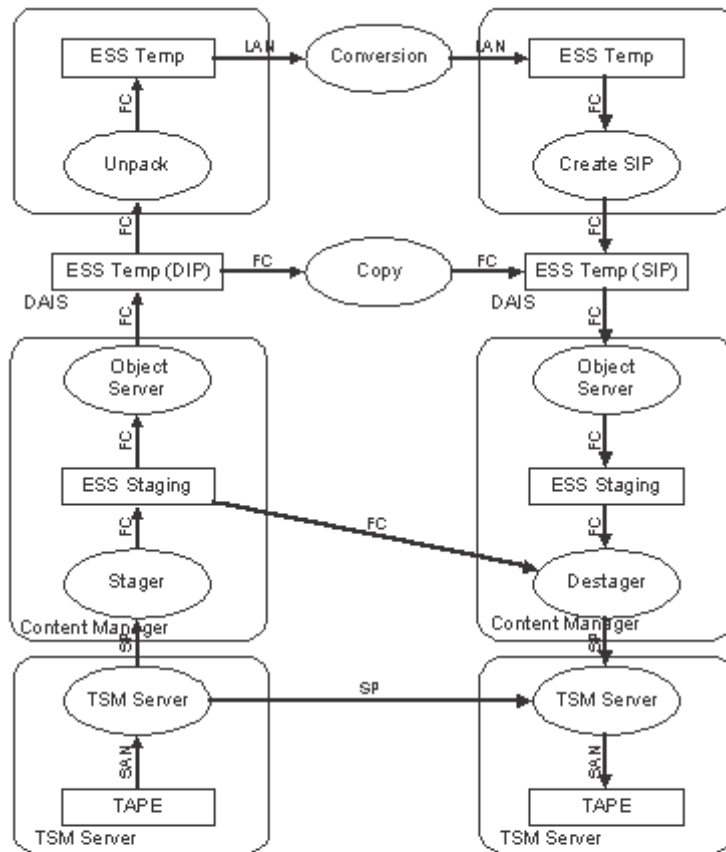


Figure 4.3 / DIAS performance model

### 4.3.1 Assumptions

Many values that are used in this performance model are based on assumptions. The following approach is suggested to validate these assumptions:

1. Measure the end-to-end performance of DIAS with actual data.
2. If these values correspond with the estimate for this model, then the Parametric Cost values used can be considered correct.
3. If these values differ substantially, more measurements are needed to detect the erroneous assumptions. First, the response time between the stop points can be measured; then the response times for the various components between the stop point where the deviation is found can be checked.

Refining a performance model is a continuous effort; the approach outlined above must be followed every time major architectural changes are implemented. The performance model has to be periodically checked against the actual operational electronic deposit system to guarantee that the model correctly reflects the actual system.

Retrieve from Tape, Migrate, Write to Optical			Processing time in hh:mm:ss,msec					
Link	Action	Function	100 kByte AIP	1 MByte AIP	10 MByte AIP	100 MByte AIP	1 GByte AIP	10 GByte AIP
<b>T-RR(18)</b>		Start						
SAN(17)	TSMserver	Read Tape	<b>2:03,029</b>	<b>2:03,286</b>	<b>2:05,857</b>	<b>2:31,571</b>	<b>6:48,714</b>	<b>49:40,143</b>
D-W(15)		TSM Server	0,200	0,200	0,200	0,200	0,200	0,200
		Save to buffer on disk	0,006	0,006	0,006	0,006	0,006	0,006
D-RC(15)		Read from buffer	0,003	0,003	0,003	0,003	0,003	0,003
FC(24)	<b>Stager</b>	Stager	<b>0,020</b>	<b>0,024</b>	<b>0,060</b>	<b>0,420</b>	<b>4,020</b>	<b>40,020</b>
D-W(15)		Save to ESS staging	0,006	0,006	0,006	0,006	0,006	0,006
D-R(15)		Read from ESS staging	0,006	0,006	0,006	0,006	0,006	0,006
FC(24)	<b>Destager</b>	Destager	<b>0,020</b>	<b>0,024</b>	<b>0,060</b>	<b>0,420</b>	<b>4,020</b>	<b>40,020</b>
D-W(15)		Save to buffer on disk	0,006	0,006	0,006	0,006	0,006	0,006
D-RC(15)		Read from buffer	0,003	0,003	0,003	0,003	0,003	0,003
SAN(17)	TSMserver	TSM Server	0,200	0,200	0,200	0,200	0,200	0,200
<b>O-SW(20)</b>		Write to Optical	<b>0,044</b>	<b>0,325</b>	<b>3,138</b>	<b>31,263</b>	<b>5:12,513</b>	<b>52:05,013</b>
<b>Total</b>			<b>2:03,542</b>	<b>2:04,088</b>	<b>2:09,544</b>	<b>3:04,103</b>	<b>12:09,696</b>	<b>1:43:05,624</b>

Table 4.2 / Sample of a performance calculation table

*Special care has to be taken in managing the staging areas used to store the intermediate results of the different processing steps. Potential deadlock situations can occur if different processes use the same staging areas.*

## 4.4 Potential Optimizing Efforts

An electronic deposit system is sometimes required to modify processes or the system architecture to facilitate the execution of a specific migration strategy. Sudden changes in one of the measurements in the different medium migration indicators (see the formula in Figure 2.2 on page 5) can result in unexpected reductions of the MEL time interval. Careful monitoring of the medium migration indicators does not absolutely guarantee that there will always be sufficient time to perform the migration. For instance, the available time in the operational window (MMW) could be substantially reduced by increased service requirements. These, in turn, could be the result of adding new user groups such as other libraries or research institutes.

First the different components that comprise the performance model have to be analyzed to identify the critical components that constitute the complete chain of activities. The

weakest link could be in any of the basic component areas: DOP, I/O bus or storage medium. Then three basic strategies can be applied to reduce the total migration time:

- € Parallel Processing.
- € Optimizing Individual Hardware Components.
- € Restructuring Process Flows.

Each of the different strategies will be discussed in the next section in the context of DIAS.

*Then three basic strategies can be applied to reduce the total migration time:*

- € *Parallel Processing*
- € *Optimizing Individual Hardware Components*
- € *Restructuring Process Flows*

#### 4.4.1 Parallel Processing

Parallel processing is an option if the bottleneck is the actual processing of the DIPs and SIPs rather than the communication bandwidth or the storage devices' read/write rates.

The relevant core processes (Retrieval, Conversion and Ingest) are executed independently, using staging areas (work queues) between the processes. The Preservation Subsystem will identify the set of AIPs to be migrated based on one of the defined migration strategies. The retrieval process will retrieve each of the identified AIPs and present the result as DIPs in a work queue. These DIPs are retrieved by the conversion process and are used to produce the converted digital object (SIP) to be ingested by the system. The SIPs are also stored in a work queue for diverted ingest processing. Finally the ingestion process converts the SIP into an AIP and stores it on the appropriate storage device.

The performance model can be used to estimate the processing time needed for each individual process. Parallel processing is a viable option for the slowest processes in the total chain. However, DIAS also controls parallel processing itself. If possible, one should rely on this feature, since it is already optimized. The specific characteristics of DIAS must also be taken into account. DIAS uses the same optical storage device for retrieval that was used to store a digital object. DIAS distributes the load by allocating the different AIPs across the multiple optical storage devices. Similarly, DIAS must use the same Object Server for retrieval that was used to store the digital object. These features should be taken into account when devices and servers are dedicated for migration.

#### 4.4.2 Optimizing Individual Hardware Components

Often it is not processing power that is the limiting factor, but rather the peripheral hardware: communication hardware, storage devices. These cases require physical alteration of the electronic deposit system hardware configuration. Communication or LAN connections can be either upgraded to connections with greater bandwidth or multiple connections can be switched in parallel. Slow storage devices can be replaced by faster ones and additional CPUs can be added to boost processing power.

Each proposed hardware improvement has to be evaluated by the performance model. Improvement of hardware components can result in other bottlenecks in the other two defined areas. For instance could work queues overflow when I/O bandwidth is being increased in turn necessitating faster processing of certain process steps.

Techniques for increasing the data quality, as described in [Feenstra 2000], using either software (CRC) or hardware (RAID) only indirectly influence the execution of a migration strategy. First and foremost, these techniques focus on increasing the data quality, but as a consequence they also increase the Medium Decay Time of the storage devices. By applying RAID technology we are able to use lower quality disks and still guarantee these disks are resilient to minor read and write errors. A better Medium Decay Time implies a longer window for medium migration. In a sense this also optimizes a hardware component. However, these techniques are generally implemented within the electronic deposit system architecture from the very beginning and are difficult to implement as a change. DIAS uses both CRC and RAID-5 technology to improve data quality.

#### 4.4.3 Restructuring Process Flows

In many cases, the migration can be performed in multiple stages. For example, if data must be migrated from disk to a relatively slow optical device, this can be done in two stages, first from disk to tape and later from tape to the optical medium. This will reduce the time the DIPs and SIPs have to be stored in the work queues, which could create a problem because other DIAS processes also use this disk space.

Since the use of work queues eliminates the direct link between the processing in the various retrieval and ingest stages, some form of flow control is needed in order to prevent buffer overflows during large-scale migration. A specific flow control component should monitor these buffers and temporarily halt the retrieval process if a buffer becomes nearly full.

If migration must be performed during production hours, this flow control must use lower thresholds in order to reserve enough buffer space for normal production. However, this flow control could also be used to make more effective use of spare processing time during operational hours. This effectively increases the medium migration window and the resulting number of digital objects that can be processed.



# 5/

## 01 conclusions



When storage technology becomes obsolete, media must be migrated. Data has to be copied from one storage medium to another on a regular basis. But the fact that a particular storage technology becomes obsolete is not the only trigger for rewriting stored digital objects. All storage media degrade over time and the data must consequently be rewritten either on the same type of medium (refreshing) or on another medium (migration).

The volumes associated with electronic deposit systems (several hundreds TeraBytes) mean that migration is no trivial task. We have shown examples where simply writing the deposit system data to optical storage devices can take almost a year, not even taking into account the processing that has to be done by the electronic deposit system. Such examples show that medium migration processes have to be carefully planned. A situation in which the time required to migrate to a new medium is more than the expected lifetime of the medium currently being used must be avoided.

*This report has shown what type of migration strategies can be adopted: conversion, distribution, migration, and refreshment. It has also explained how a performance model and the resulting medium migration indicators are used to monitor the electronic deposit system's ability to execute a migration strategy.*

This report has shown what type of migration strategies can be adopted: conversion, distribution, migration, and refreshment. It has also explained how a performance model and the resulting medium migration indicators are used to monitor the electronic deposit system's ability to execute a migration strategy. Finally, the report described the basic components involved in calculating these indicators.

The medium migration indicators alone cannot always prevent a situation in which there is insufficient time to migrate. Changes in the system architecture or load characteristics can drastically change the values of these indicators. This report has also identified the potential strategies that can be applied to reduce the time needed in these

circumstances. First the critical components comprising the complete chain of activities have to be identified. The weakest link can be in any of the basic component areas: DOP, I/O bus or storage medium. Based on the assessment, three basic strategies can be applied to reduce the total migration time:

- € Parallel Processing.
- € Optimizing Individual Hardware Components
- € Restructuring the Process Flows.

Together they provide a framework for continuously monitoring an electronic deposit system's capacity to perform potential media migrations that are triggered by medium decay or obsolescence.





# □ appendix a: references



[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](http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html).

[Diessen 2002]

Diessen, R.J. van, *Preservation Requirements in a Deposit System*, IBM / KB Long-Term Preservation Study Report Series Number 3, December, 2002.

[Diessen and Steenbakkens 2002]

Diessen, R.J. van and Steenbakkens, J.F., *The Long-Term Preservation Study of the DNEP Project - an Overview of the Results*, IBM / KB Long-Term Preservation Study Report Series Number 1, December, 2002.

[Feenstra 2000]

Feenstra, B., *Standards for the Implementation of a Deposit System for Electronic Publications*, NEDLIB Report Series, September 2000.



# 01 appendix b: 1000010 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.

**Cyclic Redundancy Check (CRC)** is a checksum that is added to the data and is the result of a checksum-algorithm applied to the data. The sender computes a checksum for the data, adds the checksum to the data (header). On receipt the receiver re-computes the checksum and checks if this matches the checksum computed by the sender; if these don't match an error has occurred during transmission.

**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.

**Digital Object Processing (DOP):** The number of time units needed to process a digital object.

**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 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.

**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).

**Medium Decay Time (MDT):** The estimated amount of time the medium should operate without substantial read and write errors.

**Medium Expected Lifetime (MEL):** The estimated amount of time the media will be supported and will be operational within the electronic deposit system.

**Medium Migration Window (MMW):** The percentage of the operational time units available to the electronic deposit system for media refreshment or migration activities.

**Medium Read Rate (MRR):** The maximum number of bytes potentially read from the medium in the chosen time unit.

**Medium Write Rate (MWR):** The maximum number of bytes potentially written to the medium in the chosen time unit.

**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.

**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.

**Operational Time Interval (OTI):** The number of time units available within one operational cycle, e.g. day, week, month or year.

**Parametric cost values** identify the usage information regarding specific IT resources for specific units of work that will be executed within the deposit system.

**Redundant Array of Independent Disks (RAID)** is used to increase the reliability of disk arrays by providing redundancy either through complete duplication of the data (RAID 1, i.e., mirroring) or by creating parity data for each data stripe in the array RAID 3,4,5). RAID-5, which distributes parity information across all disks in an array, is among the most popular means of providing RAID parity since it avoids the bottlenecks of a single parity disk.

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





IBM  
long

KB  
term

preservation  
study

