Project no. 269977

APARSEN
Alliance for Permanent Access to the Records of Science Network

Instrument: Network of Excellence

Thematic Priority: ICT 6-4.1 – Digital Libraries and Digital Preservation

D23.1 STORAGE SOLUTIONS SUMMARY OF INPUTS

Document identifier:	**APARSEN**-REP-D23_1-01-1_0
Due Date:	2013-02-28
Submission Date:	2013-03-04
Work package:	WP23
Partners:	BL, CINES, CINI, DN, FRD, IBM, CSC–IT, KB
WP Lead Partner:	ESA
Document status	Final
URN	urn:nbn:de:101-20140516133

Abstract:

This report summarises the work done with respect to options available for storage, including in-house and use of external storage providers, and a questionnaire prepared which will be analysed in the next phase of this work package.

Delivery Type Report

Author(s) ESA, BL, CINES, CINI, DN, FRD, IBM, CSC–IT, KB

Approval David Giaretta

Summary

Keyword List Digital preservation, storage

Availability ☒ PUBLIC

Document Status Sheet

Issue	Date	Comment	Author
0_1	2012-09-25	First draft, bringing together inputs from IBM, CINI, CSC,,DNB, KB	ESA
0_2	2012-11-14	Updates from BL	Kirnn Kaur (BL)
0_3	2013-01-18	Addition of info about survey questionnaire	Jeffrey van de Hoeven (KB) Silvio Salza (CINI)
0_4	2013-01-21	Updates to consistency	Veronica Guidetti (ESA)
0_5	2013-01-21	Update information from CINES	Olivier Rouchon (CINES)
0_6	2013-02-12	Minor corrections	Stefan Stein (DNB)
0_7	2013-02-15	Internal reviews – corrections on style, vocabulary and aspects of organisation	Maurizio Lunghi (FRD), David Giaretta (APA)
1_0	2013-02-23	Final version ensuring APARSEN style is followed	David Giaretta (APA)

Project information

Project acronym:	**APARSEN**
Project full title:	**Alliance for Permanent Access to the Records of Science Network**
Proposal/Contract no.:	**269977**

Project coordinator: Simon Lambert/David Giaretta

Address:	STFC, Rutherford Appleton Laboratory Chilton, Didcot, Oxon OX11 0QX, UK
Phone:	+44 1235 445716
Fax:	+44 1235 446362
Mobile:	+44 (0) 7770326304
E-mail:	simon.lambert@stfc.ac.uk / david.giaretta@stfc.ac.uk

CONTENT

EXECUTIVE SUMMARY

The overarching objective of WP23 is to provide recommendations towards adopting storage solutions that can serve digital preservation. We follow the general plan adopted by the research work packages as described in the Description of Work namely:

A. Define the scope and requirements
   - a good idea of the real drivers which determine the solutions needed, and also an idea of how to evaluate options

B. Evaluate the options available
   - understanding the efficacy and applicability of the various options with respect to digital preservation

C. Overlapping with B, develop an integrated view of the best option or the areas of applicability, or a realisation that no existing option is satisfactory. In this last case we would define the solution needed, for later development
   - integrated view in this area and available to be further integrated into the wider digital preservation vision – which may require some adjustments.

D. Plan for sustaining beyond the project and spreading excellence outside the consortium
   - plan which slots into the overall dissemination and exploitation plan.

This document addresses steps (A) and (B) by identifying and comparing WP23 partners' needs in terms of digital objects preservation and their available and currently used storage approaches and technical solutions. D23.2 carries out steps (C) and (D).

Results shall also identify the *conditions* under which *external storage services* would be used by WP23 partners. A final part of WP23 activity is to attempt to *compare different external storage services and their fitness to preserve digital objects in terms of cost, legal, policy and business needs*.

Overall, this document summarises the *options available at WP23 partners' premises* for storage, including in-house and use of external storage providers.

The content of this document and of D23.2 is produced in concert with other project WPs: WP25, WP32, WP21, and WP27.

1 INTRODUCTION

This report summarises the work done with respect to options available for storage, including in-house and use of external storage providers, and the questionnaire prepared.

The boundaries of the current report's scope is limited to the investigation of the storage solutions adopted at a key set of WP23 partners' premises, including two representative use cases of research activities in the domain of digital preservation at storage level.

Overall the purpose is to raise the reader's awareness about some of the current best practices regarding the application of storage solutions to digital preservation.

This report closes by drawing a preliminary set of recommendations and illustrating the work that WP23 partners will carry on during phase II of WP23, ending with the publication of D23.2 in June 2013 and will contain the analysis of results of a survey on digital preservation across storage solutions run among several key project partners, analysing the different options against cost items, reliability, preservation capabilities, and technology lock-in and giving recommendations about adoption of storage solutions that can serve digital preservation.

1.1 PARTNERS CONTRIBUTING TO THIS DELIVERABLE

The project partners that contributed to this deliverable are all WP23 partners, i.e. British Library, CINES, CINI, Deutsche Nationalbibliothek, European Space Agency, Fondazione Rinascimento Digitale, IBM, CSC – IT Center for Science, National Library of the Netherlands.

It is worth noting the very different mandate, work, status, and application domains of WP23 partners. For this a plain and straightforward comparison between the storage solutions adopted at their premises is neither realistic nor meaningful at this stage. Therefore Chapter 2 and Chapter 3 list the main characteristics of storage solutions as perceived and highlighted by each partner. However, the harmonised structure of this document (see Table of Contents) attempt to reflect common issues across the diverse solutions presented.

1.2 DOCUMENT STRUCTURE

Chapter 2 is the main body of this document and provides a description of the technological approaches at storage level adopted at some key WP23 partners' premises. A standard structure is adopted for reports from each partner covering data kind and volume, technical details, choices and tread-offs, use of external resources and considerations about envisaged system evolution

Chapter 3 gives two significant examples of research activity in the domain of storage solutions enabling digital preservation.

Chapter 4 illustrates the next step towards the completion of WP23 work and the publication of D23.2.

Chapter 5 provides initial conclusions and some final remarks.

2 EXAMPLES OF STORAGE SOLUTIONS

Storage is a central component in any preservation solution, and requires special functionalities in order to adequately address the need of a preservation system. Partners' needs for storage may vary substantially, e.g. in required capacity, number of objects, size of a typical object, geographical locations. Furthermore, depending on the nature of the data and its usage pattern, performance needs may vary greatly.

New technological approaches are required that meet the legal, business, cost, and scalability requirements of the 'digital age' for long-term digital information retention. Recent advances in solutions for Storage-as-a-Service, whereby storage is supplied by a storage provider in a pay-per-use mode, raise the option to adopt new technical solutions such as Cloud Storage as the basic storage for preservation systems. Such solutions are typically not enterprise in nature.

A key aspect is to identify whether such technologies can not only address the cost and scalability needs, but also become a platform for trusted storage, addressing auditability, provenance and a desired level for data integrity.

This chapter reports a description of the technological approaches at storage level adopted at some key WP23 partners' premises.

Data Types and Volumes sections focus on both data kind and data volume managed by the system. Technical details, choices and tread-offs are described within *Architecture* sections. Considerations about external additional resources are addressed in *Requirements for third party storage solution* sub-sections. Finally, *Challenges* sub-sections detail considerations about envisaged system evolution.

2.1 EUROPEAN SPACE AGENCY: EO MISSIONS GROUND SEGMENTS

2.1.1 Introduction

In 2003 the European Space Agency has launched a strategy for the evolution of the several Earth Observation (EO) missions' ground segments (handled and/or to be developed) into an open multi-mission architecture, which includes as main goals:

- Adoption of a common architecture for all missions;
- Decomposition of the facility architecture into functional block elements;
- Harmonization and standardization of interfaces;
- Evolution of current missions payload data segments into the common architecture;
- Re-utilization of already available and tested elements;
- Development of a generic multi-mission infrastructure where the elements specific to each EO mission can be plugged in.

For this purpose, the Agency has already or is in the process of harmonizing within one coherent frame some of the facilities basic functional elements as long-term and on line archive, order handling, systematic processing, product distribution including online delivery and inter-facility product exchange. Support features like monitoring and control and software management complement the framework. The long-term goal of this effort is to define a harmonized European infrastructure.

The resulting architecture, based on the ISO 14721:2003 OAIS (Open Archival Information System) standard (see [1] and [2]) is named Multi-Mission Facility Infrastructure (MMFI) and forms the common infrastructure over which ESA intends to build its future payload data ground segments. The

MMFI, via its Multi-Mission Facility infrastructure Elements (MMFE), provides data producer oriented services for data archiving, data processor oriented services for data retrieval and processing management and consumer oriented services in support of the ESA central infrastructure services.

It provides advanced features for long-term preservation and adding value.

2.1.2 Data types and volumes

The European Space Agency today manages the payload data operation of a number of Earth Observation satellites since 1975. The activity includes acquisition, archive, processing and product distribution of data from ESA and third party missions, for which several PetaBytes of data is presently archived.

The ESA EO activity started in 1975 and is continuously progressing with the operation of various ESA and Third Party satellites: ERS-1/2, Envisat, Landsat 5/7, NOAA 16/17, SeaWiFS, Spot 2/4, Terra, Aqua, Proba, SMOS, Cryosat, GOCE, ADM and the upcoming Sentinel and Earth Explorer missions.

2.1.3 Architectural overview of the storage solution

The harmonization of the Payload Data Ground Segments (PDGS) starts from the analysis of the commonalties that can be exploited among them. As schematically shown in the next Figure 1, the analysis of the elements composing a typical EO ground segment shows that many services are in common among all missions and can thus be operated by common elements.

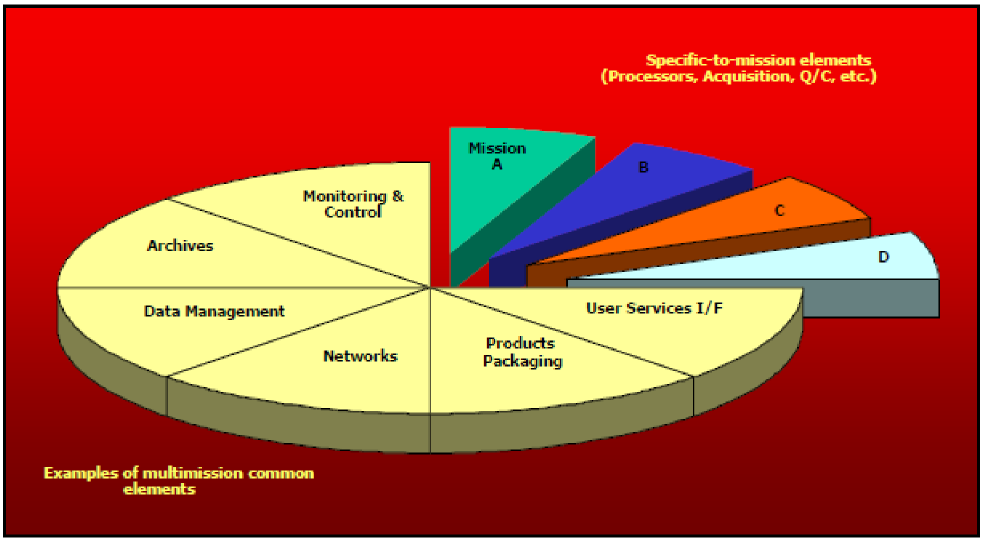

Specific-to-mission elements
(Processors, Acquisition, Q/C, etc.)
Mission A
B
C
D
Monitoring & Control
Archives
Data Management
Networks
Products Packaging
User Services I/F
Examples of multimission common elements

Figure 1: PDGS Decomposition in multimission and mission-specific elements

The archiving system, the data management system, the network, generally the systems that perform the data packaging for the users, and the user services, including the catalogues and the user access functions, are all elements that can be used as part of a common infrastructure. Present and future missions can benefit from a common approach for the development and deployment of these elements and achieve savings in the development during the project phases B/C and the following operations during phase E. Into this infrastructure, mission specific elements like processors, acquisition systems, quality control systems and other elements can be plugged in, provided that they conform or adapt to standard interfaces. The main activities and tasks to be realized in order to develop and implement this common architecture were:

- Decomposition of the facility architecture into functional block elements;

- Harmonization and standardization of interfaces;
- Standardization of products and formats across missions;
- Evolution of current missions payload data ground segments into the common architecture;
- Re-utilization of already available and tested elements;
- Priority-based integration and operational phase in;
- Develop a generic multi-mission infrastructure where the elements specific to each EO mission can be plugged in;
- Harmonization and rationalization of archives.

The Figure 2 represents the logical model of the generic ESA PDGS, based on the OAIS Reference Model.

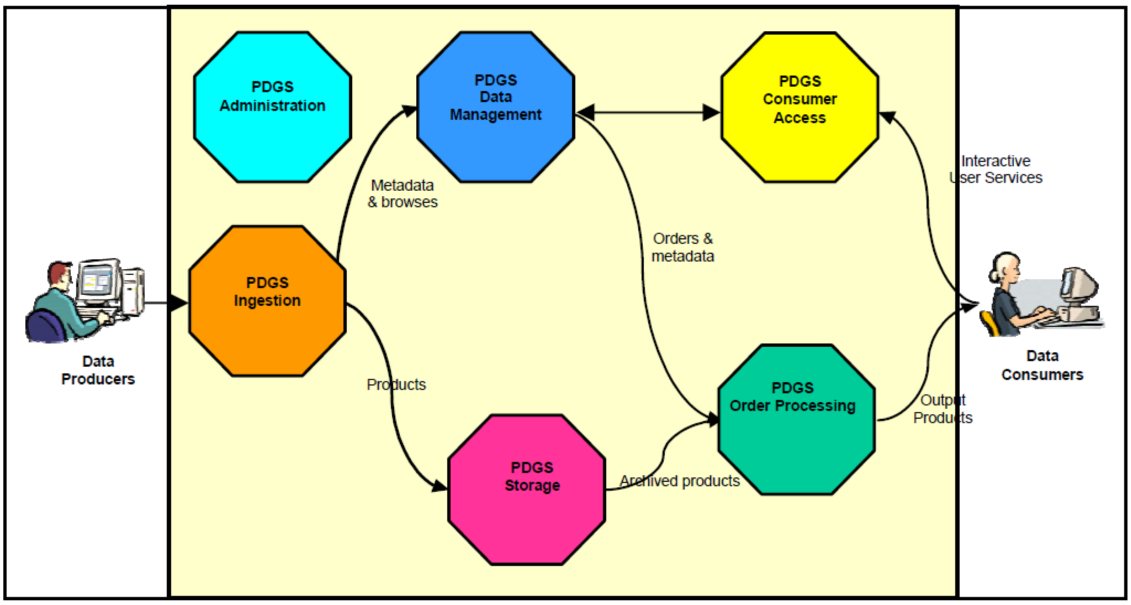


Figure 2: Payload Data Ground Segment Logical Model (OAIS-based)

Due to the distributed nature of the ESA's ground segments, composed of many acquisition and archiving centres, a PDGS for a generic mission is composed of (see Figure 3):

- a Multi-Mission Central Infrastructure component, consisting of all elements required to provide User Services (cataloguing, user access, data ordering, etc.), and Quality Assurance services (payload data quality control, sensor performance assessment, etc.)
- a distributed Multi-Mission Facility Ground Segment (FGS) component, consisting of all elements necessary for the acquisition, ingestion, long-term archive, order processing and data disseminations to end users of a specific mission. As highlighted before, a generic FGS generically requires elements that are mission-specific like processors and quality control systems, but most of its services can be provided by means of common multi-mission elements. Multi-mission centers typically have one “logical” FGS for each mission operated, but the sub-systems used can be shared if they have built-in multi-mission characteristics.

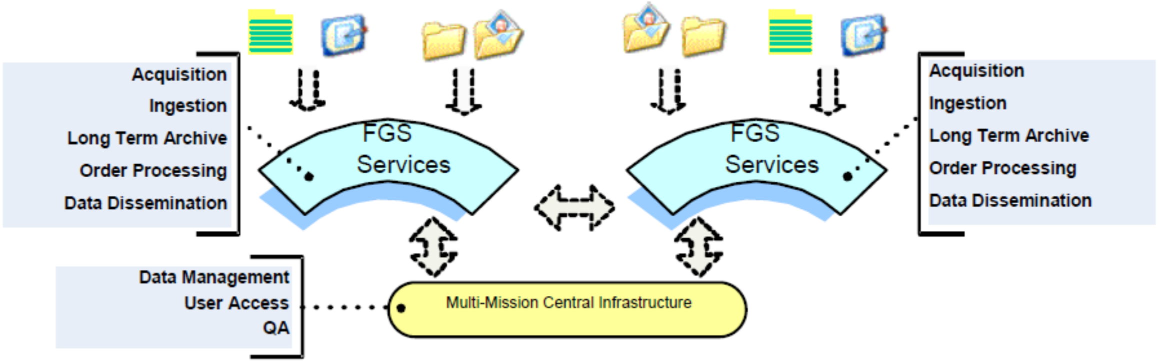


Figure 3: PDGS Generic Decomposition

The services offered by the Facility Ground Segment, can be provided by a mixture of common elements shared among multiple FGSs, and few mission-specific elements. In designing a harmonized solution, the specific characteristics of the ESA PDGSs had to be taken into account. ESA operates its ground segments in distributed centres within Europe and additional acquisition stations around the globe. For higher level data processing and in specific cases data consolidation, the centres are required to exchange some of their data assets.

The common system elements deployed in each centre constitute the ESA's Multi-Mission Facility Infrastructure (MMFI).

In terms of the OAIS reference model, such a distributed system could be called a federated, co-operating archive system. In order to fully cover and highlight the complex aspects of value adding through processing, the OAIS model was extended by a model element "Processing" that, in essence, is both a data consumer and data producer in terms of OAIS (see Figure 4).

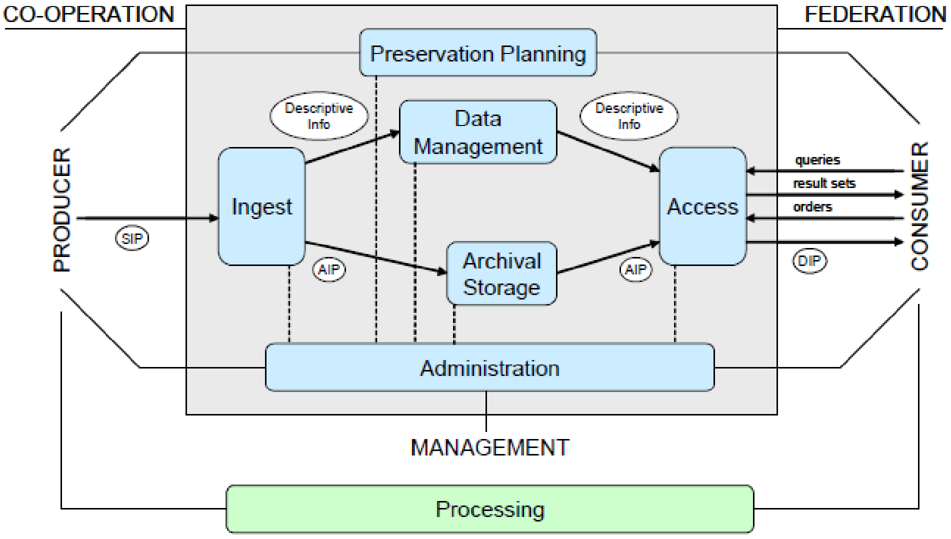


Figure 4: OAIS Reference Model Extended by a Processing Element

Combining the two elements into new model elements highlights the control loops that are introduced when systematic processing or re-processing is done and allows distinguishing it from standard access functionality for on-request value-adding processing. The dynamics of the distributed archive with federation and co-operation also triggered the introduction of explicit interoperability model elements for data and metadata exchange that cover the synchronization with the central catalogue for metadata and browse images as well as data exchange between sites. The resulting logical model for the requirements process is shown in the Figure 5.

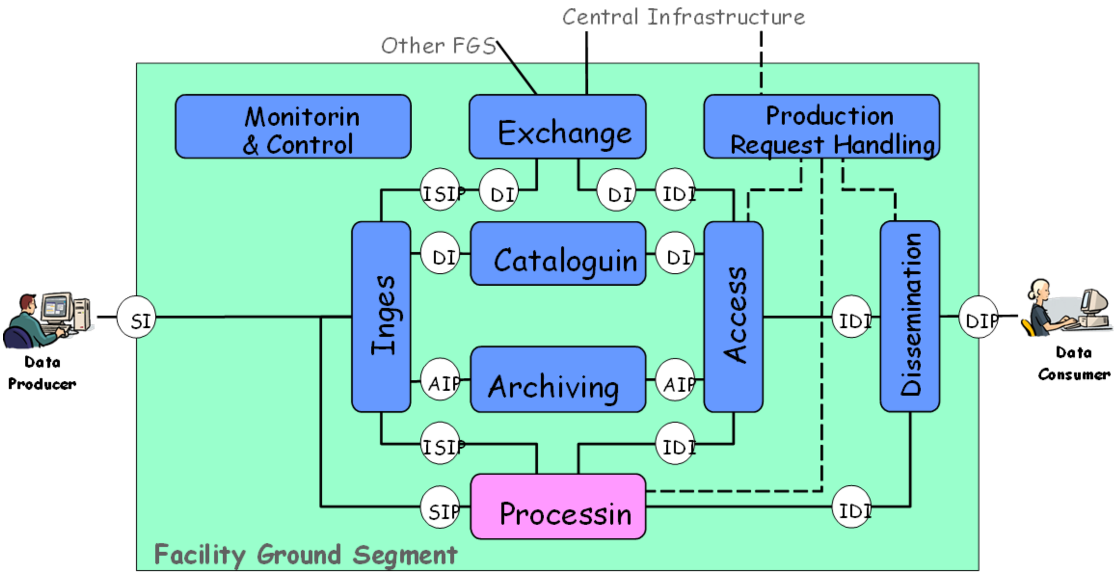


Figure 5: Facility Ground Segment Logical Model

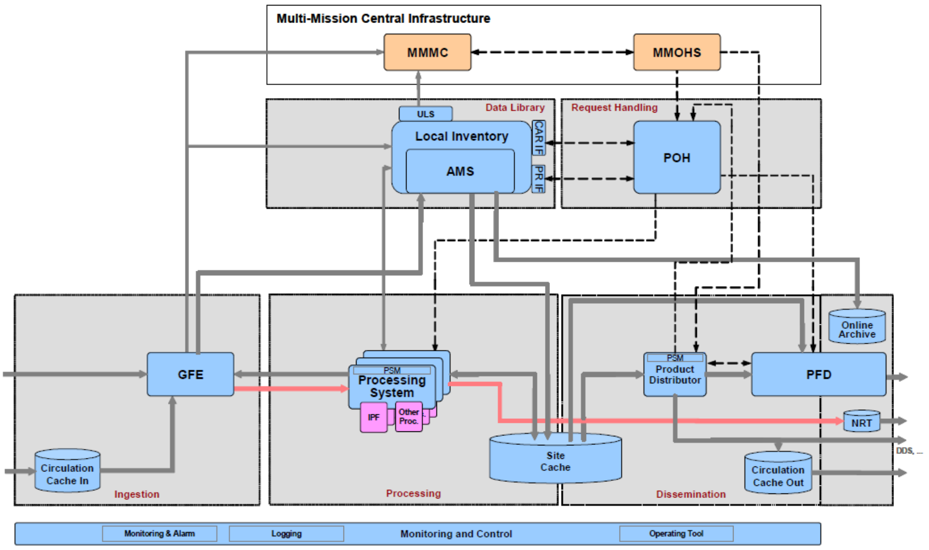


Figure 6: Generic MMFI Architecture at an ESA centre

The final generic architecture of the MMFI is shown in Figure 6, where the MMFI elements are shown in blue and the mission-specific elements in purple. The main building blocks of the MMFI are:

Ingestion

The ingestion workflows are controlled by the "Generic Front End" that provides a configurable workflow engine and a standard set of re-usable plug-ins relevant to ingestion. For data product dependent functionality for metadata extraction and browse generation the ingestion assembly relies on the functionality provided by the Data Request Server that allows accessing data (including binary data) in a configurable manner.

Data Library

The data and metadata holdings of a centre are stored in a data library assembly that consists of the data storage component "Archive Management System" and an inventory component "Local Inventory". The AMS manages the actual archiving of the data product and as an abstraction layer allows to change the underlying storage technology without affecting the architecture of the overall system. It is worth noting that, in line with the goal of achieving the maximum harmonization and standardization of the FGSs, a unique storage technology has been adopted by ESA for its EO missions.

Request Handling

Interfacing ESA’s Multi-Mission Order Handling System is a request handling subsystem, the Product Ordering System that handles production and dissemination requests and organizes the required workflow based on the product type and output medium requested in the order. The POH is supported by a set of auxiliary components that interface other MMFI elements and provide specific functionality for workflow management.

Dissemination

ESA centres need to support a variety of dissemination scenarios and dissemination channels. The key component for dissemination is the "Product Formatting and Delivery", a dissemination workflow management component that can handle different dissemination channels in a configurable manner, and "Product Distributor", a component that manages systematic delivery procedures for data circulation between centres, subscriptions and standing orders.

Processing

The MMFI support processing by providing a generic "Processing System Management" framework. The framework allows integrating mission and sensor specific processing facilities with minimal effort and offering a choice of protocols for the definition of the MMFI - processing facility interface.

Adopted Technologies

- Gbit Ethernet for production network;
- NFS, FTP;
- CORBA, SOAP;
- Limited SAN usage in archiving system (no global FS);
- Wide use of XML (Metadata files, Schemas, XQuery);
- DRB (Data Request Broker);

HW / SW

- Sun/Storage Tek SL8500/93100 Tape libraries;

- Sun/Storage Tek T1000B (1TB/tape);
- SAM-FS HSM → 40 PB distributed license
- Linux, Solaris OS

2.1.4 Requirements for third party storage provider

Additional there are utility components for service monitoring, logging and automated monitoring and alarm. The MMFI is used to implement Facility Ground Segments for all current ESA missions including ENVISAT, ERS and past and current Third Party Missions. The MMFI based ground segments are presently deployed at a number of sites across Europe.

2.1.5 Challenges

The MMFI provides comprehensive functionality for processing and archiving centres by the use of modular components as described in the previous section. Within this section we will especially focus on the features of the MMFI covering preservation and value adding concepts.

The MMFI architecture follows basic principles within the fields of preservation and value adding:

- It supports operational scenarios for preservation strategies, like the periodic migration of digital products to new information technology.
- It provides encapsulation by self-describing items as defined by the information packages in the OAIS model. The maintenance of metadata is performed and means for data consistency are supplied.
- It supports automated production by means of sophisticated data access and processing management.
- It follows a modular design, has an open architecture and streamlined interfaces that permit an easier substitution of one or more of its elements, if the need will arise in the future to ensure the long-term preservation of its data holdings and of its services.

Special attention was turned on the architecture for a well-balanced assignment of functionality-to-components which covers both concepts in an effective way. The result is a number of advanced system design and functional features which are described below.

In the functional area a comprehensive set of features or services is needed for preservation and value adding. Several MMFI components contribute to a specific service but at the same time retaining their individual autonomy. Table 1 tries to give a complete list of the functional features with a short description.

Cataloguing and archiving	Cataloguing and archiving is the basic feature to consistently manage data and metadata for long-term. In addition advanced access capabilities are available to search and retrieve products for automated production.
Automated Product Ingestion	During automated product ingestion data products are checked for consistency before archival. Metadata are extracted and where applicable browse images are generated for catalogue applications.
Order driven processing and delivery	Order driven delivery is the classic dissemination workflow initiated by a user order. Optionally a value adding processing step may occur before delivery.
Systematic data driven processing	Systematic processing describes the capability to initiate automatically processing workflows for higher level products upon the reception of a lower level product.
Systematic re-processing	Systematic re-processing is used to generate a new revision of a product

	collection due to processing algorithm or configuration update. It’s a processing schema with data from/to archive
Systematic dissemination	Subscription type systematic dissemination is similar to systematic data driven processing with the difference that the newly arrived products are not processed but delivered to one or more customers. The dissemination process may include optional post-processing or product reformatting steps.
Online archive access	Online archive access allows to directly retrieving the product data in a file based transfer protocol
Data circulation	Data circulation is concerned with the distribution of data between centres to serve data migration purposes incl. auxiliary products for remote processing.

Table 1: Summary of functional features

2.2 EUROPEAN SPACE AGENCY: RESEARCH AND SERVICE SUPPORT FOR EO

2.2.1 Introduction

The ESA Earth Observation Ground Segment Department operates the so called Earth Observation Research and Service Support (RSS). RSS primary mission is to support the EO data user's community, to ease the development of applications adding value to raw data. The RSS environment also serves the ESA ground segment harmonisation activities, collecting and classifying ground segment technology development needs [9]

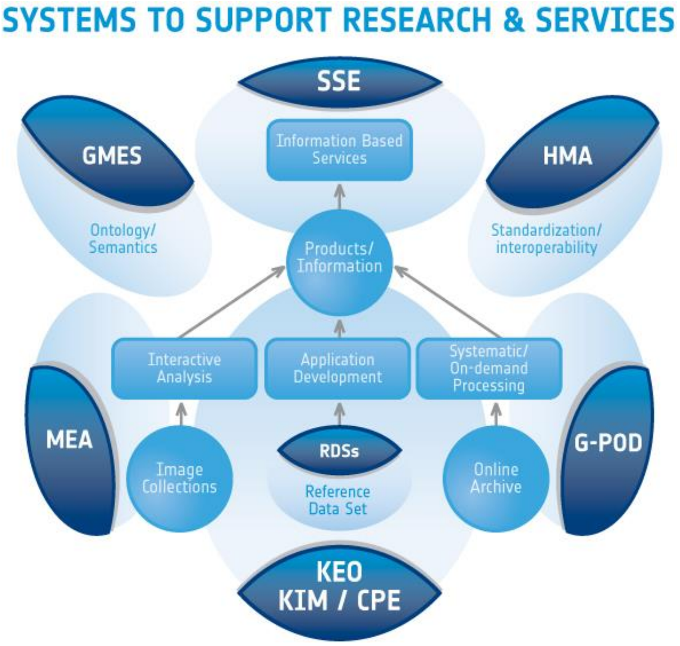


Figure 7: RSS overall architecture

From Satellite data to value added information, RSS offers specific solutions answering the EO data users community needs (see Figure 7). Presently, RSS solutions include:

- E-Collaboration Environments: Research, develop and make available environments for distributed collaboration, systematic or ad-hoc extraction of information, service chaining and provision
- Applications and Services: Support application development and service provision, also as result of cooperation among scientist, value adders and service providers
- Information Based Services: Support the development and provision of services based on the information extracted from the real time or archived EO data
- Test-beds and Reference Systems: Support the development and provision of test-beds, reference systems and data-sets

RSS also aims to offer a front-end for data access and processing, relying on software tools made available by ESA for local data analysis at the user desk. ESA provided EO software toolboxes are available at http://earth.esa.int/resources/softwaretools/. This approach is particularly adequate for those users who have started their work locally on their workstations with some data samples and want to “scale up” to massive data processing or wide service exposure. Also, many users rely on the RSS for the pre-processing of large data amounts, leaving their experimental tasks to run locally at their own labs.

G-POD Environment

ESA’s Grid Processing on Demand (G-POD) is a processing platform where EO data handling applications can be plugged for processing EO data into value added products. It integrates high-speed connectivity, distributed processing resources and large volumes of data to provide science and industrial partners with a shared data processing platform fostering the development, validation and operations of new earth observation applications.

The access to the environment in ensured by the G-POD web portal. It is a flexible, secure, generic and distributed plat-form, where the user can easily manage all its tasks. From the creation of a new task to the result publication, passing by the data selection and the job monitoring, the user goes through a friendly and intuitive interface accessible from everywhere (see Figure 8).

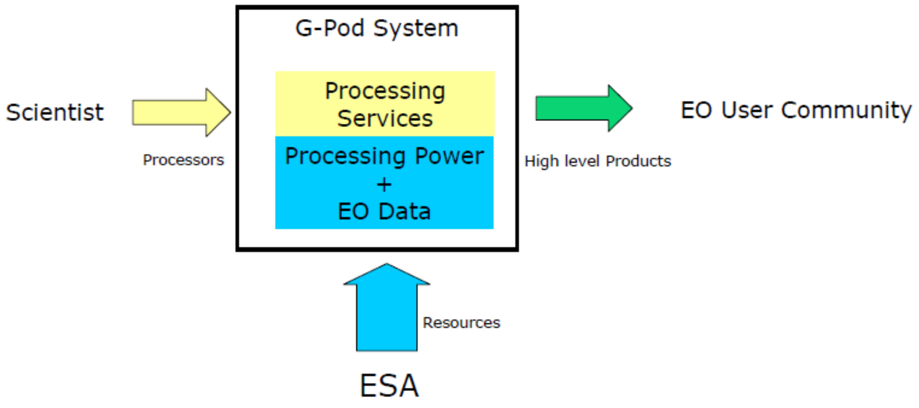


Figure 8: Objectives and Concept

2.2.2 Data types and volumes

G-POD Resources

The G-POD system hosts local GRID computing resources, including:

- Over 350 CPU in about 70 Worker Nodes
- Over 410 TB of local on-line Storage plus 180 TB of EO data accessed directly from the PACs
- Access to Cloud processing and data resources on demand (from Amazon and other providers)
- Internal dedicated 1 Gbit LAN at ESRIN and at UK-PAC archives
- 100 Mbps external connection
- HiSEEN WAN to PACs
- Software Resources on-line: IDL, Matlab, BEAT, BEAM, BEST, CQFD, NEST, BRAT
- System: GRID Globus on Linux

- User Portal

G-POD nominal resources belong to ESA and are located at ESA's computing facilities; mostly at ESRIN but also at UK-PAC. However, thanks to the flexibility of the GRID architecture, G-POD can easily federate additional computing and storage resources (see Sec 2.2.4).

G-POD Storage Data

The ESA Grid Processing on Demand system is a generic GRID-based operational environment where specific data handling applications can be seamlessly plugged into system. One of the goals of G-POD is to provide processing facilities closer to the data, in order to do so, G-POD has more than 300TB of EO data locally stored.

G-POD services use ESA and non-ESA EO missions data, Table 2 lists a selection of G-POD datasets, with details about the collection extent and coverage maps. Due to its dynamic nature, it has to be notice that G-POD hosts more data than the ones here reported, as new dataset are continuously ingested in the framework of new G-POD services. The temporal span, of the sensors data acquisition in the table, starts since 1991, for the ERS-1 missions, up to now, for the SMOS on-going mission. Depending on the requests submitted by the scientific community, even older data can be stored, temporary of permanently, depending on the project’s relevance and purpose.

Dataset	Files Number	Data Storage [TB]
ENVISAT	3384200	313
ERS	265785	55
SMOS	88390	11
MODIS	173667	0.2
MSG	170260	0.67
Total	**3816517**	**325**

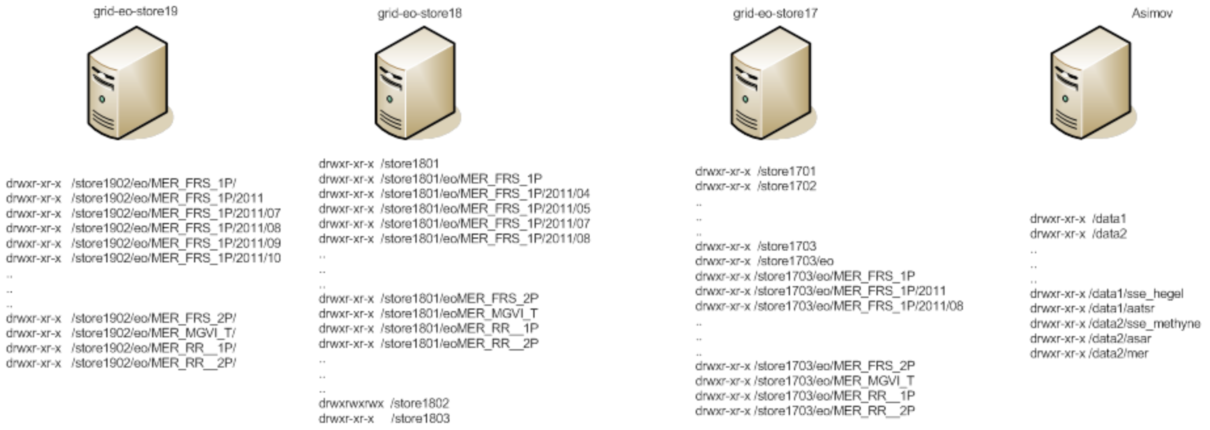
Table 2: Datasets stored in ESA’s G-POD facility. The data in the table is updated on monthly basis, therefore all the data refers to the latest month. Last update: 11 Jan 2013 09:06:24 GMT

2.2.3 Architectural overview of the storage solution

RSS Data Farm Project

In order to permit the end user to get direct access from any RSS environment, to the complete RSS data repository, the RSS Data Farm project has been put in place. The new approach, designed and implemented within the project, ensures a higher degree of flexibility in accessing data. It is now possible to ingest data directly from the former G-POD dedicated storages into the RSS WebMap Server, with no need to copy data on a local storage. The same applies to RSS environments depicted in Figure 9. The current architecture offer further benefits such as: Optimized storage space utilization, easy access control and easy scalability, as depicted in Figure 10 .

Before the RSS Data Farm project, the storage solution was based on 17 storages with different sizes for G-POD, for a total amount of about 410 TB.



```
drwxr-xr-x   /store1902/eo/MER_FRS_1P/
drwxr-xr-x   /store1902/eo/MER_FRS_1P/2011
drwxr-xr-x   /store1902/eo/MER_FRS_1P/2011/07
drwxr-xr-x   /store1902/eo/MER_FRS_1P/2011/08
drwxr-xr-x   /store1902/eo/MER_FRS_1P/2011/09
drwxr-xr-x   /store1902/eo/MER_FRS_1P/2011/10
..
..
..
drwxr-xr-x   /store1902/eo/MER_FRS_2P/
drwxr-xr-x   /store1902/eo/MER_MGVI_T/
drwxr-xr-x   /store1902/eo/MER_RR__1P/
drwxr-xr-x   /store1902/eo/MER_RR__2P/
```

```
drwxr-xr-x  /store1801
drwxr-xr-x  /store1801/eo/MER_FRS_1P
drwxr-xr-x  /store1801/eo/MER_FRS_1P/2011/04
drwxr-xr-x  /store1801/eo/MER_FRS_1P/2011/05
drwxr-xr-x  /store1801/eo/MER_FRS_1P/2011/07
drwxr-xr-x  /store1801/eo/MER_FRS_1P/2011/08
..
..
..
drwxr-xr-x  /store1801/eoMER_FRS_2P
drwxr-xr-x  /store1801/eoMER_MGVI_T
drwxr-xr-x  /store1801/eoMER_RR__1P
drwxr-xr-x  /store1801/eoMER_RR__2P
..
..
..
drwxrwxrwx  /store1802
drwxr-xr-x     /store1803
```

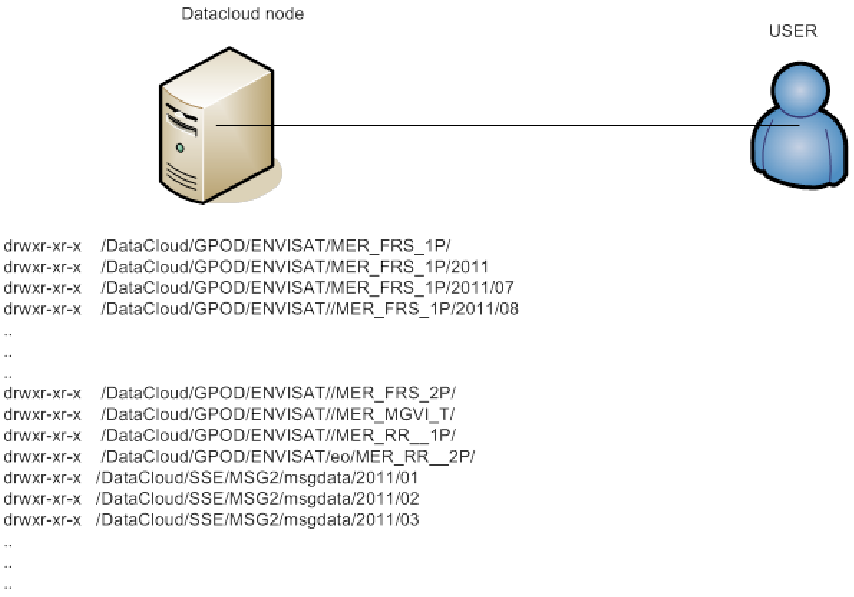
```
drwxr-xr-x  /store1701
drwxr-xr-x  /store1702
..
..
..
drwxr-xr-x  /store1703
drwxr-xr-x  /store1703/eo
drwxr-xr-x /store1703/eo/MER_FRS_1P
drwxr-xr-x /store1703/eo/MER_FRS_1P/2011
drwxr-xr-x /store1703/eo/MER_FRS_1P/2011/08
..
..
..
drwxr-xr-x /store1703/eo/MER_FRS_2P
drwxr-xr-x /store1703/eo/MER_MGVI_T
drwxr-xr-x /store1703/eo/MER_RR__1P
drwxr-xr-x /store1703/eo/MER_RR__2P
```

```
drwxr-xr-x  /data1
drwxr-xr-x  /data2
..
..
..
drwxr-xr-x /data1/sse_hegel
drwxr-xr-x /data1/aatsr
drwxr-xr-x /data2/sse_methyne
drwxr-xr-x /data2/asar
drwxr-xr-x /data2/mer
```

Figure 9: GPOD directory example – before Data Farm project

The objective has been to join all the storages as one single mount point, respecting specific conditions, such as impossibility to format the available storage, to adopt specific TCP Ports for security reasons, to ensure encrypted communications and grant access control list's for user to access specific data.

About the file system choice, after having analysed and tested any feasible solution, it has been adopted GlusterFS file system, as it is a POSIX based filesystem, can be deployed on top of an pre-existent filesystem, supports mirroring, replication, load balancing, storage quotas, ACLs for user access to data and run in kernel user mode, being administrable as a middleware.



```
drwxr-xr-x    /DataCloud/GPOD/ENVISAT/MER_FRS_1P/
drwxr-xr-x    /DataCloud/GPOD/ENVISAT/MER_FRS_1P/2011
drwxr-xr-x    /DataCloud/GPOD/ENVISAT/MER_FRS_1P/2011/07
drwxr-xr-x    /DataCloud/GPOD/ENVISAT//MER_FRS_1P/2011/08
..
..
..
drwxr-xr-x    /DataCloud/GPOD/ENVISAT//MER_FRS_2P/
drwxr-xr-x    /DataCloud/GPOD/ENVISAT//MER_MGVI_T/
drwxr-xr-x    /DataCloud/GPOD/ENVISAT//MER_RR__1P/
drwxr-xr-x    /DataCloud/GPOD/ENVISAT/eo/MER_RR__2P/
drwxr-xr-x   /DataCloud/SSE/MSG2/msgdata/2011/01
drwxr-xr-x   /DataCloud/SSE/MSG2/msgdata/2011/02
drwxr-xr-x   /DataCloud/SSE/MSG2/msgdata/2011/03
..
..
..
```

Figure 10: GPOD directory example – after RSS Data Farm project

2.2.4 Requirements for third party storage provider

G-POD can easily federate additional computing and storage resources. This capability has been used to temporary federating external storage (and computing) resources during specific processing campaigns.

The potential of cloud computing concept for G-POD is very high as responds to the need to employ a high number of resources on-demand. However, the usage of external resources is limited by the burden of moving high data volumes to the external resources over the WAN. For that reason, usage of external resources is only employed for some project with a positive trade-off between the costs of transferring data against the value of large cloud computing resources. In the past years external resources have been attached from G-POD from the following sources:

- Amazon Web Services (http://aws.amazon.com/)
- Hetzner server services (http://www.hetzner.de/)
- HP private cloud demonstrator (locally deployed at ESA)

2.2.5 Challenges

Data Storage Future Evolution

Data Management is a critical issue in order to ensure a good service and requires fast access to (almost) all the data products generated. This is currently achieved by three sourcing modalities: copying rolling archives, caching data from on-line archives and moving off-line data on-line. Nowadays, the proportion is roughly 70-10-20 for the three modalities. Considering the data volumes from future Sentinel operational missions, in the order of 2TB a day, the proportion is expected to evolve towards a 20-70-10. That is, off-line archives will tend to be replaced by on-line archives and, due to the data volumes involved, copy from other archives will not be affordable. Instead, intelligent caching over large on-line archives is to be foreseen. Therefore topics currently under research and development are:

- Optimization of data circulation for caching purposes: data granularity, network configuration, seeding strategy;
- Caching strategies based on use-patterns, trying to “guess” which data will be requested next.

Simple precursors have been developed to bridge the ESA G-POD environment with MERIS Catalogue and Inventory data source and various FTP repositories at ESA facilities.

2.3 NATIONAL LIBRARY OF THE NETHERLANDS

2.3.1 Introduction

The primary task of the National Library of the Netherlands (KB) is to preserve Dutch publications and publications about the Netherlands for the long term and give access to them. Since 2003, the KB broadened that task by including digital publications, also on an international level.

Storage forms an essential part in the KB’s ICT infrastructure. The KB manages a large range of different types of digital collections and its main function is to safely store digital objects, to ensure the integrity of these objects and their very existence, and to offer functionality that these objects can be retrieved any time throughout their lifecycle.

Due to the diverse nature and purpose of the digital collections of the KB (e.g. scientific e-journals, e-books, high quality digital still images, and archived web sites) and their use for library services, the storage solution has to deal with diverse storage requirements from different processes handling digital data.

To accommodate all digital collections and services the KB has its own data centre comprised of various types of storage systems and computing. The KB operates around 230 servers (150 virtual, 80 physical) which host services to business and consumers as well as back office. Most of these servers are following Blade architecture to increase availability, reliability and decrease cost and power consumption.

To ensure all data is processed and securely stored the solution has to be scalable. Data meant for long-term preservation will go through several checks before it is stored. This ingest process will generate a peak of 5.5 million files a day with 500 KB per file on average. The average processing rate is expected to be around two hundred thousand (200,000) files a day with average file size of 5 MB. These files will be transported from ingest workspace to Archival Storage.

Apart from ingest and long-term preservation, the KB also provides access to their collections. Various web sites hosted by the KB provide access to digital collections (e-Journals, digitised newspapers, e-books, archived websites and more) and their metadata. In most cases, derivatives for access are made from the original files to lower the load on the infrastructure and to give better user experience. The KB counts around 400,000 unique site visitors per Month and expects a growth of 10% each year. Most of these users request metadata from KB's catalogue or browse indexed text and images.

2.3.2 Data types and volumes

It goes too far to describe in detail the 90+ digital collections the KB currently preserves. Instead, Table 3: Top five digital collections at KBTable 3 shows the top five major digital collections of the KB with their current volume and expected growth per year.

Type	Total number of collection items	Total number of files	Storage volume (TB)	Expected annual growth (TB)
Digital scientific journals	17 million	170 million	20 TB	2 TB
Digitized newspapers	1.1 million	143 million	297 TB	34 TB
Digitized books	1,000	152,000	26 TB	100 TB
Digital images	Unknown	Unknown	104 TB	50 TB
Web archive	17,000	731,000	7 TB	2 TB
Total			454 TB	188 TB

Table 3: Top five digital collections at KB

2.3.3 Architectural overview of the storage solution

The KB defined a generic storage policy in 2010 which is applied to all data types stored at the KB and off-site. In principle, all primary data is backup-ed/replicated twice (one on location, the other remote) and should make use of at least two different types of storage technology (e.g. disk, tape). Per type of data it should be considered what the usage characteristics it will have (access, archiving). Based on this, a decision is made which tier to apply to it.

To ensure that required storage fits the quality requirements and costs for a certain service, the KB has defined several **Storage Tiers** that group similar storage systems with predefined characteristics. The following tiers are defined (with labelling):

Platinum (Tier 0) type of storage of which data is immediately accessible, acting as directly accessible memory. It is suitable for indexing and databases with only few mutations. Examples of media are RAM or Solid State. This tier is currently not implemented and reserved for future needs.

Gold (Tier 1) type of storage which offers high performance access and minimal chances of failures. It is used for services such as databases, indexing and running operating systems. The KB has implemented this tier as a SAN with RAID 1 using Fibre Channel disks and SAS disks operating in HP EVA P6500, 8000 and 8400 storage arrays. The current total capacity is about 100 TB.

Silver (Tier 2) type of storage which is comparable to *Gold* but with higher capacity and slightly lower performance. It is used for back office applications, web hosting and workspace. The KB has implemented this tier as a SAN with RAID 5 using FATA and SAS disks in HP EVA P6500, 8000 and 8400 storage arrays. The current total capacity is about 400 TB.

Steel (Tier 3) type of storage suitable for long-term storage of large volumes of data at low costs and power consumption. It is capable of preserving these volumes safely for long periods of time (long-term archiving). Such storage could be classified as near-line storage which allows access to data with a short delay. It is used for archival storage of digital collections. The KB has implemented this tier by using secure NAS-based storage from Fast LTA. The solution is called Silent Cube offering NAS with WORM capability and four-time redundancy per storage module. The current total capacity is 50 TB on the primary location and has a replicated storage environment elsewhere.

Bronze (Tier 4) type of storage meant for long-term archiving and back-up. Access to this data is very slow. The KB has implemented this tier by using IBM tape storage LTO3, LTO4 and LTO5 in two IBM Total Storage 3584 tape libraries managed by IBM Tivoli Storage Manager (TSM). One of the tape libraries is situated at the KB, the other off-site connected to each other via a 2 Gbit/s fiberglass link. Total capacity per site is about 500 TB.

Hierarchical Storage Management (HSM) between the tiers is very important, but not completely implemented yet. To manage storage tier 3 (Steel) and parts of tier 2, the KB is currently implementing Quantum StorNext storage manager and StorNext's shared File System SNFS. Management of tier 4 (Bronze) is done successfully by using IBM's TSM software suite.

2.3.4 Requirements for third party storage providers

In 2011, the KB defined several mandatory requirements to which third party storage providers must comply if they want to offer a long-term archiving solution. A subset of these requirements is listed hereafter:

1. The storage solution must be capable to deliver any file smaller than 100 MB within 45 seconds.
2. The storage solution must offer high capacity storage volume.
3. Storage must be done on file level.
4. No software-based reduplication must be applied.
5. The storage solution offers WORM-functionality.
6. The storage solution must offer redundancy to prevent data loss.

7. Storage components must be hot-swappable.
8. A regular bit integrity check must be applied.
9. As checksum algorithm 256-bit SHA-2 or better must be used.
10. After detection of data corruption, automatic repair must be done.
11. The storage solution must be resilient in case of malfunctions in the system itself.
12. The storage solution has an option to operate as near-line storage by turning on low power consumption.
13. Automatic monitoring and error reporting to maintenance in case of a problem.
14. Replication to another storage solution must be possible.
15. The storage volume can be expanded without loss of availability.
16. The solution must be scalable to at least 2 PB.

The overall consensus of these requirements is that preservation is more important than access. The KB is reluctant to apply cloud storage solutions especially for archival storage. The KB sees several important risks that first have to be solved before it would consider using cloud storage:

- Archival storage: most cloud storage solutions are focused on high availability and less (or not) on long-term archiving. It often lacks functions such as checks for fixity, errors and workflows for coping with file format obsolescence.
- Legislation: data is stored in one or more countries which could have different (even conflicting) legislation on openness of information & privacy.
- Vendor lock-in: once the data is in the cloud, it is difficult to migrate to another cloud or download it locally, because of the volume size and lack of standardized interfaces between clouds.
- Control: storing data in the cloud at a third party decreases your own control over the data.
- Financial: when your organisation is not capable anymore in paying the monthly storage costs, you run the risk of losing all data. When you have your own ICT infrastructure, you can always turn off the power while still having the data on it.

More on cloud storage risks, see http://www.ncdd.nl/blog/?p=2347

2.3.5 Challenges

Although the KB is in the middle of a transition from a physical to digital library it still faces some important challenges for the next years.

The first challenge is to keep storage affordable. With the current growth rate and limited budgets, organisations such as the KB need to be inventive. Well-thought hierarchical storage management is an important tool as storing data on the right tier saves money. To make valid choices, the KB computed the Total Cost of Ownership for each storage tier, as can be seen in the **Table 4**.

Partner	Tier	Cost (TB/yr)	Availability	Reliability	Long-term archiving capabilities	Tech lock-in
KB	0	-	Not available reserved for	-	-	-

			future use			
KB	1	€ 7.340,-	Very high	Moderate	None	HP EVA
KB	2	€ 1.029,-	High	Moderate	None	HP EVA
KB	3	€ 1.033,-	Moderate	Very high	WORM, 4x redundant, random disks, strong case, journaling metadata, no compression, checksums	Fast LTA Silent Cubes
KB	4	€ 435,-	Low	High	WORM, archive function, long retention period of tape	IBM tape

Table 4: Total Cost of Ownership for each storage tier

Out of a list of 34 TCO components, described in the white paper *Four Principles for Reducing Total Cost of Ownership* (June 2011, Hitachi)[1], the KB selected fourteen of the most important of them and incorporated these into their TCO for storage:

1. Hardware depreciation
2. Software purchase
3. Hardware maintenance
4. Software maintenance
5. Storage management labour
6. Back-up and disaster recovery labour
7. Data migration
8. Power consumption and cooling
9. Monitoring
10. Data centre Floor space
11. Cost of waste
12. Cost of duplicate data
13. Local and remote data circuits
14. Storage Area Networking

When compared to other (cloud) storage solutions the TCO for archival storage (tier 3) of the KB is in the mid-range (**Figure 11**) being comparable to Duracloud basic storage. The storage volume is based on expected data growth at the KB of 200 TB per year on archival storage, having its peak on 1 PB in 2018.

[1] Four Principles for Reducing Total Cost of Ownership", June 2011, Hitachi, available at: http://www.hds.com/assets/pdf/four-principles-for-reducing-total-cost-of-ownership.pdf

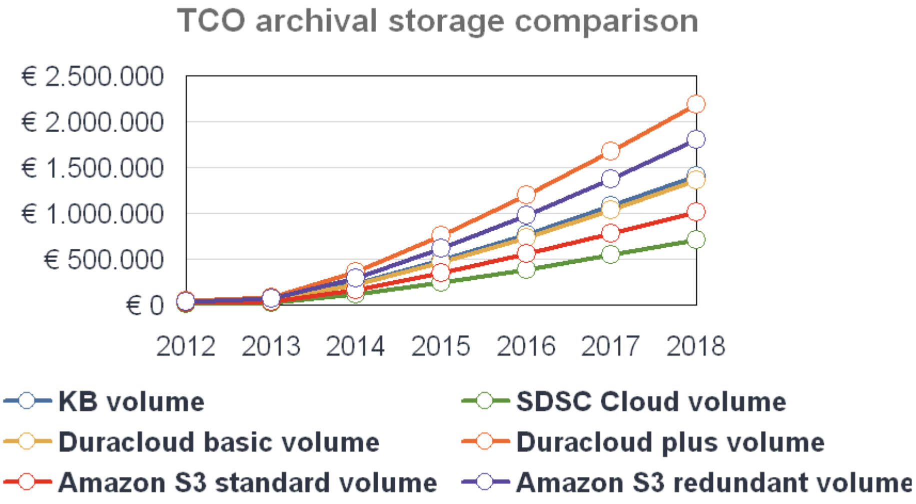


Figure 11: TCO archival storage comparison

The second challenge is data management. With hundreds of millions of computer files stored in the ICT infrastructure, it becomes a real challenge to keep track of them, to be sure the files are still correct and can be accessed by people that have the right authority. To improve this, the KB is currently applying a storage management solution and developing a new digital archiving system as replacement for the e-Depot system.

2.4 THE BRITISH LIBRARY

2.4.1 Introduction

The British Library[2] is the national library of the United Kingdom and one of the world's greatest research libraries. The Library's collection has developed over 250 years and exceeds 150 million separate items representing every era of human history and includes books, journals, manuscripts, maps, stamps, music, patents, photographs, newspapers and sound recordings in most known languages. It is a collection ranging from 3,000 year-old Chinese oracle bones to the latest e-journals.

Legal deposit for printed books and papers has existed in English law since 1662. It helps to ensure that the nation's published output is collected systematically, and as comprehensively as possible, both in order to preserve the material for the use of future generations and to make it available for readers within the designated legal deposit libraries. These are the British Library, the National Library of Scotland, the National Library of Wales, the Bodleian Library Oxford, the Library of Cambridge University and the Library of Trinity College Dublin. Legal deposit in the UK is governed by the Legal Deposit Libraries Act 2003.

Publications deposited with the British Library are made available to users in its various Reading Rooms, are preserved for the benefit of future generations, and become part of the national heritage. The Legal Deposit Libraries Act 2003 introduced a framework in which the Secretary of State may make regulations for the deposit of non-print works.

In the meantime, and pending the implementation of the formal regulations, the British Library has encouraged voluntary deposit of offline, online or "pure electronic" items, which can be managed through its digital storage system. For UK publishers of scholarly electronic journals, a voluntary deposit scheme was initiated in 2007 by the British Library and representatives from publishing trade associations and the other legal deposit libraries. Its aim is to develop processes for supporting the deposit of e-journal publications and for providing access within specified constraints. Website owners and rights holders are encouraged to give copyright permission for the British Library to make snapshot copies of their website at regular intervals for the UK Web Archive. Selective archiving of UK websites (with permission) was started in 2005 in collaboration with the Joint Information Systems Committee, the National Archives, the National Library of Scotland, the National Library of Wales and the Welcome Library.

2.4.2 Data types and volume

As well as the deposited collection, the British Library's digital collection includes donated items such as personal archives, purchased items, and large numbers of items created by digitising the Library's "analogue" collection such as manuscripts, books, newspapers and sound.

Digitisation helps to preserve the originals by providing a surrogate and thus reducing handling of the original, but also enables increased access to the collection. The Library has been engaged in several mass-digitisation projects, for example 19th Century Books, which digitised 65,000 books, and the British Newspaper Archive project[3], which is expected to digitise up to 40 million newspaper pages.

Figure 12 shows the projected increase in digital object storage volume per node over the coming years. By 2020 it is expected that there will be between 4.5 and 5.0 Petabytes per node, totalling around 20 Petabytes for all four nodes.

[2] http://www.bl.uk

[3] http://www.britishnewspaperarchive.co.uk/

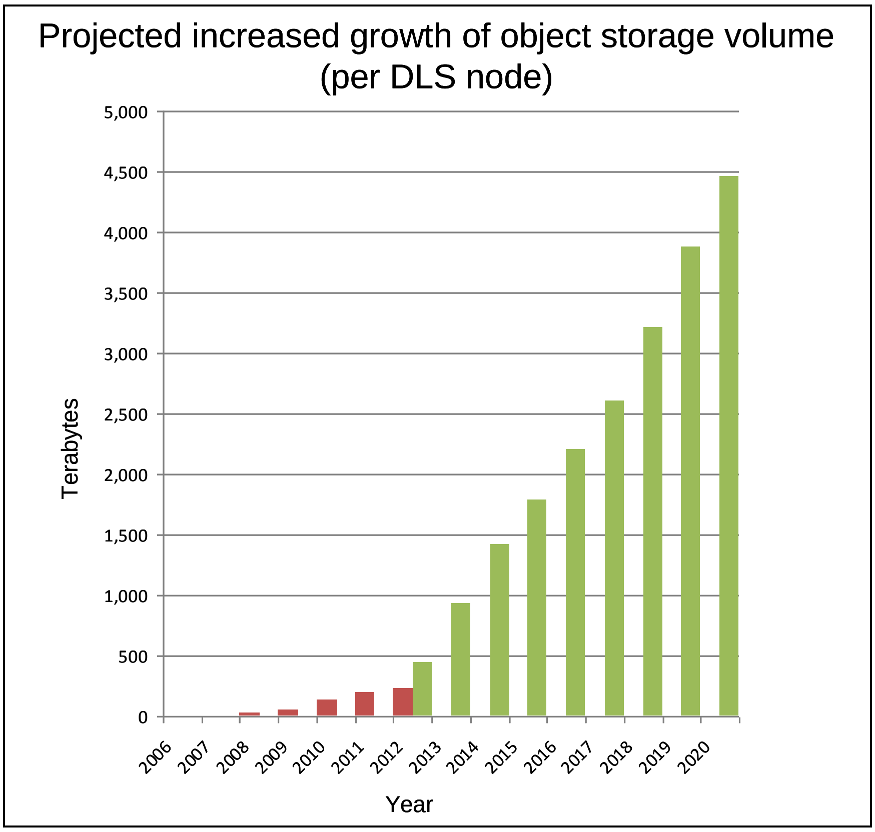


Figure 12: Projected increased growth of object storage volume

The digital objects in the DLS are in a wide variety of file formats. Some of these are container formats such as zip or gz, which contain other files of various formats. For some ingest streams the file formats are controlled, for example from internal digitisation projects. For others the Library has no control, such as with voluntary deposit or web archives. Table 5gives examples of the file formats held:

Example Digital Object File Types
xlsx
xls
xml
jp2
gz
wav
jpg
pdf
tif
zip
doc

Table 5: Supported file formats

However, as mentioned, some of these files are containers. The zip files currently mainly contain jp2 and xml files and are generally from digitisation ingest streams. The gz files are generally web archive files, and although their contents are dominated by htm and gif file formats, there could be any type of file that can be part of a web page. The complexity of the objects and the range of formats held therefore have the potential to be huge. The current distribution of digital object sizes is shown in Figure 13.

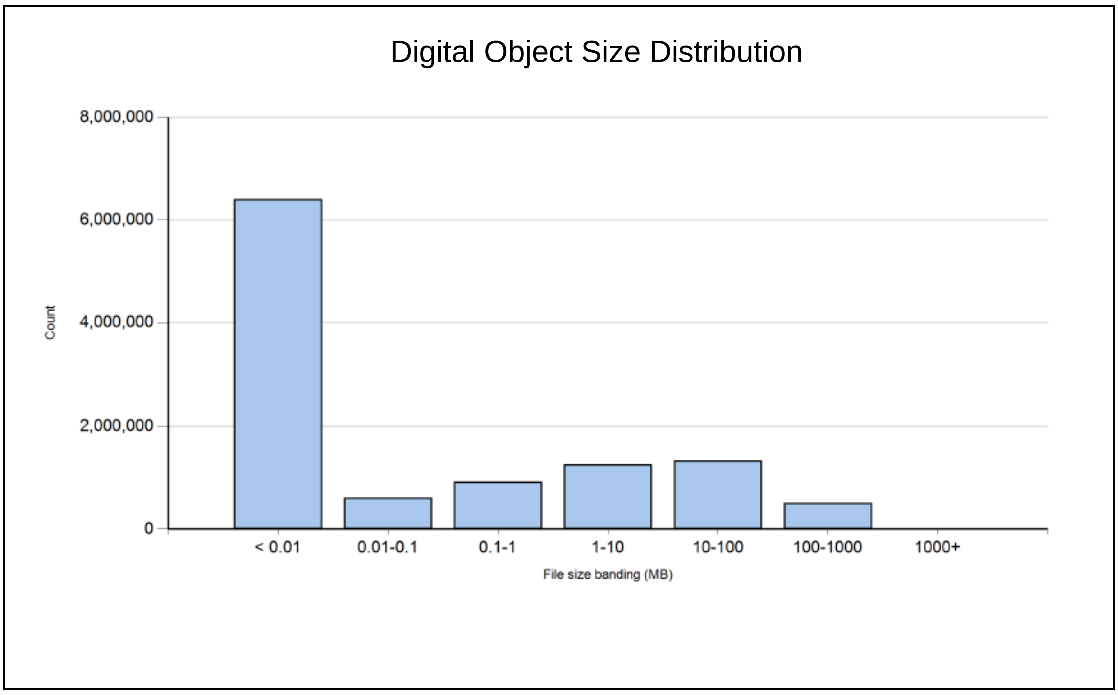


Figure 13: Current distribution of digital object sizes

Most of the very small objects are individual e-journal articles. The size distribution will change over time as the number of web archives in the collection increases with legal deposit, because these are typically around 1 GB per object.

2.4.3 Architectural overview of the storage solution

The vision behind the Digital Library System is to have shared technical infrastructure for non-print legal deposit in the UK. The UK Legal Deposit Libraries share the costs and share the services.

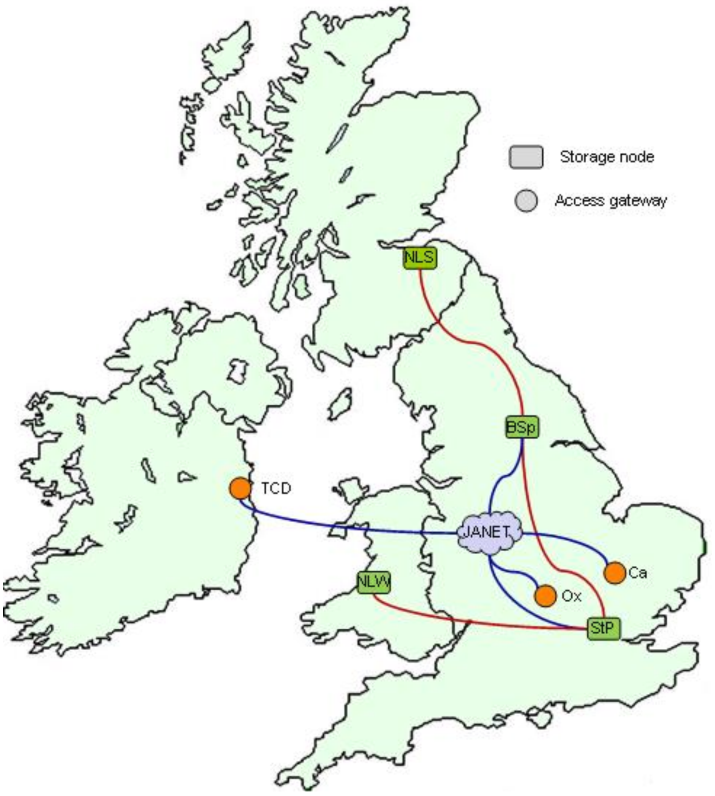
Storage nodes: British Library, St Pancras (STP) British Library, Boston Spa (BSP) National Library of Wales (NLW) National Library of Scotland (NLS) **Access gateways:** Bodleian Library, Oxford (Ox) Cambridge University Library (Ca) Trinity College Library, Dublin (TCD)	 Storage node Access gateway NLS BSp TCD JANET NLW Ca Ox StP

Figure 14: Storage nodes

Four national centres hold full copies of the system. These are the Storage nodes depicted in the Figure. The British Library has two nodes, one at its St Pancras site in London and the other at its Boston Spa site in Yorkshire. The university legal deposit libraries are entitled to legal deposit but they don't hold full copies of the system. They access the content from another storage node across Janet[4], the UK's research and education network.

The diagram below gives an overview of the architecture of the Digital Library System. It shows the four storage nodes, connected via an inter-site virtual private network (VPN).

Ingest of digital objects to the DLS takes place using ingest systems at either the BL's St Pancras or Boston Spa sites. There is a variety of ingest streams at each site; for example the BSP systems deal with e-journals, voluntary deposit items, digitised newspapers ,web archive content, and others. During the ingest process, each digital object is assigned a storage identifier called a DOM id. This is written into the METS metadata file that accompanies each object. Metadata about objects is held in the Metadata database.

Each digital object has a signature file, created during ingest by a Digital Signing Engine, or Signer (DSE in diagram). The signature file contains a hash value for the object, a secure timestamp and uses certificated cryptography.

The ingested digital objects, with their signatures, are stored at the local DLS node. (These digital objects are represented by the yellow "Preserved Digital Objects" stores in the **Figure 15**. They are then replicated at the other three nodes. Each of these yellow stores are equal in terms of the actual objects that they hold, but the physical storage servers and storage volumes may not necessarily be identical. (In fact, by design, they are not all identical.) MDB in the diagram represents the Management Information Database which holds information about the ingested objects.

[4] https://www.ja.net/

Each node verifies the signature of each object on receipt (either for ingest or replication). The signature file can be used to detect corruption or tampering in objects and every object is checked on a regular basis. If any defect is found, automatic recovery from another node takes place.

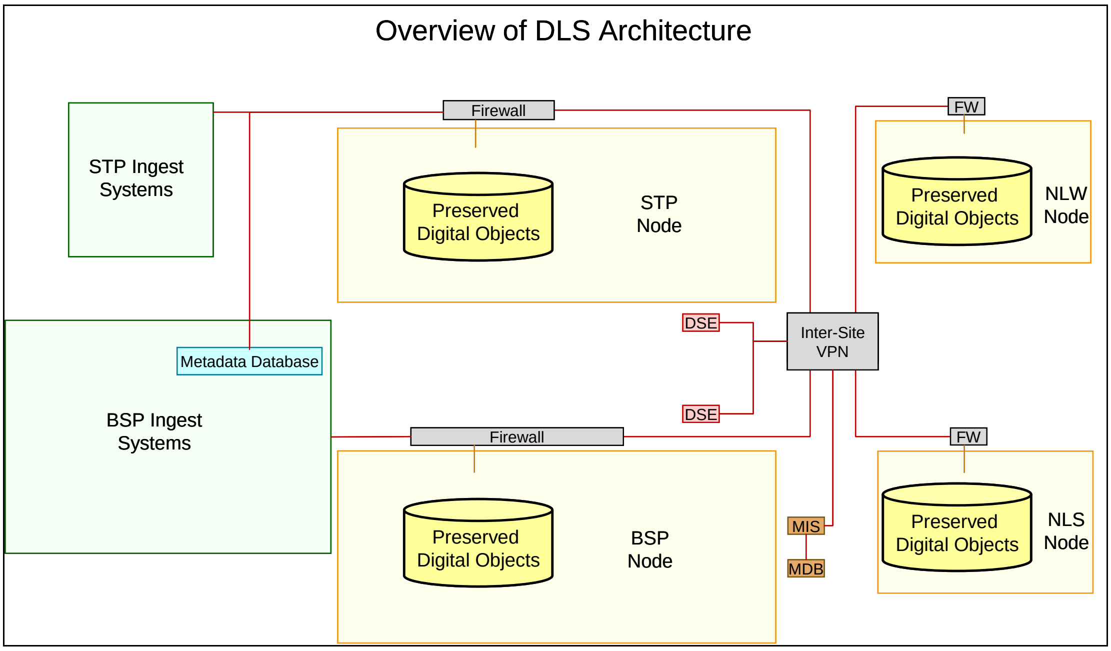


Figure 15: Overview of DLS architecture

Access to the digital objects is via an Access Gateway at the firewall of an individual node. The legal deposit libraries will usually access their nearest node, but resilience is built in to the system so that in the event of a disaster at one node, another node will be able to provide services (in terms of functionality, if not performance) whilst restoration is taking place.

The diagram below gives an overview of storage within the DLS. The ingest systems and DLS use thin-provisioned, virtualised storage for temporary storage.

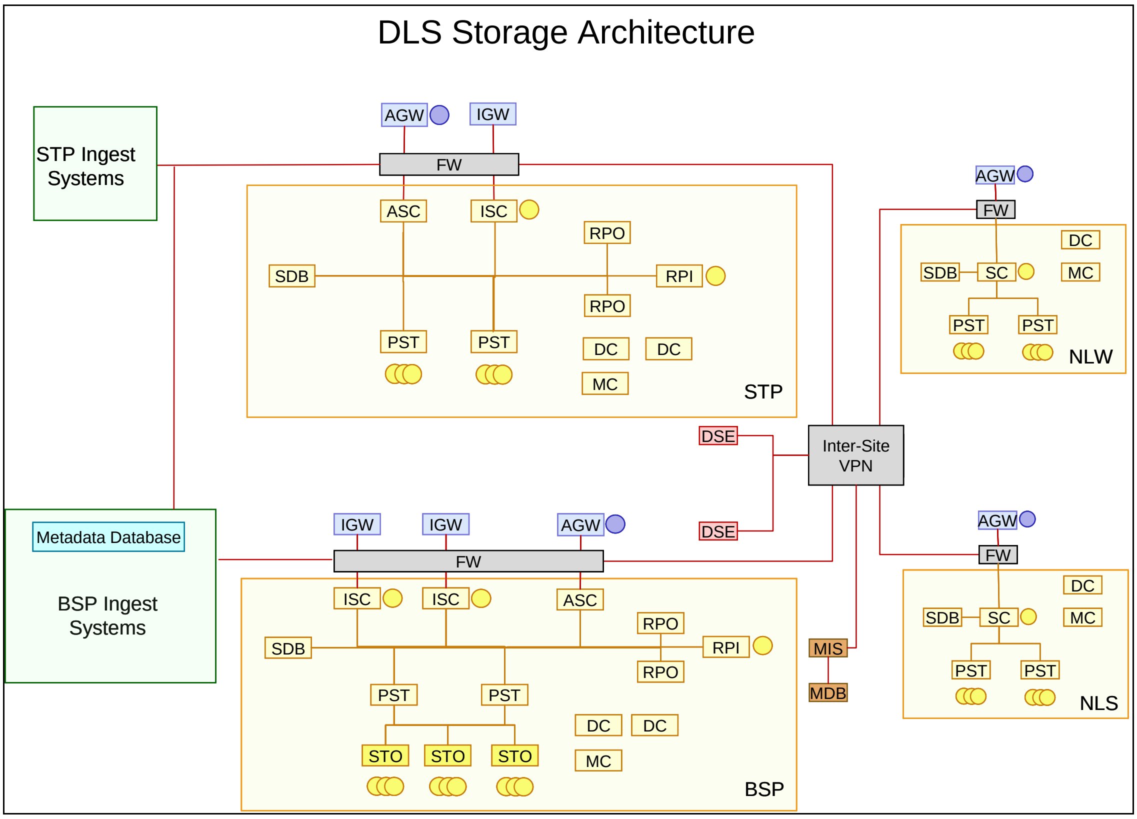


Figure 16: DLS Storage Architecture

In the DLS Storage Architecture diagram:

🔵	=	Access cache storage for Access Gateway
🟡	=	(Beside ISC, SC, RPI) Temporary storage, each could be up to 2Tb at present but thin-provisioned so not always fully utilised. (Under PST, STO) Storage for preserved digital objects. Scalable as digital collection grows.
AGW	=	Access Gateway
ASC	=	Access Site Controller
DC	=	Domain Controller
DSE	=	Digital Signing Engine
FW	=	Firewall
IGW	=	Ingest Gateway
ISC	=	Ingest Site Controller
MC	=	Management Controller
MDB	=	Management Information Database (currently occupies less than 100Gb)
MIS	=	Management Information Server

PST	=	Preservation Store Server
RPI	=	Replication In
RPO	=	Replication Out
SC	=	Site Controller (both Ingest and Access)
SDB	=	Site Database The site database tracks the managed objects and their location on disks. The location will change over time, as disks go out of warranty and data is migrated to new volumes. SDB maintains a mapping from the DOM id to the physical path on disk as it is made available over a network share.
STO	=	Linux Storage Server
VPN	=	Virtual Private Network

Table 6: DLS Storage Architecture diagram acronyms

2.4.4 Requirements for third party storage provider

There are various principles guiding the design of the Digital Library System, as follows:

Security

- Independent system nodes.

  This means that there is both a geographical separation between the nodes and a technical separation. Administrators at one node can only access that node. There is deliberately different hardware and operating systems between nodes to minimise the risk of common mode failure.

- Layered system security.

  There is a full separation between the public-facing systems and the private secure systems. Access to digital objects is only provided via Access gateways.

- Provable authenticity of digital objects.

  Digital objects are check summed during ingest and have signature files.

Preservation

- Assumed that bit-loss in inevitable.

  Ingested digital objects are replicated across all four storage nodes.

- Never be down to last copy

  Objects/data can be restored from other nodes.

- The system is self-checking and self-healing. There is periodic checking of objects' integrity, and if an object is found to be damaged, it is replaced by a good copy from another node. Preserve metadata.

  Preserves meaning and context of the digital object in addition to the object itself.

Scalability

- Infinitely scalable.

  The digital objects are stored "forever" and the collection will continue to grow.

- Extensible.

  The system is able to support new services and requirements as they emerge, without major re-engineering.

Storage

- Vendor independence.

  The system is for the long-term and it is important to not be tied into any particular vendors. It is designed to support heterogeneous storage sub-systems.

- Commodity hardware.

  Replication of objects/data means that high-end enterprise storage is unnecessary and so less expensive storage may be used.

- Purchased as needed.

  Since the system is scalable, additional storage may be added as required, as the collection grows.

Software

- Transactional and re-entrant.

  Assume that it is business as usual to have things go wrong. The system handles failure gracefully so that, for example, every stage of the ingest of an object is completed before the object is treated as preserved, and the system can "continue where it left off" when problems are resolved.

2.5 DEUTSCHE NATIONALBIBLIOTHEK

2.5.1 Introduction

As the central archival library for the Federal Republic of Germany, DNB has a legal mandate to collect and permanently archive German and German language publications including online materials. Although DNB collects some research data (e.g. as appendix to doctoral theses), such data is rather not the main focus of DNB. Consequently, DNB's policies are designed for (legal deposit) publications, and do not necessarily address data specific aspects.

The DNB hosts also the German Music Archive[5], which has the legal mandate to undertake the state-wide central collection of sheet music and sound recordings and serves as the music bibliography information centre for German.

DNB takes care that all digital publications can be utilized in accordance with legal regulations. Depending on the rights that the content producer grants us during the submission process, some publications can be provided in-house only, while others are remotely accessible.

The data which are deposited by the DNB are mainly "nonphysical media works" (online / electronic publications). Format and way of delivery is determined in the Law regarding the German National Library (DNBG[6]), the Legal Deposit Regulation and in accordance with the Law on copyright and related rights of legal protection.

All registered DNB users and deliverers accept a data protection statement.

Deliverers must confirm that they are entitled to deliver the publication as a deposit copy to DNB with the rights necessary for the legal mandate. Beyond the exceptions for use of the German Copyright Act the right holder may grant following rights: a) access for registered users via internet, b) unlimited access.

According to its legal mandate the DNB takes preservation actions like migration on archived publications. Where the law, e.g. Copyright Act, inhibits DNB thereby, agreement with the relevant right holders is sought. As for instance there is an Agreement with the German Publishers Association and the German Association of the Phonographic Industry for sound and image-sound data to remove technical measures for the protection of publications.

However, given the masses of delivery of deposit copies the DNB is not obliged to scrutinize all media on violations of data and personal privacy or other criminal deeds. As soon as an offence comes to our knowledge we do all to restore compliance with legal regulations and contracts.

Such violations may give reason for restricted access of the publication; however, the deposit copy will not be deleted. Furthermore DNB leads nestor[7], the German network of expertise in digital preservation, and engages in several national and international research activities and projects to promote progress in digital preservation solutions and enhance its own capability.

2.5.2 Data types and volumes

Since 2006, the legal mandate (see Sec 2.5.1) includes online publications. So far, DNB has focused mostly on static net publications that are collected mainly in PDF or, recently, in ePub format. We accept and archive any PDF version that is submitted to us. Also we advise our delivers to use PDF/A if it is possible.

[5] http://www.dnb.de/EN/DMA/dma_node.html

[6] http://www.dnb.de/EN/Wir/Recht/recht_node.html

[7] http://www.langzeitarchivierung.de/eng/index.htm

We are in the process of gradually digitizing all audio CDs of the German Music Archive since 1983 by using the uncompressed Broadcast Wave format.

Data types

1. Doctoral theses and teaching theses of German universities
2. DNB digitized print media
3. eBooks
4. eJournals
5. ePapers
6. Digitized audio CDs
7. Perspectively: digitized media of other institutions
8. Perspectively: websites

Table 7 gives a short current overview about the data volume and its annual increase for the most common data types.

Type	Annual quantity[8]	Total stock	Total size in TB	Average size in MB
Theses	12.980 titles	107.110 titles	~ 2,4	22,8/object
eBooks	53.000 titles	174.000 titles	~ 8,3	50/object
ePapers	90.000 editions	133.638 editions	~ 2,7	21/object
eJournals	3.119 magazines/ articles	15.786 magazines / articles	~ 0,75	50/object
DNB digitized print media	300.000 pages	~ 3 million pages	~ 34	11,5/page
Digitized audio	~ 100.000 CDs	450.000 CDs	~ 150	355/CD

Table 7: Data volume and annual increase

The table shows that roughly 200 TB of data is split in categories with high divergences in storage size and object quantity.

[8] The quantities of theses, eBooks, ePapers and eJournals are figures from 2011

2.5.3 Architectural overview of the storage solution

DNB's data management is based on in-house data storage (SONAS[9]), which serves access, and on external data storage (DIAS), run by a third party provider, for archival purposes. The in-house data storage and the outsourced data storage constitute de facto mirror sites for all content data.

Figure 17 gives an overview of the long term preservation (LTP) infrastructure of the DNB and shows that this infrastructure overlaps with the net publication (NP) infrastructure.

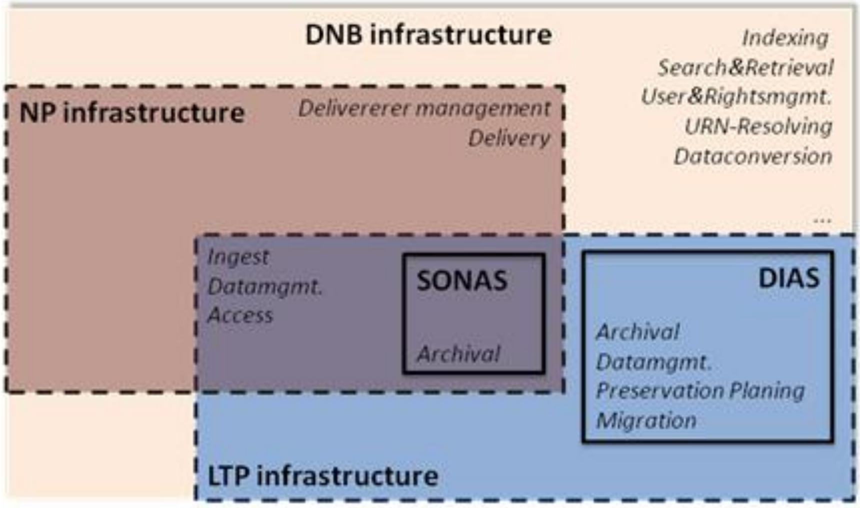


Figure 17: LTP infrastructure of DNB

The in-house data storage SONAS is highly scalable and similar to a file system. It provides a redundant data storage that also uses the IBM Tivoli Storage Manager (see description below). The main difference between SONAS and DIAS is that SONAS is mainly responsible for the provision of the collected objects in a current usable format. In detail SONAS is the storage backend of the DNB repository infrastructure and thus it is optimized for the daily access by the users of the DNB. DIAS by contrast is designed as a dark archive, which means that it holds archival copies of the SONAS access copies. DIAS is responsible for performing the long term preservation operations. If SONAS suffers data loss, or preservation actions like format migration are performed, SONAS is updated with the archival copies respectively migrated versions from DIAS.

From a more technical point of view, both DNB and the named provider use a RAID array for data storage. The hardware in use is regularly checked and monitored. Policies for managing data storage are in place. They envisage maintenance and detection of data integrity and backups in separate locations. The service provider additionally provides cost controlling, and regular reporting.

A data recovery policy is in place. If necessary, DNB and the provider can recover the status of (maximum) 24 hours ago.

In terms of security, objects delivered to DNB do not go directly into the in-house data storage, but are stored in a separate storage area before they are processed by an Import Service and transferred into the in-house data storage. Only registered DNB staff can administer the DNB in-house storage. The provider's data storage can only be accessed by DNB as the contracting body. The access is secured and encrypted.

The Architecture of the in-house data storage SONAS in more detail

[9] http://www-03.ibm.com/systems/storage/network/sonas/

IBMs high scalable storage solution SONAS (Scale Out Network Attached Storage) is designed to deliver a usable capacity of storage up to 14.8 petabyte (PB). This amount is reachable by operating up to 7200 hard disks with a capacity of 3 TB per disk in one interconnected system.

Information about the stored data are recorded separately in a data management system that gives answers to question like which files are rarely used or which files were altered since the last backup. Depending on the selection of fast but small or quick but large hard disks, SONAS supports the implementation of a hierarchical storage management (HSM) that also works with a connected Tivoli Storage Manager. The replication of data is possible within one SONAS system, but also between two systems via a WAN backbone for example.

Based on a base rack the SONAS system could be expanded with extension racks up to 20 racks. The base rack is offered in three variants from which the “2851-RXA, Configuration 3, feature code 9005” variant was installed in the data centre of the DNB. Figure 18 shows the components of this base rack. All mandatory components and one second storage controller are currently installed and used.

In particular, this includes:

- 2 Ethernet 1 gigabit (Gb) switches each with 50 ports for the component management
- 2 infiniband switches each with 36 Ports for internal data traffic
- 1 management node
- 2 interface nodes for external data traffic (internet) with two 1Gb and four 10 Gb connections
- 2 storage nodes
- 2 storage controller each with 60 hard disks of type SAS or Nearline SAS

This rack is expandable by two Disk Storage Expansion Units each with up to 60 hard disks. In addition, four further Interface Nodes can be accommodated.

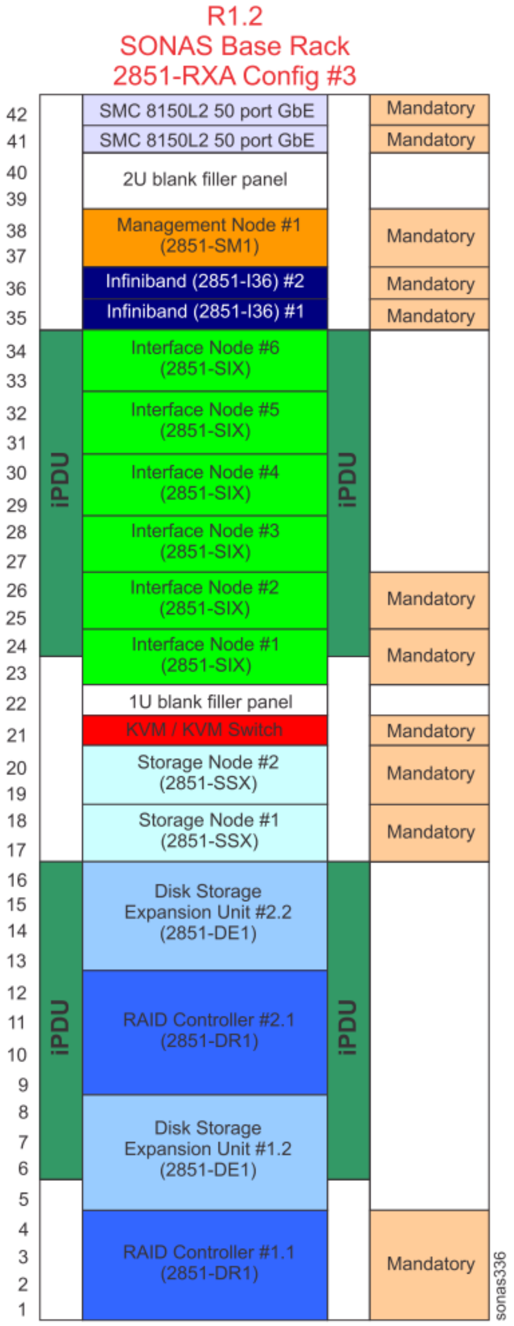

R1.2
SONAS Base Rack
2851-RXA Config #3
SMC 8150L2 50 port GbE
SMC 8150L2 50 port GbE
2U blank filler panel
Management Node #1 (2851-SM1)
Infiniband (2851-I36) #2
Infiniband (2851-I36) #1
Interface Node #6 (2851-SIX)
Interface Node #5 (2851-SIX)
Interface Node #4 (2851-SIX)
Interface Node #3 (2851-SIX)
Interface Node #2 (2851-SIX)
Interface Node #1 (2851-SIX)
1U blank filler panel
KVM / KVM Switch
Storage Node #2 (2851-SSX)
Storage Node #1 (2851-SSX)
Disk Storage Expansion Unit #2.2 (2851-DE1)
RAID Controller #2.1 (2851-DR1)
Disk Storage Expansion Unit #1.2 (2851-DE1)
RAID Controller #1.1 (2851-DR1)
iPDU
iPDU
iPDU
iPDU
Mandatory
42
41
40
39
38
37
36
35
34
33
32
31
30
29
28
27
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
sonas336

Figure 18: SONAS Base Rack[10]

The Architecture of DIAS in more detail

The following paragraph gives an insight into the design of DIAS, the IBM solution for an OAIS-compliant long term archive system, which is used by DNB. Figure 19 shows that the architecture of DIAS consists of numerous types of components. It also shows that DNB transfers its packed Submission Information Packages (SIPs) via the internet to DIAS.

[10] http://pic.dhe.ibm.com/infocenter/sonasic/sonas1ic/index.jsp

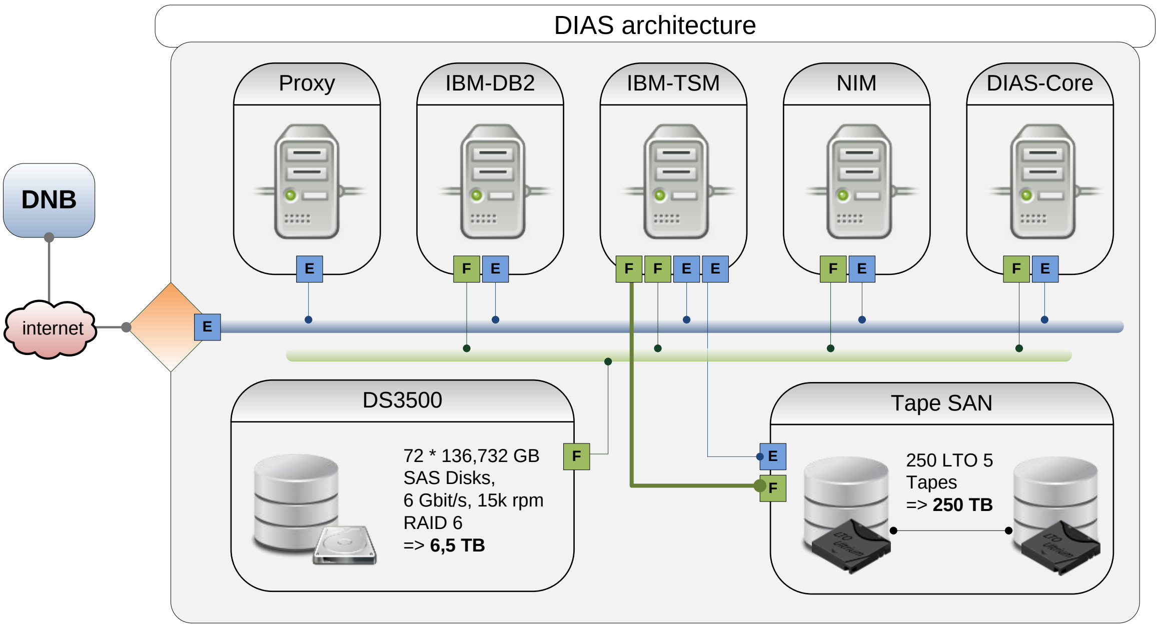


Figure 19: DIAS architecture

All DIAS components are hosted by the third party provider named GWDG in Göttingen. The GWDG is a corporate facility of the Georg-August University of Göttingen and the Max-Planck-Gesellschaft. It serves the purpose of a data processing and IT competence centre for the Max Planck Institutes and a data processing service centre for the University of Göttingen.[11] Göttingen and Frankfurt am Main (where the datacentre of DNB is located) are separated by a distance of about 230 kilometres. Thereby, DNB is able to store its data in geographically different places, which is considered to be an important requirement for a reliable long term archive.

In the following, the several DIAS components are briefly described.

Proxy

The proxy server is responsible for assigning DIAS-service requests to the corresponding server. In detail service addresses (URLs) are mapped to IP addresses and port numbers. Some additional services like monitoring software are hosted on this server, too.

IBM-DB2

IBM-DB2 hosts the IBM DB2 databases, which hold all internal administrative data and provide the OAIS-compliant *Metadata Management*.

IBM Tivoli Storage Management (TSM)[12]

The IBM Tivoli Storage Manager (TSM) is a centralized, automated data protection to help reduce the risks associated with data loss. The TSM automates data backup and restore functions, and itcentralizes storage management operations. In the present design the TSM is directly linked to disk and tape storage via fibre channel.

[11] http://www.gwdg.de/index.php?id=ueber_uns&L=1

[12] http://publib.boulder.ibm.com/infocenter/tivihelp/v1r1/index.jsp

NIM – Network Installation Management

The Network Installation Management Server gives the system administrator the capability to maintain and administrate the DIAS components. For example the NIM provides an interface to install new software or updates.

DIAS-Core

The DIAS Core hosts the essential DIAS software to run workflows (ingest, retrieval) and process search requests to the metadata management. Here, a number of steps like unpacking, analysing and validation the SIP need to be executed during the ingest process.

DS3500 – Data storage

Although every component has own disk storage for its operating system, the processing objects are stored in a separated disk storage. The disks are combined to several storage areas (LUNs), which are able to allocate to different tasks. For example there is one 500GB LUN for processing the date during the ingest process. The whole DS3500 is RAID-6 configured and extendable.

Tape SAN

The extendable storage area network (SAN), in 2-3 km distance from the GWDG main datacentre, hosts a library of LTO-5 Tapes. Currently the Tapes redundantly store 250 TB.

Internet Connection

A conventional internet connection and the secure file transport protocol (SFTP) are currently used to deliver SIPs to DIAS.

The following illustration (Figure 20: Ingest-Workflow) shows how DIAS stores digital objects across the different named storage technologies.

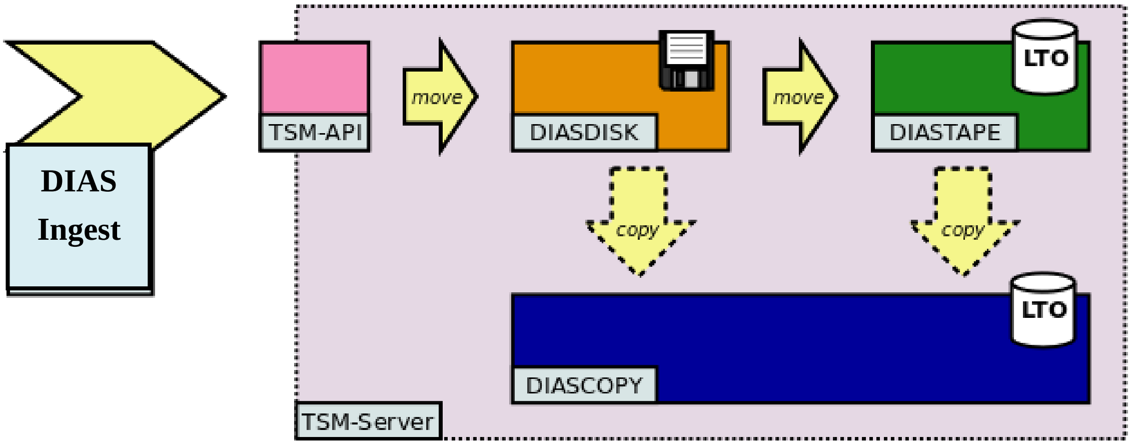


Figure 20: Ingest-Workflow

After DNB delivers a new SIP via a SFTP connection, the DIAS-Core performs the ingest process. This process ends by giving a newly built Archival Information Package (AIP) to the TSM. The TSM moves the data to the disk storage (DIASDISK) first. DIASDISK contains a 500 GB LUN that is provided by DS3500. When DIASDISK begins to reach its maximum capacity, the TSM initialize the migration of data to DIASTAPE, which is one node of the TapeSAN library.

In addition the TSM schedule copying new data from DIASDISK and DIASTAPE to DIASCOPY - the second node of the TapeSAN library. This mechanism ensures that each object is saved twice.

All components except the disk storage are connected via Ethernet. The actual content data is transferred by using a fibre channel connection with a bandwidth up to 8 Gbit/s.

2.5.4 Requirements for third party storage provider

The cooperation between DNB and its third party storage provider GWDG led to the result that it would be useful to extend the current contract with very precise regulations for purposes related to long term preservation. Therefore, aspects related to long term storage, which are parts of the DNB long term preservation policy, are going to be a part of the next underpinning contracts between the DNB and the GWDG.

The following paragraph will expose some of these key aspects that the DNB considers as very important. These points should be interpreted as a proposal for concluding contracts in this context.

Bitstream Preservation

The DNB defines bitstream preservation as the protection of loss of data integrity. It is a requirement across all steps of the digital preservation processes ingest, curation and access – rather everywhere the data is touched and processed.

The bitstream preservation should address the following issues:

- random changes (by machines or humans),
- data carrier errors and system failures,
- corrupted backups,
- intentional manipulations by a 3d party (hacker, virus infection),
- disaster like fire or air crash.

So a contract with a third party storage provider should contain reporting- and controlling arrangements that inform the customer about the data integrity regularly.

Concerning the **ingest** process the reporting should include:

- immediately feedback of incorrectly transferred packages,
- confirmation of accepted objects.

Concerning the **curation** activities the customer should get regular reports about:

- completeness of archived data,
- performed corrective actions,
- current corrupted files which are not corrected yet.

System integrity

Besides the data integrity it is also important that the provider is aware and reports on the integrity of the entire storage system and its components. For example the contract should define how long an exchange of a defective system component should be allowed to take. The spectrum here could range from the "on the fly"-exchange of a single hard drive or tape to the full restore of a crashed database server with all its configurations, data recovery and physical network connections and settings.

A regularly report should give time-related information about:

- utilization of bandwidth and data line,
- utilization of hardware (storage, CPU, …) and software resources (TSM, JVM…),
- maintenance activities (e.g. exchanged hard disks).

The provider should provide such information by using monitoring tools like Nagios[13] or Ganglia[14].

Recovery

According to quality assurance it's also recommended to consider tests for recovery scenarios. This should give answers to questions like:

- How long does a full recovery take?
- Is the recovered data valid?
- In what way is the data integrity testable?
- Is the recovery workflow efficient enough?

Availability

A further aspect in concluding contracts with a third party storage provider are regulations about service availability. In general this is defined as a part of the Service Level Agreements (SLA).

For example, the DNB expect a 24/7 service availability of 99 %. Thereby the service maximum downtime may not exceed 9 hours a year. It's also important to define a single point of contact and times for first, second and third level support.

2.5.5 Challenges

Besides the typical challenges like new file formats, DRM protection or dynamic content, the increasing data volume is one of the biggest challenges that the DNB is facing.

The reasons for the increase are various. For example, the number of objects to be archived is increasing, because publishers and other content producers publish their publications more and more digitally. Furthermore, highly resolved multimedia formats and digitalization imply a higher file size. The digitization of books and audio CDs as well as the web harvesting initiative push up the requirements on the infrastructure and storage.

In association with this, the time for processing a SIP and transmitting it via the internet to the 3rd party storage provider is longer than before. It turned out that the internet connection to the 3d party storage provider is a definitive bottle neck. Therefore the bandwidth must be appropriately dimensioned; otherwise the current data quantity is not handled. Also the retrieval and provision of archived objects (DIP) with big file size is costlier. Against this background DNB has to make regular efforts to keep its own storage technology and the out sourced DIAS components up to date.

[13] http://www.nagios.org

[14] http://ganglia.sourceforge.net

2.6 CENTRE INFORMATIQUE NATIONAL DE L'ENSEIGNEMENT SUPÉRIEUR

2.6.1 Introduction

CINES operates state of the art computer services in high-performance computing (Jade supercomputer ranked 24 in the Top500 as of June 2011), long term data and digital document preservation (CINES is the national centre for the archive of digital PhD theses from French universities and PAC, the CINES preservation platform, stores 15To of scientific data as of November 2011) as well as the hosting of computer equipment.

The centre provides researchers from universities and public research institutes with high performance parallel computing platforms. The staff of the HPC department at CINES makes hardware and software available to the users and provides specific assistance for research modelling.

2.6.2 Data types and volumes

Research domains are classified in thematic committees such as Environment, fluid dynamics, biomedical simulation, astrophysics and geophysics, quantum chemistry and molecular modelling, etc.

On a yearly basis, roughly 750TB of data are generated as part of the HPC research activity, which is the more than the maximum storage capacity available online at the moment on the CINES infrastructure which will be described below.

2.6.3 Architectural overview of the storage solution

The main HPC cluster – Jade – is a SGI Altix ICE 8200, a scalar, parallel supercomputer capable of a Rmax of 267 Tflops/s. Jade is made of 23040 cores spread on 2880 nodes (each of them being equipped with 2 processors Intel Quad-Core E5472 or X5560). Each node has 30 (resp.34 GB) of available memory, which is a total of 91 TB. The 46 racks of the cluster are connected to the 12 racks of the shared file system that has a capacity of 690 TB.

On Jade, the storage system (Narcisse) available is as follows: two disc volumes are available for scientific data, i.e. /scratch and /data, on a SGI ALTIX 450 (16 Montecito processors) with 64 GB memory, running Linux. Another disc space, work, which is fairly smaller, is only used by nodes (binaries, etc.). In addition to the online discs infrastructure, two libraries allow the storage of data on tapes. The /scratch repository is used for temporary data, whereas /data allows the storage of data in a longer term – in any case shorter than a couple of years.

On a daily basis, the data stored on /data are backed up on tapes. There are two tape libraries for security purposes, but the same information will be duplicated on both. The technologies of the two robots are slightly different: the first one (Jaguar, IBM 3592) is faster and more expensive than the second one (LTO4), which is rather used for redundancy and security purposes. The data is thus eventually available on disc as well as on tape.

The /data environment is generally filled at 95%. When a user needs to add more information and room has to be made, the data which has not been accessed recently is deleted (as long as they have been backed up on tape, of course).

The /data environment is made of five disk drives running in parallel, and has a capacity of 500TB. When data are requested again after their removal from the discs, they are restored from the tapes to the discs (average restore speed is 200Mb/s).

The online discs are based on SATA technology, with a disc capacity of 750GB and 1TB configured in RAID6, operated by Lustre, a massively parallel distributed file system commonly used for large scale cluster computing, and NFS as a transfer protocol.

The current architecture has a throughput limitation: Lustre limits the I/O at 200MB/s per node, and there is only one node in place at the moment to restore the data. But this could be improved at a better rate, up to 600MB/s, by implementing a new technology.

The overall architecture is described in Figure 21 here below:

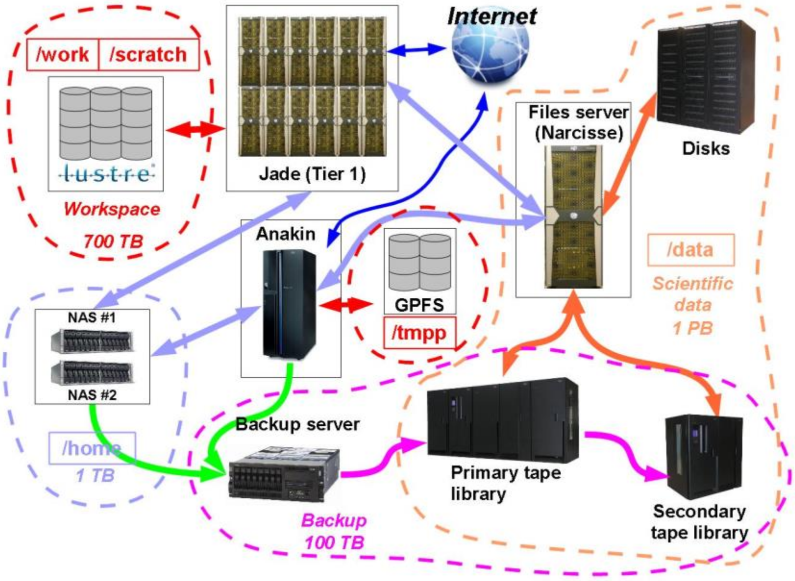
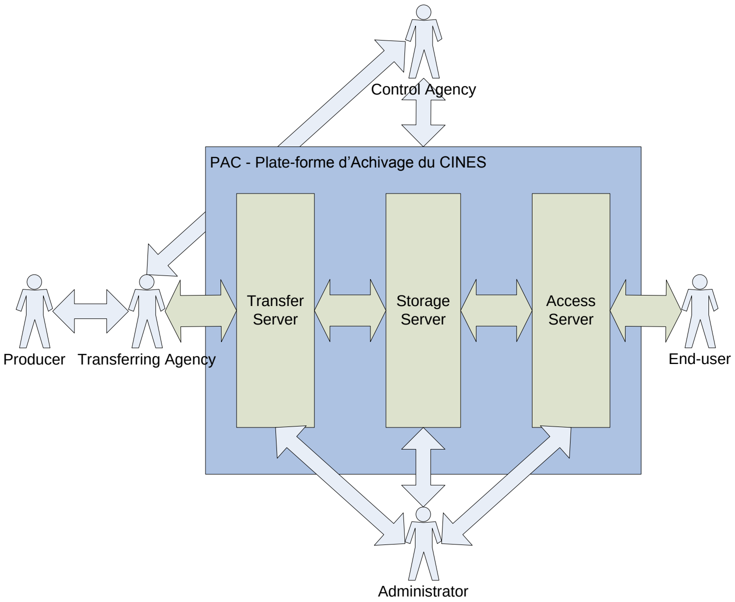


Figure 21: CINES Overall Architecture

In 2004, the French ministry for Higher Education and Research (MESR) assigned the CINES the mandate to assure and certify the long term preservation of the French electronic PhD theses, as well as the digitized publications of the Persée web portal, which is focused on social sciences and humanities. After three years of conception and development, a first version of the PAC system – acronym for “Plateforme d’Archivage au CINES”, or CINES Archive Platform, v1.0 – was deployed in spring 2007, with the objectives of proving the feasibility of the concept, validating the architecture and processes orientations and integrating the STAR application developed by the ABES (French national agency for university libraries) for the transfer, access, reference and archiving of electronic PhD theses.

A year later, the CINES decided to proceed to an “invitation to tender” for the purchase of a platform based on the industry standards, with high volume capabilities i.e. more than 40 TB. The PAC v 2.0 system has been implemented as a set of three logical servers, as defined in the OAIS model (ISO 14721, see below), and is now running under the same generic approach, thus avoiding specific developments when rolling new projects out, as they all share the same processes, infrastructure and storage platform :

- An ingest server, where the transferring agency can transfer the archives collected from the producer ;
- A storage server, where the archives are maintained ;
- An access server, where authorized users can search, consult and retrieve on-line the archives.



The PAC system is physically comprised of a hardware platform and its software layers.

- The hardware platform currently includes the transfer and storage servers (two redundant Oracle Fire X4150 application servers, two Oracle STK 6140-4G 40 TB storage units and a 2 PB IBM tape library shared with the HPC infrastructure); four copies are made of the archival packages, two being stored locally on disks, and the two others being stored on tape, one of which is actually done in another datacentre 400 kms away for disaster recovery purposes.

- The software layer is a specific Java application. Its components have been developed in order to be compliant as much as possible with quality standards. The archive file format validation and characterization is done by interfacing Jhove, the open source tool validation format developed by JSTOR and the Harvard University library, DROID, from the UK National Archives, ImageMagick, and several other tools (mPlayer, ODF validator, etc.) to the PAC software.

This solution is made available to the whole Higher Education and Research community, as long as the corpus to archive is acknowledged as national value/interest.

2.6.4 Challenges

The digital documents currently preserved at CINES are mainly literature, i.e. text, images, sounds and videos. But there is a strong focus on the raw data as well, and a project is currently being run – ISAAC (acronym for “Information Scientifique Archivée Au CINES” or Scientific Data Preserved at CINES), with the objective to offer preservation services for primary data to the research community. The infrastructure put in place as part of this project will be based on widely used technologies such as Java, iRODS, etc. The purpose is to allow researched to preserve the data they produce as part of their use of HPC systems (among which the CINES one) during a longer period than the year during which they have access to the datacentre storage facilities. This will help them to perform the necessary appraisal, and identify the data which has value and needs to be preserved, and by extension the data than can be deleted.

2.7 SAMENWERKENDE UNIVERSITAIRE REKEN FACILITEITEN / STICHTING ACADEMISCH REKENCENTRUM AMSTERDAM

2.7.1 Introduction

SURFsara HPC Central Archive (CA) is a mass storage facility for long-term preservation of large datasets produced in academic research projects that are of particular relevance to European and Dutch academic communities. The CA currently stores data that were produced on several incarnations of the Dutch supercomputing facility and various other academic HPC facilities at SURFsara. In addition, the SARA CA stores data of Dutch academic groups from European-funded projects such as DEISA and PRACE that were not necessarily produced at a SURFsara facility, but on tier-0 or tier-1 systems from a PRACE partner site.

2.7.2 Data types and volumes

The datasets stored at the CA are quite heterogeneous in size and retrieval patterns, while rare or occasional retrievals dominate. The archive facility was deployed as a dedicated archive in 1991, for the Dutch national supercomputer – then a CRAY YMP system. The notion of large datasets has considerably evolved since 1991, along with the implementation of the archive facility, but until 2007 the infrastructure remained dedicated and technically very tightly coupled to incarnations of the Dutch national supercomputer. Since 2007, the solution is indeed "central" in the sense that all HPC computers and visualization facilities at SURFsara can access the CA to store or retrieve data. When the Dutch supercomputer joined DEISA in 2004, it happened occasionally that datasets produced by Dutch groups in DECI projects running on an HPC system at another site were added to the SURFsara archive using the Dutch super computer as an intermediate step.

Since the advent of PRACE and the integration of tier-0 and tier-1 facilities across Europe into an HPC "ecosystem" the number projects with off-site produced data that will be archived at the SARA CA is expected to grow substantially. One aspect of the recent restructuring of the CA in 2011 was the already obvious need to better facilitate and streamline external access. A service node for external access, "archive.sara.nl", independent of any of the HPC facilities, was added. Users can access the CA directly from an offsite location, through a number of protocols, ssh/scp/sftp and GridFTP, to store and retrieve data, or to logically reorder or physically relocate already stored data. The number of methods for external access will be extended with project specific https-based interfaces and other suitable generic protocols.

The SURFsara CA still contains a considerable number of old datasets that are fairly small by today's standards next to more recent datasets produced in Dutch "Data Challenge" projects and DEISA/PRACE projects which have sizes ranging from 30 to 110 TB.

A large quantity of the data in the archive must be preserved for a long time but are accessed rather rarely. Although there are some exceptions, many data in the CA are not entered as datasets available to the general public, but remain the intellectual property of the research groups that produced them. Typically, and as an example, data are produced by a PHD student working on a particular project that fits in a program of a research group in the longer term.

In most cases, the biggest portion of the data is retrieved when an article by the PHD student is reviewed or published and some additional questions rise. Once answered, the data remain usually untouched until they are retrieved at the start of follow-up projects in research programs, and access rights have been arranged for one or more new PHD students of the group. The data or one of their subsets that is needed more frequently during the course of follow-up projects are generally staged onto the HPC compute platform where the follow-up project runs, and thus don't stay on-line in the CA during the course of the project.

Recently, a gradual shift in the population of archive users has been observed, from data-owners archiving their data essentially for themselves, their close collaborators and immediate successors, to data-owners archiving datasets for limited, moderated, public availability. The data retrieval is still expected to be fairly rare, though slightly more frequent than before, because a limited number independent parties - rather than one – have interest in the data. It is believed that the continuation of this trend will have a limited impact on storage strategies, but may have more consequences for the retrieval interfaces that can be offered.

The data volume of the SURFsara grid-enabled storage infrastructure is currently over 3 PB – with a substantial part of the CA.

2.7.3 Architectural overview of the storage solution

The current CA facility is based on a tiered storage environment built around a small cluster of seven relatively simple nodes using clustered XFS (CXFS) file systems. All data are ingested or accessed in the CA as files on the on-line disk storage of one these file systems. A parallel version of SGI's Data Migration Facility (DMF) software provides automated, complete, hierarchical storage management for these file systems, i.e. migration to and from other media, such as tapes or secondary disk storage.

CXFS is the SGI implementation of a SAN-shared file system. XVM, SGI volume manager software, and is used to create a single, aggregate, logical volume from multiple hardware RAID5 LUNs on the SAN, for each CXFS file system. Striping rather than concatenation of hardware LUNs is generally used as the method of aggregation, as striping increases significantly the performance of the file system. RAID5 or mirroring (RAID1) configuration options on the level of the XVM volume manager are not used, since RAID5 redundancy is already applied on the storage hardware level.

The diagram below shows a more detailed architectural overview of the current CA environment. All nodes of the CXFS cluster have multiple SAN fibres to the disk storage of the CXFS file systems. Some nodes of the CXFS cluster, labelled "cdm1" to "cdm3" in Figure 22, act exclusively as NFS servers exporting the CXFS file systems over several 10Gbit/s and 1Gbit/s network links to NFS clients, i.e. nodes of SURFsara HPC and visualization facilities as well as the external service node. Automated policies for hierarchical storage management are in place, but ultimately, the user knows best which files are required on-line and when. It is standard policy to give the end user control over this matter and to provide some tooling to overrule the effect of automated policies. However, it cannot be communicated over the NFS protocol whether a file in an exported file system has its data blocks offline or on-line. Therefore all NFS client hosts are also equipped with SURFsara rdmf ("remote DMF") client software, which brings elementary hierarchical storage management primitives, functionality to list and manipulate the on-line/off-line status of files, to the NFS client hosts.

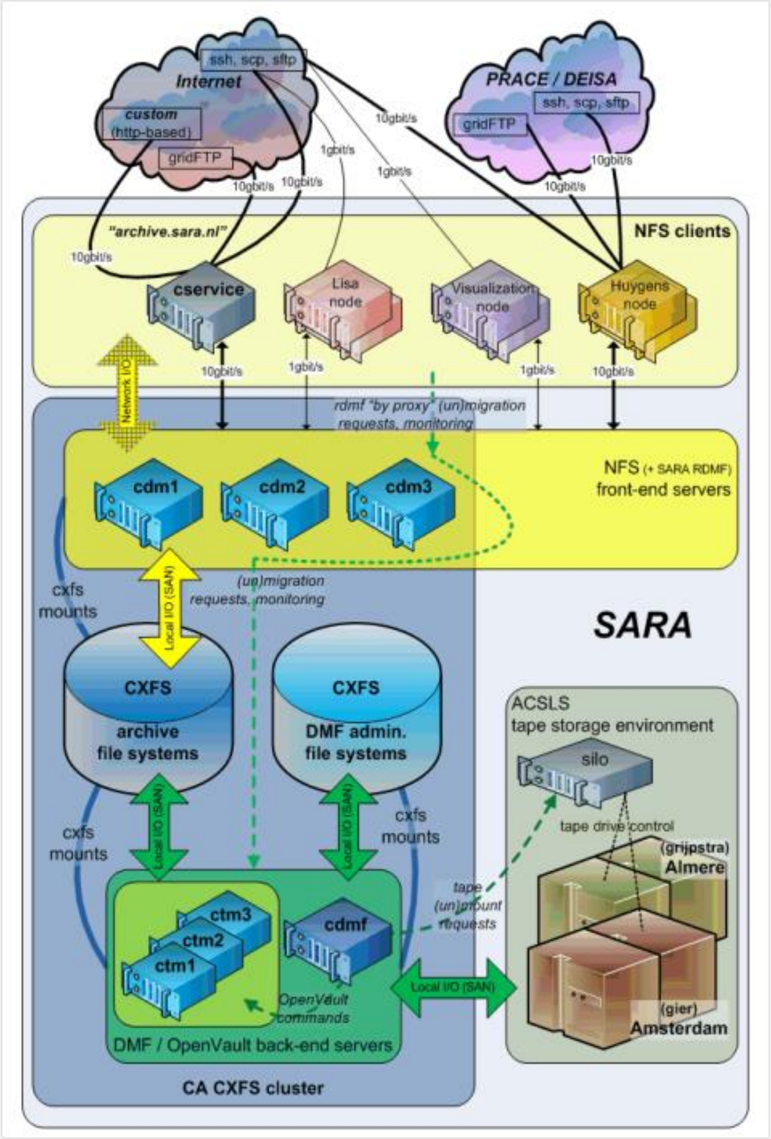


Figure 22: SARA Overall Architecture

Other nodes of the CXFS cluster act exclusively as "tape servers" – labelled "ctm1" –"ctm3" - and have 4Gbit/s SAN connections to a number of tape drives for migrating between the tape backend and the disk storage of the file systems. A single node – labelled "cdmf" - is meta-data server for the file systems and assumes various other orchestrating and communication roles.

The functionality provided by the seven nodes of the CXFS cluster can in principle be provided by a large monolithic server node that combines all functions. In previous incarnations of the CA this was actually the case. The combination of clustered XFS + parallel DMF together is used to implement a migration facility that is distributed and more scalable in several independent dimensions.

An important policy of the CA is to store dual copies of data on tapes in two distinct "volume groups" that are kept at geographically different datacentre locations, in order to ensure a high level of reliability and availability. The "local" SAN fabric, which is depicted in the picture above, connects the tape servers to the tape drives, and actually spans two datacentre locations that are about 35 kilometres away from each other. The tape drives are concentrated in two tape libraries: one is in Amsterdam, the other is in Almere. The core switches in the datacentres in Amsterdam and Almere are connected via long distance links over two independent dark fibres. The connection between the dark fibres and the rest of the SAN at each datacentre is based on coarse wavelength-division multiplexing (CWDM) technology. The SAN switches on either side effectively “see” four 4 Gbit/s fibre channel connections to the other side. Both in Amsterdam and in Almere the tape back-end is a StorageTek SL8500 Library, currently equipped with StorageTek T10KB drives.

The grid mass storage service also uses the same software technology but it has its own DMF instance, its own CXFS cluster, its own separately organized front-end with dedicated light paths to CERN, etc. The remote backup service is based on IBM TSM software, and thus organized very differently. Both services do not fall within the scope of this document. They should however be briefly noted here, because it must be granted that their sharing of the same tape back-end infrastructure has made for some flexibility and economics of scale effects that have been beneficial for the total cost of ownership of all services, including CA. It is for instance possible to temporarily shift a tape drive from the CA to the LHC tier-1 environment, or vice versa.

Long-term bytestream preservation and backup are two distinct services for users. But users expect that backups are available to prevent the CA from disasters. The dual copies on tape do not constitute a complete backup for situations where the file systems on disk and/or the DMF database are lost. Reliable backups are provided by exploiting some characteristics of the hierarchical storage management implementation and complementing it with the necessary metadata backups:

1. A daily backup of the DMF database
2. A daily “metadata only” full backup of all file systems with hierarchical storage management
3. A daily backup of those files that are stored in the archive but are either considered too small to migrate to dual tape copies or have not been migrated yet.

The DMF database and journals are backed up separately. A procedure has been developed to create a backup of items 2 and 3 above in a single run per file system. To keep the backup overhead small as well as the backup window short, the number of files to be included with the metadata must be kept at a minimum. Rather than migrating to tape only when on-line disk space becomes tight, the hierarchical storage management software is configured to scan frequently for files that are newly added to the archive. It schedules the new or updated files immediately for dual copying to tape, but without necessarily discarding the on disk copy when the instances on tape have been created. This ensures that data that have put in the archive indeed have a backup well within 24 hours. It also ensures that the overwhelming majority of new files in the archive have been pre-migrated before the complementary procedures run and that they can indeed be complemented with i-node only data. Once created, both datacentre locations store their own copies of the complementary DMF database backups and meta-data backups.

The “metadata only backup” of a file system is actively used as a tool that allows extensive restructuring of disk resources, i.e.: for splitting or merging file systems, and/or migrating to new hardware in a fairly small maintenance window, in addition to being a necessary backup complement,. The backup contains the bfids of all migrated files and the DMF database is structured in such a way that it is void of any relative or absolute path assumptions. Thus a “metadata only backup”, or one of its subsets, can be restored to any file system that is brought under DMF control. When the “metadata only restore” is complete, the file system is directly usable - although it will initially have a performance penalty, since 100% of its file data-blocks are initially still off-line. As soon as files are accessed, the “un-migration” starts.

2.7.4 Challenges

As mentioned before, a gradual shift to archiving of data that is kept for public availability, or at least for availability to several parties rather than one, has been observed. This implies that the archived data can also be retrieved by people that have had little or no prior contact with the original producer, that are not likely to have the opportunity to engage in informal communications about, which particular details are stored where, in what files, etc. With this shift, the flexible but informal file system interface, with its supposedly intuitive and descriptive tree-ordering, possibly augmented with an occasional “README file”, becomes less appropriate. The standardization of content catalogues and file formats, formalization of the description of the datasets and methods of indexing are thus required between these parties. On top of that, new access protocols are needed that maximize the use of such well-defined metadata structures, to implement efficient methods of remote pre-retrieval browsing, and subsequent partial retrievals.

The use of the platform-independent and self-descriptive NetCDF format, rather than just archiving the raw originally produced data files, is becoming more widespread to store and preserve files, particularly in the geophysical sciences community. This format allows efficient description and storing of data subsets within a file. One of the Dutch primary investigators in this field has also recently expressed an interest in OpenDAP as a new, complementary access protocol for the SARA CA, since this protocol directly facilitates the remote retrieval of subsets from files. A pilot enabling OpenDAP access to one of the recent geophysical datasets stemming from a PRACE tier-0 simulation will be set up.

2.8 COMPUTATION-BASED SCIENCE AND TECHNOLOGY RESEARCH CENTER

2.8.1 Introduction

CaSToRC is an initiative of the Cyprus Institute which objective is to rollout a Tier-1 HPC facility and adequate computing resources to enable Cyprus and the Eastern Mediterranean research community to pursue forefront computing-related research.

2.8.2 Data types and volumes

CaSToRC provides support for HPC and data intensive computing in fields such as climate and environmental modelling, fluid dynamics and heat transfer, high-energy and plasma physics, materials science, chemistry, 3D visualization, computational biology and financial and economic modelling. As of yet, the storage facility has a capacity up to 1,5PB spread over various storage resources and technologies. The classification of data determines the strategy under which it is kept. This classification is appraised at the time of granting access to systems (e.g. through project/preparatory access calls). While classification is constantly revised, the general classes of data fall under the following:

1. **“Scratch**” is data that are either needed only during job-execution, or data that are easy to re-generate or transfer from other locations while reusable; they are retained for the lifetime of the job unless special arrangements are made to accommodate users having many jobs with short inter-arrival times, or concurrent execution of jobs requiring the same data-sets ;

2. “**Short-retention**” is data that is planned to be kept between 6-24 months, or the lifetime of a project; they are retained for the duration of the project to which they were granted, with an added (small) grace period of access to the data. This includes preparatory access with a typical active period of around six months and project access with a typical lifetime of 12-24 months ;

3. “**Long-retention**” is data that is planned to be kept for periods longer than 24 months; they data are retained for longer periods of time and may not be associated with one specific project. This data comprises data-sets that are reusable across projects or belong to user communities that closely collaborate with CaSToRC ;

4. “**Essential Data**” a replica is maintained for redundancy purposes; data are flagged as such after agreement with the associated users. Since replicated data take up more space, the extra storage usage is accounted to the users. In the case that the data is flagged “Essential”, one copy is maintained in Tier1b storage and another one is maintained in Tier2 storage. Several other factors also affect whether a dataset is flagged as “Essential”, such as the volume of the data and frequency of updates. Out of the two replicas, one copy is designated as the “master” copy, and depending on the frequency of update of the dataset, as well as other project-specific reasons, the “master” copy is placed either in Tier1b or Tier2. Replication is automated via standard tools (cron, rsync) when conditions permit. “Long-retention” data is usually also classified as “Essential” unless it is prohibitively large.

5. “**Critical Data**” are low-volume data that are systematically and regularly backed up. They include *service-data*, such as data catalogues, and other important metadata.

It is important to note that these categories are not necessarily mutually exclusive and, in fact, they usually are not. For example, both Short and Long-retention data can be classified as “Essential” so that a replica is maintained.

2.8.3 Architectural overview of the storage solution

Storage at CaSToRC consists of a loosely tiered storage configuration, comprised of two main tiers: *Tier1 (a, b)* and *Tier2*. *Tier1* consists of the recently added high-performance GPFS storage over QDR Infiniband. Tier2 consists of fibre-channel based storage as well as network-attached storage belonging to specific user communities. The diagram in Figure 23 outlines the storage configuration and the associated interconnects.

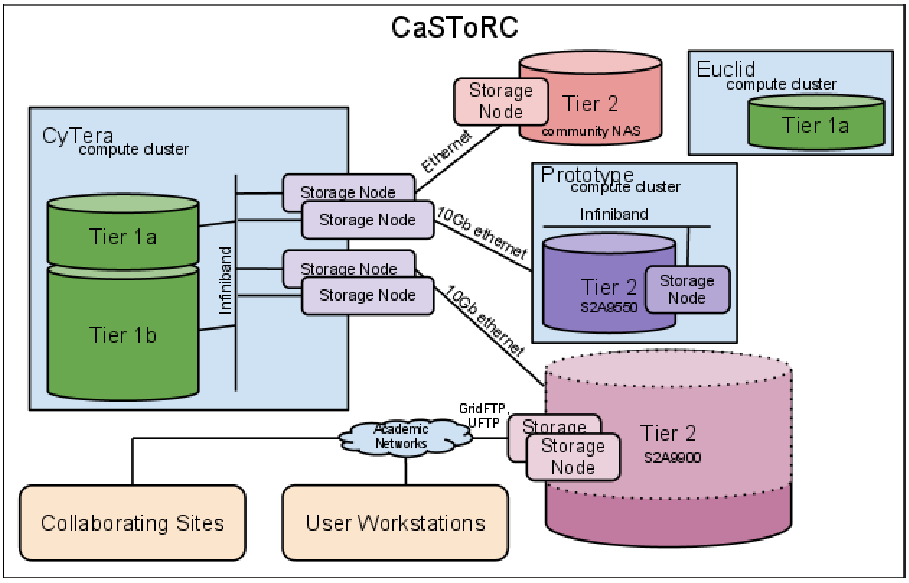


Figure 23: CaSToRC storage configuration

The storage resources at CaSToRC include the following:

- Tier 1(a,b): CyTERA Storage (Infiniband - GPFS - 300TB)
- Tier 2: DDN9900 Storage (FC - Lustre - 200TB - 1PB)
- Tier 2: DDN9550 Storage (FC - Lustre/NFS - 50TB)
- Tier 2: Smaller project-owned network-attached storage systems (NFS - varying capacities)

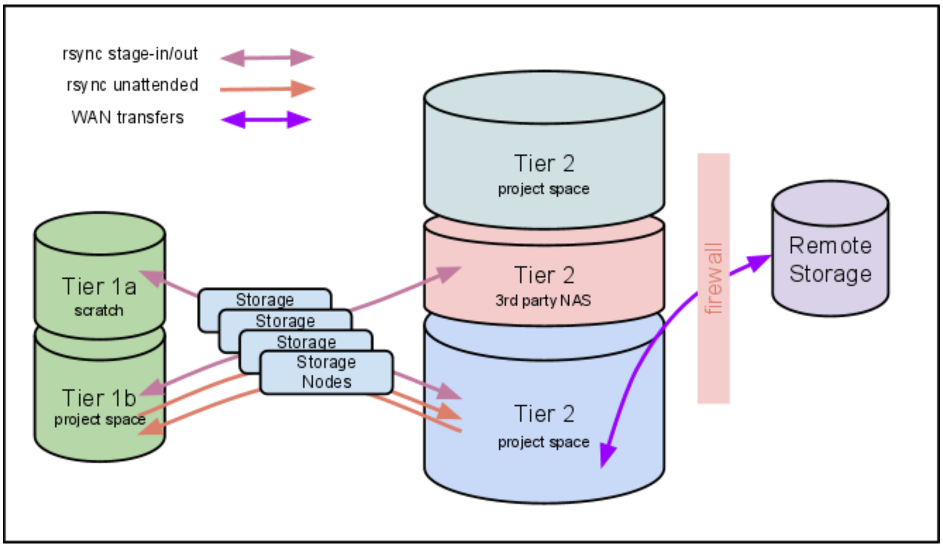


Figure 24: CaSToRC storage interconnection

Maintaining a balance between cost, flexibility and data safety, and taking into account the aforementioned classifications, CaSToRC follows a loosely-tiered storage policy. In contrast to integrated HSM systems, the transition of data between tiers is not fully automated or transparent.

Data transfer between tiers relies on either:

1. Job-associated automation scripts / work-flows, or
2. Interactive user commands, or
3. Semi-automated tools that handle replication for redundancy.

Storage is allocated to the tiers based on classification. “Scratch” is allocated on Tier1a. Tiers 1a and 1b are built of high-performance disks over high-performance interconnections and are directly connected to the CyTERA compute cluster. Tier 1a holds the user’s $SCRATCH directory but does not contain any $HOME, $WORK, $MASTER or $REPLICA directories. Tier1b is attached in the same way as Tier1a, i.e. directly to the compute cluster. Tier1b holds the user’s $HOME and $WORK directories and mostly short-retention data. Tier 2 holds many of the $REPLICA directories and data classified as long-retention.

Data that does not reside on Tier1 needs to be staged prior to execution. The data transfer is both triggered and monitored manually by users, performed using pre-processing scripts (e.g. Torque “prologue” scripts), or handled via work-flows/job-dependencies. After job completion the data is staged out via the same mechanisms.

CaSToRC provides helper scripts to assist the user in transferring files between Tier2 and Tier1a. Taking into account the size of the data, as well as other factors, users can employ some of the helper-scripts in their jobs or job-descriptions in order to handle the staging-in and staging-out of data.

WAN transfer mechanisms are becoming more and more important as resources attract users from remote locations. The main requirements for performance, security and even integrity are handled by widely-used Grid technologies such as GridFTP and UFTP. Via the employment of the UNICORE middleware, these wide-area transfers can be integrated into workflows to facilitate automation. As shown in figure 11, WAN transfers need to traverse firewalls and provide security and tools such as GridFTP (and globusonline.org which is under evaluation) make this possible and even user-friendly.

Data that originates from remote locations is initially stored on Tier2 disks, due to its easy accessibility from the middleware servers over NFS. The data are then relocated if necessary to their final destination.

While network bandwidth really needs to be handled as a finite resource (and is quite restricted at times), it is currently dealt with on a best-effort basis. The fact that bandwidth is shared with other non HPC-related traffic also presents problems with scheduling transfers.

2.8.4 Challenges

Longer-term preservation of data presents several serious challenges, such as the obsolescence of hardware, media faults and disasters. Further challenges, but just as important, include the disassociation of data from metadata as well as human factors (deliberate and otherwise). Besides the technical issues, there is also the issue of cost of ownership and operation. In order to provide long-term preservation, all of these issues and potentially several others need to be effectively addressed. To address these issues it is imperative that specific documented policies are put in place. While it is important for an HPC facility to have documented policies and procedures, the long-term nature of data-retention makes it all the more necessary.

For example, in order to tackle obsolescence of media and hardware, mid-to-long-term media migration plans need to be put in place, that take into account data and hardware life-times, availability of replacement parts for disk storage systems etc. The *initially* low-volume of data to be preserved and the online requirements of core applications at CaSToRC (such as Digital Cultural Heritage, and Climate Modelling) that require longer-term storage, have not raised demand of a tape system (which could potentially cause a drop in the overall cost or retaining data). CaSToRC maintains a basic high-level “data-catalogue” of projects, the projects’ datasets, their data-lifetime, storage allocation(s) and storage locations (i.e. which system the data reside on). This “data-catalogue” is routinely maintained by CaSToRC staff whenever data is migrated. This data-catalogue also keeps limited information on the different systems’ lifetimes (e.g. product end-of-life for disk arrays).

One other important issue for long-term data retention is the issue of associated metadata. For example, a Digital Cultural Heritage project stores a high volume of images and 3D data in a folder structure, but also maintains several indices and metadata in relational databases. These relational data form the basis of several services provided by this community and their preservation is perhaps just as important as the original data itself. Care is taken to keep associated data and metadata preserved with special considerations for data integrity.

2.9 SWEDISH NATIONAL INFRASTRUCTURE FOR COMPUTING

2.9.1 Introduction

Swestore is a Swedish national storage service on the infrastructure level, making storage space available over the internet for academics through an application procedure to the Swedish Research Council (VR). Swestore also aims to provide the means for convenient sharing of research data at the discretion of the scientists owning the data. This initiative addresses the increasing need for long term archival storage brought on by mandated retention time on availability of raw research data, in combination with the ever increasing size and rate of research data produced. Swestore also aims to address data sharing problems in joint research efforts where large data sets needs sharing.

The aim of the Swestore storage is also to build a nationally accessible, robust, flexible and expandable system of general use in most cases where access to large scale storage is required. From a user prospective, it has the appearance of a single large logical file system, while the data is physically distributed across all six SNIC centres. The data distribution brings the advantages of locality and cache effects, among other things. SweStore is also intended to become the building block for future Swedish participation in the EUDAT Collaborative Data Infrastructure.

As a project, Swestore national storage is in its early production phase with several things being in a state of flux or yet to be decided. However, the current status of the SweStore National Storage project is described. Storage costs, strategies for data lifetime management and data integrity and availability issues are discussed.

2.9.2 Data types and volumes

Five scientific communities share the SweStore infrastructure, among which environment and climate, humanities, bio-imaging or proteomics. The SweStore National Storage repository currently holds 300 TB of replicated user data, using a total of 600 TB raw surface and the current total capacity is 1200 TB.

2.9.3 Architectural overview of the storage solution

The issues of file format changes as well as changes in the underlying software and hardware stack during the life span of data or transitions to emerging storage technologies need to be addressed by any long-term data storage solution. Since these aspects are not dependent, they can be decoupled to allow the experts of scientific disciplines to cater for the different needs of their respective scientific communities as they evolve in terms of content, file formats and meta data, as well as leaving the technical considerations involved in storage to storage experts. Reflecting this, scientific expert community groups have been put in place along with SweStore, in addition to a previously existing SNIC storage experts group.

The Swestore storage group consists of storage specialists from all six Swedish academic HPC centres within SNIC. The focus of this group is storage technology, both on the hardware and software side. The list of responsibilities ranges from operating disk and tape storage pools within the dCache back-end to the access protocols, clients and associated services of the front-end.

As mentioned before, some scientific expert community groups have been created, in addition to the storage group within SNIC, which responsibilities lie in gatekeeping, i.e. deciding which data formats are allowed to be stored, as well as gathering metadata about the data allowing systematic book keeping providing a fall back route for orphaned data, i.e. data with no owner. The data itself is always owned by the individual scientist storing data in SweStore.

The current sharing of responsibilities within Swestore leaves virtually all aspects of long term accessibility of data outside the purely technical for the scientific community expert groups to decide. The storage technical expertise within SweStore has taken on a purely advisory role. The advice given so far bearing consideration have mainly been regarding best practices in terms of data integrity and validation and data availability strategies in terms of file formats. Basically, a line is drawn in terms of responsibilities between the core storage services and the outward facing services. All decisions regarding data vetting and metadata book keeping and publication is the realm of the scientific expert communities, while the technical decisions are handled by the SweStore storage group. The intended data and information flow is that the scientist needing storage applies for this to an allocation committee. Provided the application is granted, metadata about the data set is submitted to the scientific expert group to the extent required by the scientific expert community group in question and data is then directly uploaded via the ARC client middleware tools to the distributed dCache storage pools. Currently the ARC client middleware is the only route supported to upload data to Swestore.

Read access is intended to go via the storage services layer, although it is currently only accessible via WebDAV or the ARC client, the possible means of access follow those that dCache provides.

The SweStore National Storage infrastructure is fully implemented using the distributed storage solution dCache. In Figure 25 below, shown in blue are currently available access protocols, while yellow marks access protocols under development to interface with the dCache back-end and orange marks candidates under evaluation or investigation for data type services.

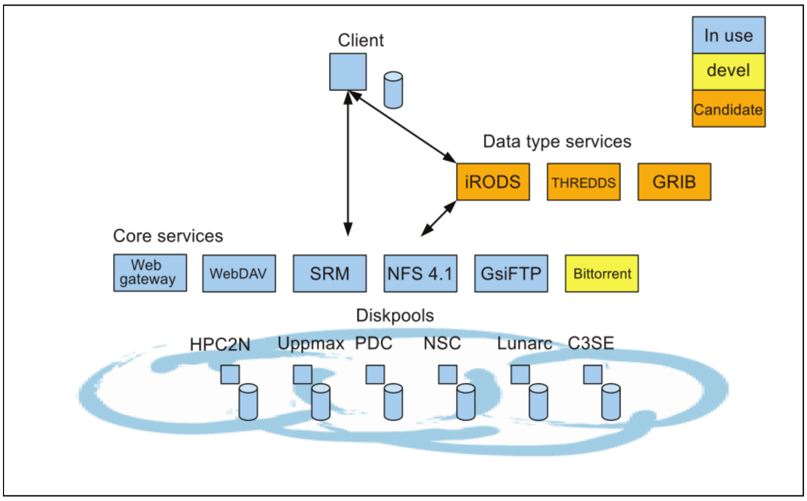


Figure 25: SWEDISH NATIONAL INFRASTRUCTURE FOR COMPUTING architectural overview

The core services of the system are located at HPC2N at Umeå University. There are over 65 online storage pools (as of November 2011) attached to the system, which are spread between the SNIC centres at Lunarc, C3SE, NSC, PDC, Uppmax and HPC2N. A file upload usually go directly from the source to one of the storage pools without passing the core service giving a high aggregated transfer performance to the system. All files in the SNIC part of storage are replicated to a different site for availability reasons. With the exception of the core services, an entire site can be offline without any loss of functionality for SweStore.

There are several access protocols for SweStore National Storage. The primary ones being the SRM and WebDAV protocols, but access can be made by any means supported by dCache, see figure 6. Authentication for Swestore write (and read) access is carried out by means of X.509 certificates and anonymous read only access data is made publically available at the discretion of the data owner.

2.9.4 Challenges

There are no backups of data on SweStore. The files are replicated between sites to minimize the risk of data loss due to hardware problems. But if the end user deletes a file it will be lost. Future improvements in the pipeline for dCache may provide a waste basket functionality which could minimise damage of end user mistakes.

There is currently no tape backend attached to SweStore, but this is subject to change in the future. Tape storage in the Swestore context would be used for archiving data not being accessed frequently as the on-line nature of disk storage is considered highly desirable and one of SweStore purposes is to act as a file system of general use.

An important feature of Swestore access is that staging data in and out from grid clusters with the ARC middleware client works out of the box. Instead of copying data to the clusters and then copying the results out again you can let ARC do that for you. Just specify the input file URLs you need and let ARC do the work for you.

2.10 FINNISH IT CENTER FOR SCIENCE

2.10.1 Introduction

The National Digital Library (NDL) of Finland is an entity within the remit of the Ministry of Education and Culture, which basis is formed by libraries, archives, museums, and other organizations in Finland storing cultural heritage material and the actors responsible for their IT solutions affecting hundreds of organizations. The NDL aims to create a nationally unified structure for contents and services ensuring the effective and high-quality management, dissemination, and digital preservation of cultural digital information resources. Nationally shared digital preservation infrastructure and services will draw the practices of memory organizations closer, reduce the costs and fragmented nature of the systems, and intensify cooperation. CSC – IT Centre for Science is responsible for the design and implementation of the NDL’s digital preservation (DP) system.

2.10.2 Data types and volumes

Almost all memory organizations under the Ministry of Education and Culture of Finland (partner organizations) are under an obligation to preserve a lot of their digital material. Most of the cultural heritage material consists of digitized documents, maps, photographs, newspapers and sound recordings. In the future, this material will be mostly born-digital, which increases the volume of the data.

To support NDL's DP planning, a survey was carried out in spring 2011 with the partner organizations to clarify the scope and various other aspects of the digital information resources managed by the partner organisations that will be stored in the NDL's DP system. The survey helped to specify the extent of the digital material of partner organisations and, in particular, it examined what kind of preservation methods they require. The survey results are summarized in Table 8. It can be seen that the data volume increases rapidly and that the space requirement with different types of objects is very heterogeneous. It is assumed that there will be approximately 12.5 PB of data to be stored to the system in 2020.

	2010		2011		2015		2020	
	Number of objects (millions)	Size (TB)	Number of objects (millions)	Size (TB)	Number of objects (millions)	Size (TB)	Number of objects (millions)	Size (TB)
Files and documents	11.6	328	15.4	394	25.6	646	48.7	1,301
Photos	1.7	18	2.1	30	3.9	68	6.1	120
Films	0.1	495	0.2	1,143	0.8	3,055	1.2	8,020
Sound recordings	1.2	606	1.5	771	2.4	1418	3.7	2,176
References	19.5	1.2	21	1.5	27	2.4	34	3.4
Online archive	496	20	646	27	1,396	59	2,300	97
Radio and TV archive	0.8	95	1.2	142	2.9	327	5.0	558

Total	**530**	**1,563**	**687**	**2,509**	**1,458**	**5,575**	**2,400**	**12,275**

Table 8: Data volume estimates in the 2011 survey.

2.10.3 Architectural overview of the storage solution

The preservation system will be built with an idea of locally distributed storage systems, see Figure 26. The core idea is to utilize common x86 bit PC-hardware to server as a storage point. In this solution, storage nodes are independent of each other, and binding component is redundant file catalogue database. The disk resources are divided equally for each server so that each server is responsible of local RAID6 array. Such kind of basic server platforms exist where a 4U height server chassis can take care of 36 of 3.5” disks. When using 4 TB SAS disks and RAID6 technology with one online-spare disk per 15 disks, each server can take care of 112 TB. In 2013, if there will be 4 PB of data to be stored, approximately 37 servers are required per data copy. This solution is a lot cheaper compared to a solution using SAN network. Currently, one server with the disks costs around 18 k€ and one terabyte costs 150€. This kind of distributed system is very scalable, and the total costs caused by adding or upgrading the hardware piece by piece remain in a reasonable level.

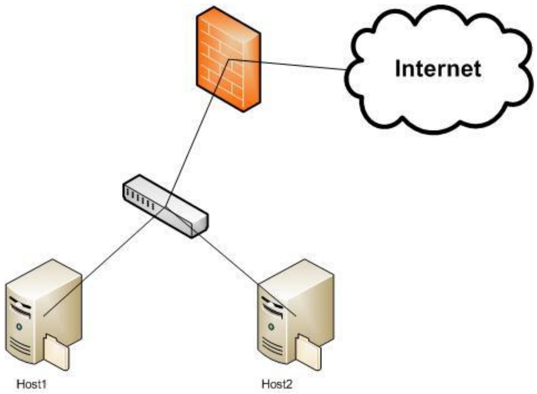


Figure 26: Distributed storage system.

In addition, the AIP is saved in a tape library. Tape library is a two phase platform where in front of actual tapes is a staging disk area. The staging disk area size is 1/4 of primary disk capacity. Tape library has its own data management software which guarantees that there is always file stub available on staging disk even when file has been written into tape. Creating a tape archive copy of 4 PB data requires approximately 1000 tapes (e.g. IBM 3592 tapes, 4.0 TB/pc) and approximately 200 days from one tape library drive, if it is assumed that the writing speed is all the time 240 MB/s. However, the staging disk and several tape library drives will reduce the load in the other parts of the system. Staging disk can also be used as an extra backup copy of the recently added data. Data consistent verification requires same amount of I/O bandwidth than its writing process. Therefore extra drives needs to be considered, if the tape library throughput is a bottleneck.

Of course, the media type (disk and tape) is depended on the current technology, and the used hardware and media type may be something else in the future. As hardware becomes obsolete, the copies will be transferred onto new hardware and storage media. The bit preservation should also be geographically decentralised into separate sites, which will include the ability to receive, store, and utilise data. The connection between the sites will be done with a fibre backbone. At least two sites will be formed. These sites are located in Espoo and in Kajaani, Finland, with about 550 km distance between.

Several databases are required for the use of the DP system, e.g.:

- A database for the properties (such as location) of each data object.
- A database for the metadata of the data separately, which makes it possible to retrieve and update the metadata.
- A database for logging the workflow and results of different preservation operations, such as different validation tasks and conversions.

To remain the reliability and scalability, each database will be in their own servers and replicated with master-slave architecture (see Figure 27).

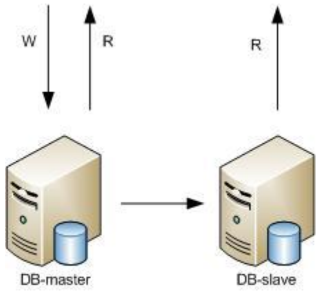


Figure 27: Metadata replication.

2.10.4 Requirements for the storage system

The selection of digital information and its preparation for digital preservation will primarily be the responsibility of the partner organisations. However, number of services to facilitate the work will be offered by the DP system. These services include pre-ingest, ingest, archival storage, usage, management and support services.

The digital preservation system will be built according to the OAIS model. The functionality and services related to the preservation of digital information are the key components of the DP system. This will guarantee the reliability, understandability, and immutability of the material.

The DP system will ensure that the original bits of the digital object remain unchanged and can be run on modern hardware. Multiple copies of the digital object and its metadata will be created to different types of media. The integrity of the copies will be monitored at regular intervals with checksum calculation processes. The packages stored in the DP system (AIPs) will be refreshed at regular intervals in line with the OAIS reference model. As storage equipment becomes out-dated, copies will be transferred onto new, modern hardware and storage media. Data reliability will be ensured through appropriate security arrangements. Preserving all essential metadata together with the digital objects in the DP system will keep the material understandable. In addition, the digital object will be migrated into the format used at any particular time in the future. Migrations will be designed and tested in collaboration with partner organizations in line with the OAIS model. A preservation plan submitted in the received package (SIP) guides, how the AIP is formed from SIP and how it is managed.

The DP system will include search functions necessary to find content. Separate data sites ensure that the data is always available (even if the other site is unreachable), the usage of the net traffic can be shared equally between the geographical locations, and the risk of data loss is minimal (e.g. data is safe and available even in case of emergency, such as fire or flood). By default, only the partner organisation that has ingested the data can collect the required version of the digital information transferred into storage from the DP system in situations where the material has been damaged within the organisation's own system or when its authenticity is uncertain. Licenses and access rights determine which other parties can access and use the content. In line with the OAIS reference model, the DP system will deliver the digital information as dissemination information packages (DIP) that are formed in compliance with common specifications.

The SIPs will be validated against the NDL’s specifications within the DP system before they are formed into AIPs. The NDL will also provide a packaging service application, which helps the partner organisations to form a SIP compliant with the NDL METS profile. The DP system will also comprise a series of complementary advisory and support services, suitable for different situations and organisations of different sizes. Data management services include functions that, for example, allow the partner organisation to upgrade the metadata of the preserved digital information and add new versions of the content. Digital information may also be removed from storage if necessary. Data management services will also include the updating of preservation plans, DP system maintenance, cost control, and risk management.

2.10.5 Challenges

The SIP inspection process and conversion to AIP may require a lot of resources, and therefore it is important to distribute the process concurrently to different low-cost servers. Since the data volume is high and includes large objects, the integrity monitoring process needs to be designed as a parallel operation for several servers. One of the main challenges is the data transmission from the client user to the DP system. For example in 2013, the amount of digital data (4 PB) in the preservation system requires over one year transmission time with using a single 1 Gbit/s connection, if the actual average speed is 80% of the theoretical maximum. It is very important to plan the processes as optimal as possible, so that useless data transmission will not occur. The heterogeneity of the data with several partner organizations gives certain challenges to specifications. Although a common specification already exists, it most likely will be updated several times in the future. Since the needs of all partner organizations must be considered, the specification update process requires a lot of discussion and collaboration.

3 EXAMPLES OF RESEARCH ON STORAGE SOLUTIONS

There are clearly many research activities on-going in the area of storage. We present here two specific research (as opposed to production and service) activities in storage being carried out by APARSEN partners.

3.1 RESEARCH CARRIED ON AT IBM HAIFA RESEARCH LAB

SIRF: Self-contained Information Retention Format

While preserving business, public and personal assets involves various stakeholders and technologies, storage has a key role in this spectrum as it is where the data resides for most of its lifecycle. Moreover, the storage should comply with standards to enable interoperability of today's preservation systems with those of the future. Recognizing the significance of standardization in storage for long term retention, the Storage Networking Industry Association (SNIA) is working towards that aim. SNIA started by conducted a survey with more than 250 participants who revealed that over 80% of respondents had retention requirements of at least 50 years **[1]** . The top external factors driving these retention requirements were mainly business risks and compliance with regulations. Consequently, SNIA formed the Long Term Retention (LTR) Technical Working Group (TWG) **[7]** in 2008 to address storage aspects of digital retention. LTR which is co-led by IBM Research-Haifa and HP received the mission to lead storage industry collaboration, and develop technologies, models, educational materials and practices related to retention and preservation.

The LTR TWG is working on Self-contained Information Retention Format (SIRF) to create a standardized vendor-neutral storage format that will help its customers interpret preservation objects in the future, even by systems and applications that do not exist today. SIRF provides strong encapsulation of large quantities of metadata with the data at the storage level, and enables easy migration of the preserved data across storage devices. In contrast to the semantics of traditional file systems, which include only limited metadata about each file, SIRF provides for the rich metadata needed for preservation and ensures its grouping with the information objects.

SIRF proposes an approach to digital content preservation that leverages the processes of the archival profession thus helping archivists remain comfortable with the digital domain. One of the major needs to make this strategy possible is a digital equivalent to the physical container - the archival box or file folder - that defines a series, and which can be labelled with standard information in a defined format to allow retrieval when needed. SIRF is intended to be that equivalent - a storage container format for a set of (digital) preservation objects that also provides a catalogue with metadata related to the entire contents of the container as well as to the individual objects and their interrelationship.

The SIRF format is intended for a mountable unit storage container but should be agnostic to the actual medium. While some data is preserved on CDs and DVDs, tape and disk storage systems are currently the predominate types of media on which large quantities of data are preserved. In many cases, the preservation data tends to be cold (inactive) and is seldom accessed over time. SIRF is agnostic to the type of storage medium as well as to any type of data. SIRF enables the mountable unit storage container to be self-describing and self-contained to the extent possible. This capability is particularly useful for offline storage media, which are often deployed for cold data, but SIRF also works well for storage that remains online.

SIRF is defined using a layered approach, with two levels. The levels differ in their catalogue. The SIRF level 1 catalogue contains unique metadata, some of which is not included within the preservation objects, but is mandatory to make those preservation objects portable into the future. Examples of such metadata are retention hold, reference counts, preservation object fixity algorithms, fixity values and fixity calculation dates. The SIRF level 2 catalogue includes additional information that may also be included within the preservation objects, and is used for fast access to the preservation objects. Examples of such metadata are links to representation information needed to assure referential integrity, metadata about the relationship among the preservation objects, packaging format, etc.

In the LTR TWG, we defined nine use cases for SIRF including four generic use cases and five workload-based use cases that are published for public review **[8]** . The generic use cases are unlinked to a specific type of data or application and consider just technological changes to the environment. In contrast, the workload-based use cases are specialized for concrete workloads and consider additional non-technological changes to the environment. All use cases derive requirements for SIRF that are used in its definition. You can find more on SIRF use cases, requirements, proposed definition, metadata and format in **[10]** .

3.1.1 The ENSURE Approach Preservation Storage Solutions

ENSURE - Enabling kNowledge Sustainability Usability and Recovery for Economic value

Guaranteeing long term usability for spiralling amounts of data produced or controlled by organizations with commercial interests is quickly becoming a major problem. Guided by real world use cases in health care, finance, and clinical trials, ENSURE extends the state of the art in digital preservation, which to-date has primarily focused on relatively homogeneous cultural heritage data information through innovative solutions considering:

- *Cost and Value*: Evaluate the cost and benefit of different quality solutions, enabling a business to choose the most cost effective solution.
- *Preservation Lifecyle Management*: Build on industry standard lifecycle management approaches to manage the preservation lifecycle, meet regulatory compliance, allow changes in the preservation approach to reflect environmental changes, address evolution of ontologies and manage the quality of digital objects over time.
- *Content-aware Long Term Data Protection*: Provide data protection over long periods of time, addressing changes to personally identifiable information, new and evolving regulations, and manage user identities over the decades.
- *Utilize Emerging ICT*: Evaluate the costs, risks and benefits and demonstrate how to use emerging, commonly available Information Technology to enable scalable solutions for digital preservation, in particular considering Cloud Storage and virtualization techniques.

ENSURE started in February of 2011, and will run until the end of January, 2014. See [12] .

The ENSURE architecture is consistent with the OAIS standard [11] and consists of:

- A set of plug-ins that provide specific functionality such as format management, regulatory compliance, integrity checks, access to specific storage clouds etc.
- A runtime SOA framework that allows composing an OAIS solution from appropriate plug-ins to meet a user's requirements including economic considerations,
- A configurator and a cost/performance/quality analysis engine which can evaluate a proposed preservation solution
- The next section provide detail on the storage services component of ENSURE.

3.1.1.1 PDS Cloud

PDS Cloud (or Preservation DataStores in the cloud) provides a preservation-aware storage infrastructure for ENSURE based on storage and compute cloud enablement technologies. It offloads OAIS-based functionality to the storage to decrease probability of data loss, and support automation of preservation processes. Thus on the one hand, PDS Cloud exposes an OAIS-based interface to provide its enhanced archival services, and on the other hand, it leverages multiple heterogeneous generic cloud storage and compute from different providers. PDS Cloud extends PDS [12] which provided storage infrastructure for the CASPAR EU project, with advanced functionality in a cloud environment.

The main objectives of PDS Cloud are to:

1. Support access to multiple various cloud storage and cloud compute, as well as enable migration of data and virtual appliances between different clouds. This includes using multiple clouds concurrently, while taking advantage of each one's special capabilities.

2. Support a flexible OAIS-compliant data model for a multi-cloud environment that has configurable data management capabilities and can adhere to diverse data aggregations with different requirements for preservation capabilities.

3. Enhance future content access by supporting accessing content with a cloud based virtual machine.

4. Support computational storage via storlets mechanism to perform data processing on the cloud. Storlets are restricted modules performed in a sandbox close to the data and utilize the data locality property to execute data-intensive functions more efficiently.

5. Support advanced flexible dynamic fixity (integrity) checks and auditing capabilities that complement the generic clouds capabilities.

6. Support complex interrelated objects and manage an object's inter-relationships and links while maintaining referential integrity. This includes the support for OAIS collections (AICs) and referential integrity among the various parts of the collection.

During the first year of the project we have completed the first phase prototype of PDS Cloud and integrated it into ENSURE use cases demo in the testbed.

PDS Cloud is a preservation-aware cloud service that interacts with a heterogeneous set of cloud storage and compute. **Figure 28** depicts the PDS Cloud architecture and its components. PDS Cloud is divided into two main layers:

- **Multi-Cloud Service --** supports access to a heterogeneous set of cloud storage and compute providers.

- **Preservation Engine --** supports preservation functionality for OAIS AIPs. This layer accepts requests via the PDS Cloud interface and services them, utilizing the multi-cloud service underneath. The preservation engine includes four main services: (1) AIP service to ingest and access various types of AIPs including AIUs, AICs, RepInfo AIPs, and storlets AIPs, (2) Migration service to support logical preservation and transformation of the data from one format to another creating new version AIPs, (3) Fixity and Audit service to support bit preservation ensuring the bits are not altered as well as providing audits to an external third party, (4) Administration service to administer tenants and configuration of PDS Cloud.

3.1.1.2 Logical Data Model for OAIS Archival Storage

PDS Cloud provides a new logical data model suitable for preservation that abstracts implementation details in the underlying clouds. The data model consists of tenants, aggregations, dockets and AIP objects. Figure 29 depicts the data model with its relationships.

Each AIP may contain references to multiple RepInfos in its manifest file; this allows several view paths to interpret the same data. A RepInfo is by itself an AIP and thus includes its own manifest, metadata and content. Each RepInfo is a shared object and may be referenced by multiple AIPs, thus the relationship between standard AIPs and RepInfos is m:n.

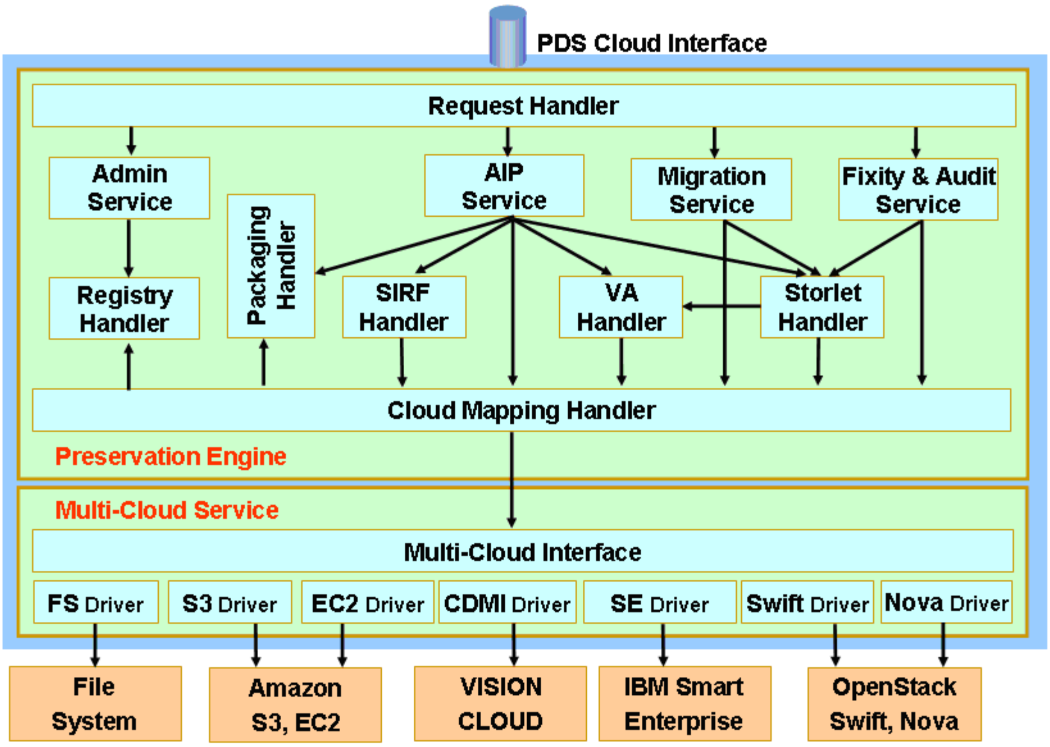


Figure 28: PDS Cloud Interface

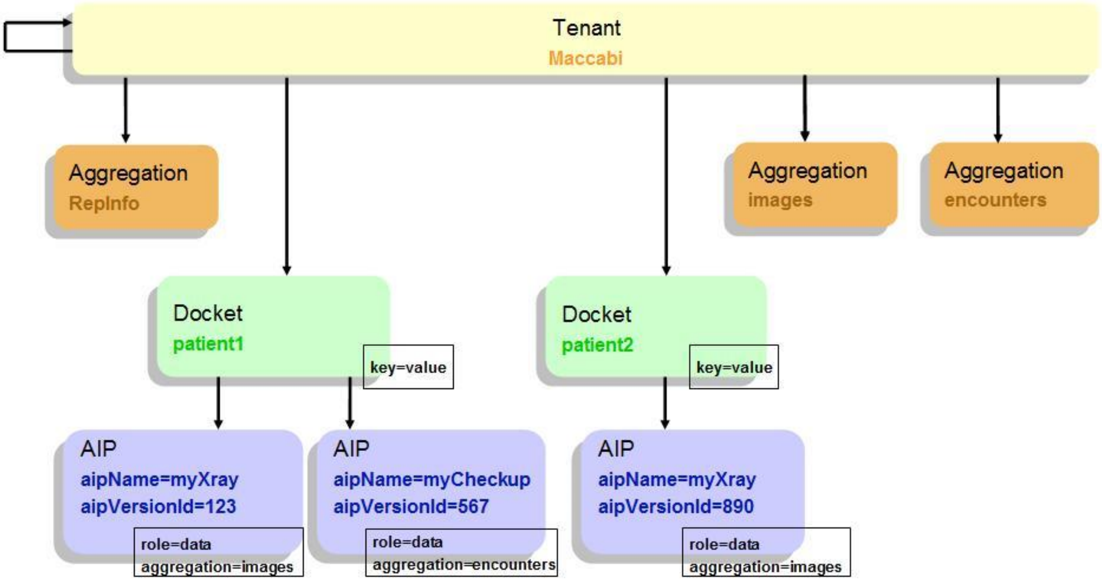


Figure 29: PDS Cloud data model

3.2 RESEARCH ON OPTICAL DEVICES CARRIED ON AT CINI

3.2.1 Introduction

CDs and DVDs are a rather unreliable recording media, and therefore one may assume that they should be considered as unfit for digital preservation, especially for long term preservation. This is indeed confirmed by several unfortunate accounts of disks that have become unreadable after a few years, with the consequent loss of data. Some of the readers may have even directly experienced the problem, by storing digital pictures and other personal documents on CDs, and being afterwards unable to access them.

However, due to their very low cost and to their widespread distribution, CDs and DVDs have been widely used, and in many cases are still used, by several organizations, including very respectable ones, as storage media in their preservation repositories. Typically they are used for secondary or backup storage copies in redundant storage schemes, while primary and/or access copies are kept on online redundant storage systems as RAIDs.

According to our experience it is quite unlikely that many of these organizations may be convinced to give up the use of optical removable disks as storage media, at least in the short and medium term. It is therefore interesting to investigate how the reliability of such devices can be assessed, and possibly to devise storage procedures that, introducing appropriate controls, may raise the reliability levels in order to meet target values that could be considered appropriate even in a long term digital preservation environment.

The problem of building a reliable system out of unreliable components has indeed already been faced and solved in several engineering contexts, notably in the design of computer systems and their components. This scheme are typically based on two elements:

- *redundancy*, which is used to reduce the failure probability;
- *continuous monitoring*, which is used to keep the situation under control and to be able to prevent failures before they happen.

Such an approach may be deployed also in the case of CDs as DVDs to achieve even high levels of reliability, as we will show in the next sections.

3.2.2 Optical removable disks as a preservation media

Currently many preservation repositories still use CD, but the use of DVD has recently increased, mostly due to the larger volume of data they can hold. Blue Ray DVD have been also taken into consideration in recent times.

The main difference between CDs, DVDs and Blue Ray disks concerns the laser wavelength, and consequently the recording density. However the physical and chemical structure of the disks, the manufacturing characteristics and the recording technology are pretty much the same.

The analysis we present in this report is mostly concerned with writable CDs, and more specifically to those recorded according to the ISO 10149 standard, but, due to the similarities in the physical and chemical structure and of the recording schemas, our conclusions can be easily extended to DVDs and Blue Rays. More specifically the operational guidelines we propose may be implemented for any kind of optical removable medium.

Recording quality and temporal preservation limits for all kind of storage media are mainly influenced by three factors:

- the quality of the recording medium;
- the quality of the recording device;

- the environmental conditions in which the disks are preserved.

3.2.3 Quality of the recording media and of the recording device

Unfortunately the quality of the optical recording media is constantly degenerating and is strongly affected by heavy market competition. Writable CDs and DVDs are substantially produced for the consumer market that is more concerned with the price than with the quality of the product. On the other hand 'professional' customers account only for a minimal share of the market, and therefore producers are less and less interested in manufacturing and selling, even at higher prices, quality items. This is, of course, a very negative factor if the disks have to be used in a digital preservation environment.

Physical structure of recordable CDs

From a physical point of view, a CD/DVD is a 1.2 mm (0.047 in) thick disc made of polycarbonate and is composed of several layers. Starting from the side opposite to the label:

- an external protective layer laid to harden the surface;
- a layer containing a spiral groove, called the *pregroove* (because it is moulded in before data are written to the disc), to guide the laser beam upon writing and reading information;
- a layer of polycarbonate resin;
- a layer of *dye*, an organic polymer, which is modified during the writing (*burning*) process;
- a reflective layer of gold or silver;
- a layer of protective lacquer (UV curable polyacrylate resin) to increase resistance to environmental stress cracking;
- the label (UV curable ink).

The writing process is based on inducing transformations on the dye layer by heating it with a concentrated high-power laser beam. When heated at a given temperature, the dye layer, which is initially transparent, gets *burned*, i.e. becomes opaque.

Information is coded as a binary sequence of transparent and opaque segments which are written by the high-power writing laser along a spiral groove. It may then be read at a later time by focusing a low-power reading laser beam along the spiral grove. The laser beam is reflected by the metal layer only when it goes through an area of the dye that has not been burned, and thus the alternate sequence of opaque and transparent segments generates a binary sequence.

All the layers in the CD structure may potentially suffer chemical and physical transformations when the CD is preserved, and may therefore pose threats to the readability of the recording. Glues, which are used to connect the different layers, add further problems.

The most delicate part, as far as preservation is concerned, is anyway the dye layer, the element that is most vulnerable and subject to degradation, mainly because of the light (UV rays) and cyclic variations of temperature and relative humidity. Recent experiments, consisting in accelerated aging procedures (see Section 3.2.5), have identified Phtalocyanine as the most resistant kind of dye.

As far as the reflecting layer is concerned, the aging tests have proven gold to be more resistant to aging than silver. Silver is often used because of a better reflection index, which allows higher reading speeds, but is more subject to oxidation processes.

Evaluating the quality of the recording media and the recording devices

In evaluating the lifetime of an optical recordable disk we should consider two components.

1. Evaluate the *quality of the media and the recording equipment*, which can be measured by recording a sample pattern and reading it immediately after it has been recorded. But, to understand what we mean by *quality* and how it can be measured, we need first to discuss the internal structure of the CD, which is presented in Section 3.2.4.

2. Evaluate the *effects of aging*, i.e. how the passing of the years affects the readability of the recording.

The second component is by far the more difficult to evaluate since direct experimentation is not feasible since it would require many years, possibly decades. However some experimentation is still possible by using *accelerated aging techniques*, a very complex procedure that can be carried out only in a few specialized laboratories with sophisticated equipment that we will discuss in Section 3.2.5.

Moreover this component needs to be seriously taken into account because the degradation of readability due to aging has been proven to be a crucial factor. Even every day's experience tell us that some CDs that were 'perfectly readable' became unreadable after a few years.

Therefore, when provisioning, the management of a preservation repository should mostly rely on the trustworthiness of the manufacturer and on the characteristics published in the product data sheets, if any.

One should be anyway very careful and make a clear distinction between the dealer and the manufacturer, since very often the same product is sold by different dealers, under different labels. On the other hand it may happen that different batches sold by the same dealer, under the same label, actually correspond to different products. It is therefore necessary to repeat the analysis on every individual batch.

However it is always possible to check the origin of a given CD/DVD, since this information (*media ID*) is coded in a standard way, recorded inside the CD/DVD and easily readable by means of specialized applications, most of them open source. But it is worth to mention that media IDs can be faked as well, and sometimes are. Therefore, it is advisable to buy CD/DVDs only from reliable and well known dealers and to avoid using blank CD/DVDs (without a manufacturer's or dealer's label), which are often sold at very attractive prices.

Only recently some international dealers have become interested in marketing high-end products, which they claim to be high quality and better suited for long term preservation. These products are generally labelled as *high quality* or *archiving* and have a golden reflective layer, special lacquers and a supposedly better manufacturing quality.

It may also be useful to consider market research studies that compare products of different dealers and are often based on the results of tests that involve hundreds of CD/DVDs. One should anyway consider that there may be different quality criteria, as for instance maximum writing and reading speed that may have nothing to do with the persistence of the recording we are interested in.

Due to the above mentioned problems, we may recommend as a wise policy, to rely only on well-known suppliers, and to buy CD/DVDs only in large batches, possibly from the same manufactured lot. Large quantities make more reasonable to perform on each batch the simple tests that we shall discuss later (see Section 3.2.4) to assess at a basic level the quality of the recording media.

Quality of the recording devices

Similar remarks can be done for the recording devices, since only consumer-oriented products are currently marketed, and the few producers of high-quality professional devices went out of business.

More specifically, there are two main problems:

- the alignment of the writing laser beam with the pregroove (see subsection 3.2.3 - **Physical structure of recordable CDs**);

- the non-uniformity of the CD/DVD rotational speed while it is written.

The first problem may cause the reading laser, that (hopefully) follows the pregroove, to mistake and opaque part for a transparent one, and thus to read 1 instead of 0, or vice versa.

The second problem, which results in a non-uniformity of the recording density with the consequent errors in reading the sequence of 0 and 1 (transparent and opaque), and is quantified by the *jitter*, a parameter that can be measured through the analysis of the recorded media (see subsection 3.2.4 - **Rating the error levels**). Of course reading devices have the same problem, and potentially, in the worst case, the two jitters may add up.

As mentioned, jitter can indeed be measured by recording a fixed pattern on the media on the device to be sampled and then reading it with a reference device, i.e. one with calibrated and (almost) uniform rotational speed. But, again, such devices are usually found only in specialized laboratories.

In trying to avoid all these problems one should of course restrict only to buying the best devices on the market. This, unfortunately, may be not enough.

A complementary approach is to constantly monitor the correct operation of recording devices. Even without reference devices, or other specialized equipment, this can be done with a comparative analysis of CD/DVDs of the same batch but recorded by different devices. By means of this procedure it is possible to detect faulty devices that should be immediately replaced.

3.2.4 Assessing the quality of the recording process

Low level structure of the optical disks

To understand how the recorded CD quality level can be assessed it is necessary to explain, even if summarily, the rather complex encoding scheme used for these optical recording media. This may seem (and possibly it is!) a bit cumbersome and annoying, but going into these details it is still necessary to properly understand how the reliability of a CD can be assessed, and what the assessment actually means.

According to the standard introduced initially for audio CDs, in a data CD data are encoded in a redundant way by exploiting special encoding schemes that allow, when reading the CD to *detect* and possibly to *correct* at least part of the errors that may have been introduced in the write/read process.

This kind of redundant encoding is strictly necessary since the probability of a mistaken bit is indeed very high. To realize that one should consider that a recording segment that corresponds to an individual bit along the recording track has a size of a fraction of a micron, quite less than a grain of dust. If there were no redundancy a single mistaken bit would make the whole disk unreadable.

Error-correcting encoding schemes used in CD are very sophisticated and, among other tricks, spread the redundant information on the disk surface, in such a way that a local damage, for instance a scratch, affects a large amount of the recorded information, but only in a limited way, thus allowing the error correcting codes to recover the damage.

More precisely, the encoding is based on a well-known scheme called *CIRC (Cross Interleaved Reed-Solomon Code)*, international standard IEC 908. The schema is based on several interleaved encoding layers, which altogether guarantee a high level of robustness. This redundancy is heavily paid, since the ratio between the net and the gross amount of information recorded on the CD is about one to three. That is to record 700 MB of information on a CD more than 2 GB are actually used.

Data in CDs are organized in groups of 33 byte called *frames*, but only 24 byte of each frame actually carry primary information. Sets of 98 frames are grouped in a *sector*, for a total of 2352 byte. Unlike audio CDs, data CD use only 2048 byte out of each sector, and use the remaining 304 byte for a further level of *ECC (Error Correction Code)*, totally independent from the CIRC, a sort of *last resort*.

Moreover the encoding exploits a technique called interleaving, consisting in distributing a 24 byte block in 109 adjoining blocks, in such a way that to destroy a single byte it would be necessary to destroy all 109 blocks. This guarantee, in the case of frequent damages like scratches, dust and fingerprints, that enough information is left to recover the damaged bytes.

The CIRC schema is based on two levels of encoding:

- the *inner encoding level*, called C1, that has the capability of recovering up to two errors (two mistaken byte) per frame; a larger number of errors cannot be recovered at this level and the recovery action is therefore delegated to the outer level;
- the *outer encoding level*, called C2 that has the purpose to recover the errors that have not been recovered at the inner level; in this case too, the encoding can only recover up to two errors per frame; a larger number of errors cannot be recovered.

In audio CDs more than two errors at level C2 in the same frame cannot be recovered. But in data CDs the error may still have a chance to be recovered by the further redundancy at the sector level (the *last resort* discussed above), provided that the same problem do not arise for other frames in the same sector. However this *third level* recovery is not handled by the firmware of all CD readers, and most readers in such a situation would rate the CD as unreadable.

Rating the error levels

The presence of errors that may be recovered is totally transparent to the user. Therefore, the mere fact that a CD could be read does not provide any clear information about the quality of the recording. The successful outcome could indeed correspond to quite different situations, ranging from no errors at levels C1 and C2 to a high frequency of double errors at level C2.

However, it is possible to give a detailed characterization of the error levels by means of special devices, called *CD analyser*, produced and marketed by several manufacturers in a range of price of several thousands of Euros. Such devices give a complete account of all errors at the different encoding levels.

But even without a CD analyser meaningful measures of the error levels can be obtained by means of software tools that interact at low level with the CD reader firmware, can detect the existence of level C1 and level C2 errors, even when they can be recovered by the firmware and therefore do not compromise the readability of the CD. Such tools are often part of CD recording software or are supplied with recording devices. Open source tools are available as well.

If the analysis is performed just after the CD is recorded, it gives valuable information both on parameters that are ascribable to the quality of the recording media, and on parameters that depend on the quality of the recording device.

For a quantitative characterization of the error levels one should refer to the following parameters:

- **E11**: rate of single errors at level C1 (errors that can be recovered at level C1);
- **E21**: rate of double errors at level C1 (errors that can be recovered at level C1);
- **E31**: rate of errors that cannot be recovered at level C1;
- **E12**: rate of single errors at level C2 (errors that can be recovered at level C2);
- **E22**: rate of double errors at level C2 (errors that can be recovered at level C2);
- **E32**: rate of errors that cannot be recovered at level C2;

Error rates are generally expressed in errors per second and are measured on a 10 second interval reading the disk at speed 1 (for higher speeds the measure should be scaled accordingly). In some cases the total number of C1 and C2 errors in the CD is given as well. As different tools may rely on different definitions of the error rates, one should always carefully check the documentation in order to give the measures the correct interpretation.

Further parameters that are often used are:

- **BLER** (Block Error Rate): is the rate per second, averaged on a 10 seconds interval, of the number of sectors where *at least* one error at level C1 has been detected, that is E11+E21+E31. High values of BLER, both average and peak values are a clear index of low quality.
- **BERL** (Burst Error Length): a measure of the number of consecutive sector where errors have been detected. High values of BERL, compared to average values of BLER indicate that errors are locally concentrated.
- **Jitter**: as already mentioned, it gives a measure of the non-uniformity of the CD rotational speed during the recording, and therefore of the recording density. High values of jitter a clear index of a low quality recording device.

Quality criteria

Up to now no threshold levels have been already officially agreed for the above mentioned parameters that is no official standard has been set for the quality of CD recordings. However, one may rely on a series of meaningful references that may be followed when defining quality criteria for recording CDs in digital archives. Such recommendations have been issued by international standardization bodies like ISO and ECMA, and by competence centres well known at international level like NIST (National Institute of Standards and Technology).

These recommendations may be summarized as follows.

- The rate of frames with **C1** level errors (expressed in faulty frames per second and averaged on a 10 seconds interval) should stay below the threshold of $3x10^{-2}$ (recommendation in the *Yellow book*, standard ECMA-130, ISO/IEC 10149).
- **BLER** should stay below the threshold of 220 errors/s (a threshold recommended for industrial CD production by the *Yellow Book* and the *Red Book*). But this level should be considered as the maximum tolerable one, since, for instance, in quality production lines of audio CDs **BLER** is usually kept below the threshold of 50. Other sources, notably NIST, do not consider BLER a definite value as an index of the quality level of the recording media.
- **BERL** should stay below the threshold of 6, that is one cannot accept more than six consecutive sectors with errors.
- **Jitter** should be as low as possible; an acceptable threshold seems to be 35ns.
- Values of **E22** different from 0 should be considered as the indication of a consistent and distributed amount of errors and a limit value not to be trespassed (NIST recommendation).
- When even a single error of **E32** type is detected, the CD is to be considered as *fatally compromised*. If, by good chance, the error is recovered at ECC level, one must immediately make a copy of the CD.

It has been clearly proved that, all conditions being equal, the values of all the above mentioned quality parameters, and most notably of **jitter**, get worse when the recording speed is increased, it is therefore recommended to keep it low, approximately below 8x.

Further practical (good sense) recommendations have been issue by UNESCO within the *Memory of the World Programme* (2006):

- always make at least three copies of every CD (*master*, *safety* and *working*, for access);
- always select CD with different types of dye, to get a second chance (unfortunately Leonardo didn't do that when painting the *Last Supper*...);
- always use single session recording processes (*DAO-Disk At Once*).

3.2.5 Effects of aging on optical recording media

General remarks

As already mentioned, the aging of optical disks may compromise their readability because of several factors:

- chemical reactions may deteriorate the dye layer, and consequently the difference between transparent and opaque segments may become less pronounced;
- the golden or silver layer may become less reflective because of metal oxidation;
- glues that are used to connect the various layers may deteriorate and become opaque to the reading laser beam.

All these factor contribute jeopardize the operation of the basic read/write mechanism which is based on detecting the different way in which the reading laser beam is reflected when it goes through transparent or opaque segments of the dye layer. Therefore, a readable CD may become unreadable since further errors are introduced and, at a given moment there may be so many errors that the error-correcting code is unable to recover them.

Of course, the whole process depends on the quality of the media and of the recording process, as discussed in Section 3.2.4, since the higher is the original quality of the recording the larger are the margins to withstand further errors. But there are further factors influencing the aging process, as for instance the environmental conditions in which the CDs are preserved, and in particular temperature and humidity.

Several interesting studies have been carried out on the subject that give us a qualitative idea of the aging process, but unfortunately we cannot yet rely on clear conclusions and precise statements on the life expectancy of recorded CDs and DVDs.

Since its foundation in 2005, within *OSTA (Optical Storage Technology Association)*, an associations of optical disks manufactures, a specific structure, the *Optical Disc Archival Testing (ODAT) Committee*, is operating with the purpose to define standards for testing high quality optical recording media for digital preservation.

All aging tests that have been suggested so far are based on *accelerated aging techniques* a methodology allows to simulate in experiments that last a matter of hours an aging process of several dozens of years.

Accelerated aging experiments: the NIST/Library of Congress study

A very interesting study, so far the most important and conclusive one, has been performed by the *NIST (National Institute of Standards and Technology)* of the US Department of Commerce and by the Library of Congress to test the longevity of recordable optical media including DVD-R, DVD+R, DVD-RW, DVD+RW, and CD-R. The results of the study have been published in 2007[15] and provide quite valuable estimates of the CD and DVD longevity.

The NIST-LC study considerers only the effects of temperature and relative humidity on the media. The standardized life expectancy is defined for discs maintained at a temperature of 25 ºC and at a

[15] NIST/Library of Congress (LC): *Optical Disc Longevity Study*. Final report, September 2007. http://www.loc.gov/preservation/resources/rt/NIST_LC_OpticalDiscLongevity.pdf

relative humidity level of 50%, but can be used to give an estimate of the life expectancy at any moderate storage condition. Discs exposed to more severe conditions of temperature and humidity are expected to experience a shorter life.

The test plan does not attempt to model degradation due to exposure to light, corrosive gases, contaminants, or mishandling, nor does it account for variations in the playback subsystem. All these further factors play an important role in determining the life expectancy of recorded CDs and DVDs, but can be kept under control if the preservation procedure of the recorded media is conveniently devised and organized.

In the NIST-LC study the *life expectancy* of a disk is defined as the length of time (usually measured in years), that a disc is expected, with 95% confidence, to store data before uncorrectable errors are imminent, when the disc is maintained at the ambient conditions of 25 °C and 50 % relative humidity. More specifically, it is assumed that a CD type media reaches the *End of Life (EOL)*, i.e. the point at which uncorrectable errors actually occur, when **BLER** (see subsection 3.2.4 - **Rating the error levels**) reaches the value of 220. Similarly, for DVD type media, a parameter called **PIE**, analogous to **BLER** is given a limit value of 280.

The aging experiments involved several media products (whose brands were not revealed), specifically seven different products for CD-R media and fifteen different products for DVD media. All CD-R and DVD discs in the study used *Phthalocyanine* dye as the recording layer, therefore the results of the study directly apply only to discs that use this kind of dye.

The experiments are based in the so-called *Eyring model* that gives a way to simulate the effects of aging over a long period of time (years or decades) by subjecting test samples to a sequence of *incubation cycles* (each lasting a few hours), where specific *stress conditions*, i.e. high values of temperature and relative humidity are maintained.

In this specific case *for each product* a sample of 90 discs was used and the related experiment was organized as follows.

- All 90 disks in the sample were recorded with a fixed test pattern. The recording had been organized in such a way to produce *good quality* recording, i.e. with a level of **BLER** (in the case of CD media) below 40 and with a level of **PI** (in the case of DVD) below 100.

- The 90 disk sample was divided in 6 groups of 15 disks each. Each group was subjected to a sequence of incubation cycles *under different stress conditions* (i.e. incubation cycles for different groups had different stress conditions). Values of temperature ranged from 60°C to 80°C, and values off relative humidity from 55% to 85%

- **BLER** (for CD media) or **PI** (for DVD media) were measured at the end of each incubation cycle for each individual disk.

- The measures *for each individual disc in each group* were used to determine the *time to failure* (usually measured in hours) *for that specific disc*.

- The times to failure for all discs in a specific group were fitted to a log-normal distribution to determine the *mean time to failure for the group*, i.e. under that specific stress conditions.

- Time to failure for the 6 groups were regressed against temperature and relative humidity according to an *Eyring acceleration model*. The model was then used to estimate the *distribution of lifetime* of the specific product when stored at the ambient condition of 25°C and 50% relative humidity.

- The distribution of lifetime for the specific product was finally used to determine the *life expectancy for the specific product*, defined as he length of time one can expect, with 95 %

confidence, that a disc is expected to store data before uncorrectable errors are imminent, when the disc is maintained at the ambient conditions of 25 °C and 50 % relative humidity.

All the above steps were repeated for each individual product to be tested. The results of the experiments can be summarized as follows:

- **Recordable CD.** Out of seven CD-R media products tested in the study, five products (71%) had an estimated life expectancy in ambient conditions of more than 45 years. The remaining two products (29%) had an estimated life expectancy in ambient conditions of 30-45 years. The results are shown in Figure 30.

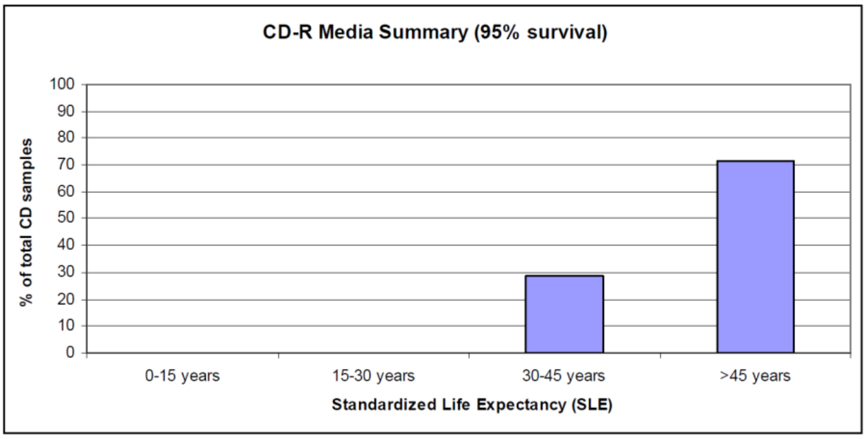


Figure 30: Life expectancy of CD-R

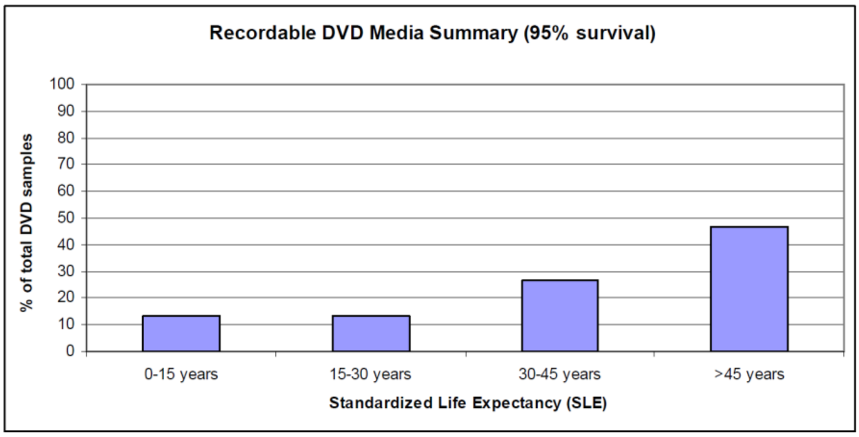


Figure 31: Life expectancy of recordable

- **Recordable and rewritable DVD.** Out of fifteen DVD products tested, including five DVD-R, five DVD+R, two DVD-RW and three DVD+RW types, seven of the products tested (47%) had an estimated life expectancy in ambient conditions of more than 45 years. Four products (37%) had an estimated life expectancy of 30-45 years in ambient storage conditions. Two products (13%) had an estimated life expectancy of 15-30 years and two products (13%) had an estimated life expectancy of less than 15 years when stored in ambient conditions. The results are shown in Figure 31.

- **Recordable DVD by type**. Recordable DVD life expectancy (95 % survival rate) by type, given as a percentage of the total number of samples of that type are shown in Figure 32. According to these results one may conclude that the different types of recordable DVD behave in a quite similar way, including the DVD-RW that one may have expected to behave worse than the other ones.

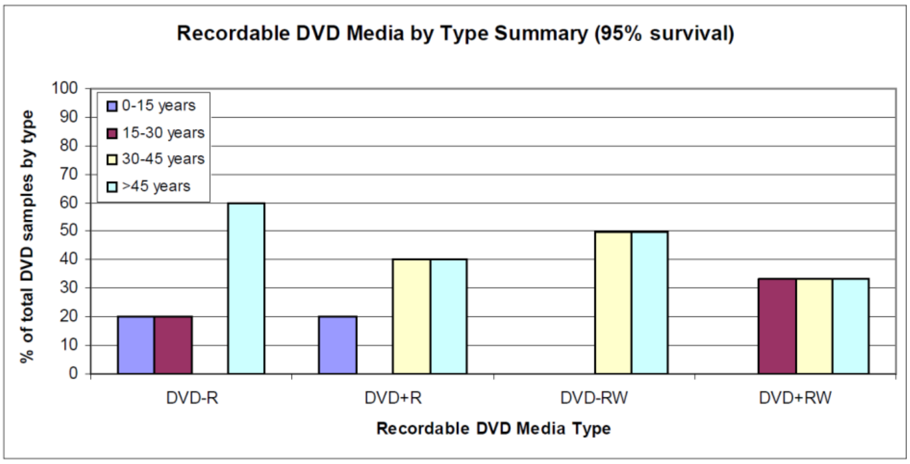


Figure 32: Life expectancy of recordable DVD by type

In order to correctly understand how the result of the NIST/LC study should be interpreted, one should clearly bear in mind the specific conditions under which these life expectancy estimates have been calculated:

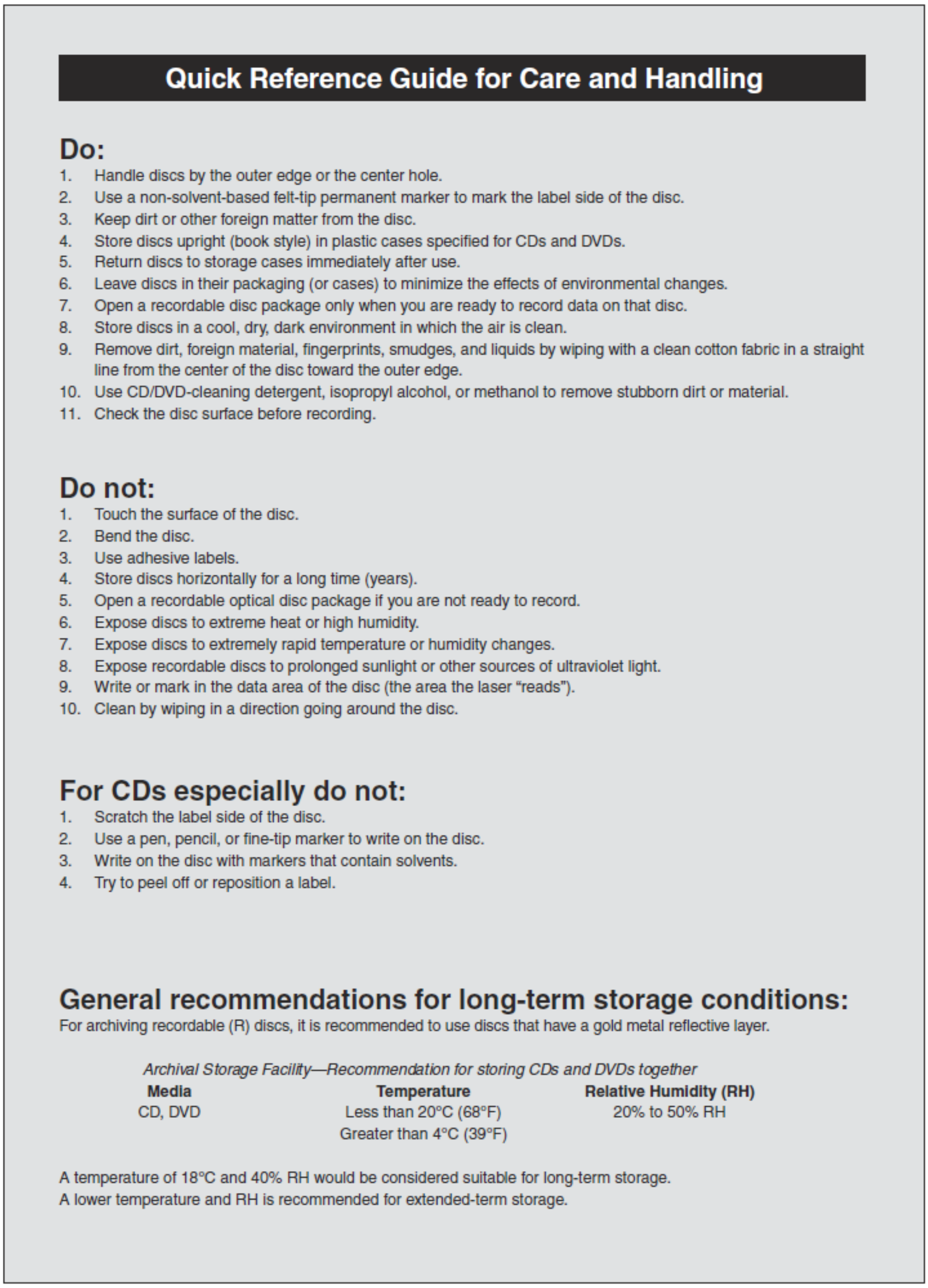
1. all discs were selected among high quality brands and products; life expectancy for products of different quality could be quite different;
2. all discs used *Phthalocyanine* dye as the recording layer; discs that use a different dye may have a different life expectancy;
3. the initial level of error (just after recording) of all discs was low (i.e. **BLER** below 40 or **PI** below 100); the preservation of disks with higher original levels of errors may lead to a considerably shorter life expectancy;
4. the estimated life expectancy values which are the result of the experiments refer to discs there are preserved in optimal conditions, i.e. at 25°C and at a relative humidity of 50%; higher values and/or cyclic fluctuations of these parameters may considerably shorten the life expectancy.

Altogether one may say that the results were not so bad, maybe better than expected. But this is most certainly due to the quality of the products and to low level of errors in the original recordings that provided a very robust starting point and a very large margin to the degradation process. Therefore one should be very careful in extending these results unless one can be sure of having these two factors (and the environmental conditions) under control.

Another interesting result is that CDs perform considerably better than DVDs. This is not surprising and is possibly due to the physical differences in CD and DVD optical media and to the maturity of the technology. In particular, the larger size of the bit markings in CD media not only reduces the relative effect of media or dye degradation, but also means that more stable, less sensitive dye may be used.

Care and handling of recorded optical media

As we have seen, many factors affect the longevity of optical media, including the stability of the materials used in the recording and substrate layers; the relative sensitivity of the recording layer to the effects of heat, humidity and light; the quality of the manufacturing process; the amount and type of markings or labels attached; the compatibility of the media with the recording device and the quality of the initial recording; the age of the media at the time of recording; the storage and handling conditions encountered and more.



Quick Reference Guide for Care and Handling

Do:

1. Handle discs by the outer edge or the center hole.
2. Use a non-solvent-based felt-tip permanent marker to mark the label side of the disc.
3. Keep dirt or other foreign matter from the disc.
4. Store discs upright (book style) in plastic cases specified for CDs and DVDs.
5. Return discs to storage cases immediately after use.
6. Leave discs in their packaging (or cases) to minimize the effects of environmental changes.
7. Open a recordable disc package only when you are ready to record data on that disc.
8. Store discs in a cool, dry, dark environment in which the air is clean.
9. Remove dirt, foreign material, fingerprints, smudges, and liquids by wiping with a clean cotton fabric in a straight line from the center of the disc toward the outer edge.
10. Use CD/DVD-cleaning detergent, isopropyl alcohol, or methanol to remove stubborn dirt or material.
11. Check the disc surface before recording.

Do not:

1. Touch the surface of the disc.
2. Bend the disc.
3. Use adhesive labels.
4. Store discs horizontally for a long time (years).
5. Open a recordable optical disc package if you are not ready to record.
6. Expose discs to extreme heat or high humidity.
7. Expose discs to extremely rapid temperature or humidity changes.
8. Expose recordable discs to prolonged sunlight or other sources of ultraviolet light.
9. Write or mark in the data area of the disc (the area the laser "reads").
10. Clean by wiping in a direction going around the disc.

For CDs especially do not:

1. Scratch the label side of the disc.
2. Use a pen, pencil, or fine-tip marker to write on the disc.
3. Write on the disc with markers that contain solvents.
4. Try to peel off or reposition a label.

General recommendations for long-term storage conditions:

For archiving recordable (R) discs, it is recommended to use discs that have a gold metal reflective layer.

Archival Storage Facility—Recommendation for storing CDs and DVDs together

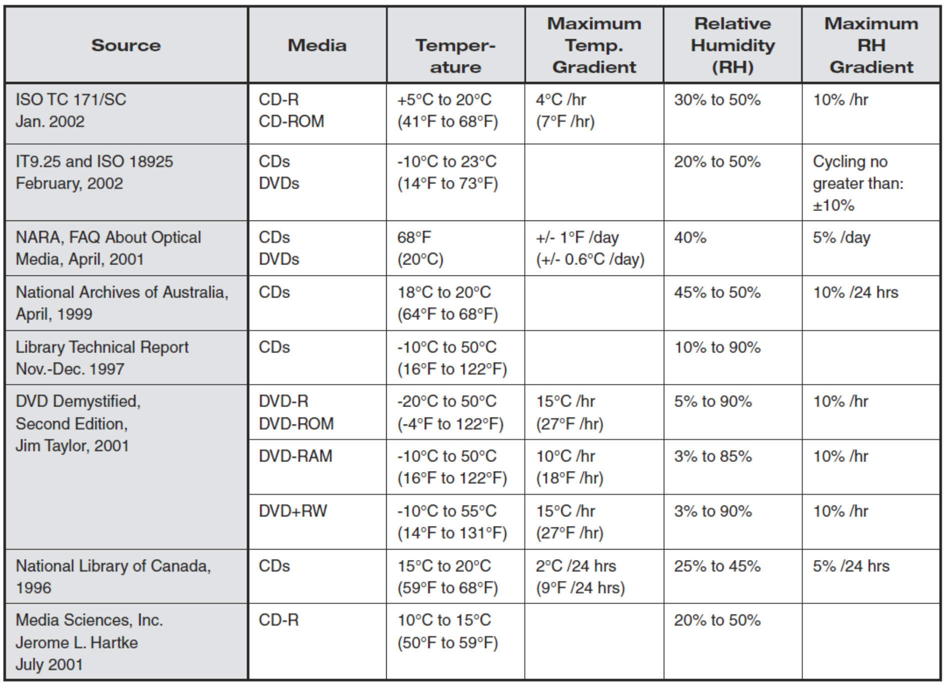
Media	Temperature	Relative Humidity (RH)
CD, DVD	Less than 20°C (68°F) Greater than 4°C (39°F)	20% to 50% RH

A temperature of 18°C and 40% RH would be considered suitable for long-term storage.
A lower temperature and RH is recommended for extended-term storage.

Figure 33: Quick reference guide from the NIST publication

All these factors are important in determining the actual life of a particular disc and there are obviously great variations possible within each factor. It is not feasible to model all of the factors and this study does not attempt to.

A quite complete guide to the care and handling of CDs and DVDs is provided in another NIST publication[16], a guide intended for archivists and librarians whose lecture is highly recommended.



Source	Media	Temper-ature	Maximum Temp. Gradient	Relative Humidity (RH)	Maximum RH Gradient
ISO TC 171/SC Jan. 2002	CD-R CD-ROM	+5°C to 20°C (41°F to 68°F)	4°C /hr (7°F /hr)	30% to 50%	10% /hr
IT9.25 and ISO 18925 February, 2002	CDs DVDs	-10°C to 23°C (14°F to 73°F)		20% to 50%	Cycling no greater than: ±10%
NARA, FAQ About Optical Media, April, 2001	CDs DVDs	68°F (20°C)	+/- 1°F /day (+/- 0.6°C /day)	40%	5% /day
National Archives of Australia, April, 1999	CDs	18°C to 20°C (64°F to 68°F)		45% to 50%	10% /24 hrs
Library Technical Report Nov.-Dec. 1997	CDs	-10°C to 50°C (16°F to 122°F)		10% to 90%	
DVD Demystified, Second Edition, Jim Taylor, 2001	DVD-R DVD-ROM	-20°C to 50°C (-4°F to 122°F)	15°C /hr (27°F /hr)	5% to 90%	10% /hr
	DVD-RAM	-10°C to 50°C (16°F to 122°F)	10°C /hr (18°F /hr)	3% to 85%	10% /hr
	DVD+RW	-10°C to 55°C (14°F to 131°F)	15°C /hr (27°F /hr)	3% to 90%	10% /hr
National Library of Canada, 1996	CDs	15°C to 20°C (59°F to 68°F)	2°C /24 hrs (9°F /24 hrs)	25% to 45%	5% /24 hrs
Media Sciences, Inc. Jerome L. Hartke July 2001	CD-R	10°C to 15°C (50°F to 59°F)		20% to 50%	

Figure 34: Recommended storage parameters from different sources

The guide contains a series of practical recommendations that are well summarized in the quick reference guide which is here reported in Figure. Moreover a full account of the storage temperature and relative humidity ranges recommended in various technical sources is given, which is reported here in Figure 34.

3.2.6 Guidelines for the use of optical storage media

According to the discussion in the previous sections, recordable CDs and DVDs are potentially a decent storage media for digital preservation, provided that they are properly managed and preserved. Moreover we have seen that the life expectancy of optical recordable media depends basically on three main factors:

1. the quality of the recording medium;
2. the quality of the recording process;
3. the effects of aging.

Importantly, for the first two factors we have analysed, in Section 3.2.4, how that they can be measured and controlled by checking the error levels when testing for quality the recording medium, and the individual discs after they are recorded.

[16] Byers, F. *Care and Handling of CDs and DVDs: A Guide for Librarians and Archivists*, National Institute of Standards and Technology and the Council on Library and Information Resources, 2003. http://www.itl.nist.gov/iad/894.05/docs/CDandDVDCareandHandlingGuide.pdf

As for the effects of aging we have a general understanding of them thanks to the NIST/LC study that we have discussed in Section 3.2.5, and, even if they cannot be accurately predicted, we may at least formulate some approximate predictions. Furthermore, the effects of aging can be kept under control by monitoring the change in the error levels as time progresses.

Therefore, if we decide to use this storage media, we may even achieve high reliability levels, provided that we manage to properly organize the storage procedure as a *quality process*. As in any quality process, it is very important to carefully define the procedure in all details, and to clearly state the rules that must be followed for its proper implementation and the controls that should be performed.

In this section we will discuss how a well devised storage procedure should be organized. The procedure is described for CD media, but it can be easily extended to DVD media, *mutatis mutandis*.

Step 1 - Defining the quality levels and the criteria for quality controls

- Define for each of the four parameters **BLER**, **E12**, **E22**, **E32** and **jitter** four different *maximum thresholds*:
    - the first threshold is to be used for quality control in when provisioning (step 2);
    - the second threshold is to be used for quality control after recording (step 3);
    - the third threshold is to be used for periodical quality control (step 4);
    - for **E22** and the **E32** all three threshold should be 0;
    - for all the others parameters the first threshold should be lower the second threshold and the second should be lower than the third.
- Define the criteria that should be used for the construction of the samples, for the controls after recording (step 3) and of the periodical controls (step 4):
    - a sample should be formed by *Homogeneous Sampling Sets (HSS)*, i.e. should be representative of a set of disks from the same manufacturer batch, and recorded under the same conditions;
    - each HSS should be selected by uniformly sampling the corresponding disk set with respect to the recording order;
    - the number of disks in each HSS should be at least 2% of the number of disks in the set, with a minimum of 5.
- Define the *minimum and maximum thresholds for the temperature and the relative humidity* in the disk storage areas, and the *maximum interval of time* for which it is accepted that the thresholds are crossed;
- Define the *maximum recording speed* to be used;
- Define the *rate of periodical controls* (see step 4);
- Define the *maximum preservation time* for the recorded disks. After that time new copies of the disks should be produced regardless to the outcome of the quality controls.

Step 2 - Provisioning

- Only dealers and manufacturers of well-known reliability should be considered.

- Only products with high quality labelling should be considered.

- Buy only large *Manufacturing Batches (MB)* of each product to allow an accurate quality control of the recording medium.

- Each MB should be accurately sampled and controlled according to the sampling criteria and the thresholds defined in step 1.

- If the threshold is crossed even for a single parameter and a single disk, the whole MB should be discarded as not fit for preservation, and possibly used for other purposes.

- Details of the MB, including the average and maximum values for **BLER, E12, E22, E32** and **jitter** for all the disks in the sample, should be recorded in the *Preservation Repository Monitoring DataBase (PRMDB)*.

Step 3 - Recording the disks

- All disks should be recorded in two copies. The two copies should be burned by different recording devices, and possibly on disks from two different batches.

- A third copy should be recorded and preserved for access, especially if the archive is expected to have significant levels of access during the preservation. An equivalent option is to have the third copy on an online storage system.

- Before each disk is recorded a hash file should be produced from the content. All the hash files should be preserved in the *Preservation Repository Monitoring DataBase (PRMDB)*.

- After the recording a hash file should be produced by reading each disk and checked against the corresponding hash file in the PRMDB.

- Recording should be performed in batches, each being homogeneous with respect to the quality of the media and the recording conditions. *Recording Batches (RB)* should be numbered and the *RB number* should be associated to each disk in the PRMDB.

- A sample should be selected for each RB, according to the criteria defined in step 1, and **BLER, E12, E22, E32** and **jitter** should be measured for all the disks in the sample.

- All the details about RB should be recorded in the PRMDB, including the MB, the composition of the sample and all the measures for all the disks in the sample.

- If the threshold is crossed even for a single parameter and a single disk in the sample, the whole RB should be discarded and the origin of the problem should be tracked. If the problem is with a recording device it should be recalibrated or replaced.

Step 4 - Periodical controls.

- Periodical controls should be performed with the periodicity defined in step 1 for all the RB in the archive.

- The controls should be performed on the *same sample of disks* that had been selected for the post recording controls (step 3). All the data about the RB and the sample should be recorded in the PRMDB.

- For each disk in the sample all the parameters should be measured and checked against the historical series of values for the given disk recorded in the PRMDB. Besides checking that the values of the parameters are below the thresholds, comparing them against the historical series should allow to control that the regular progress of the aging process.

- If the threshold is crossed even for a single parameter and a single disk in the sample, the whole RB should be discarded, and replaced by a new copy. The origin of the problem should be tracked as well. If the problem is found to be with a recording device it should be recalibrated or replaced.

- If a given RB is discarded, the periodical controls of all the RB that were recorded on disks from the same MB should be anticipated and performed as soon as possible.

- All the data about the periodical controls should be recorded in the PRMDB.

Step 5 - Preservation of the disks

- The two copies of the same disk should be preserved in different preservation areas, and possibly in different locations.

- Temperature and relative humidity in the preservation areas should be accurately monitored; their values should be periodically measured and recorded in the PRMDB.

- If for the temperature and/or the relative humidity in a given preservation area the minimum or maximum threshold is crossed for an interval of time longer than the maximum value defined in step 1, then the periodical controls of all the RB preserved in that area should be anticipated and performed as soon as possible.

- Access to the preservation areas should be restricted and all individual accesses to the preservation copies of the disks should be explicitly authorized and recorded in the PRMDB.

Step 6 - Periodical copies

- A copy of all the disks in each RB should be produced within the maximum preservation time defined in step 1, unless the copy of the RB had been anticipated (see step 4 and step 5);

- When the copy is produced the two copies of each disk should be joined, to check their integrity between them and with the hash file of the original content which is preserved in the PRMDB.

- The copy is performed, periodically checked and preserved according to the rules in the previous steps.

Normative references

- ISO 18927:2002, Imaging materials – Recordable compact disc systems – Method for estimating the life expectancy based on the effects of temperature and relative humidity, first edition

- ECMA-379, 2007, Test Method for Estimation of the Archival Lifetime of Optical Media
- IEC 60908 (1999-02): Compact disc digital audio system. This document including amendments approximates the Philips-Sony Red Book
- ISO/IEC 10149:1995, Read-Only 120 mm Optical Data Disks (CD-ROM)
- EMCA130 2nd Edition - June 1996, Data interchange on read-only120 mm optical data disks(CD-ROM)
- Orange Book, part B, - Recordable Compact Disc System, November 1990 (SONY and Philips Corp.)
- ISO/IEC 16448:2002: Information technology, 120 mm DVD -- Read-only disk
- ISO/IEC DIS 23912:2005: Information technology, 80 mm (1.46 Gbytes per side) and 120 mm (4.70 Gbytes per side) DVD Recordable Disk (DVD-R)
- ISO/IEC 17344:2005: Information technology, Data interchange on 120 mm and 80 mm Optical Disk using +R format - Capacity: 4.7 and 1.46 Gbytes per side
- ISO/IEC 17342:2004: Information technology, 80 mm (1,46 Gbytes per side) and 120 mm (4,70 Gbytes per side) DVD re-recordable disk (DVD-RW)
- ISO/IEC 17341:2005: Information technology -- Data interchange on 120 mm and 80 mm optical disk using +RW format - Capacity: 4,7 Gbytes and 1,46 Gbytes per side
- IEEE 101-1995, Guide for the statistical analysis of thermal life test data

4 NEXT STEPS AND PLANNED ACTIVITY

In the previous chapters of this report we have documented the activity that has been carried out in WP23 to analyse the storage solutions currently in use at some WP23 partners' premises. Moreover, in section 3 studies are presented on technical issues and the reliability of storage media, an important topic, since, according to the experience of our preliminary analysis, an adequate quantification of the risks and the definition of appropriate management and monitoring policies are quite seldom undertaken.

The analysis of storage solutions and of storage management policies allowed us to gain a better understanding of the problems, and to select a set of relevant issues which need further investigation.

The outcome of this first phase has indeed provided the appropriate framework for a second round of analysis, that we are currently carrying out on a wider scope, that involves a substantially larger number of digital preservation repositories.

The tool to support this investigation is an on-line questionnaire. WP23 will invite all project partners to fulfil the questionnaire, now in preparation. Replies will allow WP23 partners to draw an accurate picture of a number of storage practices and digital preservation.

The questionnaire in point will be published together with questionnaires from WP21 'Preservation Services' and WP27 'Scalability', for the very similar nature and synergy of activity with WP23.

We provide below a preliminary draft structure of the questionnaire to support WP23 topics: questions will be sent out to project partners during March 2013.

1. Global data of the organization managing the repository

- General profile of the repository:
- mission;
- ownership (public/private);
- legal regulations to comply with;
- experience (for how long has the repository been working);
- volumes (total number and total size of preserved digital objects);
- annual quantity (number and size of new digital objects preserved every year);
- expected growth rate (understanding the scalability challenge).

2. Type of preserved digital objects

Profile of the preserved digital objects, possibly several sets of data, if more than one type is preserved:

- description;
- designated community;
- ownership and digital rights;
- access scheme (registered/unregistered users)
- format;

- volume (total number and total size of preserved digital objects);
- annual quantity (number and size of new digital objects preserved every year);
- access rate (number of accesses per object per year);
- retention time (how long digital objects of this type must be preserved);
- legal regulations (specific regulations for this type of object, if any);
- storage solution (referring to the next sections).

3. Storage management policies

- Is there a formalized and declared policy on storage of data in place?
- Key criteria of the policy:
  - redundancy;
  - monitoring;
  - risk assessment;
  - data recovery.

4. In-house storage solutions

Possibly several sets of answers if several storage solutions are managed:

- description;
- purpose (access/preservation)
- type (HD, RAID, CD/DVD, WORM, tape ...);
- experience (for how long this solution has been working);
- lifespan (how long is this solution going to last);
- redundancy (level and mode);
- controls (type and kind of routine controls);
- staff (number of employees devoted to the management of this storage solution);
- cost estimate (total cost per TB/year);
- risk assessment (reliability and availability);
- perceived weakness (and possible improvements).

5. Outsourced storage solutions

Possibly several sets of answers if several storage services are currently purchased:

- description;
- provider;

- type (server farm, private cloud, public cloud ...);
- experience (for how long this solution has been working);
- redundancy (level and mode as declared by the provider);
- terms of agreement (what the provider guarantees: access, security, persistence);
- storage cost (total cost per TB/year);
- access cost (per access and/or per MB transferred);
- risk assessment (reliability and availability);
- perceived weakness.

6. Moving to new storage solutions

Did you consider moving to an outsourced storage solution?

- type (certified storage service, private, private shared, public cloud ...);
- motivation (cost, reliability, availability, scalability);
- fears:
    - lack of control;
    - fear of unauthorized access to data;
    - fear of loss of data 1: technical problems;
    - fear of loss of data 2: provider bankruptcy;
    - fear of loss of data 3: becoming unable to pay the fees because of financial or cash problems (the service cannot be temporarily shut down);
    - vendor lock-in: problems in moving data to a different provider.

Running that survey we aim at getting a deeper understanding of issues arising and possibly an estimate of their statistical incidence.

The analysis of replies shall serve drawing a set of guidelines and recommendations to combine digital preservation capabilities with storage solutions.

5 CONCLUSIONS AND FINAL REMARKS

This section provides a brief summary of the ideas of best practice currently underpinning ways of supporting digital preservation solutions through storage capabilities, based on the cases presented in this document (Chapter 2 and Chapter 3).

- store your data in geographically different places
- use RAID technology to store your data redundantly
- do not use only one storage technology and medium
- be aware of vendor lock-in
- choose storage technologies according to your requirements (regular user access vs. dark archive), hierarchical storage management could be a good solution
- establish a disaster recovery policy and test this scenario regularly
- the used hardware should regularly checked and monitored
- data integrity should regularly checked for example by performing checksum checks
- implement policies for data storage management
- every change of the data (for example by repair actions) has to be documented (provenance)
- use secure and encrypted connections to protect your data during the transfer

This work will be developed further in D23.2, where results from surveying the community will be presented in detail and the final steps in the APARSEN research process will be undertaken.

The survey prepared (see section 4) has the overarching objective to broaden the community feedback on the topic ‘relevance of storage solutions with respect to digital preservation of the data’, and so to provide a deeper insight on the topic than what already acquired through WP23 partners.

In a second step, the survey is meant to promote APARSEN results on such a topic, preparing the ground for future developments.

6 REFERENCES

[1] ISO 14721:2003, “Space Data and Information Transfer Systems - Open Archival Information System - Reference Model”, Edition 1, February 2003.

[2] CCSDS 650.0-B-1., “Reference Model for an Open Archival Information System (OAIS)” – CCSDS Blue Book, Issue 1, January 2002.

[3] G.M. Pinna, P. Mougnaud, G. Petitjean, C. Demange – “ADAR, a Study for an Advanced Data ARchiving System for Earth Observation” – SpaceOps2004, Montreal, 17-21 May 2004.

[4] Kiemle, S., Schroeder-Lanz, A.-K., Reck, C., Object Query Language – “Enabling Service for Earth Observation Product.

[5] Processing, Access and Dissemination” – Ensuring the Long Term Preservation and Adding Value to the Scientific and Technical Data, PV 2004, Frascati, 5-7 October 2004, ESRIN, ESA, WPP-232, p. 17-24, (2004).

[6] Michael Peterson, Gary Zasman, Peter Mojica, and Jeff Porter. 100 year archive requirements survey. Technical report, Storage Networking Industry Association, January 2007.

[7] SNIA Long Term Retention (LTR) Technical Working Group https://www.snia.org/apps/org/workgroup/ltrtwg

[8] Self-contained Information Retention Format (SIRF) use cases and functional requirements, working draft – version 0.5a, SNIA, September 2010, http://www.snia.org/tech_activities/publicreview/SIRF_Use_Cases_V05a_DRAFT.pdf

[9] P.G. Marchetti, G. Rivolta, S. D’Elia, J. Farres, N. Gobron and G. Mason, “A Model for the Scientific Exploitation of Earth Observation Missions: the ESA Research and Service support”, IEEE Geoscience and Remote Sensing Society Newsletter, Vol. 162, pp. 10-18, March 2012

[10] Simona Rabinovici-Cohen, Mary G. Baker, Roger Cummings, Sam Fineberg, and John Marberg, "Towards SIRF: Self-contained Information Retention Format", Proceedings of the Annual International Systems and Storage Conference (SYSTOR), May 30-June 1, 2011, Haifa, Israel.

[11] Reference Model for an Open Archival Information System (OAIS) - Draft Recommended Standard, CCSDS 650.0-P-1.1 (Pink Book) Issue 1.1,” tech. rep., The Consultative Committee for Space Data Systems (CCSDS), August 2009. http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS6500P11/CCSDSAgency.aspx

[12] Rabinovici-Cohen, S., M. Factor, D. Naor, L. Ramati, P. Reshef, S. Ronen, J. Satran and D. Giaretta, “Preservation DataStores: New Storage Paradigm for Preservation Environments,” IBM Journal of Research and Development, vol. 52, pp. 389–399, July/September 2008. http://www.research.ibm.com/haifa/projects/storage/datastores/papers/rabinovici.pdf

[13] "ENSURE" http://ensure-fp7.eu