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

D23.2 STORAGE SOLUTIONS ANALYSIS REPORT

Document identifier:	**APARSEN**-REP-D23_2-01-1_2
Due Date:	2013-12-31
Submission Date:	2014-05-05 **(for this revised version)**
Work package:	WP23
Partners:	CSC, DNB, BRITISH LIBRARY, ESA, KB, CINI, IBM, CINES
WP Lead Partner:	ESA
Document status	Released
URN	urn:nbn:de:101-20140516144

Abstract: This report provides an integrated view of the options for storage solutions for preservation of digital resources, based on the experience of partners in APARSEN and a survey conducted on the subject.

Delivery Type	Report
Author(s)	ESA, CINI, DNB, KB
Approval	Simon Lambert, David Giaretta (STFC)
Summary	
Keyword List	
Availability	☒ PUBLIC

Document Status Sheet

Issue	Date	Comment	Author
0.1	2013-05-27	Draft first version by ESA, CINI, DNB	A DellaVecchia, V Guidetti (ESA)
0.2	2013-06-13	Inclusion of KB's input	M Ras (KB)
0.3	2013-05-14	Inclusion of CINI's input	S Salza (CINI)
0.4	2013-06-25	Add remarks and conclusions section	A Della Vecchia (ESA)
0.5	2013-06-27	Inclusion of KB's input	M Ras (KB)
1.0	2013-06-30	Format update and finalisation	David Giaretta
1.1	2014-01-10	Revision of deliverable to take account of different perspectives and of findings of 27.1 on scalability	Diana Pasquariello (ESA)
1.2	2014-04-23	Revision of Chapter 3 to provide conclusions and lessons learnt	A Della Vecchia (ESA), K. Kaur (BL)

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. "*D23.1 Storage Solution Summary of Inputs"* (see [1]) identified and compared WP23 partners' needs in terms of digital objects preservation and their available and currently used storage archives and technical solutions. The present document aims at collecting information about gaps of storage solutions adopted by third parties, and hence providing recommendations to address these gaps. A questionnaire has been prepared and sent to the interviewees, asking them to provide information on:

1. Basic information about the institution and the repository
2. Types of digital objects
3. In house storage solutions
4. Outsourced storage solutions
5. Moving to an outsourced storage solution

Analysis of the returned questionnaires suggest to significantly improve the design process, by stressing: *i) clear and formal collection of the requirements* by the management, who is aware of the mission of the repository; *ii)* better *cooperation* between the management and the technical staff in *selecting, designing and evaluating* the storage infrastructure; *iii) quantitative analysis*, the key to understand and improve weak points and to produce reasonable estimates of cost, reliability and availability.

1 INTRODUCTION

APARSEN is a Network of Excellence funded by the EU (2011-2014) with the goal of overcoming the fragmentation of the research and of the development in the digital preservation area by bringing together major European players. Among other activities, APARSEN is investigating the storage solutions currently used in digital archives and preservation repositories, in order to compare the different approaches and technical solutions, and their fitness to preserve digital objects in terms of cost, legal, policy and business needs. Moreover, the project is also analysing the scalability problems that arise from the increase in number and size of preserved digital objects that many repositories are currently experiencing.

This report contains the analysis of results of a survey on digital preservation. It analyses different infrastructures against cost items, reliability, preservation capabilities, and technology lock-in and giving recommendations about adoption of storage solutions that can serve digital preservation.

The questionnaire, used as source of information for the analysis, has been specifically prepared (see ANNEXES 4) and sent to a number of repositories across Europe, having a variety of profiles that range from large repositories of scientific data, to national libraries and national archives, up to minor digital libraries and repositories run by small organizations with a limited budget.

1.1 PARTNERS CONTRIBUTING TO THIS DELIVERABLE

The project partners that contributed to this deliverable: European Space Agency (ESA), Consorsio Interuniversitario Nazionale per l'Informatica (CINI), Deutsche National Bibliothek (DNB) and National Library of the Netherlands (KB) are all WP23 partners.

It is worth noting the very different mandate, work, status, and application domains of the involved WP23 partners. Due to such heterogeneous and complementary competences, similarly to what done in the framework of WP23 part I, a plain and straightforward comparison of the analysis, on the collected questionnaires, is neither realistic nor meaningful. 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 an analysis of the answers to the questionnaires received from actors identified by the involved WP23 partners. In order to permit the reader easy comparison among the following heterogeneous surveys, the chapter adopts a standard structure.

Chapter 3 provides conclusions and some final remarks.

2 RESULTS

The Chapter presents in a systematic way the analysis, carried out by the WP23 part II partners, of the answers to the questionnaire distributed to external relevant actors managing with storage solutions. The following sections, sharing a common structure in order to permit a cross-comparison, provide information about:

1. *General Remarks*
   Generic statistics about the questionnaires (e.g., number of involved institutes/companies, proposed vs addressed questions, relevance of (not) answered questions, quality and completeness of the answers, feedback and/or comments of the interviewees).
2. *Basic information about the repositories*
   Description of institutions/companies involved in the survey (e.g., name, background, expertise, international relevance, needs, description of the repository, etc…), on the basis of the questionnaire section *1-Basic information about the institution and the repository*.
3. *Preserved digital objects and storage solutions*
   Report about digital object the interviewees cope with and about the storage solutions they propose, on the basis of the questionnaire sections *2-Types of digital objects*, *3-In house storage solutions*, *4-Outsourced storage solutions* and section *5-Moving to an outsourced storage solution.* The analysis also includes, where possible, relation among needs, proposed solutions, costs and identified technology limitation.
4. *Analysis and proposal for improvements*
   Identification and proposal about topics of interest, which could be significantly improve the interviewees infrastructures and services, on the basis of the point 3 outcomes.

2.1 EUROPEAN SPACE AGENCY

2.1.1 General remarks

The questionnaire has been submitted to agencies, institutes and private companies with attested expertise in the Earth Observation (EO) domain. It is here assumed that the heterogeneity of the know-how of the involved parties will lead to a better and wider comprehension of the EO domain real needs about storage solutions.

The quality of the collected answers is generally good, but most of the questionnaires were returned with the section *4-Outsourced storage solutions* completely empty. This indicates that even if the outsourced storage solution is an available and mature option, it is not still completely accepted by the EO domain community. The possible reasons will be analysed in the next sections. Nevertheless, for most of the interviewees, the possibility to move to an outsourced storage solution might be an option compatible with their needs.

Table 1 lists the involved parties which have returned the questionnaires, detailing, if any, responses to specific sections.

Sections Participants	*1-Basic Information*	*2-Type of digital objects*	*3-In house storage solutions*	*4-Outsourced storage solutions*	*5-Moving to an outsourced storage solution*
1 - ESA/MMFI	Complete	Complete	Complete	Incomplete (0/15)	Complete
2 - ESA/RSS	Complete	Complete	Complete	Incomplete (0/15)	Complete

3 - VITO	Complete	Complete	Complete	Incomplete (0/15)	Incomplete (0/15)
4 - DLR	Complete	Complete	Complete	Incomplete (0/15)	Complete
5 - INGV	Complete	Complete	Complete	Incomplete (0/15)	Complete
6 - SISTEMA	Complete	Complete	Complete	Complete	Complete
7 - GISAT	Complete	Complete	Complete	Incomplete (0/15)	Incomplete (0/15)
8 - EPISTEMATICA	Complete	Complete	Complete	Complete	Complete

Table 1: Questionnaire sections filled up by involved parties

2.1.2 Basic information about the repositories

This section will provide information about interviewed contacted by ESA, in terms of scope, mission and adopted repository technology solutions.

1. *ESA Multi-Mission Facility Infrastructure (MMFI)*

   In 2003 the European Space Agency 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 [2] and [3]) is named Multi-Mission Facility Infrastructure (MMFI) and forms the common infrastructure over which ESA intends to build its future payload data ground segments. 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:

   - 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. 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.

2. *ESA Research and Service Support (RSS)*

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 (see http://rssportal.esa.int/rss-portal.php). The RSS environment also serves the ESA ground segment harmonisation activities, collecting and classifying ground segment technology development needs.

From Satellite data to value added information, RSS offers specific solutions answering the EO data users community needs. 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.

RSS data repository directly relies on the ESA gound segment, this makes available all the EO products collected in the framework of ESA missions and third party missions, for which an agreement with ESA has been set-up, as well.

3. *VITO, Flemish institute for Technological Research*

VITO is a leading independent European research and consulting centre developing sustainable technologies in the area of energy, environment, materials and remote sensing.

Since 1998, VITO has hosted the image processing, archiving and dissemination centre for the spaceborne SPOT-VEGETATION data. In April 2013, the PROBA-V mission was launched, with the aim of continuing the VEGETATION time series at 1 km resolution, and making available also 1/3km products, with a daily near global base distribution policy. VITO, which developed the user segment, is responsible for the processing, archiving and distribution of all products. In addition to the VEGETATION operational activities, VITO hosts several other processing facilities, which e.g. offer hyperspectral images from the airborne APEX instrument or bio-geophysical parameters in the frame of GIO Global Land.

These heterogeneous processing facilities can all ingest the data into a common multi mission archiving facility which acts as the Long Term Data Archive (LTDA) for data preservation. The VITO LTDA is a generic archiving system that utilizes Hierarchical Storage Management (HSM) software for the data lifecycle management.

As an independent and customer-oriented research organisation, VITO provides innovative technological solutions as well as scientifically based advice and support in order to stimulate sustainable development and reinforce the economic and social fabric of Flanders. In the domain of Earth Observation, the centre tries to enable society to access reliable geo-information, based on remote sensing and in situ data, to monitor our changing environment, at local, regional and continental scale.

4. *DLR, German Aerospace Centre*

   The German Remote Sensing Data Centre (DFD) is an institute of the German Aerospace Center (DLR) with facilities in Oberpfaffenhofen near Munich and in Neustrelitz in the state of Mecklenburg – Western Pomerania. DFD and DLR's Remote Sensing Technology Institute (IMF) together comprise the Earth Observation Center EOC, which has become a center of competence for earth observation in Germany.

   DFD supports science and industry as well as the general public. With its national and international receiving stations DFD offers direct access to data from earth observation missions (e.g., TerraSAR-X), derives information products from the raw data, disseminates these products to users, and safeguards all data in the National Remote Sensing Data Library for long term use. Its geoscience research related to the atmosphere, global change and civil security facilitates access to products and approaches based on remote sensing and consolidates their utilization in scientific and commercial domains. DFD operates thematic user services, in particular the World Data Center for Remote Sensing of the Atmospheric (WDC-RSAT), and the Center for Satellite-based Crisis Information (ZKI).

5. *INGV, Istituto Nazionale Geofisica e Vulcanologia*

   The Istituto Nazionale di Geofisica e Vulcanologia was born in September 1999 through a merger of former Istituto Nazionale di Geofisica, Osservatorio Vesuviano and three other institutions: Istituto Internazionale di Vulcanologia, Istituto di Geochimica dei Fluidi and Istituto di Ricerca sul Rischio Sismico. INGV was meant to gather all scientific and technical institutions operating in Geophysics and Volcanology and to create a permanent scientific forum in the Earth Sciences. INGV cooperates with universities and other national public and private institutions, as well as with many research agencies worldwide. The new institution, currently the largest European body dealing with research in Geophysics and Volcanology, has its headquarters in Rome and important facilities in Milano, Bologna, Pisa, Napoli, Catania and Palermo.

   The main mission of INGV is the monitoring of geophysical phenomena in both the solid and fluid components of the Earth. INGV is devoted to 24-hour countrywide seismic surveillance, real-time volcanic monitoring, early warning and forecast activities. State-of-the-art networks of geophysical sensors deliver a continuous flow of observations to the acquisition centers of Rome, Naples and Catania, were the data are analyzed around the clock by specialized personnel. In addition to being analysed for research and civil defence purposes, the data supplied by numerous monitoring networks are regularly distributed to the public institutions concerned, to the scientific community and to the public. INGV operates in close coordination with the Ministry of University and Research and with Civil Protection authorities, both at national and local level. INGV also cooperates with the Ministry of Environment, the Ministry of Education, the Ministry of Defense and the Ministry of Foreign Affairs in the frame of large research programs of national and international relevance.

   INGV data repository mainly consists in multimission Synthetic Aperture Radar (SAR) data used for seismic and vulcanic deformation studies. The archive is mostly composed of Level 0 (raw data) products, with a lesser amount of Level-1 (focused complex) products. A non exhaustive list of sensors, used in past, and some of them currently in use, include ERS-1 SAR, ERS-2 SAR, ENVISAT ASAR, ALOS PALSAR, COSMO-SkyMed and TerraSAR-X.

6. *SISTEMA*

   SISTEMA is a SME privately-held company mainly focused on the implementation and development of products and services based on satellite observation of the Earth-atmosphere system. It provides a wide range of commercial products and services "off the shelf", based on satellite data analysis for agriculture, soil management, environmental monitoring and cloud and precipitation estimation based on single sensor, multi-sensor, multi-spectral and multi-temporal data analyses.

SISTEMA mainly relies on the data centre of sister company MEEO Srl, Italy, which consists in a relevant collection of multi-sensors EO products (e.g., LANDSAT5/7, MODIS, (A)ATSR, AVNIR-2, etc…).

7. *GISAT*

   GISAT is a SME sized company comprised of small core team and extensive network of supporting multidisciplinary experts for specifically oriented tasks. The educational background of the GISAT team is in agriculture, cartography, geography and natural sciences, land management, surveying and informatics. The highly qualified experts are experienced in project management, consultancy and wide range of remote sensing and GIS applications in different thematic areas.

   GISAT's internal repository, consists in EO images and GIS data sets, collected in the framework of European projects during the last 20 years. It provide its clients with wide range of value added, complete, high quality and 'state-of-the-art' geo-information services based on the Earth Observation technology. Data repository is an essential element of the overall internal structure, and special value is recognised to the Long Term Data Preservation for historical mapping assessment, especially in change detection.

8. *EPISTEMATICA*

   Epistematica is a company providing IT services for designing and developing knowledge-based software systems. Its repository manages the ontologies developed during the last five years projects, both for external clients and for internal R&D activities. The company aims to help companies and institution to maximize the value of their data, by making them accessible as knowledge via the Semantic World Wide Web.

The Table 2 summarised the main features of the surveyed repositories. Most of the involved interviewed have shown an advanced expertise about storage solutions/e-data preservation. They can be preliminarily grouped in public/governmental agencies and private companies. The first ones, on the one hand, typically governmental research institutes or space agencies, have direct access to or directly manage satellite missions. This implies that they cope with big/very big data repositories, for which they apply agreed and well-defined policies about storage solutions (e.g., Long Term Data Preservation initiative, see http://earth.esa.int/gscb/ltdp/). The private companies, on the other hand, might have preferential access to data repository of the agencies. They typically do not strictly apply any particular policy for storage solution, or even none at all. About the access data policy, both public and private parties typically own reserved data collections, restricted to internal authorised users, and public collections, for which a direct access is possible once registered (e.g., https://earth.esa.int/web/guest/missions/content?p_r_p_564233524_assetIdentifier=revised-esa-earth-observation-data-policy-7098).

REPOSITORY	ACTIVITY		POLICY				ACCESS	
	Type	Years of activity	Declared policy	Public link	Regular checks	Data recovery	Local access only	Registered users only
1 - ESA/MMFI	Public	> 20	Y	Y	Y	Y	Y + N	Y
2 - ESA/RSS	Public	5-10	Y	Y	Y	Y	Y + N	Y
3 - VITO	Public	10-20	Y	Y	Y	Y	Y + N	Y

4 - DLR	Public	10-20	Y	Y	Y	Y	Y + N	Y
5 - INGV	Public	5-10	N	N	Y	N	Y	Y
6 - SISTEMA	Private	<5	N	N	N	N	Y	N/A
7 - GISAT	Private	10-20	N	N	Y	Y	Y	Y
8 - EPISTEMATICA	Private	<5	N	N	Y	Y	Y	N/A

Table 2: Main feature of the surveyed repositories

2.1.3 Preserved digital objects and storage solutions

All the interviewed, participating to the survey, are deeply involved into the Earth Observation domain (see Table 3). This implies that the most common type of Digital Objects (DO), they typically work with, is the satellite image. Although the DO in object is conceptually the same entity for all the interviewed parties, it is typically provided in many different formats, depending on the sensor it has been acquired from. A short and non-exhaustive list of the most used formats includes: N1, HDF4/5, ENVI, GeoTIFF, CF-NetCDS, SAFE, and many others. The formats adopted for storing and disseminating metadata and in-situ-data are even more diverse.

A further common aspect, that all the involved interviewed report in the returned questionnaires, is the expected increase of the managed DOs, which they will have to cope with in the next five/ten years. Public/governmental agencies expect an increase higher than 500%, but for the private companies, the increase should be in the range of 100% - 500%. The main reasons can be identified in:

1. *Constellation Configuration*

   Some of the current available missions, and even more frequently by upcoming ones, adopt a constellation configuration. This kind of satellite mission implies a volume of data double (e.g., Sentinel-1/2, Tandem-X/L) or four time bigger (e.g., Cosmo-SkyMed) than a classic mono-satellite mission (e.g., ENVISAT, ERS-1/2).

2. *Systematic Global Acquisition*

   One of the most relevant achievement of Sentinel-1/2/3 missions, in the framework of GMES initiative (see http://copernicus.eu/), is to provide a systematic global acquisition, increasing as consequence the storage resources, same for TerraSAR-X, Tandem-L and Cosmo-SkyMed.

3. *Very High Resolution Sensors*

   In the last decade the sensors incredibly increased in terms of radiometric, spectral and spatial resolutions. Synthetic Aperture Radar (SAR) sensors moved from 20-30 meter spatial resolution (e.g., ASAR ENVISAT) to sub-meter resolution (e.g., Cosmo-SkyMed, TerraSAR-X). Even better are the resolution of optical sensors, which easily get to 45 cm (even lower) spatial resolution. As a consequence, the DOs archive, increase exponentially.

The distinction between public/governmental agencies and private companies is clearly evident here also, when the questionnaire answers related to total amount of data are processed, both in term of volume and size. The first ones, providing Payload Data Ground Segment (PDGS) infrastructure, need of extremely capacious storage solutions, able to host a huge volume of data (as long as possible for any possible future reuse of them). Furthermore, acting under the auspices of the Long Term Data Preservation initiatives (see http://earth.esa.int/gscb/ltdp/), they must ensure a quite long retention period. On the other hand private companies typically host very specific datasets, with specific areas of interest and times of interest, acquired from a few complementary sensors, making the storage solution issue not a critical issue. This also motivate the shorter retention period.

DOs versioning is a transversal issue. On the one hand, referring to EO data Level 0 and 1, the availability of several versions typically depends by specific activities such as calibration/validation (e.g., http://calvalportal.ceos.org/cvp/web/guest), performed by agencies and/or data provider. On the other hand, the availability of several versions for Level 2 data (Value Added Products – VAPs), may instead depend by the availability of alternative scientific algorithms (e.g., retrieval of soil moisture, land temperature, vegetation index from satellite radiometry), which can be developed and/or made available by public/governmental agencies, universities, private companies and scientists.

REPOSITORY	DIGITAL OBJECTS					
	Retention period	Total number	Total size	Yearly increase	Size of metadata	Versions
1 - ESA/MMFI	> 20 y	> 10^6	> 1 PB	> 100%	5% - 10%	> 5
2 - ESA/RSS	10 – 20 y	> 10^6	< 1 PB	20% - 100 %	< 5%	> 5
3 - VITO	10 – 20 y	> 10^6	< 1 PB	> 100%	5% - 10%	1
4 - DLR	10 – 20 y	> 10^6	> 1 PB	> 100%	< 5%	1 - 5
5 - INGV	> 10 y	< 10^5	< 100 TB	20% - 100 %	< 5%	1
6 - SISTEMA	5 – 10 y	10^5 - 10^6	< 100 TB	> 100%	5% - 10%	> 5
7 - GISAT	5 – 10 y	< 10^5	< 100 TB	> 100%	< 5%	1 - 5
8 - EPISTEMATICA	< 5 y	< 10^5	< 100 TB	20% - 100 %	< 5%	1

Table 3: Volumes and types of preserved digital objects

Table 4 summarises the answers collected in sections *2-Type of digital objects* and *3-In house storage solutions* of the questionnaire. On the one hand, private companies, which have to manage with small internal data repository, adopt easier solutions based RAID 1 configuration, permitting to both contain the costs and the maintenance of the infrastructure. On the other hand, public/governmental agencies, whose mission is also acting as data provider, put in place much more complex data storage infrastructures. Data access and (bit) preservation are typically ensured by more expensive RAID 5/6 configurations, and backups are periodically performed on TAPE media, permitting data redundancy at each layer of the infrastructure. Furthermore, in line with Long Term Data Preservation (LTDP) guidelines, with the purpose of minimising as much as possible the possibility of losing stored data, ESA, VITO and DLR also perform geographical backup (on local facility and on remote locations as well). These archives are also periodically tested, in order to be sure about restoring procedures (e.g., check of tapes readability, checksum tests and periodic replacement of the tapes).

About the backup period, it can be said that this is strictly dependent on the data the agency must cope with (e.g., level of EO data). Satellite EO data acquisitions (e.g., Raw and Level 0) typically do not need to be frequently reprocessed, hence an incremental backup is sufficient. On the other hand, level 1 and 2 of EO data, and Value Added Products VAP (e.g., geophysical parameters retrieved by

satellite measurements), might change more frequently in time, due to the availability of new and more enhanced scientific algorithms.

None of the interviewees provided information about Cost, availability and reliability of the infrastructures.

REPOSITORY	STORAGE FOR PRESERVATION		STORAGE FOR ACCESS		BACKUPS			COST & RELIABILTY		
	In house	Type	In house	Type	In house	Type	Backup period	Cost TB/y	Availability	Reliability
1 - ESA/MMFI	Y	TAPE	Y	RAID5/6	Y	TAPE	ns	ns	ns	ns
2 - ESA/ RSS	Y	RAID5/6	Y	RAID5	N	ns	ns	ns	ns	ns
3 - VITO	Y	RAID5/ TAPE	Y	RAID5	Y	TAPE	1-7 days	ns	ns	ns
4 - DLR	Y	TAPE	Y	RAID5	Y	TAPE	ns	ns	ns	ns
5 - INGV	Y	RAID1	Y	RAID1	Y	RAID1	> 180 days	ns	ns	ns
6 - SISTEMA	Y	RAID1	Y	RAID1	N	ns	ns	ns	ns	ns
7 - GISAT	Y	RAID1	Y	RAID1	Y	RAID1	1-7 days	ns	ns	ns
8 - EPISTEMATICA	N	Ns	Y	RAID1	Y	RAID1	7-30 days	ns	ns	ns

Table 4: Storage solutions, cost and reliability (ns: not specified)

ESA/MMFI adopts for all its distributed Facility Ground Segment components a Hierarchical Storage Management system (HSM). DLR, as ESA/MMFI, proposes a COTS Hierarchical Storage Management (HSM) system for the storage management (see http://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-6888/11391_read-22884/). Both the agencies adopt a tiered storage solution. All the remaining ones rely on in-house storage solutions, which, for advanced and complex architectures, may imply higher costs of maintenance and needed effort (e.g., VITO).

None, apart from one, of the public/governmental agencies entities put in place, neither in the past nor in the present, an outsourced storage solution. ESA/RSS is the only one that occasionally uses external storage solutions, to cope with unusual and unplanned peak of data to preserve. The possibility to

improve reliability, availability and scalability would be the most tempting factors which would justify a migration. Anyway, the burden of moving high data volumes to the external resources over the WAN, the possibility of losing control of the data and the fear of unauthorised access to them, seem to be the main reason because this solution is not broadly adopted. Among the private companies, both SISTEMA and EPISTEMATICA adopt outsourced storage solutions, with the aim of reducing maintenance costs and increase reliability and availability.

2.1.4 Analysis and proposals for improvements

The returned questionnaires have shown interesting information, confirming some reasonable assumptions about the storage solutions in the context of EO domain. The survey shows that mission, scope and available budget, of an organisation, largely determine the data storage solutions adopted.

Public/governmental agencies, whose main mission is managing EO missions and related ground segments (e.g., ESA/MMFI, DLR, VITO), make use of the most advanced technologies for storage solutions nowadays available. They take care about the data preservation policy as well, for example the LTDP initiatives at European level (http://earth.esa.int/gscb/ltdp/), representing in the domain of the EO, the state-of-the-art about storage solutions. The survey points out that for this specific mission the outsourced storage solution does not represent a viable option. Main reasons are: a) fear of unauthorised access to data repositories, b) it is needed the direct control of the data repositories, c) the cost for renting huge external repositories would be comparable or even higher than putting in place specific in house solutions. None of them provided information about reliability, availability and costs. These are information that should be easily formulated by these kind of involved entities, and made available to end user.

ESA/RSS and INGV, even if still part of the public/governmental agencies group, have limited mission and scope. They respectively support external scientific research initiatives, and perform scientific activities on very specific thematic area. As a consequence, storage solution requirements are not so stringent as for previous entities. Typically, a tread-off between available technologies and budget is a priority. Here the outsourced storage solution might be an effective alternative in case of unplanned peaks of data to preserve, for which, the purchase of new hardware would not be justifiable in the long period. 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.

Private companies (e.g., SISTEMA, GISAT, EPISTEMATICA), whose mission is to make business with the EO data, typically offer high level services to the end-user. For this purpose, they do not need very capacious and/or sophisticated data repositories. Here an outsourced storage solution might be the right choice (at least temporary), as it would permit the private company to rent on-demand resources, reducing both management and personnel costs.

2.2 DEUTSCHE NATIONAL BIBLIOTHEK

2.2.1 General remarks

The investigation carried out by DNB has directly addressed 5 institutional repositories:

- 3 have answered and returned the questionnaire;
- 2 did not answered;
- 1 institution that was addressed indirectly (by using a mailing list), has doubt about fitting into the questionnaire.

All repositories are managed by public administrations and are preserving records in the areas e-Government and Cultural Heritage (CH) on regional and national level. The quality of answers has been generally good. But the completeness as shown in Table 5 differs greatly. In the case of incompleteness the ratio of answered/unanswered questions is indicated. Furthermore the table indicates a section as complete even though some questions about costs have not been answered.

Sections Participants	*1-Basic Information*	*2-Type of digital objects*	*3-In house storage solutions*	*4-Outsourced storage solutions*	*5-Moving to an outsourced storage solution*
1 - Bundesarchiv	Complete	Incomplete (0/15)	Incomplete (0/13)	Incomplete (0/9)	Incomplete (0/4)
2 - HBZ	Complete	Complete	Complete	Incomplete (6/9)	Complete
3 - State Archives of Hessia	Complete	Incomplete (8/15)	Complete	Incomplete (7/9)	Incomplete (2/4)

Table 5: Completeness of the survey

Section *5-Moving to an outsourced storage solution* is considered as complete even though an outsourcing solution is not planned. In one case only section *1-Basic Information* has been given attention. That may be because page 4 of the questionnaire is over 90% not used. So the survey participant probably had the impression that the questionnaire ends on that page.

2.2.2 Basic information about the repositories

The Survey carried out by DNB concerns the following repositories:

4. *Bundesarchiv*

   The Bundesarchiv houses records from across the German central government and, in smaller numbers, from the central courts with the legal responsibility of permanently preserving it and making them available for use. The Bundesarchiv is home to 300 kilometres of paper documents, 12 million images and about 10 million born digital files that cover about 200 years of German history, in the case of the born digital files about 40 years. The Bundesarchiv is the central advisory body on the care of records and archives, in all media, from creation to long-term preservation.

5. *Hochschulbibliothekszentrum (hbz)*

   edoweb is a repository for online resources (electronic documents and web sites) with a regional background. The software was developed by the Hochschulbibliothekszentrum Nordrhein-Westfalen (hbz). Web resources for the archive are selected and collected by the Landesbibliothekszentrum / Rheinische Landesbibliothek Koblenz. Special emphasis is laid on PDF-documents and websites which are either legal deposit materials or relevant for the history and culture of Rhineland-Palatinate. The web resources are stored on the project's own servers; long term preservation is intended. The resources are catalogued and can also be retrieved from the edoweb system itself.

6. *State Archives of Hessia*

   The State Archives of Hessia (Hessisches Landesarchiv) are responsible for the preservation of documents produced by the administration of Hessia. This includes digital born documents or - in a broader sense - all kinds of digital information that could be of historical value. The "Digital archives of Hessia" (Digitales Archiv Hessen), that are part of the State Archives of Hessia, is responsible for the ingest and preservation of those materials (in an archival sense: for ever).

The main features of the surveyed repositories are summarized in Table 6. One participant has more than 10 years of experience. All have a declared but not public policy on storage. All performs regular integrity checks and only one has no established and tested workflows for data recovery. In one case the collection is accessible under multiple conditions (Y+N). That means for example that there are types of documents which are unlimitedly accessible and others are only accessible for registered users.

REPOSITORY	ACTIVITY		POLICY				ACCESS	
	Type	Years of activity	Declared policy	Public link	Regular checks	Data recovery	Local access only	Registered users only
1 - Bundesarchiv	e-gov	10-20	Y	N	Y	Y	--	--
2 - HBZ	CH	5-10	Y	N	Y	Y	Y + N	Y + N
3 - State Archives of Hessia	e-gov	<5	Y	N	Y	N	Y	Y

Table 6: Main features of the surveyed repositories

2.2.3 Preserved digital objects and storage solutions

Table 7 shows volumes and types of preserved digital objects of the surveyed repositories.

REPOSITORY	DIGITAL OBJECTS							
	Retention period	Total number	Total size	Yearly increase	Size of metadata	Type	Formats	Versions
1 - Bundesarchiv	--	> 10^6	< 100 TB	< 10%	--	born digital	--	--
2 - HBZ	> 10 y	< 10^5	< 100 TB	< 10%	< 5%	electronic documents, web sites, e-prints, electronic theses/dissertations	HTML, Images, PDF	> 5
3 - State Archives of Hessia	--	< 10^5	< 100 TB	< 10%	< 5%	electronic documents (born digital), photographs, audios, videos, statistical microdata, extracted information from databases	PDF, txt, csv, xml, mpeg, tiff, svg and other	1-5

Table 7: Volumes and types of preserved digital objects

With more than 1 million preserved objects the Bundesarchiv stands out. But for all the total size is lower than 100 TB. In general type and format of the digital objects are typical for the corresponding collection.

The expected increases of digital objects are shown in Table 8. In general the participants expect an increase from 100% to 500% for the near future. Afterwards they guess that this amount will increase to 1000% and more. One indicates that the required storage size for that time will grow at a slower rate.

REPOSITORY	Increase in the number		Increase in the total size	
	next 5 years	next 10 years	Total next 5 years	next 10 years
1 - Bundesarchiv	100%-500%	> 1000%	100%-500%	> 1000%
2 - HBZ	100%-500%	> 1000%	100%-500%	100%-500%
3 - State Archives of Hessia	100%-500%	100%-500%	100%-500%	100%-500%

Table 8: Expected increase of preserved objects

The following table, which is based on data from sections *2-Type of digital objects* and *3-In house storage solutions* of the questionnaire, summarizes the structure of the storage systems. Unfortunately only two participants have been given attention for this topic.

REPOSITORY	STORAGE FOR PRESERVATION		STORAGE FOR ACCESS		BACKUPS			COST & RELIABILTY		
	In house	Type	In house	Type	In house	Type	Backup period	Cost TB/y	Availability	Reliability
1 - Bundesarchiv	--	--	--	--	--	--	--	--	--	--
2 - HBZ	Y	HD/ RAID1	same	HD/ RAID1	Y	HD, Tape	7-30 days	ns	7	7
3 - State Archives of Hessia	Y	RAID6	same	same CD/ DvD	Y	RAID	7-30 days	250€ + 1250€	ns	ns

Table 9: Storage solutions, cost and reliability (ns: not specified)

The figures for availability and reliability seem to be not realistic, possibly a misunderstanding of using a scale as a power of ten.

As outsourced storage solutions were indicated

- the German LOCKSS network as an private shared cloud (jointly managed by a consortium of several repositories)
- the Hessische Zentrale für Datenverarbeitung (HZD) as an IT-Service-provider for the administration in Hessia

In both cases the staff that is devoted to the management of the storage system was lower than 5. One indicates costs for one TB of 250€ for the in house storage and 1250€ for an outsourced storage solutions. One indicates fears of using an outsourced solution. The given reasons are:

- lack of control,
- unauthorized access and
- loss of data because of a provider bankruptcy.

One indicates weaknesses concerning the currently used in house storage solution. An indicated improvement is the extension of backup generations (week, month, year).

2.2.4 Analysis and proposal for improvements

The following **conclusions** can be drawn:

- The interest to use outsourced storage solutions exists, but fears lead to the fact that repositories prefer to relay on in house solutions.
- All participants are aware that is important to have an institutional, formal and clear statement (policy) for storing their data.
- It seems to be a best practice to preserve more than one distinct version for each digital object.
- It has to be noted that outsourced storage solutions are used only as third backup while storage for preservation, access and backup is still managed in house.
- All given solutions are based on redundancy and store their data geographically distributed.
- The majority consider regular integrity checks as fundamental for preserving the bit stream.
- It should be investigated further whether the difference of costs (of 1000€) between using an in house and outsourced solution is typical.

The following proposals for **improvements** should be considered:

- Backups should be generated on a daily to weekly basis instead in the worst case on a monthly basis.
- Regarding to the policy it is recommended to include also statements about workflows for data recovery.
- The repositories and IT-service provider should be able to indicate figures for reliability and availability of their systems.
- Furthermore figures for costs should be identifiable.

2.3 CONSORSIO INTERUNIVERSITARIO NAZIONALE PER L'INFORMATICA

2.3.1 General remarks

The investigation carried out by CINI has involved 9 different repositories, and all but one have answered and returned the questionnaire. All the surveyed repositories are managed by Public administrations and institutions, and are mostly preserving digital records in two different areas: e-Government and Cultural Heritage (CH). All but one have agreed to a full disclosure of the information they have provided. With the one that has asked us for non-disclosure (number 3 in Table 10) we have agreed that the repository should only be denoted in our reports by means of a generic description, which, nevertheless, gives a clear idea of the profile and of the area of activity.

The quality of the answers has been generally good and the data provided interesting, but in most cases questionnaire sections *4-Outsourced storage solutions* and *5-Moving to an outsourced storage solution* have been given little or no attention. This could be acceptable for sect. 4, as one may not currently adopt such a solution, but everyone was expected to answer sect. 5, which is just meant to investigate the interviewee's attitude toward these new interesting possibilities. When compelled, some have

finally answered, but they have mostly roughly dismissed the possibility of outsourcing their storage, as if a wise person should not even consider the idea. We shall discuss later this prejudice.

The implementation of all the surveyed repositories has been strongly influenced by the Italian regulations on the preservation of digital records, which were issued in 1998 and that became in 2001 mandatory for all Public administrations, as well for as private companies that preserve legally relevant records, such as invoices, balance sheets etc. These regulations, that unfortunately have not been updated in the last decade, demand that the records are preserved by collecting them in large *preservation volumes*, digitally signed and time stamped. The consequence of this has been twofold: *i)* on the negative side, the general attitude is that the main (only) thing to care about is to formally comply with the regulations, regardless of the actual robustness of the implementation; *ii)* on the positive side, one generally needs to maintain separate copy of the records for access, since providing access to the preservation volumes (large digital mummies) is often unrealistic. Moreover, the regulations unhappily mention, as the primary choice for storage media, optical disks, a notoriously unreliable solution.

2.3.2 Basic information about the repositories

Our survey concerns the following repositories:

1. *Toscana regional preservation repository, Firenze*

   Toscana is a large Italian region (administrative district) with a population of almost 4 million and includes about 300 municipalities. As with a few other regions in Italy, Toscana has decided to set up a single large preservation repository at regional level, which is meant to provide preservation as a service to all the branches of the regional administration (including the Health Care System), as well as to all the local administrations within the region (10 provinces and about 300 municipalities). Beside other problems, such an approach has been compelled by the strict and rather complex Italian national regulations on the preservation of digital records that make it difficult and unreasonable to efficiently manage such a repository on a small or medium scale. The individual administrations remain owners of their own records, and are granted continuous access to them.

2. *Local e-government preservation repository, undisclosed*

   As the previous one, and according to the same principles, this repository provides preservation as a service to all local administrations within an Italian district with a population of 500.000 inhabitants that includes about 200 municipalities. The repository is actually part of a large ERMS designed to support all the e-gov activities within the district. Most of the records relate to paperless administrative e-gov procedures, and largely rely on digital signatures and the Electronic Registered E-mail Service (PEC), two procedures to which formal regulations grant full legal value in Italy.

3. *Preservation repository of Vicenza Public Health Care System, Vicenza*

   This repository preserves the medical records (test results, physicians' reports, etc.) from all the Health Care facilities in the province of Vicenza (about 900.000 inhabitants), according to the Italian National regulations, both the ones on the preservation of medical records (very strict), and the more general ones on the preservation of digital records. Records are in a variety of formats: DICOM formats for test results, and PDF and PDF/A with digital signature for medical reports. The preservation of the records starts shortly after their creation, generally within 24 hours. Access to the repository is granted both to the staff and to the patients, limited to their own records.

4. *Registry of Banca d'Italia (the National Bank of Italy), Rome*

   Banca d'Italia is the Italian Central Bank, and is a statutory agency in charge, among other duties, of controlling Italian banks. As every Italian public administration, it is compelled by law to manage all incoming and outgoing mail through the so-called *Electronic Registry System*, that includes the preservation all the mail and the attachments. This is what this repository deals with, according to the above mentioned regulations. All records, including

digitized images from paper, are converted to PDF/A format and then preserved as such. Access to the repository is granted only to registered users within the Bank.

5. *Cultural Heritage preservation repository at Cineca, Bologna*

   This repository is managed by CINECA, a large consortium of Italian Universities whose mission is to provide a variety of ICT services, ranging from supercomputing to the management of digital libraries and preservation repositories. This specific repository provides services to several prominent cultural institutions, to manage and preserve several kinds of digital objects in the Cultural Heritage area. Legal regulations include the ones about DRM, about the preservation of items relevant as Cultural Heritage, and the general ones about the preservation of digital records. Digital objects are in PDF and in a variety of formats for images and digitized text (TIFF, JPEG, GIF, PNG). Access to the preserved objects is then granted to a large audience, including remote and unregistered users.

6. *Preservation repository of Sapienza Digital Library at Cineca, Bologna*

   This repository, which is also managed by CINECA (see above), is connected to SDL (Sapienza Digital Library) a large infrastructure supporting the management and dissemination of all the digital resources of the Università di Roma "La Sapienza", that ranges from the collections of the university libraries and museums to the scientific literature published by the departments. As in the above case legal regulations concern DRM, the preservation of relevant Cultural Heritage items. The variety of digital objects is quite large and includes book scans, images, audio, audio-videos, documents, hi-res images (Tiled Pyramidal images). The variety of formats goes along (JPEG, TIFF, RAW, JPEG2000, MP4, 3GP, PDF, and more). Access is granted, with different rights, to users from both inside and outside the university.

7. *Staging repository of Sapienza Digital Library, Rome*

   This repository is strictly connected to the previous one. Its main purpose is to collect the digital objects from the departments within the university, and to prepare them for the ingestion in the main repository. However, it has been decided that the digital objects that go through it should also be preserved locally, and therefore, appropriate procedures, albeit very limited, have been set up. Legal regulations and formats are just the same specified above for the main repository. Access is granted only to the staff.

8. *Magazzini Digitali at the Intalian National Library, Florence*

   The Central National Library of Florence (BNCF) is the institution that in cooperation with the National Library of Rome and Venice, is devoted implement the National Legal Deposit i.e. a repository were, according to the Law, all documents digitally published in Italy have to be preserved. This includes several kind of documents, including PHD thesis that are harvested form the institutional repositories of Italian universities. "Magazzini Digitali", literally Digital Stacks, is an infrastructure that has been setup by BNCF, with the cooperation of the Rinascimento Digitale foundation, to collect and preserve these documents, and that became operational a few years ago.

The main features of the surveyed repositories are summarized in Table 10. Most of them have only a limited experience: only two have more than 5 years of activity. Only four repositories have a declared policy, not surprisingly the ones in the e-gov area, but none of them has provided a link to a public document stating the policy. In all cases the policy demands for regular integrity checks, but in one case only a workflow for data recovery is specified.

REPOSITORY	ACTIVITY		POLICY				ACCESS	
	Type	Years of activity	Declared policy	Public link	Regular checks	Data recovery	Local access only	Registered users only

1 - Toscana Regional Repository	e-gov	<5	Y	N	Y	N	N	Y
2 - Public e-gov Local Repository	e-gov	<5	Y	N	Y	N	N	Y
3 - Vicenza Health Care System	e-gov	5-10	Y	N	Y	Y	N	Y
4 - National Bank of Italy	e-gov	<5	Y	N	Y	N	Y	Y
5 - Cultural Heritage at Cineca	C.H.	5-10	N	---	---	---	N	N
6 - Sapienza Digital Library (Cineca)	C.H.	<5	N	---	---	---	N	N
7 - Sapienza Digital Library (staging)	C.H.	<5	N	---	---	---	Y	Y
8 - Magazzini Digitali at BNCF	C.H.	<5	Y	N	Y	N	Y	Y

Table 10: Main features of the surveyed repositories

2.3.3 Preserved digital objects and storage solutions

All the surveyed repositories have declared to be each preserving a single type of Digital Object (DO), and therefore in Table 10 and Table 11, which are based on data from sections 1 and 2 of the questionnaire, we report some features of the DOs directly as features of the repository. As for the volumes, all the repositories are of about the same size, with a number of preserved DOs that ranges from less than 100.000 to more than 1.000.000, with an expected yearly increase that ranges from less than 20% to more than 100%, and an average size of the DO of about 10 MB, getting to hundreds of MB in one case only. All repositories but two allow remote access, but usually restricting it to registered users, with a single exception. In all cases the retention period is more than 10 years.

REPOSITORY	**DIGITAL OBJECTS**						
	Retention period	**Total number**	**Total size**	**Yearly increase**	**Size of metadata**	**Versions**	**Formats**
1 - Toscana Regional Repository	> 10 y	$<10^5$	< 100 TB	20% - 100 %	5% - 10%	1 - 5	PDF/A, XML
2 - Public e-gov Local Repository	>10 y	$> 10^6$	< 100 TB	20% - 100 %	< 10 %	1 - 5	PDF/A, TIFF
3 - Vicenza Health Care System	>10 y	$> 10^6$	< 100 TB	20% - 100 %	5% - 10%	1	DICOM, PDF, doc
4 - National Bank of Italy	>10 y	$> 10^6$	> 1 PB	< 10%	5% - 10%	1 - 5	PDF/A
5 - Cultural Heritage at Cineca	> 10 y	$>10^6$	< 100 TB	20% - 100 %	< 5%	1	PDF, TIFF, JPEG, JPEG 2000, PNG, GIF
6 - Sapienza Digital Library (Cineca)	> 10 y	10^5 - 10^6	< 100 TB	> 100%	< 5%	> 5	PDF, TIFF, JPEG, PNG, MP4, 3GP

7 - Sapienza Digital Library (staging)	>10 y	< 10^5	< 100 TB	> 100%	< 5%	> 5	PDF, TIFF, JPEG, PNG, MP4, 3GP
8 - Magazzini Digitali at BNCF	> 10 y	10^5 - 10^6	< 100 TB	100% - 500 %	5% - 10%	1	HTML, PDF, EPUB

Table 11: Volumes and types of preserved digital objects

In all the repositories in the e-gov area, that manage mostly digital or digitized text documents, the ubiquitous formats are PDF and PDF/A, with sometimes the addition of TIFF for digitized paper documents, the only exception being the Vicenza HC system, where DICOM formats are used for medical tests results. The situation is instead more complex for the three repositories in the Cultural Heritage area, because of the variety of preserved multimedia objects. On the other hand the relative size of metadata is apparently larger for the repositories in the e-gov area (sometimes over 10%), than in the ones in the CH area, where it is always less than 5%.

Table 12, which is based on data from sections 2 and 3 of the questionnaire, summarizes the structure of the storage systems, that is typically on three levels:

- a *front-end level* meant to provide access to the preserved objects;
- a *core level* devoted to preservation;
- a *back-end level* where incremental backups of the core level are preserved.

The two upper levels are systematically based on RAID5 systems or similar, so providing within each level a redundant storage schema. Therefore, at least theoretically, it takes something worse than two independent single failures at the two upper levels of storage, to lose some information. And even then there is the backup level, even if one should expect it being not completely up to date. But one should always consider that failures at the different storage levels may not actually be independent, especially if all the equipment is in the same room or in the same building/town (floods, fires, earthquakes).

REPOSITORY	STORAGE FOR PRESERVATION		STORAGE FOR ACCESS		BACKUPS			COST & RELIABILTY		
	In house	Type	In house	Type	In house	Type	Backup period	Cost TB/y	Availability	Reliability
1 - Toscana Regional Repository	Y	WORM	same	same	Y	RAID	1-7 days	ns	ns	ns
2 - Public e-gov Repository	Y	RAID 5	Y	RAID5	Y	TAPE DVD	1-7 days	ns	99.5%	ns
3 - Vicenza Health Care System	Y	RAID 5	Y	RAID5	Y	TAPE DVD	1-7 days	ns	ns	ns
4 - National Bank of Italy	Y	RAID 5	Y	RAID5	Y	RAID TAPE	1-7 days	ns	90%	100% (?)
5 - Cultural Heritage at Cineca	Y	RAID 5	same	same	Y	TAPE	1-7 days	600 €	95%	ns

6 - Sapienza Digital Library (Cineca)	Y	RAID 5	Y	RAID1	N	ns	ns	ns	99.98%	ns
7 - Sapienza Digital Library (staging)	Y	HD, RAID	Y	RAID1	Y	HD RAID	1-7 days	ns	ns	ns
8 - Magazzini Digitali at BNCF	Y	HD, RAID	same	same	Y	RAID	7-30 days	ns	70%	ns

Table 12: Storage solutions, cost and reliability (ns: not specified)

Backups are generated on a daily to weekly basis, and are actually *incremental backups*, since the preserved DOs do not get any update after the preservation begins. Therefore, each backup generates only the backup copies of the DO ingested after the previous one. In some cases, consumer grade optical disks, i.e. CDs and DVDs, are still used for backup copies, a quite questionable choice, as we discuss in detail in another section of this report.

In our survey, all repositories relied on in-house storage solutions, and four out of seven are structured on three levels. In the remaining three cases a single storage level is used both for access and preservation. This, besides reducing the redundancy, may potentially raise a problem for the authenticity of the preserved DO, since access is allowed to the preservation copy. This remark may hold for *5-Cultural Heritage at Cineca*, especially since access is granted to remote unregistered users (see Table 10 and Table 12). Instead, in *7-Sapienza Digital Library (staging)* the choice is acceptable, both because of the limited mission of the repository and because access is granted to staff people only. Finally, in *1-Toscana Regional Repository* the authenticity is granted since redundant WORM (Write Once Read Many) devices (EMC^2 Centera) are used to implement a single access/preservation storage level. However this choice may be questioned on another regard, since access to the data goes through proprietary firmware, something that can potentially fail and/or become unsupported over time.

As for cost, availability and reliability estimates, the answers have been rather disappointing. A cost estimate has been given in one case only (*6-Cultural Heritage at Cineca*), and the figure could be a realistic one only because of the limited size of the repository. But why then do not even consider an outsourced solution? Four estimates of availability have been given, and all seem to be reasonable. However, the impression is that these figures are available only because these repositories are managed by third parties, and therefore the figures were somewhere on the contract.

Reliability, i.e. the probability of losing part of the preserved data, was certainly the hottest issue, but the estimate has been provided only in a single case, and the figure was unrealistic (100%), possibly a misunderstanding between expressing a very low probability *as a negative power of 10* (e.g., $P=10^{-15}$) and *using a scale of 10* for rating it.

2.3.4 Analysis and proposals for improvements

From the above analysis we can draw a few interesting conclusions, and we can formulate a series of recommendations to improve the effectiveness and the quality of the storage infrastructure.

a) Improve the design process

The storage solutions implemented in most of the surveyed repositories appear to be reasonable, and, in general, adequate to the mission of the repository and to the specific requirements for the preserved DOs. However, the answered questionnaires reveal indirectly some kind of weakness in the design and quality assessment process:

- all but one of the interviewees were unable to specify the *storage cost* (TB/y); a good design should instead evaluate TCO (*Total Cost of Ownership*) and use it to compare different alternatives;
- practically all the interviewees were unable to provide any estimate of the *reliability* of their system, i.e. of the probability of losing some of their data; 100% reliability exists only in fairy tales, so one should like (pretend) to have at least a rough estimate of it;
- *availability* estimates have been given in only about half of the cases, and this apparently indicates that the management of the repositories were mostly concerned with providing good access to consumers;
- all the interviewees have been unable to reveal the *weak points* of their system; but even in the best system there must be some weak point, and the owner should be aware of it.

What should be improved is therefore the design process, by stressing: *i) clear and formal statement of the requirements* by the management, who is aware of the mission of the repository; *ii)* better *cooperation* between the management and the technical staff in *selecting, designing and evaluating* the storage infrastructure; *iii) quantitative analysis*, the key to understand and improve weak points and to produce reasonable estimates of cost, reliability and availability.

b) Improve reliability

All the proposed storage solutions are based on *redundancy*. This is, of course, a very good point for *reliability* and *availability*. But in all surveyed repositories redundancy is achieved at *device* or *local* level, and this is a weak point, since, if all the devices are sitting in the same room or building, a single accident (flood, fire, earthquake) may be a fatal one.

Reliability can be greatly improved by introducing redundancy on a *geographical* level. The 2009 earthquake destroyed a good deal of the Computing Center of the University of l'Aquila (Italy), but all student careers' records were safely recovered since the repository used to perform daily incremental backups to the Cineca Computer Center in Bologna, *hundreds of miles away*. Any level of local redundancy would have instead proved useless. In a two or three level storage structure, as in all surveyed repositories, geographical redundancy can be easily achieved by moving one level of storage to a remote location. But this involves *outsourcing*, and is often prevented by prejudices, as we shall discuss in a moment.

c) Provide a clear statement of the storage management policy

In any kind of enterprise, the formal statement of the policy, the procedures and the responsibilities is the first step towards quality management. In our survey three repositories out of seven had no declared policy, and the other ones, have apparently an insufficient one, and did not make it public. Integrity checks are mostly made at ingestion time only, and one has declared to have a data recovery workflow.

Considering the peculiar mission of a preservation repository, and the legal regulations it often has to comply with, setting up and declaring a policy, by making it public at least within the Designated Community, appears as a necessary requirement. As for the implementation and the management of the storage infrastructure, the policy should specify at least:

- regular integrity checks, performed not only at ingestion time, but along the whole preservation process;
- which levels of redundancy are provided and which controls are performed on a regular basis;
- which procedures have been established to migrate to new (and possibly still unspecified) storage system;
- an established and tested workflow for data recovery.

d) Take into due consideration outsourced storage solutions

Section 5 of the questionnaire (*Moving to an outsourced storage solution*), was targeted at investigating the interviewee's attitude toward these new interesting possibilities. Unfortunately, besides not exploiting this possibility in their own repositories, all the interviewees did not want to pay any attention to it. Like all prejudices, this one mostly depends on being biased and on a substantial lack of knowledge. Unfortunately, most people are still naively convinced that if they implement something at home, they are in full control of it, and therefore it is going to be of better quality. This may often be an unfounded belief.

Most people think only of public clouds, and consider this an intrinsically unreliable service often marketed by unreliable people. Instead, as well a *public cloud*, there is a variety of outsourcing schemes to be taken into account: *i) storage farms*: managed by reliable providers, in known locations according to strict procedures and security criteria; *ii) private shared cloud*: jointly managed by a consortium*; iii) redundant cloud*: data stored in several clouds, managed by different vendors; and plenty of variations and combination of these.

Of course disadvantages must be carefully weighted, as *lack of control*, *unauthorized accesses*, *vendor lock-in*, *provider bankruptcy*, all accounted for in our questionnaire, and in some cases these may prove to be a decisive negative factor. But advantages can be very attractive, especially for small to medium size repositories, above all *cost effectiveness*, *scalability*, *reliability* and *availability*. And, as discussed above, outsourcing some level of the storage infrastructure, allows one to introduce redundancy at a geographical level, and therefore to greatly improve reliability and availability.

2.4 NATIONAL LIBRARY OF THE NETHERLANDS

2.4.1 General remarks

The results of the KB survey are not based on the APARSEN survey as set up under WP23. KB has carried out a survey for the National Coalition on Digital Preservation (NCDD) in The Netherlands in the 1st half of 2012. Goal of this survey was to get an overview of the current situation of storage solutions within The Netherlands. The survey was sent out to over 200 Dutch organizations in four different domains. KB received 169 surveys of which 97 were fully completed. Furthermore, some respondents did not give us permission to use their data for other purposes, therefore we left these specific results out of this analysis. It finally left us with 89 useful results. These 89 were used for this analysis.

All the surveyed organizations are managed by Public administrations, and are mostly preserving digital records in four different areas: e-Government, e-Science, Media and Cultural Heritage. In some questions additional remarks were requested. Some of these are used in this analysis. Main result of the KB/NCDD survey was in the number of responses. This gives us a fairly good picture of the Dutch situation concerning storage and storage solutions.

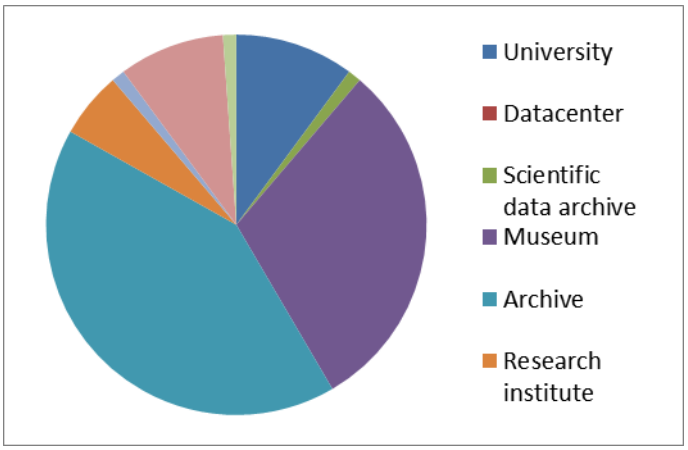
The Dutch questionnaire had a different design as does the APARSEN questionnaire. However, we managed to match most of the questions and answers in the KB/NCDD questionnaire to the questions in the APARSEN questionnaire.

The quality of the answers has been generally good and the data provided interesting. However some questions turned to be difficult for many organizations to answer. This is mainly due to the fact that there are a lot of small and medium size organization approached in the survey. The conclusions in the in sections 2.4.2 and 2.4.3 are mainly built on the answers we received from the large organizations. Especially for the small and medium size organizations it is difficult to answer questions on storage solutions in use or out sourced storage solutions. Probably most of these SMO's do not have yet any solution in place. Also questions on size and diversity of digital collections to be preserved for long term turns out to be not that easy to answer. And when we ask for expected growth, it really becomes a problem for SMO's. So the results of the survey depend heavily on the large scale organizations.

2.4.2 Basic information about the repositories

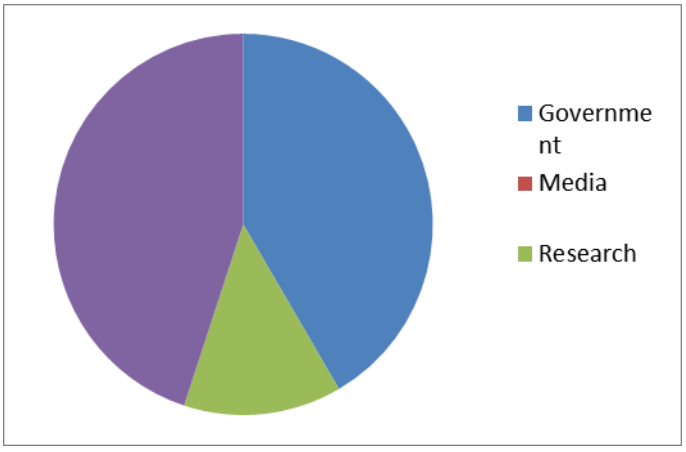
The types of organisation contacted for the survey are reported in Table 13, the pie chart depicts the representativeness of each of the contacted organisations.

All of them are publicly funded organizations. Majority of respondents represent an archive (local, regional or national) or a museum. Other major stakeholders are universities and governmental organizations, the domains which are represented most are culture and government, as shown in Table 14. Research is a good third, but the media domain is greatly underrepresented. That should be taken into account when analyzing the following answers. Note that this survey was not targeted towards libraries in the Netherlands. That's why they are missing in this survey.



Type of organization	Count
University	9
Datacenter	0
Scientific data archive	1
Museum	27
Archive	37
Research institute	5
Audio/visual institute	1
Government	8
Other	1
Total	**89**

Table 13: Number of involved organisations



Domain	Count
Government	37
Media	0
Research	12
Culture	40
Total	**89**

Table 14: Interviewed wrt application domains

The majority of respondents (53%) noted that they do not have a preservation policy for digital information in place, see Figure 1. Of the respondents to the survey less than half of them declare to have a written policy on digital preservation. Over 50% declare to have some sort of quality requirements for long-term management for digital objects. Examples of the requirements are: back-up policies, validation of data to be ingested, periodical checks on data stored, checks on file formats.

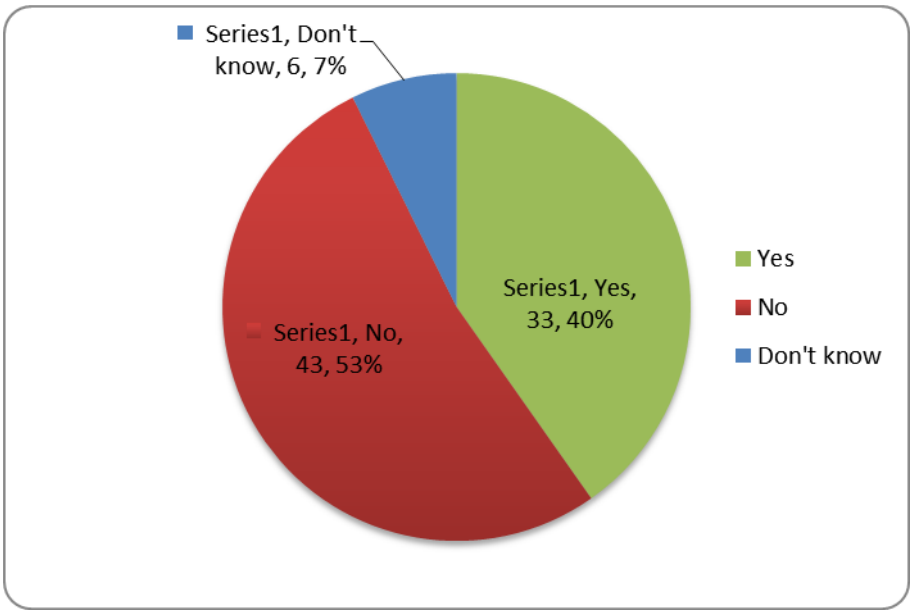


Figure 1: Preservation policy (N=82)

The majority of respondents (53%) noted that their organization does have defined certain quality measures for preservation of their data. However, it is not said what these measures are, but it shows that apart from having a policy many organizations do have the notice that measures must be taken to retain the data over time. On the other hand, still 40% does not set quality measures which is a high number and might lead to significant loss of data over time.

About 42% of the respondents acknowledge that their organization is able to effectuate the defined policy or preservation measures on maintaining its data quality, see Figure 2. However, a large group (58%) does not have that knowledge or is not convinced their organization can do that.

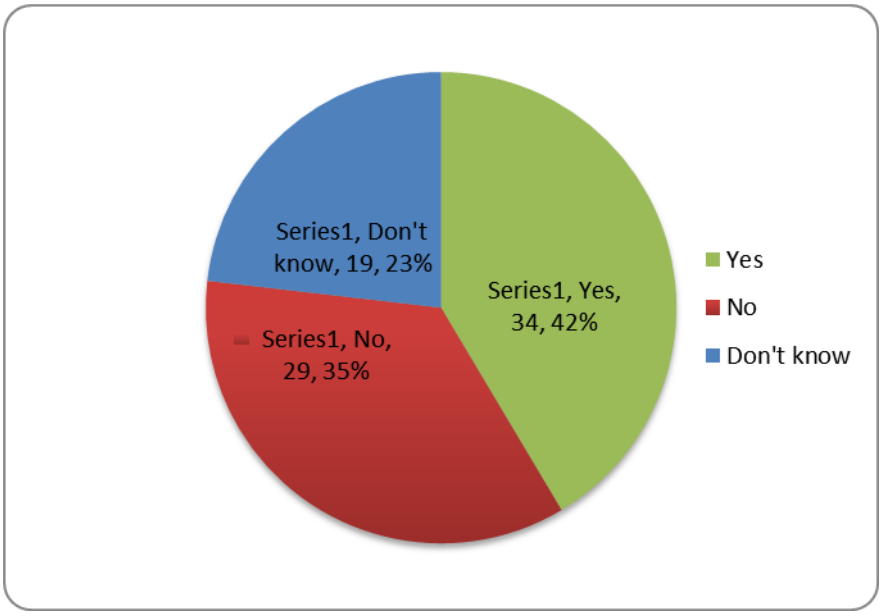


Figure 2: Effectuation of policies and/or measures about data quality control (N=82)

2.4.3 Preserved digital objects and storage solutions

Of the 89 respondents we used in this analysis all collected a very diverse set of collections and information, as depicted by Figure 3. Almost all respondents manage texts and images. Furthermore, databases, audio and video and websites are managed. However this does not always mean that these objects are stored and managed for the long term. These collections could also be stored for short-term access or, as e-mail, software and source code, being part of the organizations internal management. As for the experience with preserving digital objects; this was not a specified question in the survey. However, we do have some information on this. Most organizations only have few years' experience with preservation of digital objects. Only some larger organizations do have a long standing experience in the field of preserving digital objects. KB, national archives, institute for sound and vision are large organization which entered the field of DP at least 10 years ago and do have a preservation solution in place or are in the middle of developing a large scale solution. Most small scale organizations do not have archival solutions, but are working closely together with the large institutes or outsource their storage.

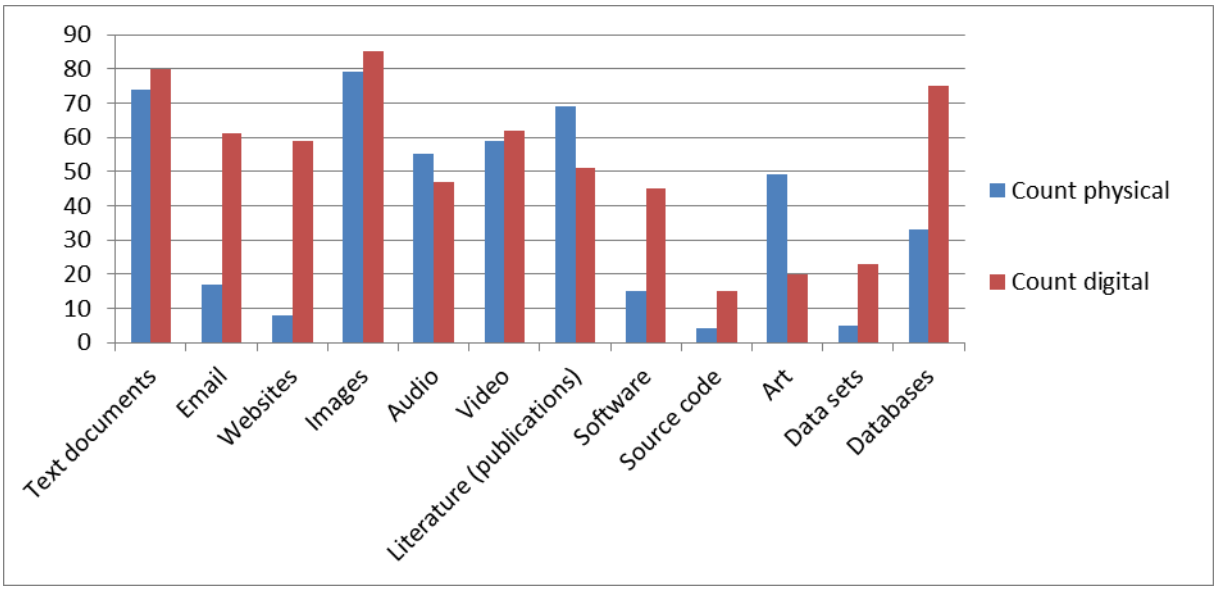


Figure 3: Type of preserved digital objects (N=89)

We collected no information on the number of digital objects preserved. What we did is ask for the number of storage capacity used on the different types of collections. As shown in Figure 4, most organizations store less than 1 PB of data within the different object types. Only few respondents reach to 1 PB or even more. Especially in the category of images a number between 1 TB and 1 PB is reached for most of the respondents. Software, arts, source code and datasets are collected and managed by only a few institutions. Most respondents manage database. In many cases these will be databases used in internal management.

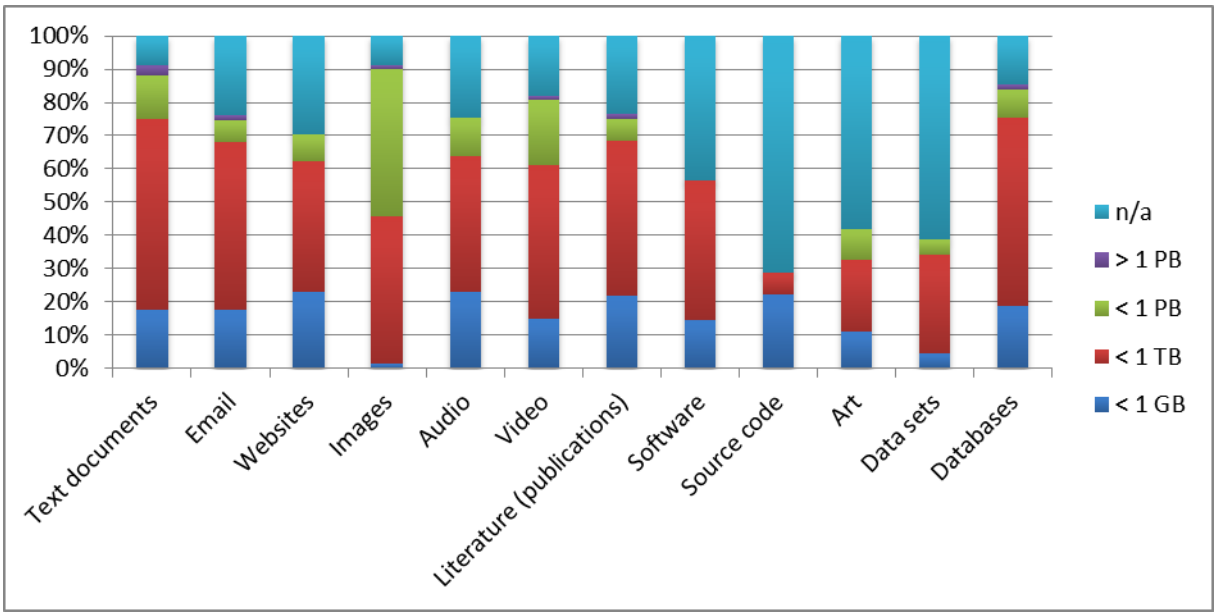


Figure 4: Size of preserved objects (N=87)

Figure 5 provides a quick view on the storage period versus data type learns that email, websites and software are seen as expandable. Whereas text documents, art, images, audio and video are assumed to be worthwhile to preserve for a much longer period.

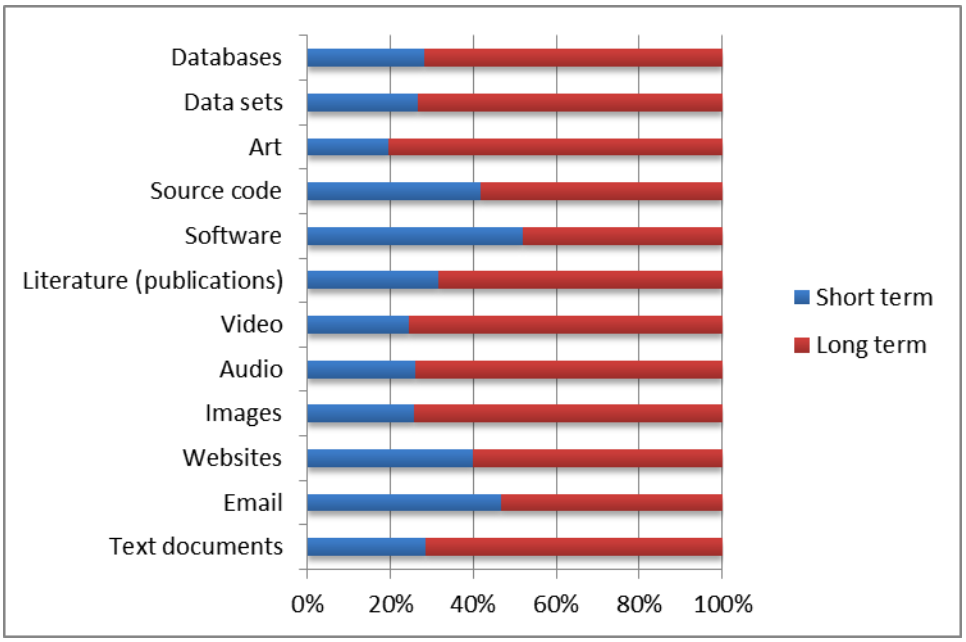


Figure 5: Short or long-term storage (N=87)

Analysing the expected increase, depicted in Figure 6, we see that a growth is expected to max 1 PB in almost all categories of digital objects. Growth is 100% - 200% for most cases: text, e-mail, websites, AV. Growth to over 1 PB is expected for AV materials and datasets.

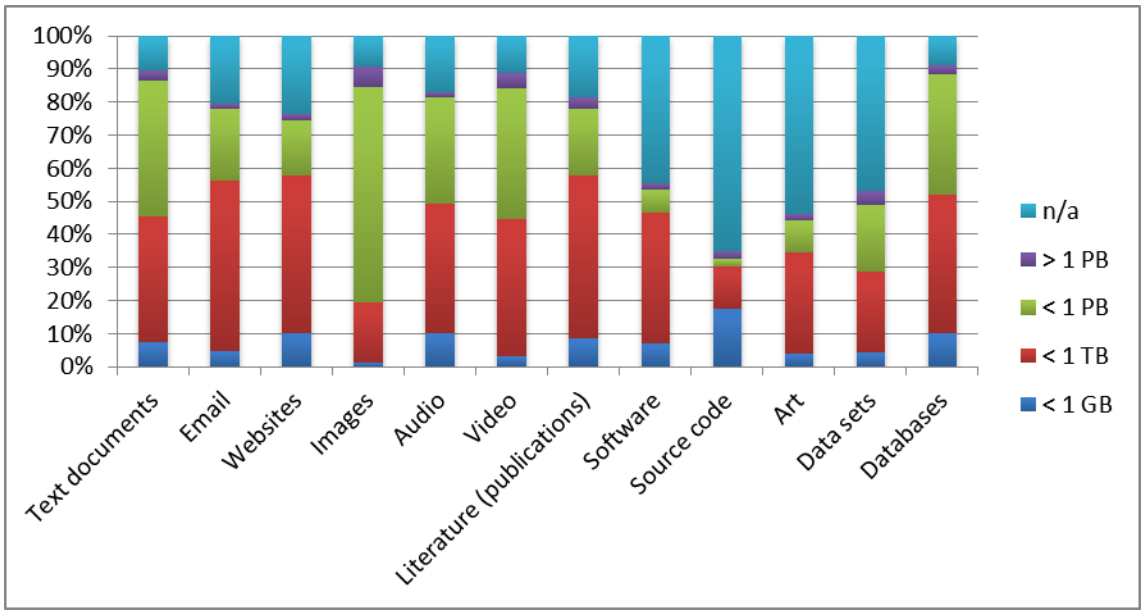


Figure 6: Expected increase of preserved objects in five years (N=87)

Figure 7 shows that all, but 2, respondents manage digital information, either for preservation, access or institutional management. Out of 87 respondents 33 manage own storage solutions, whereas 39 partly manage own solutions and have storage partly outsourced. Own storage solutions are managed mainly for internal management systems; for e-mail and company systems. Back-up facilities are partly outsourced, as is long-term storage of images and AV materials. 9 out of 87 have their storage completely outsourced to a third party solution.

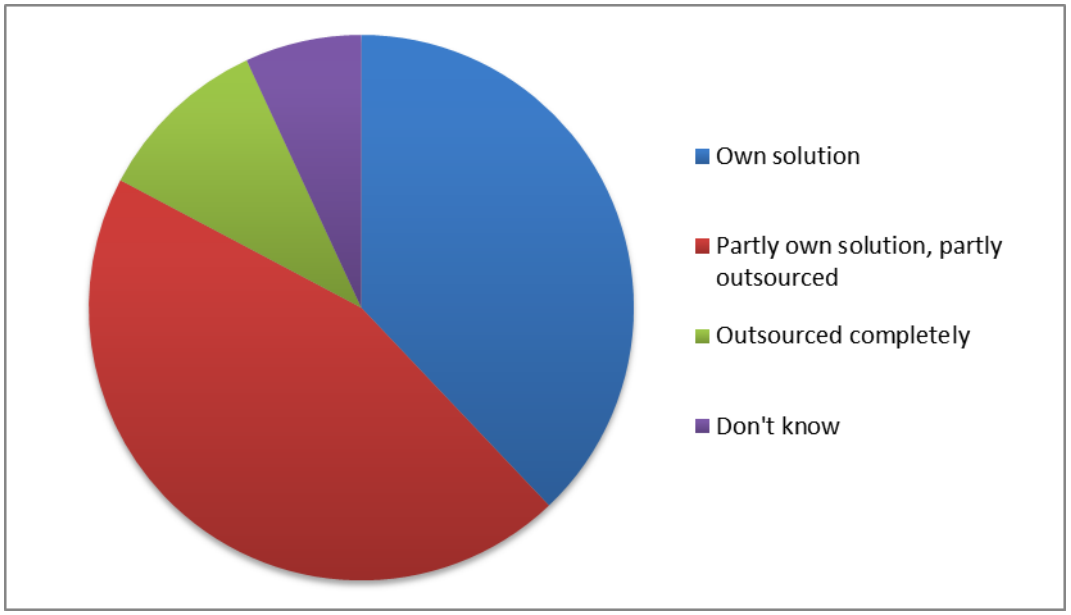


Figure 7: Storage solutions (N=87)

General impression about outsourcing solutions, is that there is no common sense on where to safely store your data. The most common places are DANS and Pictura (storage for digitized images), but after these the Figure 8 blurs into many different outsourcing partners. There is a strong tendency

towards minimizing the number of data centres for storing safely cultural and scientific data, partly stimulated by the Dutch Government.

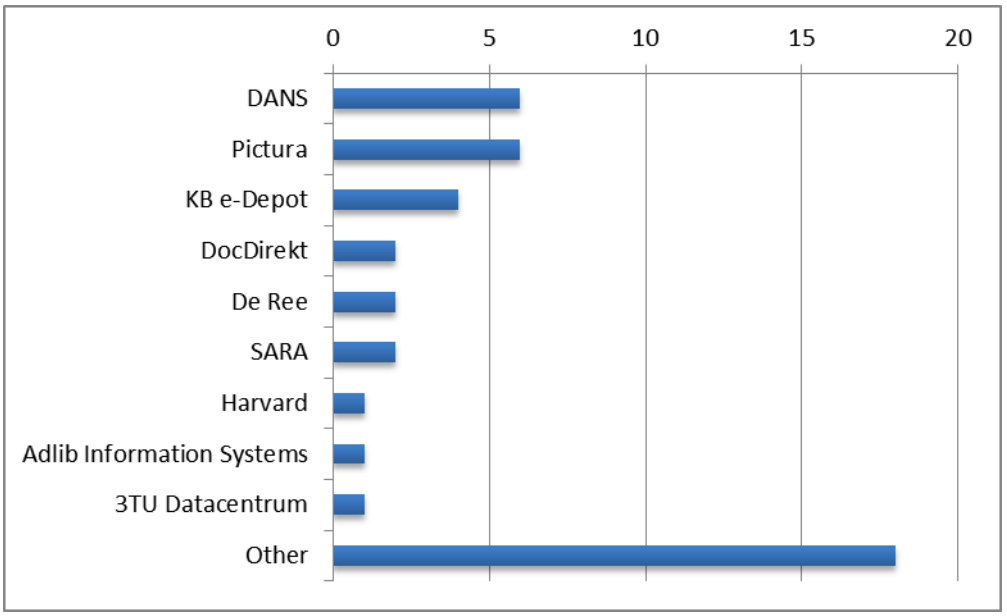


Figure 8: Outsourcing solutions (N=43)

By far, most respondents do not store their own data outside the jurisdiction of The Netherlands, as depicted in Figure 9. About 11% do, mostly only for a part of their data: research data of international research projects and research teams, collaboration with the LOCKSS network, web based content.

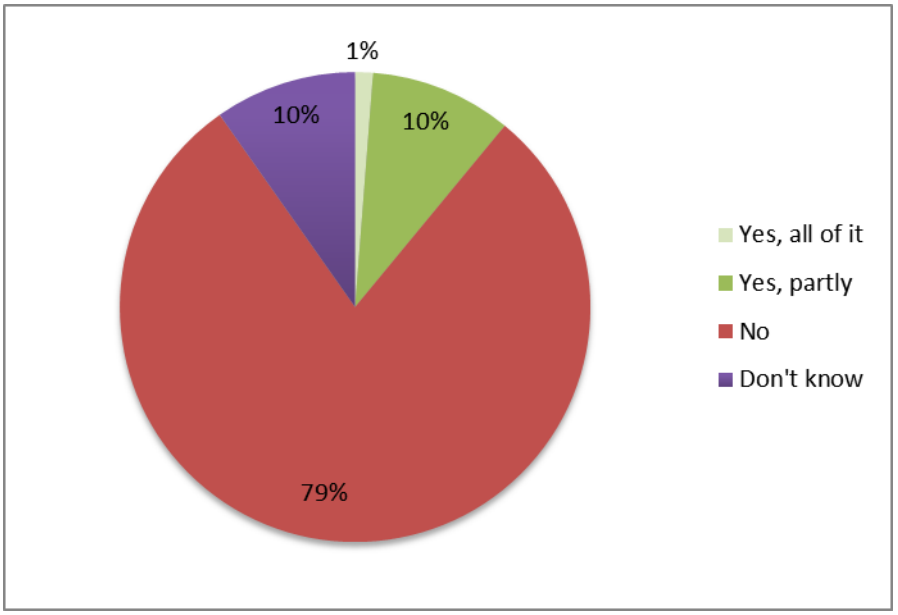


Figure 9: Data storage outside Nederland (N=43)

2.4.4 Analysis and proposals for improvements

Consider to centralise long-term storage

The general outline is that most Public administrations in the Netherlands that did take part in the survey do have their own ICT infrastructure or have a mixture of an in-house and outsourced storage solution. For smaller organisations it may be wise to combine forces and work towards a shared or outsourced storage solution as the amount of data to be stored in the future will grow significantly, as pointed out by the survey results. Further development of long-term preservation archives such as DANS and KB e-Depot in the Netherlands might help in achieving a more clustered preservation landscape. This line can be extrapolated on European level as certain long-term storage hubs, also outside the Netherlands, can ensure that digital cultural heritage and research results Europe-wide can be safely stored for the long term. With several of those larger hubs across Europe, there will be a confident level of duplication and persistency while costs can be kept under control as nations can save expenses on maintaining numerous digital archives across Europe.

Define a preservation policy

Organisations that maintain a storage solutions meant for data that need to be stored for a longer period of time, work is still to be done as a majority does not have defined a preservation policy. This is a real concern and points out that awareness on this point still needs to be raised. Apparently, the common line is still to start with storage on technical level while defining a policy is of later concern. But if preservation is taken seriously, also by management, it starts with a policy. From then on, each activity on preservation can be put into context of that policy, giving it structure and priority.

Legal implications recognised

A positive outcome is that most organisations seem to be aware of the risks it brings when data is stored outside their own jurisdiction. It might also be that outsourcing outside their own country is not yet common and therefore scores low. In that case, more awareness needs to be created.

On the other side, it would certainly help if local jurisdiction would be brought in line with European or pan-European regulation. Diffusion on regulations and different national interpretations of European directives makes it difficult to come to larger storage hubs as suggested in the first item.

Preserve the software as well

A subtle but important outcome of the survey results achieved in the NCDD survey is that software (and source code) is very often classified as expendable because it is stored only for the short term. As we find out more often today, software is the linking pin between storage and retaining access to digital information. Without the correct software, the meaning of data cannot or only partly be understood. It is therefore important that software preservation becomes more mainstream just like any other digital object.

3 CONCLUSIONS, RECOMMENDATIONS AND LESSONS LEARNED

The present chapter aims to summarise and harmonise conclusions and remarks provided by the WP23 part II partners on the basis of the returned questionnaires. Common gaps will be proposed first, and recommendations will be suggested accordingly. The section ends with statements on lessons learned.

Gaps and Recommendation Analysis

The questionnaires received from the involved interviewees have been collected, analysed and reported following a common structure, with the aim of merging and comparing such heterogeneous results (i.e., different expertise, application domains and missions of the communities the interviewees belong to). The main issues, identified through the analysis of the returned questionnaires, are mainly related to:

a) Costs

Vague or missing information has been provided about costs (TB/y), only few interviewees have been able to do that. Anyway, the definition of a cost modelling (see SR3) has to be considered a mandatory task.

Recommendation: Costs should takes into account any cost factors like personnel, storage tiers, backup/redundancy needs and maintenance (mid-/long-term). The availability of a precise cost model permits the evaluation of Total Cost of Ownership (TCO), cost per TB per year (TB/y), and forecasting of storage in view of scalability activities for future years. Furthermore, it permits the comparison of the costs of local storage vs outsourced storage.

b) Reliability and Availability

Information collected about reliability and availability has been scarce and very poorly detailed. Most of the interviewees were unable to provide any estimate of the reliability (e.g., probability of losing some of their data) and availability of their systems.

Most of the proposed storage solutions are simply based on local redundancy concepts, through several techniques (e.g., RAID, tapes, tiered architecture). This can be surely a good starting point to address reliability and availability, but for long-term digital data preservation, it cannot be considered sufficient nor a reliable solution. In fact, when redundancy is achieved only at device or local level, based on RAID configuration or periodically performing data integrity (e.g., SR6c), with devices sitting in the same room or building, a single accident (flood, fire, earthquake) may be a fatal one.

Recommendation: Reliability, in particular, can be greatly improved by introducing redundancy on a geographical level which would permit to cope with disaster recovery (see SR6c).

A possible architecture solution, to guarantee high level of availability and reliability, might combine tiered storage structure and outsourcing. In this case, local redundancy solutions simply based on RAID would be adopted in local facilities, while geographical redundancy of massive backup would be ensured by external cloud resources.

c) Weak points

None of the interviewees has been able to reveal weak points of their systems.

Recommendation: A clear picture about system weak points should be mandatory, as every system has some weak point, and the management should be always aware about it. As for costs, reliability and

availability issues, also this gap points out limitation concerning requirement collection and design phase of the storage architectures.

d) Digital Object Preservation Policy

It appears, from the returned questionnaires, that most of the interviewees approach to the storage solutions, first technically, and at later stage defining a preservation policy. Furthermore, it is frequently performed at best effort, without a well-defined context, hence limiting structure and priorities of the long-term preservation activities (see 2.1.4, 2.3.4 and 2.4.4).

Recommendation: Digital preservation of digital objects should rely on a set of preservation actions properly planned and documented by data holders and archive owners, and applied to the data themselves and to all the associated information necessary to make those data understandable and usable by the identified user communities. The data preservation policy should specify at least:

- regular integrity checks (e.g., checksum), performed not only at ingestion time, but along the whole preservation process, also including periodic verification of the used hardware;
- which levels of redundancy are provided (e.g., hierarchical storage joint to geographical backup) and which controls are performed on a regular basis;
- which procedures have been established to migrate to new (and possibly still unspecified) storage system (e.g., outsourced solutions);
- provenance of every change of the data (e.g., by repair actions, or re-processing), also keeping trace about adopted software and tool, in order to permit long-time access and management of the digital objects
- use secure and encrypted connections to protect your data during the transfer (e.g., geographic backup)
- an established and regularly tested workflow for data recovery, including disaster recovery

e) Outsourced Solutions

The returned questionnaires show caution and scepticism in adopting outsourced storage architectures, which typically lead the interviewees to rely on in house solutions. Very frequently, these solutions are considered unreliable only on the basis of preliminary feelings, instead of detailed and documented technical and market analysis. This kind of approach suggests that the criticisms are mainly due to scarce knowledge about the available outsourced storage solutions. On the other hand, as a matter of fact the following issues are perceived as serious limitations:

- Fear of unauthorised access to data repositories
- Lack of direct control of the data repositories
- Vendor lock-in
- Provider bankruptcy
- Rental cost comparable or even higher than in house solutions
- Burden of moving high data volumes to the external resources over the WAN
- Legal implications when data is stored outside their own jurisdiction

It is reasonable to suppose that a quantitative analysis of what the market offers in terms of outsourced storage solutions, combined with a detailed analysis of the specific requirements might cut down, or at least reduce, most of the unfounded belief against them. Small to medium size repositories, in particular, should take care about the promising advantages of the outsourcing, as it can offer to them the right trade-off about cost, availability, reliability and scalability on-demand. The outsourcing can also be proposed for architectures where redundancy at a geographical level is a key aspect, and also when they aims to propose centralised long-term storage solutions (see 2.3.4 and 2.4.4).

An example of complex architecture, moving toward the outsourcing, is provided by ESA. This is already implemented for the online data access, in particular for the access to the so-called Explorer (http://www.esa.int/Our_Activities/Observing_the_Earth/The_Living_Planet_Programme/Earth_Explorers) missions, as well as for Third Party Missions (https://earth.esa.int/web/guest/missions/3rd-party-

missions/overview). In the next twelve months a similar move will be performed for the archiving services for the same missions listed above. Same philosophy is going to be adopted in the framework of the Copernicus Space Segment Architecture (introduced here below).

LESSONS LEARNED

The interviewees, as shown through the answered questionnaires, fully recognise the relevance of the digital preservation activity. Furthermore, all of them, independently from their missions and/or domains, must undertake digital preservation activities for ensuring continued access to their digital archives. In most of the situations, the adopted storage solutions appear to be reasonable, and in general, adequate to the specific addressed tasks. However, a deeper analysis of the returned questionnaires pointed out that requirements collection, design phase and quality assessment process of the storage architectures, frequently show major gaps, sometime partially considered or in some situations even neglected. In particular, it is expected the design phase will address specific needs, first collected from involved user communities, and hence translated in system requirements. Typical examples of user and system requirements are listed here below:

General community needs

- UR1. Enhanced operability, reliability and availability of the data
- UR2. Easy exploitation of data
- UR3. Access to historical data collection (e.g., long term data preservation), including associated information (e.g., metadata, processing software)
- UR4. Clear statements about data policy, legal aspect (e.g., free or privileged access) and confidentiality (e.g., Authentication, Authorization and Accounting)
- UR5. Large user communities networking
- UR6. Cross-fertilisation among heterogeneous domains

General system requirements

- SR1. Modular, flexible, portable and scalable storage architecture, that can be tailored as necessary to meet the specific domain goals, through standardised technologies and interfaces
- SR2. Versioning of the data in order to manage the historical, current and future data
- SR3. Definition of the cost modelling for the storage architecture
- SR4. Reduction of operating and functional costs to meet budget constraints, also leveraging technological evolution and continuous decrease of HW costs
- SR5. Responsiveness and performance compliant with the user access rate
- SR6. Definition of the reference operations concept, based on the following principles:
  - SR6a. Services defined in Service Level Agreements (SLAs) and use of Key Performance Indicators (KPIs)
  - SR6b. Interoperability among sub-systems
  - SR6c. Backup (e.g., RAID configuration), data integrity (e.g., data checksum and test) and disaster recovery (e.g., geographical redundancy)
  - SR6d. Reporting and monitoring (e.g., automatic warning/alert services)
  - SR6e. Optimisation of the HW procurement (e.g., tiered architecture) and maintenance plan (e.g., check of tapes readability and periodic replacement)
  - SR6f. Access control
  - SR6g. Maintaining corporate values of quality of service

Copernicus Space Segment Architecture

An example of architecture, filling most of the gaps identified and analysed within the document, is going to be provided in the framework of Copernicus EC program (http://www.copernicus.eu/). It will be a shared, federated, outsourced and regulated environment, based on a very accurate activities of requirements collection and design phase. It will consist in a number of federated systems permitting access and exploitation of satellite data (https://sentinel.esa.int/web/sentinel/home), and in-situ sensors. In particular, ESA has been taking care about the development of the space component of the overall infrastructure (http://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus). Each sub-system will be in charge of storing and making available a piece of the overall dataset, that Copernicus aims to provide to the user community. In this context, the choice of a suitable storage solution represents a critical aspect. On the one hand, the program will have the purpose of permitting the policymakers and public authorities to take critical decisions about environmental regulation, emergency, natural disaster or humanitarian crisis. On the other hand, there is the need to make available value added services responding to public and/or commercial requirements.

The Copernicus Ground Segment (GS) will provide facility for satellite data downlink, processing, storage, data/satellite quality control and distribution. It will consist of a Core Ground Segment, focusing on Copernicus functions and elements, and a Sentinels Collaborative Ground Segment (CGS), providing supplementary access to Sentinel Missions data and the frame for cooperation. The Sentinels CGS will provide an infrastructure permitting to reply to specific needs about Sentinel data exploitation, including: data acquisition and near real time production (e.g., local station), collaboration in data product and algorithms definition, data product dissemination and access (e.g., mirror sites). Formal agreements will regulate both any form of collaboration and definition /implementation of dedicated operational interfaces.

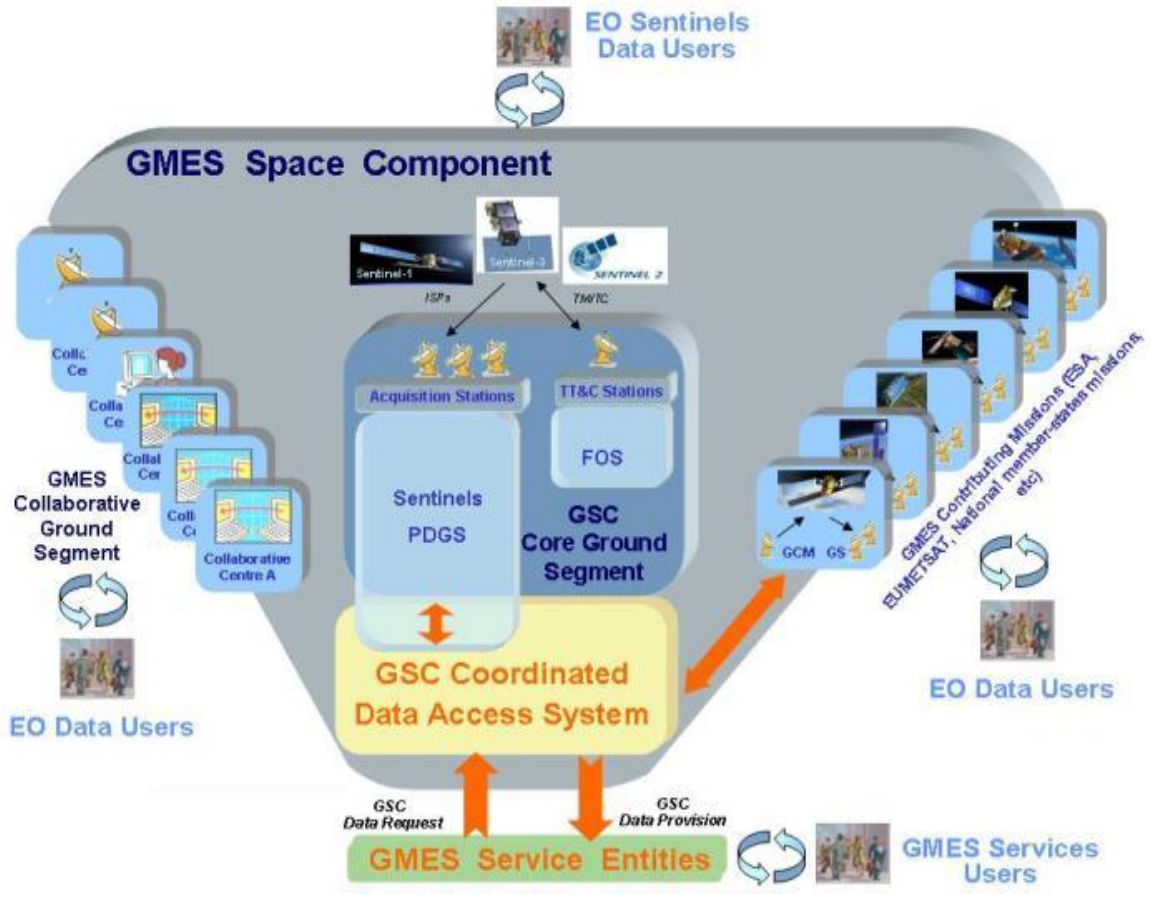


Figure 10: Copernicus Ground Segment Architecture

The Copernicus GS architecture, as depicted in

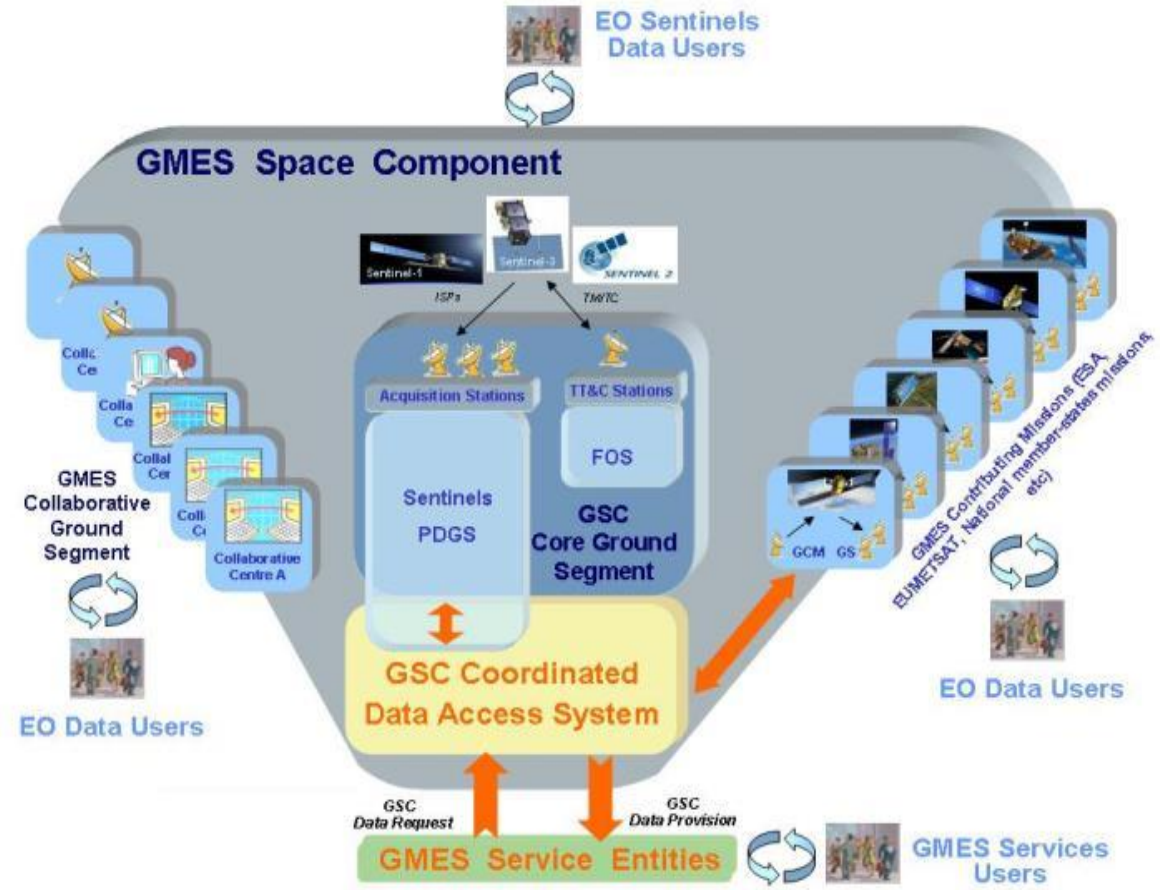


Figure 10, is based on federated, distributed and independent GSs operating both Sentinels and Third Party Missions (TPMs). The data they singularly manage will be made available transparently to the end user through a single virtual access point, shared and based on standardised interface, the Space Component Data Access System. The system intends to be the hub of an interoperable network of distributed European ground segments contributing to Copernicus, culminating in a harmonised, one-stop-shop for users.

The GSs that will be part of the Copernicus Ground Segment Architecture, will be aligned to a common Operation Concept, which implicitly will address most of the gaps identified by this document, aiming to:

- Better serve a broader user community, improving the availability and facilitating the exploitation of data *(b)*
- Improve the overall missions operability, reliability and availability *(b)*
- Sustainably reduce operating cost to meet budget constraints *(a)*
- Ensure preservation and access of data and associated information in the long term *(d)*
- Leverage technological evolution and continuous decrease of HW costs, considering outsourced technologies and infrastructures *(e)*, following a service approach with appropriate cost modelling *(a)*, avoiding duplications and reducing HW procurement.
- Use commercially available (European) products whenever possible
- Standardised, documented interfaces to support modular concepts
- Leverage on competences developed historically in European scientific centres for the different types of sensors and scientific applications. Variety of missions and sensors and availability of the relevant expertise in different organizations in Europe

4 ANNEXES

4.1 ANNEX I: QUESTIONNAIRE

APARSEN is a Network of Excellence funded by the EU (2011-2014) with the goal of overcoming the fragmentation of the research and of the development in the digital preservation area by bringing together major European players. Among other activities, APARSEN is investigating the storage solutions currently used in digital archives and preservation repositories, in order to compare the different approaches and technical solutions, and their fitness to preserve digital objects in terms of cost, legal, policy and business needs. Moreover, the project is also addressing analyzing the scalability problems that arise from the increase in number and size of preserved digital objects, that many repositories are currently experiencing.

This questionnaire has been produced as part of this effort, in order to analyze the current practices and to gather comments and suggestions from all the involved parties, based on their individual experiences. The questionnaire is being sent to a number of repositories across Europe, having a variety of profiles that range from large repositories of scientific data, to national libraries and national archives, up to minor digital libraries and repositories run by small organizations with a limited budget. The results of the analysis we plan to perform by investigating such a large audience will hopefully provide an interesting feedback to the digital preservation community, and will allow sharing different experiences.

We do hope that you may give your contribution and cooperate with us by filling in this questionnaire, and we will be grateful to you for doing so. To minimize your effort we have tried to keep the number of questions within reasonable limits, and we have designed the form to speed up the fill in process, by allowing most answers to be given just through checkboxes.

The questionnaire is structured in five sections:

1. Basic information about the institution and the repository
2. Types of digital objects
3. In house storage solutions
4. Outsourced storage solutions
5. Moving to an outsourced storage solution

Sections 1 and 2 are appropriate for any kind of repository, and we expect everyone to complete them. Sections 3 and 4, instead, may be alternative, at least for those repositories relying on a single kind of storage solution (in house or outsourced). As for section 5, due to the relevance of the topic, we will really appreciate if you can answer these questions as well, even if you have currently no intention to move to an outsourced storage solution.

On the other hand, since you may have several types of digital objects and several storage solutions within your repository, additional forms are provided to you for these sections, if you need to fill them.

Though almost all questions can be reasonably well understood without the need of specific instructions, we nevertheless provide you with a full set of instructions, that you will find enclosed in the distribution bundle. You do not have, of course, to read them, but just to refer to them, if you need to. The instructions include also a list of selected terms from the Digital Preservation Glossary that is being produced by the APARSEN project.

For any further question, please do not hesitate to ask for assistance the person who has contacted you to send you the questionnaire, or send an e-mail directly to: **salza@dis.uniroma1.it.**

1. BASIC INFORMATION ABOUT THE INSTITUTION AND THE REPOSITORY

1.1. Name of institution:

1.2. Contact person

Last name:

Name:

e-mail:

1.3. Description (the institution and the repository)

1.4. Mission

1.5. Ownership:

- ☐ public
- ☐ private

1.6. Legal regulations (specific national and international regulations to comply with)

1.7. Experience: how long have you been preserving digital objects?

- ☐ less than 5 years
- ☐ 5 to 10 years
- ☐ 10 to 20 years
- ☐ more than 20 years

1.8. Total number of digital objects preserved:

- ☐ less than 100.000
- ☐ 100.000 to 1.000.000 more
- ☐ than 1.000.000

1.9. Total size of the preserved objects:

- ☐ less than 100 TB
- ☐ 100 TB to 1 PB (1000 TB)
- ☐ more than 1 PB

1.10. Annual quantity (number of new objects per year):

- ☐ less than 10.000
- ☐ 10.000 to 100.000 more
- ☐ than 100.000

1.11. Annual size (total size of new objects preserved every year)

- ☐ less than 1 TB
- ☐ 1 TB to 100 TB
- ☐ more than 100 TB

1.12. Expected increase in the number of preserved digital objects (refers to 1.8)

- Expected increase in the next 5 years less than 100%
  - ☐ from 100% to 500%
  - ☐ more than 500%
- Expected increase in the next 10 years less than 100%
  - ☐ from 100% to 500%
  - ☐ more than 1000%

1.13. Expected increase in the total size of preserved objects (refers to 1.9)

- Expected increase in the next 5 years less than 100%
  - ☐ from 100% to 500%
  - ☐ more than 500%
- Expected increase in the next 10 years less than 100%
  - ☐ from 100% to 500%
  - ☐ more than 1000%

1.14. Policy

- Is there a declared policy on storage of data in place?
  - ☐ Yes
  - ☐ No
- Public link to the policy:
- What are the key criteria of the policy? *(check one or several)*
  - ☐ Regular integrity checks?
  - ☐ Determinations for controlling / monitoring and redundancy
  - ☐ Established and tested workflow for data
  - ☐ recovery? Other (please specify)

1.15. Additional comments

2. TYPE OF DIGITAL OBJECTS

(if appropriate fill a separate form for each object type)

2.1. Description

2.2. Producers and Designated Community

2.3. Ownership of digital rights:

☐ internal
☐ external

2.4. Access policy:

a) ☐ unlimited ☐ registered users only

b) ☐ from anywhere ☐ only from specific locations

c) Are there specific requirements to be met on access time?

☐ Yes ☐ No

2.5. Format(s)

2.6. Versions (number of distinct versions preserved for each digital object):

☐ 1
☐ 1 to 5
☐ more than 5

2.7. Size of metadata (ratio between the size of metadata and the size of data)

☐ less than 5%
☐ 5% to 10%
☐ more than 10%

2.8. Connections between objects (are there objects groups that are often accessed together?)

☐ Yes
☐ No

2.9. Annual quantity (number of new objects of this type per year):

☐ less than 10.000
☐ 10.000 to 100.000 more
☐ than 100.000

2.10. Total number of objects of this type preserved in the repository:

- ☐ less than 100.000
- ☐ 100.000 to 1.000.000 more
- ☐ than 1.000.000

2.11. Total size of the of objects of this type preserved in the repository:

- ☐ less than 100 TB
- ☐ 100 TB to 1 PB
- ☐ more than 1 PB

2.12. Average access rate (average number of accesses per object per year):

- ☐ less than 1
- ☐ 1 to 100
- ☐ more than 100

2.13. Retention period (how long objects of this type must be preserved):

- ☐ up to 5 years
- ☐ 5 to 10 years
- ☐ more than 10 years

2.14. Storage levels used for this type of object (referring to sect. 3 and 4)

- Storage for access
  - ☐ RAM/Solid state
  - ☐ Hard disk
  - ☐ RAID or similar
  - ☐ Tape
  - ☐ WORM disk
  - ☐ CD or DVD
- Storage for preservation (if different)
  - ☐ RAM/Solid state
  - ☐ Hard disk
  - ☐ RAID or similar
  - ☐ Tape
  - ☐ WORM disk
  - ☐ CD or DVD
- Storage for further backup copies
  - ☐ (if any) RAM/Solid state
  - ☐ Hard disk
  - ☐ RAID or similar
  - ☐ Tape
  - ☐ WORM disk

☐ CD or DVD

2.15. Additional comments

3. IN HOUSE STORAGE SOLUTIONS

(if necessary fill a separate form for each different storage solution you are currently managing)

3.1. Description

3.2. Purpose:

☐ access only
☐ preservation only
☐ both

3.3. Technology:

☐ RAM/Solid state
☐ Hard disk
☐ RAID or similar
☐ Tape
☐ WORM disk
☐ CD or DVD

3.4. Experience: how long have you been using this storage solution?

☐ less than 5 year
☐ 5 to 10 years
☐ more than 10 years

3.5. Redundancy (not including backups)

- **Level**: ☐ none ☐ RAID1 ☐ RAID5 ☐ other
- **Mode**: ☐ device ☐ local ☐ geographical (more than 50 km)

3.6. At which intervals are backups performed for this level of storage?

☐ 1 to 7 days
☐ 7 to 30 days
☐ 30 to 180 days
☐ more than 180 days
☐ never

3.7. Staff (number of employees devoted to the management of this storage system)

☐ less than 5
☐ 5 to 10
☐ more than 10

3.8. Cost estimate (total cost per TB/year): €

3.9. Which controls are you regularly performing, and at which intervals, to check the integrity of stored data?

3.10. Perceived weakness: why do you think the solution is not completely satisfactory? How could it be improved?

3.11. Time horizon: how long do you plan to reasonably carry on with this solution before replacing the devices and/or moving to a different storage solution?

☐ less that 1 year
☐ 1 to 5 years
☐ more than 5 years

3.12. Risk assessment (with regard to this storage level)

- *Reliability*, i.e. probability of loosing (part) of your stored data:
  - ☐ I cannot provide any reliable estimate
  - ☐ Yes, my estimate of system reliability is (express as power of 10) :
- *Availability*, fraction of time during which access to the preserved objects is
  - ☐ granted: I cannot provide any reliable estimate
  - ☐ Yes, my estimate of system availability is (express as power of 10):

3.13. Additional comments

4. OUTSOURCED STORAGE SOLUTIONS

(if necessary fill in a separate form for each different storage service you are currently purchasing)

4.1. Description

4.2. Name of the provider:

4.3. Classification:

- ☐ public cloud
- ☐ private shared cloud: jointly managed by a consortium of several repositories redundant cloud: data stored in several clouds, managed by different vendors
- ☐ other (specify)

4.4. Experience: how long have you been using this storage solution?

- ☐ less than 1 year
- ☐ 1 to 5 years
- ☐ more than 5 years

4.5. Redundancy (as declared by the provider)

4.6. Terms of agreement (what does your provider guarantee you, and in which terms):

Access

Security

Persistence

4.7. Storage cost (TB/year): €

4.8. Access cost. a) per access: € b) per MB transferred: €

4.9. Risk assessment (with regard to this storage level)

- *Reliability*, i.e. probability of loosing (part) of your stored data:
  - ☐ I cannot provide any reliable estimate
  - ☐ Yes, my estimate of system reliability is (express as power of 10)

- *Availability*, fraction of time during which access to the preserved objects is
  - ☐ granted: I cannot provide any reliable estimate
  - ☐ Yes, my estimate of system availability is (express as power of 10)

5. MOVING TO AN OUTSOURCED STORAGE SOLUTION

5.1. Which outsourcing scheme are you willing to consider, if any? (Check one or several)

☐ public cloud
☐ private shared cloud: jointly managed by a consortium of several repositories
☐ redundant cloud: data stored in several clouds, managed by different vendors
☐ other (specify)

5.2. Motivations: why could you consider moving to a different storage solution? (Check one or several)

☐ improve cost
☐ improve reliability
☐ improve availability
☐ improve scalability
☐ other (specify)

5.3. Fears: why would you refrain from moving to an outsourced storage solution? (Check one or several)

☐ 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 vendor lock-in: problems in moving data to a different provider

5.4. Additional comments

4.2 ANNEX II: QUESTIONNAIRE INSTRUCTIONS

1. BASIC INFORMATION ABOUT THE INSTITUTION AND THE REPOSITORY

1.1 Name of the institution

Name of the institution that manages the repository.

1.2 Contact person

Person that should be contacted, if necessary, for further information.

1.3 Description

Short description of the repository and of the institution that manages it.

1.4 Mission

Mission that the institution has been given as a whole, and specifically with reference to the repository.

1.5 Ownership

Is your institution a private or public organization?

1.6 Legal regulations

Are there any specific legal regulations (national or international) that apply to the objects preserved in the repository? For instance, special regulations apply, in general, to medical records.

1.7 Experience

How long have you been preserving digital objects, and providing this a reliable service?

1.8 Total number of preserved digital objects

By digital object we mean every single object that is individually preserved in the repository, and that can be individually retrieved.

1.9 Total size of preserved digital objects

Total size does not include backup copies or replications, created or managed within the repository for whatever purpose. It just accounts for the original size of the preserved objects.

1.10 Annual quantity

Estimate of the average number of new objects that are ingested by the repository every year.

1.11 Annual size

Estimate of the average yearly increase in the total size of preserved digital objects (see 1.9).

1.12 Expected increase in the number of preserved digital objects

How much do you expect the total number of preserved object (see 1.8) to increase in the next 5 and 10 years

1.13 Expected increase in the total size of preserved objects

How much do you expect the total size of preserved object (see 1.9) to increase in the next 5 and 10 years

1.14 Policy

Is there a formal and declared policy on storage of data the repository must conform with. Such a policy should be stated in a formal document. Specify the key criteria according to the list or add additional ones.

1.15 Additional comments

Add any further information you may like to provide.

2. TYPES OF DIGITAL OBJECTS

If your repository preserves several kinds of digital objects that are substantially different, you may want to fill in several copies of this section. You may find in the distribution bundle a separate file with a copy of this section. You may duplicate it as you need.

2.1 Description

Short description of this type of digital objects, to clearly identify what they are.

2.2 Producers and Designated Community

Specify the organizations that provide the information to be preserved, and community of users that is interested in preserving the digital objects, and in being able to continue accessing them (see glossary).

2.3 Ownership of digital rights

Who owns the digital rights on the preserved objects? Is it the institution managing the repository or not?

2.4 Access scheme

a) Is access limited to registered users be registered users.?

b) Is access limited to local users?

c) Are there specific requirement on access time (e.g. maximum access time)?

2.5 Formats

Format of the digital objects an/or of their components if they have a complex structure.

2.6 Versions

Are several versions of the same object typically preserved in the repository as a single object?

2.7 Size of metadata

Quantify the impact of metadata on the total amount of stored data.

2.8 Connections between objects

Are there connections between objects, such that groups of objects are preserved and often accessed together?

2.9 Annual quantity

Average number of new objects of this type ingested by the repository every year.

2.10 Total number of objects preserved

Total number of objects of this type currently preserved in the repository.

2.11 Total size of the preserved objects

Total size of the objects of this type currently preserved in the repository.

2.12 Average access rate

Average number of accesses per year and per preserved object.

2.13 Retention period

How long must the objects of this type be preserved, according to the regulations the repository must comply with.

2.14 Storage levels

Referring to the storage solutions described in section 3 and 4 of the questionnaire, specify which ones are used for access copies, for preservation copies and for backups.

2.15 Additional comments

Add any further information you may like to provide.

3. IN HOUSE STORAGE SOLUTIONS

In house solutions are opposed to outsourced storage solutions, as for instance service provided by storage farms or by the cloud. If your repository manages several kinds of storage solutions that are substantially different, you may want to fill in several copies of this section. You may find in the distribution bundle a separate file with a copy of this section. You may duplicate it as you need.

3.1 Description

Short description of this type storage system the form refers to.

3.2 Purpose

Are the copies of digital objects stored in this storage system used only for access, only for preservation or for both?

3.3 Technology

Which technology exploits this storage system? If appropriate, if for instance several layers of storage are included, you may check several.

3.4 Experience

How long is your experience in using this specific solution?

3.5 Redundancy

Some storage solutions involve storing several replications of the same object. This may be achieved by replicating the data within a single device, e. g. a RAID 5 NAS, or by storing different copies in several devices in the same location, or in different locations.

3.6 Backups

Specify at which intervals are backup performed. By backup we mean generating additional copies of the preserved objects to be stored and preserved elsewhere.

3.7 Staff

How many of your employees are entirely devoted to manage this storage system?

3.8 Cost

Provide even a rough estimate of the Total Ownership Cost of this storage system. Includes everything: purchase, maintenance, personnel, rental cost etc.

3.9 Controls

Specify which kind of controls (if any) are regularly scheduled to check the integrity of the data stored within this storage system.

3.10 Perceived weakness

Do you rate this solution as completely satisfactory, or you perceive some weakness in it?

3.11 Time horizon

How long are you planning to continue using this solution, before obsolescence or other problems may force you to move to a different one.

3.12 Risk assessment

The probability of losing the data refers to this level of storage, i.e. means you become unable to access them within this storage system. This does not means that you have lost your data, since you may still be able to access redundant copies or backups stored in different systems. The same goes with availability.

3.13 Additional comments

Add any further information you may like to provide.

4. OUTSORCED STORAGE SOLUTIONS

If your repository relies on several storage services, you may want to fill in several copies of this section. You may find in the distribution bundle a separate file with a copy of this section. You may duplicate it as you need.

4.1 Description

Specify the kind of service and how it is provided

4.2 Name of the provider

Name of the vendor or of the organization that is providing you the service?

4.3 Classification

Check the appropriate type or provide your own, by giving a concise definition of it.

4.4 Experience

How long have been using this storage service?

4.5 Redundancy

Does the provider specify which kind of redundancy is implemented to guarantee the persistence of the stored data? Sometimes it does.

4.6 Terms of agreement

Persistence, access and security (confidentiality of the data) are the three main issues. What your provider guarantees you and in which terms.

4.7 Storage cost

There may be different billing schemes. Some include everything, some bill access independently from storage. This should be the basic cost, including storage, but not necessarily access.

4.8 Access cost

According to different schemes, one may pay per access and/or per MB.

4.9 Risk assessment

The probability of losing the data refers only to this level of storage, i.e. means you become unable to access them through this storage service. This does not means that you have lost your data, since you may still be able to access redundant copies or backups stored in different systems. The same goes with availability.

5. MOVING TO AN OUTSORCED STORAGE SOLUTIONS

By outsourced storage solution we mean buying storage as a service commercially sold by a third party, instead of setting up and managing directly a private storage. Even if you have currently no intention to move to such storage solution, it may still be very interesting that you fill in this section of the questionnaire, at least to specify which outsourced storage solutions you have considered (or you are considering) and which could be, in your opinion, the pro and cons of such a move.

5.1 Which outsourcing schemes are you willing to consider?

By public cloud we mean networked online storage commercially sold as a service by a vendor. Instead, by private cloud we mean a similar service set up by a consortium, with access limited to the member of the consortium. Setting up a private cloud allows typically to meet specifications that are not met by public clouds (at least in the opinion of consortium members). Finally by redundant cloud we mean a scheme where data are stored in several public clouds, managed by different vendors, so providing further redundancy, and potentially avoiding vendor lock-in.

5.2 Motivations

Specify which advantages you may expect from moving to an outsourced storage solution. The most likely motivations are listed, but, if you can see a different one, please indicate it.

5.3 Fears

Specify which disadvantages you may expect from moving to an outsourced storage solution. The most likely motivations are listed, but, if you can see a different one, please indicate it.

5.4 Additional comments

Add any further comment you may like to provide.

4.3 ANNEX III: QUESTIONNAIRE GLOSSARY

Access Policy

- Written statement authorized by the repository management that describes the approach to be taken by the repository for providing access to objects accessioned into the repository. The Access Policy may distinguish between different types of access rights, for example between system administrators, Designated Communities and general users. OAIS
- A set of rules and guidelines that determine how the institution's collections, services, products and databases are accessed. ANZ

Archive

- An organization that intends to preserve information for access and use by a Designated Community. OAIS
- An organisation (or part of an organisation) responsible for appraising, acquiring, preserving and making material available. ANZ

Designated Community: An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities. A Designated Community is defined by the archive and this definition may change over time. OAIS

Digital Archive: A repository for the long-term maintenance of digital resources and making them available. ANZ

Digital Object

- An object composed of a set of bit sequences. OAIS
- Digital objects are units of content managed by a digital archive. Digital objects have as one of their attributes an Identifier. They can be seen to be the atomic level of content. Smaller units can be contained within them, but the repository manages the digital object in a singular fashion. ANZ

Digital Preservation

- Refers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary. Digital preservation is defined very broadly for the purposes of this study and refers to all of the actions required to maintain access to digital materials beyond the limits of media failure or technological change.Those materials may be records created during the day-to-day business of an organisation;"born-digital" materials created for a specific purpose (e.g. teaching resources); or the products of digitisation projects.This handbook specifically excludes the potential use of digital technology to preserve the original artefacts through digitisation. DPC
- Principles, practices, methods, strategies and managed activities that ensure long term preservation for continued access to digital materials for as long as necessary. ANZ

Digital Rights Management: An umbrella term referring to any of several technical methods used to control or restrict the use of digital content. ANZ

Format

- Specific, pre-established structure for the organisation of a file or bitstream. ANZ
- "The physical medium in which information is recorded or carried, e.g. paper files, computer printout, photographs, microfilm, [electronic] records, plans, cards, volumes, etc. A selection of descriptive elements set out in a prescribed manner and sequence so that the resulting description will be standardised for all types of records." ANZ

Metadata

- Data about other data. OAIS

- Information which describes significant aspects of a resource. Most discussion to date has tended to emphasise metadata for the purposes of resource discovery. The emphasis in this handbook is on what metadata are required successfully to manage and preserve digital materials over time and which will assist in ensuring essential contextual, historical, and technical information are preserved along with the digital object. DPC
- Structured information that describes and/or allows users to find, manage, control, understand or preserve information over time. Source: ICA req OR "Data describing context, content and structure of records and their management through time." (ISO 15489-2001) ANZ

Preservation: The processes and operations in ensuring the technical and intellectual survival of objects through time. Source: derived from Continuum ANZ

Producer: The role played by those persons, or client systems, who provide the information to be preserved. This can include other OAISs or internal OAIS persons or systems. OAIS, ANZ

Repository: "The building or room, or part thereof, set aside for the storage of archives and/or intermediate records. Archival repositories are often constructed to meet specific environmental standards designed to ensure the longevity of the records." ANZ

Retention Period: "The period of time, usually based on an estimate of the frequency of current and future use, and taking into account statutory and regulatory provisions, that records need to be retained before their final disposal. Sometimes used to indicate the length of time records are to be retained in offices before being transferred to intermediate storage." (KA, p.479). The retention period usually commences from the time of the disposal trigger. ANZ

Version: An attribute of an AIP whose information content has undergone a transformation on a source AIP and is a candidate to replace the source AIP. OAIS

Glossary References

OAIS - Open Archival Information System - Reference Model - ISO 14721:2012
http://www.alliancepermanentaccess.org/index.php/knowledge-base/member-resources/digital-preservation-glossary/

DPC - Digital Preservation Coalition
http://www.dpconline.org/advice/preservationhandbook/introduction/definitions-and-concepts

ANZ - Archives of New Zealand
http://archives.govt.nz/advice/continuum-resource-kit/glossary/definitions-full-list

5 REFERENCES

[1] *"D23.1 Storage Solution Summary of Inputs"*, Doc. Identifier APARSEN-REP-D23_1-01-1_0 http://aparsen.digitalpreservation.eu/pub/Main/ApanP2ReviewDocumentsForReviewers/APARSEN-REP-D23_1-01-1_0.pdf

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

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