DELIVERABLE

Project Acronym: **DCH-RP**

Grant Agreement number: **312274**

Project Title: **Digital Cultural Heritage Roadmap for Preservation - Open Science Infrastructure for DCH in 2020**

Deliverable D3.5 Final version of the Roadmap

Revision: final v1.0

Authors:

Borje Justrell (RA)
Antonella Fresa (PROMOTER)

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	X
C	Confidential, only for members of the consortium and the Commission Services	

Revision History

Revision	Date	Author	Organisation	Description
0.1	28/06/2014	Borje Justrell	RA	Initial structure (outline)
0.2	19/09/2014	Borje Justrell Antonella Fresa	RA PROMOTER	Information added in all chapters
0.3	01/10/2014	Börje Justrell	RA	New text mainly in the Executive Summary, Conclusions and Annexes
1.0	01/10/2014	Claudio Prandoni Antonella Fresa	PROMOTER	Added Section “A web-space for the Roadmap” and final check

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

TABLE OF CONTENTS

LIST OF FIGURES

GLOSSARY

Specific terms and the definitions used in this deliverable:

Born Digital - Digital materials which are not intended to have an analogue equivalent.

Cloud computing - a phrase used to describe a variety of computing concepts involving a large number of computers connected through a real-time communication network such as the Internet.

Digital archaeology – the process of retrieving a digital resource which has become inaccessible and unusable due to technological obsolescence and/or poor preservation of metadata about its format, structure and content (for digital records also its appearance).

Digital asset – the material produced as a result of digitisation or digital photography; the term includes also more complex accumulations such as online learning resources, web pages, virtual reality tours and digital/visual files.

Digital curation - has wider coverage than digital preservation and involves maintaining, preserving and adding value to digital data throughout its life-cycle.

Digital preservation - a set of activities required to make sure digital objects can be located, rendered, used and understood in the future.

Digital record – any information that is recorded in a form that only a computer can process and that satisfies the definition of a record as stated in the formal regulation and/or the policy for the cultural institution in mind.

Digital resources – encompasses both digital records and digital assets.

Digitisation – the process of converting analogue data carriers (parchment and paper records, microforms, photos, film and audio and video tapes) into digital form using scanning, digital photography, or other conversion methods.

E-Infrastructure - the term used for the technology and organisations that support research undertaken through distributed regional, national and global collaborations enabled by the Internet. It embraces networks, grids, data centres, and collaborative environments; it can also include supporting operations centres, service registries, single sign-on, certificate authorities, training, and help-desk services.

Grid computing - the collection of computer resources from multiple locations to reach a common goal.

Hub - a common connection point for devices in a network (could be of different kind).

Memory institutions - a metaphor used about a repository of public knowledge; a generic term used about institutions such as libraries, archives, museums, clearinghouses, electronic databases, and data archives, which serve as memories for given societies or mankind as a whole.

Metadata – information about data required to manage, search, understand, use, and preserve it.

Mashup - in web development, a web page, or web application, that uses content from more than one source to create a single new service displayed in a single graphical interface.

NUMERIC Study – a study on statistics on digitisation of cultural material in Europe; built on the results of this study a EC- funded project, ENUMERATE led by Collections Trust in the UK, has the task to create a reliable baseline of statistical data about digitisation, digital preservation and online access to cultural heritage in Europe.

Ontology – a structural framework for organising information; used in artificial intelligence, the Semantic Web, systems engineering, library science, information architecture etc as a form of knowledge representation about the world or some part of it.

Persistent identifier - a long-lasting unique reference to a digital object, which could be a single file or set of files.

Virtualisation - refers in computing to the act of creating a virtual (rather than actual) version of something, including a virtual computer hardware platform, operating system (OS), storage device, or computer network resources.

Visualisation - any technique for creating images, diagrams, or animations to communicate a message. Visualisation today has ever-expanding applications in science, education, engineering (e.g., product visualisation), interactive multimedia, medicine, etc.

ABBREVIATIONS

AAI	Authentication and Authorization Infrastructure
AIP	Archival Information Package
API	Application Programming Interface
AQuA	Automated Quality Assurance Project
CHI	Cultural Heritage Institution
COPTR	Community Owned Preservation Tool Registry
CLARIN	Common Language Resources and Technology Infrastructure
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DIP	Dissemination Information Package
DCH	Digital Cultural Heritage
DC-NET	Digital Cultural Heritage NETwork
DCH-RP	Digital Cultural Heritage – Roadmap for Preservation
DP	Digital preservation
EC	European Commission
e-IRG	e-Infrastructure Reflection Group
EU	European Union
EUDAT	European Data Infrastructure
GRID	See Grid computing
ICT	Information and Communication Technologies
HPC	High Performance Computing
HW	Hardware
IaaS	Infrastructure as a Service
INDICATE	International Network for a Digital Cultural Heritage e-Infrastructure
MW	Middleware
NGI	National Grid Initiative
NARA	National Archives and Records Administration (US)
NREN	National Research and Education Network
OAIS	Open Archival Information System
PaaS	Platform as a service
PB	PetaBytes
PEST	Political, Economic, Scientific, Technological
PoC	Proof of Concept
PraaS	Preservation as a Servic
PSNC	Poznań Supercomputing and Networking Center
RAID	Redundant array of independent disks (earlier: Redundant array of inexpensive disks)
SaaS	Software as a Service
SCAPE	SCAlable Preservation Environments
SIP	Submission Information Package
SOA	Service Oriented Architecture
SW	Software
TDR	Trusted Digital Repository
TB	TeraBytes
VPN	Virtual Private Network
VRC	Virtual research Community
VRE	Virtual Research Environment
VRO	Virtual Research Organization

1. EXECUTIVE SUMMARY

This deliverable presents the final version of the roadmap for digital preservation that the DCH-RP project is tasked to design. The aim of the DCH-RP project is to develop a roadmap to implement a preservation infrastructure for digital cultural heritage, primarily targeting two main communities in order to help them plan ahead:

- policy makers on different levels and
- owners of digital preservation programmes at cultural heritage institutions

The aim is also to assist managerial teams of cultural heritage institutions in taking decisions related to digital preservation.

Along the way, the project has put more and more emphasis on targeting also the providers of e-Infrastructure services.

The DCH-RP roadmap is built on two basic assumptions: firstly, that existing e-Infrastructures for research and academia are efficient channels also for the delivery of advanced services to be used by the digital cultural heritage sector for distributed digital preservation and, secondly, that it will be possible to establish common policies, processes and protocols which will allow digital DCH organisations to access e-Infrastructures, despite the fact that NRENs and NGIs are national entities, sometimes with different policies and procedures for access and usage.

This deliverable aims to provide a practical instrument to decision makers, offering an overview of the principal problems and challenges that the digital preservation poses, a range of references to existing solutions and a critical synthesis of the steps that memory institutions and policy makers should be ready to take. It consists of:

- An *introduction* to the many aspects associated with digital preservation and the main challenges to be faced while planning the implementation of a digital preservation system;
- A *list of services* that should be taken into consideration as priorities for memory institutions addressing preservation of their digital objects;
- A *Roadmap for digital preservation* focusing on the preservation landscape and its major areas to address; this part also include condensed versions of the roadmap short-term as well as medium- and long-term;
- An *action plan* which proposes how to establish a value chain and specific actions to take in the major areas of the Roadmap.
- A *web-space* which has been created to ensure the sustainability of the Roadmap.

A separate section concludes with a general review of the results in this deliverables.

A total of four annexes are also included:

- A trust model suitable for the use of e-Infrastructures
- Prioritised actions in selected areas of the Roadmap
- License agreements and terms of usage
- Best Practices

2. INTRODUCTION

2.1 STRUCTURE OF THE DOCUMENT

This deliverable is the final version of the roadmap for digital preservation that the DCH-RP project has been committed to design. The deliverable is organised as follows:

Section 2 (Introduction) - sets out the structure of the document and the objectives of the deliverable, including a short review of the main input provided in other deliverables;

Section 3 (Setting the Scene) - offers an overview of the general context for the deliverable;

Section 4 (**Services to Address**) – presents services to address and requirements to set up when planning for distributed digital preservation;

Section 5 (A Roadmap for Preservation) – presents the different parts of the Roadmap including condensed versions short- , medium- and long-term focusing on what to do and when;

Section 6 (An action Plan) – propose how to establish a value chain and actions to take in identified major areas of the Roadmap when planning for the use of distributed digital preservation services;

Section 7 (A web-space for the Roadmap) – presents the dedicated web page that has been created in the DCH-RP showcase on Digitalmeetsculture to preserve, maintain, update, discuss and keep alive the Roadmap;

Section 8 - summarises on a general level the results in previous sections;

Annex 1 - sets forth a trust model suitable for the use of e-Infrastructures;

Annex 2 – contains a selection of prioritised actions in selected areas of the Roadmap

Annex 3 – describe issues on license agreements and terms of usage

Annex 4 – gives an overview of best practices captured during the project.

2.2 OBJECTIVES OF THE DELIVERABLE

Unlike digitisation, where common approaches and best practices are well developed, digital preservation is still an area where workflows and easily applicable universal toolkits are not widely available, although the toolbox is constantly being topped up. Current solutions normally require adaptation to the specific mandate of the individual cultural heritage institution, its existing technological infrastructure and the competences of its staff. The cultural heritage sector is also producing a large volume of digital content that needs to be safely stored, permanently accessed and easily re-used over time by different end-user groups. Improving digital preservation practices in cultural heritage institutions is, without any doubt, a complex task.

The need to address this situation and to offer concrete and robust support to cultural heritage institutions efforts in digital preservation was identified by the former INDICATE project.[1] To get an understanding of the magnitude of the situation, an initial survey of existing digital preservation tools and services was

[1] http://www.indicate-project.eu/

commissioned by its sister-project DC-NET.[2] Therefore, the DCH-RP project can be seen as a logical follow-up of both the INDICATE and DC-NET projects.

The aim of the DCH-RP project is to develop a roadmap to implement a preservation infrastructure for digital cultural heritage. The roadmap targets primarily two main communities in order to help them plan ahead:

- policy makers on different levels and
- owners of digital preservation programmes at cultural heritage institutions

The aim is also to assist managerial teams of cultural heritage institutions in taking decisions related to digital preservation.

Along the way, the project has put more and more emphasis on targeting also the providers of e-Infrastructure services.

The design of the roadmap has been supported by practical experiments (proofs of concept) in the project partners' countries. The fact that the volume of DCH data produced is continually increasing, implies a substantial annual investment in preservation which is demonstrated by the figures presented in the NUMERIC study.[3] This study outline the findings of a survey conducted among cultural institutions in EU member states during 2007-2009. The value of annual budgets for digitisation at European cultural heritage institutions was at that time estimated to be in total 80 million euro (staff time devoted to digitisation work only partly included).

In addition to the challenge of the growth of digital resources, the DCH sector also has the challenge of the complexity of the information itself. Common procedures and workflows, shared internationally, would reduce the cost both in terms of time and money to be allocated to this task and would contribute to the general interoperability and openness of scientific DCH data. The so-called 'hard sciences' are already demonstrating that research can advance its capability by the use of e-Infrastructures offering high-speed connections, shared computing and storage resources, sophisticated authentication and authorisation mechanisms etc. A basic assumption is, therefore, that existing e-Infrastructures for research and academia (including NREN, NGI and other data infrastructures) could also be efficient channels also for the delivery of advanced services that can be used by the digital cultural heritage sector in the field of digital preservation.

Another foundation of the work is the assumption that it will be possible to establish common policies, processes and protocols which will allow digital cultural heritage (DCH) organisations to access e-Infrastructures, despite the fact that NRENs and NGIs are national entities, often with different policies and procedures for access and usage.

A first step in the development of the DCH-RP roadmap for preservation was presented in deliverable D3.1 *Study on a Roadmap for preservation*, which provides an analysis of key characteristics and requirements of digital preservation in cultural heritage institutions and how they could be linked with e-Infrastructure services, and a framework and a preliminary action plan for the development.

Deliverable D3.1 also looks at types of analysis that are required and propose a possible timeline for the roadmap.

[2] See Digital Preservation Services: State of the Art Analysis by Raivo Ruusalepp and Milena Dobreva (for the DC-NET project) at http://www.dc-net.eu

[3] http://cordis.europa.eu/fp7/ict/telearn-digicult/numeric-study_en.pdf

A second step was deliverable D3.4 *The Intermediate Roadmap*, which aims to provide a first description of what the roadmap of preservation will look like. It took into account the feedback of all other activities in the DCH-RP project during the first year of its life-time. This deliverable was broadly disseminated and presented at work shops, conferences and other meetings in order to get as much feed back as possible (see deliverable D2.3.2 *Report on dissemination activities*). The outcome of these activities has been handled by the DCH-RPs work packages and set out as parts of their provided input to the roadmap.

This input has been as follows:

- WP2 (Dissemination and Sustainability) has produced deliverable D2.2.2 *Report on community building, stakeholders consultation and sustainability strategy* which describes also also the results of a survey based on deliverable D3.4 *The Intermediate Roadmap;* the aim of it was to fine-tune the final version so it aligns as closely as possible with the priorities of cultural heritage institutions that are either implementing digital preservation programmes or are planning to do so.
- WP3 (Preservation Roadmap) has, besides the above mentioned deliverables D3.1 *Study on Roadmap for Preservation* and D3.4 *The Intermediate Roadmap*, produced
    - deliverable D3.2 *Standards and Interoperability Best Practices Report* which presents (and refer to) standards, best practices, and identifiers of interest for the Digital Cultural Heritage (DCH) sector;
    - deliverable D3.3 *Registry of Services* which presents the structure and initial content of a Registry of Services for digital preservation purposes; the registry collects and describes information and knowledge related to tools, technologies and systems applicable for the purposes of digital cultural heritage preservation, and reviews existing and emerging services developed and offered by R&D projects, public organisations and commercial solution vendors.
    - an analysis of IaaS and future DCH preservation opportunities (published as Annex 2 in deliverable D3.4).
- WP4 (Case Studies and Best Practice) has produced
    - deliverable D4.1*Trust Building Report* which reports on trust and trust building, an issue identified as a key one for the DCH-RP roadmap;
    - deliverable D4.2 *Engagement with commercial publishers* which documents factors that may make publisher partnerships with the DCH community more or less likely, and review experiences from e-journal preservation.
    - deliverable D4.3 *Report on e-Infrastructure Concertation which* documents the outcome of a concertation workshop with the main topic to discuss deliverable D3.4 *The Intermediate Roadmap*; in this workshop participate DCH-RP partners, representatives of DCH organisations from several European countries and representatives of e-infrastructures.
- WP5 (Proofs of Concept) has conducted two proofs of concept using the SCRUM methodology and with WP3 acting in the role of the product owner. The overall objective of these PoCs was to validate in concrete experiments assumptions and concepts expressed in the DCH roadmap to preservation. The main results are reported in
    - deliverable D5.3 *Report on the First Proof of Concept* which reports of conducted Proofs of Concept handling seven out of 14 identified scenarios covering the

following fundamental concerns of digital preservation: (1) Operational challenges,(2) End user concerns and (3) New services and infrastructure integration;

- deliverable D5.4 *Report on Second Proof of Concept* which, focused on integrated solutions and services, reports of four conducted experiments covering a wide variety of solutions with the potential to implement parts of the DCH roadmap to a satisfactory level and with reasonable integration effort.

As a spin-off from in the second Proof of Concept, WP5 published the report *Digital Preservation at Cultural Heritage Institutions in Sweden: Analysis of the Current Situation and of Future Needs and Requirements.*

3. SETTING THE SCENE

3.1 PRESERVING DIGITAL OBJECTS

3.1.1 Definitions and strategies

The importance of preserving digital objects is well understood in today's society. Hardware and media obsolescence, lack of support for older computer formats, human error as well as malicious software can all lead to loss of digital objects. If several of these factors are at hand, the higher is the probability that it will occur. Preservation, however, is not concerned only with sustaining single digital objects. To be used meaningfully in the future, digital objects should be preserved in contexts which make them understandable to future users.

Digital preservation is defined by the DigitalPreservationEurope project as "a set of activities required to make sure digital objects can be located, rendered, used and understood in the future".[4] A more comprehensive term 'digital curation' is often used in parallel with digital preservation. It has a wider meaning and involves "maintaining, preserving and adding value to digital data throughout its life-cycle".[5]

The key challenge in preserving usability of digital objects over time is to overcome technology obsolescence, but a set of other issues around managing collections of digital objects is also involved.

During the past two to three decades, focus has moved from finding the 'ideal' long-term storage media to weighing the advantages and risks of different digital preservation strategies, and to define practical solutions based on standards that may use a number of strategies concurrently. Today, there are several strategies available for sustaining the use of digital objects in the future. The main ones are shown here:

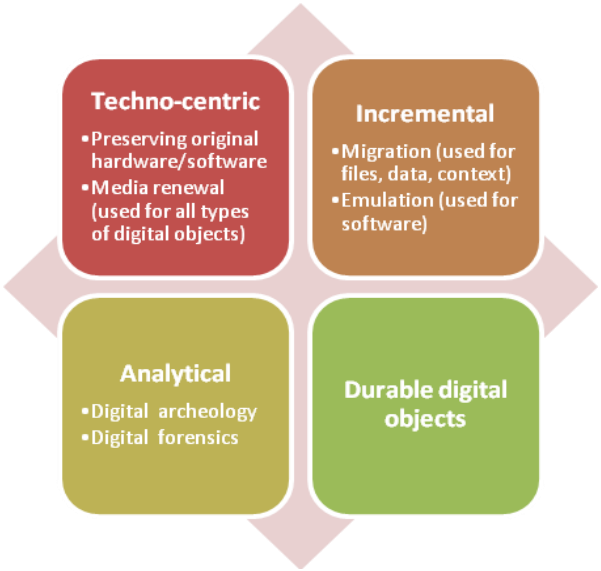


Figure 2: Strategies for sustaining the use of digital objects

Source: *Digital Preservation Services: State of the Art Analysis* (Raivo Ruusalepp and Milena Dobreva)

The *techno-centric strategy* aims to preserve original hardware and software in a usable state in the future. It involves regular storage media renewal to make sure that the physical digital objects are not corrupted.

[4] http://www.digitalpreservationeurope.eu/what-is-digital-preservation/
[5] http://www.dcc.ac.uk/digital-curation/what-digital-curation

Incremental change relies on either migration of digital objects into new formats or preserving the formats of the digital objects and using emulation to be able to use them. The migration strategy normally uses standardised file formats which are repeatedly converted to keep up with present technical generation. The emulation strategy preserves the original file formats and uses emulation at alternative levels. During technical generation changes either to the original software, to the original operating system or to the original technical platform are emulated into the new technical environment, in the latter cases combined with preserved original software.

Analytical strategies are currently based on techniques used in computer forensics. The underlying logic for this strategy is to apply specialised methods for recovery of objects which are in demand in the future instead of 'mass preservation' which does not seem realistic, having in mind the volume of digital information involved.[6] This is basically a strategy for selecting digital objects to be stored long term and methods most suitable for preserving them.

Yet another strategy seeks for methods of changing the formats of the digital objects in a way which allows the objects themselves to invoke preservation actions. Such objects are some times called *Durable digital objects*.

The first three strategies require rigorous organisation of processes in organisations; the fourth one is still under development. All these strategies outline the principles of preservation; in practice they are implemented within archival lifecycles that integrate various tools and/or services. These lifecycles can be specific to organisations, depending on organisational mandate, the types of object they hold, and their target users.

Of the strategies mentioned here, the migration strategy has for a long time been the dominant one. Combined with the OAIS model - see below - it is used by most institutions working with digital preservation. Standardised file formats are normally used for the digital objects to be preserved. To avoid technical obsolescence the digital objects are converted to new standardised file formats at the point of technical generation changes. These conversions are expected to be carried out without information loss. In the foreseeable future the migration strategy will probably continue to be the most used one, at least for in-house preservation. In a longer perspective, increased use of distributed preservation services like e-Infrastructures may change this situation.

Regardless of chosen strategy or combination of strategies, cultural heritage institutions often make a distinction between the master version of digital data and at least one surrogate delivery version. The master version should contain as much intellectual, visual or audio content as possible, be saved in a standard (non-proprietary) file format, and preferably be duplicated across multiple locations. Delivery versions of data may be re-sized, compressed, and saved in whichever format is suitable for delivery to the user. Delivery versions are typically of lower quality (more compressed) than their original master files.

3.1.2 The OAIS model and the analysis of preservation layers

The diversity of both digital objects and types of cultural heritage institutions that are responsible for their preservation creates variations in the level of tools used in practice, but the underlying process could be described as universal. The pivotal standard in the domain, *ISO 14721:2003 Space data and information transfer systems – Open archival information system – Reference model*, widely known as the OAIS model, is a functional framework that presents the main components and the basic data flows within a digital preservation system. It defines six functional entities that synthesis the most essential activities within a digital archive: ingest, preservation planning, archival storage, data management, administration,

[6] The pioneering work in this domain was called *digital archaeology*

and access. Recently, some major European libraries have proposed to combine these six stages into a smaller number of use-cases that preservation systems address.[7]

The OAIS model looks at data stored in the digital archive as a fluid object that can (co-)exist as three types of information packages:

- Submission (SIP) is used to transfer data from the producer to the archive;
- Archival (AIP) is used for the archival storage and preservation;
- Dissemination (DIP) is used within the access function when consumers request archived materials.

As a reference model, the OAIS standard does not imply a specific design or formal method of implementation. Instead, it is left to users to develop their own implementation by analysing existing business processes and matching them to OAIS functions.

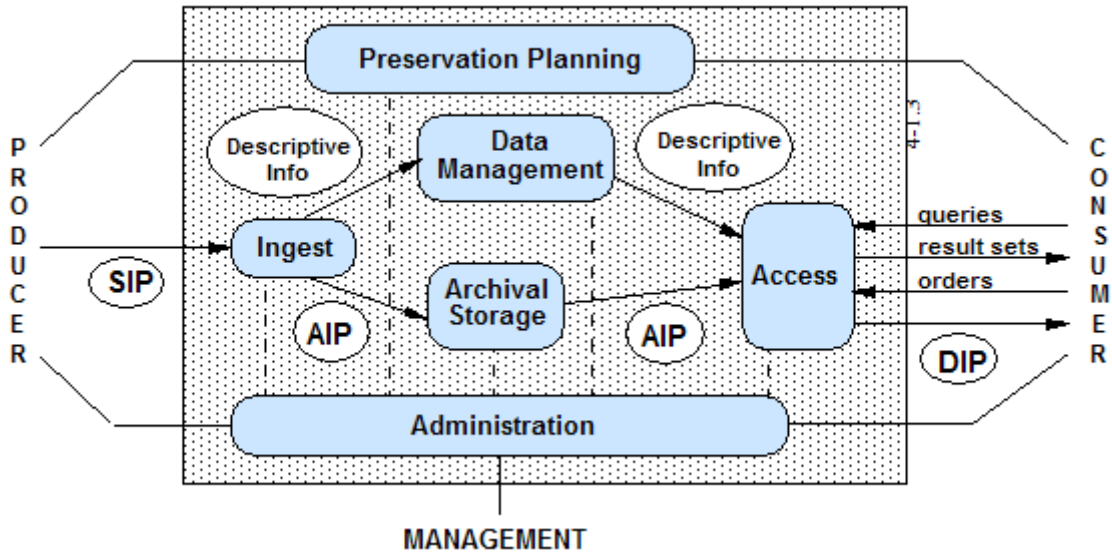


Figure 3: The OAIS functional model

In maintaining the accessibility and usability of digital objects over time, an often used method for analyzing them is built on the presumption that every digital object consist of three layers: a physical, a logical and a conceptual. All three layers and their relations have to be considered and understood in order to get proper preservation actions. These actions are often identified and referred to as “bit preservation”, “logical preservation”, and “semantic preservation”.

Bit presentation is seen as a number of basic actions ensuring the integrity of the 0:s and 1:s (the sequence code) over time and serves as the ground pillar for any other preservation actions.

Logical preservation focus on the representation of the digital object and activities in this field has the aim to ensure the quality of being able to retail the object and maintain accessibility over time. File format sustainability is of course one major issue here. Much effort has over the year been made on setting up requirements and recommendations for file format sustainability.

[7] A report of four major national libraries in Europe looks at three core functions – ingest, retention, and access. See BL, KB, DNB, NB, 2010. See also http://public.ccsds.org/publications/archive/650x0m2.pdf

Semantic preservation includes activities focusing on long-term understandability of the content but also on capturing contextual information about the domain/environment in which the digital object was created.[8]

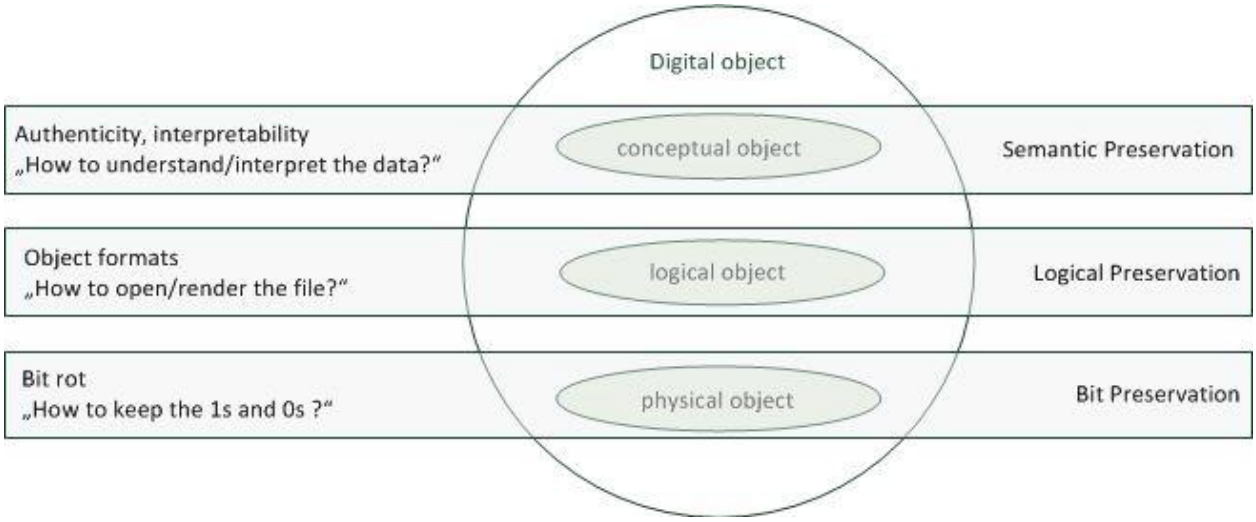


Figure 4: The layers of a digital object

Source: EU project DURAARK, deliverable D6.6.1 Current state of 3D object digital preservation and gap-analysis report, see http://www.duraark.eu/deliverables

Over the past decade, automation of preservation functions has mainly been seen within the context of holistic software solutions that provide digital collection management as well as digital preservation tools. The digital repository software and digital archive software solutions have dominated the preservation software market while not always providing support for active digital preservation. Since digital repository software has been available as open source, it has become very popular, especially for research libraries acting as 'institutional repositories'. Companies, like IBM,[9] Tessella,[10] ExLibris[11] and others, have developed dedicated software systems for digital archive management. While very practical as digital collection management tools, not all repository software solutions offer support for long-term digital preservation.

The core challenges addressed by DCH-RP are in the first place targeted towards the OAIS preservation functions, but they are interconnected with a number of other functions that together form the digital archive.

3.1.3 Digital preservation and roadmaps in an European context

Member States of the EU have taken the position that the preservation task should be their responsibility. Therefore, each Member State is developing and implementing its national preservation strategy, which includes the preservation of digital master copies that takes place at national memory institutions or at other public institutions which are the direct responsibility of governments. National frameworks that

[8] See EU project DURAARK, deliverable D6.6.1 Current state of 3D object digital preservation and gap-analysis report, (http://www.duraark.eu/deliverables) and references referred to there; concerning research in file formats , see for example the InterPARES project (E. Peters McLellan, General study 11: Selection digital file format for long-term preservation. Online, March 2007, http://www.interpares.org/ip2/ip2_general_studies.cfm)

[9] http://www-935.ibm.com/services/nl/dias/
[10] http://www.digital-preservation.com/
[11] http://www.exlibrisgroup.com/category/RosettaOverview

regulate this area, like rules on legal deposit and the handling of public records, exist and the publishing sector is also involved (especially with regard to born digital material).

However, there are many commonalities that exist among the national preservation strategies which have to be addressed in common and in a coordinated manner among memory institutions, the Member States of the EU and more generally internationally in order to share solutions and to contribute to interoperability and openness. Common procedures and workflows, shared internationally, would reduce the cost both in terms of time and money to be allocated to digital preservation and would contribute to the general interoperability and openness of scientific data (including research data from the DCH sector) which is stated as the priority for the global knowledge society.

The importance of long-term preservation and its complementarities to digitisation efforts was highlighted in the report of the Comité des Sages (Reflection group on bringing Europe's cultural heritage online) that clearly stated the digital preservation mandate of memory institutions.[12] Also important is the EC Recommendation on digitisation and online accessibility of cultural material and digital preservation[13] published by the EC on 28/10/2011.

The attention and commitment of the EC to research and development in the domain of digital preservation was highlighted at the Commissions expert workshop The Future of the Past, held in Luxembourg in May 2011.[14] This workshop discussed previous research agendas in the domain of digital preservation and formulated a number of potential research topics of high relevance to the future development of the domain, among them digital preservation infrastructure – an area where DCH-RP will provide contributions for the digital cultural heritage domain.

Roadmaps are useful instruments for presenting the scope and coverage of an e-Infrastructure. They are also frequently used within projects and institutions in the digital preservation domain. Some roadmaps can be very detailed as for example the roadmap developed for the UK Parliamentary archives (2008),[15] which presents environmental, policy, preservation, presentation, standards, skills, and communication developments over time. The Open Planets Foundation developed a *Tools and Services Roadmap*[16] to outline their software development plans. The APARSEN project roadmap[17] presents research topics and larger themes; preservation services are a research topic under the theme of sustainability. Some projects use roadmaps to present various formats, e.g. the PrestoSpace[18] project which presents formats for the audio-visual material. There are also a number of national roadmaps, especially in the area of research infrastructures that address arts and humanities.[19]

However, there is not an existing roadmap that the DCH-RP project could build on or progress further. The project has to develop its own roadmap for the specific domain and task that it is addressing. This roadmap will be supplemented by practical tools which will help on one hand the monitoring of activities and thus would be of benefit in a political context, but will also offer knowledge instruments to

[12] The New Renaissance, 2011: 6

[13] Full text of the recommendation is available online at: http://ec.europa.eu/information_society/activities/digital_libraries/doc/recommendation/recom28nov_all_versions/en.pdf

[14] Billenness, C. (2011) The Future of the Past, Report on the Proceedings of the Workshop, European Commission, Luxembourg, 4 – 5 May 2011. Available: http://cordis.europa.eu/fp7/ict/telearn-digicult/future-of-the-past_en.pdf

[15] http://www.parliament.uk/documents/upload/strategy-road-map-final-public.pdf presents the roadmap diagram and http://www.parliament.uk/documents/upload/digital-preservation-strategy-final-public-version.pdf - the justification.

[16] http://www.openplanetsfoundation.org/community/tools-and-services-roadmap

[17] http://www.alliancepermanentaccess.org/index.php/current-projects/aparsen/aparsen-roadmap/

[18] http://wiki.prestospace.org/pmwiki.php?n=Main.Roadmap

[19] See for example the Danish roadmap for RI http://en.fi.dk/publications/2011/danish-roadmap-for-research-infrastructure-2011/uk-roadmap.pdf; Large research (Czech roadmap, 2010) http://www.infrafrontier.eu/docs/national_roadmaps/Roadmap_CR.pdf; Australian humanities infrastructure http://www.paradisec.org.au/blog/2011/03/australian-humanities-research-infrastructure-funding/

stakeholders from the DCH domain (cultural heritage institutions) to make informed decisions on digital preservation.

3.2 MAIN CHALLENGES

The main challenges that the DCH-RP project has to meet have, of course, their roots in difficulties to maintain accessibility and usability of digital objects over time. But they are also connected to questions like the benefit of using e-infrastructure for preservation, what to preserve, sustainability issues and how to raise awareness about the roles of different actors in the implementation of a distributed digital preservation infrastructure.

3.2.1 Making current and future digital information accessible and usable over time

Deliverable D3.4 (The Intermediate Roadmap) concludes that the cultural heritage sector is faced with a number of general challenges in making current and future digital information accessible and usable over time. These challenges are closely related to a number of noteworthy differences between digital and paper-based material. But, it is not only the changing form of the objects that is new. The changing way of work that follows the introduction of objects in digital form, force cultural heritage institutions to integrate new concepts, methods and tools for digital preservation to be carried out in parallel with traditional paper-based preservation.

The Digital Preservation Coalition has pointed out six differences between digital and paper-based material: [20]

Machine Dependency - digital materials require specific hardware and software in order to access them.

Technology obsolescence - the speed of changes in technology means that the timeframe during which action must be taken is very much shorter than for paper. It is measured in a few years compared to decades or even centuries when preserving traditional materials. Technology obsolescence is, therefore, generally regarded as the greatest technical threat to ensuring continued access to and use of digital resources.

Fragility of the media - the media digital materials are stored on are inherently unstable and can detoriate very quickly without suitable storage conditions and management, even though it may not appear to be damaged externally.

Loss of integrity - the ease with which changes can be made and the need to make some changes in order to manage the material means that there are challenges associated with ensuring the continued integrity, authenticity, and history of digital materials.

Doing nothing is not an option - the implications of allocating priorities are much more severe than for paper. A digital resource which is not selected for active preservation treatment at an early stage will very likely be lost or unusable in the near future.

Preservation prior to creation - the nature of the technology requires a life-cycle management approach to be taken to the maintenance of digital resources. A continual programme of active management is needed from the design and creation stage of a computer system and onwards, if preservation of that system is to be successful.

[20] http://www.dpconline.org/advice/preservationhandbook/digital-preservation/strategic-overview

All differences are interconnected, and together they clearly indicate that a radically different approach is required in managing digital objects compared with paper-based materials. It is also important to have in mind that the greatest asset of digital information, the ease with which it can be copied or transferred, is paralleled by the ease with which the information can be corrupted or deleted.

3.2.2 Showing the benefit of using e-Infrastructures for preservation

Deliverable D3.4 summarises the work on the DCH-RP projects road map so far, by saying that the use of e-Infrastructure in meeting the demands mentioned above looks promising. During the assessment of D3.4 it was made very clear (see deliverable D4.3) that there are still no proofs presented, that cooperation between the DCH sector and e-infrastructures for the delivery of advanced services digital preservation gives better value for money than other solutions. The value chain that DCH and e-Infrastructures can create together is still to be defined.

E-infrastructures are not for free, and there are not many commercial distributed preservation systems in place today. E-infrastructures are normally not especially skilled in preservation but have great knowledge about data management. So, if e-infrastructures have to develop preservation system it will be costly, but they can on the other hand benefit from economies of scale.

The DCH-RP project has looked into other domains, to see if there are experiences in digital preservation field that are transferrable to the DCH domain (see deliverable D4.2). In the e-journal preservation community, much has been achieved in terms of evolving mechanisms and organisations to look after digital preservation. The technical, organisational and financial challenges have been proved to be solvable, given strong commitment from the communities involved. The key issue seems to be the ways in which these communities have organised themselves to bring about long-term agreements and infrastructures for preservation

3.2.3 Models for what to preserve

What needs to be preserved, and what can be preserved - and how to make the choice? The question of a selection process arises sharply by the huge amount of digital material that are produced and waiting to be preserved Memory institutions might have a public commitment to preservation (including legal deposits and state archives duties), but one cannot talk about digital culture preservation without talking about value. What is important to some potential users (contemporary and future) is not necessarily important to others.

Criteria for disposal of cultural heritage information and objects are in place but differ between domains and professions. The archival community has, for example, for long used methods to classify archival objects with respect to their worth, mainly based on evidential and informational aspects. Handling this process (called appraisal) is a fundamental part of an archivist's professional duties.

Different types of data and digital objects require also different types of preservation methods and activities. This has, within all sectors in the cultural heritage domain, had an influence on expected areas of expertise and the resulted ina growing number of professionals with hybrid competences (library science/archival science/museology and IT).

Another aspect on what to preserve, is that different countries have different rules about preserving data. It is for example not always possible to preserve valuable data outside a particular country or on a server owned or controlled by a commercial service provider.

3.2.4 Sustainability issues

There are different dimensions of sustainability that the roadmap has to tackle: the sustainability of the roadmap itself and the sustainability (political, financial, organisational, and technological) of the preservation infrastructure to which the roadmap leads.

For the sustainability of the roadmap and thus to ensure a clear direction, the roadmap needs an endpoint and should be maintained as long as this endpoint has not been reached.

For the sustainability of the preservation infrastructure, funding and opting for a pan-European solution is by many regarded as a must (see deliverable D4.3). Broadly recognised and accepted standards are a basic part of the concept. It seems also to be a widespread opinion that digital preservation cannot be realized without funding at national (storage, software, etc) and to some extend European level. The way to a digital preservation arrangement that "runs itself" is too long and too winding. Actions are needed in shorter terms.

To get a sustainable preservation infrastructure there must also be willingness and cooperation in place between DCH organisations at national as well at pan-European level. Otherwise the e-Infrastructure will get a too diffuse and shuttered costumer market for their distributed preservation services, and the roadmap will run the risk to remain an abstract document.

Deliverable D4.2 points out that in the e-journal domain, the governance arrangement developed around e-journal preservation have generally been carefully designed to involve and keep onboard a variety of potentially competing interest – in order to achieve a shared common good that can be sustainable into the future.

3.2.5 Awareness raising

When using distributed preservation services from e-Infrastructure providers, raising awareness is an important part:

Firstly, the owners of digital collections have to understand the importance of preserving their content; training and learning resources should, therefore, be made available for this purpose.

Secondly, cultural managers (museums, libraries, archives directors) have to support the owners in designing the correct workflow and understanding the digital preservation phase as an integral part of the digitisation process. The dramatic speed of technological change has led to a situation where many organisations have not been able to fully articulate their needs in this field, much less employ or develop staff with appropriate skills.

Traditional training in the cultural heritage sector does not always provide knowledge of skills and tools necessary to deal effectively with emerging information technologies. Neither does individual self-improvement by staff members, which smaller institutions sometimes rely on. The step from in-house digital preservation to distributed solutions makes it also crucial to understand the concepts and procedures used in information system applications from a preservation perspective. Even if today's situation has resulted in a growing number of professionals with hybrid competences (library science/archival science/museum science and IT) within all sectors in the cultural heritage domain, these professionals are too few and not equipped enough to fully embrace the preservation implications of digital objects created and embedded in different kind of information system applications.

Thirdly, the policy makers have to understand the need for a plan for investments in preservation as long as they plan for investments in digital assets. Member States have recognised their responsibility in digital preservation, and national preservation strategies are implemented all over Europe. But these strategies have to be followed by concrete efforts in terms of resources, human as well economical. To create digital

objects, born digital or converted from analogue originals, without taking into account how to preserve these objects is a very risky approach.

Fourthly, the users of the digital resources have to be aware of the various contributions they can provide to the digital preservation process (annotations, inputs for content selection, etc.). User involvement is still on a primary stage, but concepts like “crowd-sourcing” and “user-interaction” is rapidly growing in the cultural heritage domain.

4. SERVICES TO ADDRESS

The following services to address and requirements to set up have been identified by the memory institutions that participated in the DCH-RP project as priorities when planning for distributed digital preservation.

Naturally, this list is a high-level summary of the areas of services and requirements to be considered, and each institution need to explore the details with regard to the individual work-flows and specific vocation of its organisation.

The services and requirements discussed in this chapter are aiming to cover the whole preservation process and is also connected to the OAIS model as illustrated in the following figure:

Figure 5: Services to address and requirements to set up

4.1 INGEST

To ingest different record types to an e-Infrastructure-based preservation system, all files

- Need to be checked for integrity and consistency with standards using automated routines that document the outcomes of these checks;
- Need fixity information to be attached to them, including persistent identifiers that will allow for identification and to check file integrity at any point in time.

Meeting these requirements makes it possible for the cultural heritage institutions to evaluate

- To what extent tools for the required ingest processes are in place;
- How well they are running;
- What are the time and effort required;

Check points**:** Tools run without failures - Processes run fast - The integrity of all files can be checked after the ingest process - The level of automation of the entire process is high - Time and effort required is manageable.

4.2 STORAGE

An e-Infrastructure-based preservation system has to store the files in such a way that they can be retained with full accessibility and usability. The authenticity of the files should also be guaranteed. Strategies for replacing obsolete technology with new technology have to be in place.

Meeting these requirements makes it possible for the cultural heritage institutions to evaluate:

- To what extent the requirements on storage are met;
- What are the time and effort required.

Check points: Requirements on formats and standards for raw data are fulfilled - Appropriate metadata standards are in place as well as a trustworthy strategy for replacing obsolete technology - Time and effort required is manageable.

4.3 ACTIVE DIGITAL PRESERVATION

An e-Infrastructure-based preservation system has to have a number of complementary curation services like

- Schedule-based integrity checking
- Dereferencing and deleting
- Migration of (and possibilities to actually move) preserved files to new versions of software and/or hardware
- Possibilities to export data
- Conversion and transformation of data
- Administering retention.

Meeting these requirements makes it possible for the cultural heritage institutions to evaluate

- To what extent an e-Infrastructure is mature enough for implementing active digital preservation;
- What additional capacity it needs to develop in case there are any deficiencies.

Checkpoints: Tools run without failures - Curation services run fast and meet the requirements - Level of transparency is acceptable - The level of automation of the entire process is high - Time and effort required is manageable.

4.4 ACCESS

Needed services are

- List items
- Find items
- Retrieve items
- Emulate
- Administer access

Meeting these requirements makes it possible for the cultural heritage institutions to evaluate how they can select services meeting their needs for access, and how to select from available offers.

Checkpoints: Tools run without failure - To what extent services for access are in place and are running well - Time and effort required is manageable - Matrix of metrics and minimum requirements for quality are in place.

4.5 ORGANISATIONAL ISSUES

There have to be clear agreements on outsourcing in place covering aspects like

- Cost reduction
- Increased effectiveness
- Increased quality
- Acceptable level of resources (technical and human)

- Minimising risks/trust building

Policies for outsourcing have also to be decided by the cultural heritage institutions.

The level of technical and human resources should be taken under control in order to keep it at the acceptable level, as fixed by the archival owner.

Meeting these requirements makes it possible for the cultural heritage institutions to evaluate how e-Infrastructures are able to handle distributed digital preservation.

Check points: Draft text of agreement that both the cultural heritage institutions and the service providers have judged to be right or commendable

4.6 SERVICE ARCHITECTURE

Agreements on standards have to be in place that covers services like

- Data resource setup interoperability
- Aggregation
- Advanced search support
- Persistent identifiers
- User authentication and access control

Meeting these requirements makes it possible for the cultural heritage institutions to evaluate to what extent an e-Infrastructure has the capacity to offer the service architecture needed.

Check points: Draft text of agreement that both the cultural heritage institutions and the service providers have judged to be right or commendable.

5. A ROADMAP FOR DIGITAL PRESERVATION

5.1 THE ROADMAP AS AN INSTRUMENT

The “map” in the roadmap draws the landscape of digital preservation for the DCH sector based on the current situation, but needs also to take into account how the situation will change in the future. Much depends on the maturity of both the preservation workflow in the DCH sector and the preservation services available from the e-infrastructures. Preservation and access needs to have a dynamic approach. It is important that the preservation process does not remain only a post-production task.

The landscape is also changing and at different levels, technical, political and legal. Distributed solutions like government clouds are becoming increasingly prevalent and some DCH institutions may be obliged to make use of them. New data infrastructures with a portfolio of services, including different levels of storage and preservation, are constantly being built. Societal changes have also to be taken into consideration.

The “road” in the roadmap points to an action plan, and actions are needed in a number of areas: tools, services, authentication, trust, governance models, user requirements, funding models and business models, skills / training / awareness etc.

It is clear that many of these areas are relevant not only for digital preservation but exist also in other domains. It should not be forgotten that DCH data is also research data, and many common problems are shared with organisation handling traditional research data.

The DCH-RP roadmap will integrate three domains of necessary intervention (business change, policy framework and better tools) with the major PEST factors (political, economic, scientific, and technological). The compilation of the roadmap will also need integration of a multitude of viewpoints and aspects, both those foreseen in the planning of the project and new ones discovered during the project’s lifetime.

Most cultural heritage institutions have in-house solutions for handling their digital objects. When comparing in-house digital preservation with distributed e-Infrastructure services, it is inevitable that some discrepancies will appear, such as incompatibility of purposes or scope, lack of technical or semantic interoperability, reliance on different standards, and jurisdictional and legal barriers, etc. Therefore, the DCH-RP roadmap has a strong focus on what to do and on the usability of services and technologies. In the first review of the DCH-RP project, made by the European Commission, the need to focus on the usability of services and technologies and on working solutions in the roadmap was underlined

In order to achieve this, the DCH- project has adapted the following structure of the roadmap:

Firstly, it gives a description of the **digital preservation landscape (“the map”),** concentrated on how to meet stakeholder needs and the selection of the main components of the roadmap;

Secondly, it identifies an **action plan (“the road”)** with challenges and advantages to target, practical actions to take up, and services to address.

5.2 DESCRIBING THE DIGITAL PRESERVATION LANDSCAPE – TO MET STAKEHOLDERS NEEDS

5.2.1 A snap shot of the current situation

When the DCH-RP project has met cultural heritage institutions to get their view on distributed digital preservation (in work shops, proofs of concept, surveys etc.), the result is a broad and scattered picture.[21] It can be summarised as follows:

It is important that the long-term preservation issues are taken into account already at the creation phase of digital information; examples given on such issues are format conversion and storage. On the other hand storage of digital objects is often defined as purely technical storage (on bit level), and preservation as securing the stored objects in a way so future users can reach, access, and understand them.

There seems to be fairly good consensus about the time span of preservation solutions:

- Short-term preservation – solutions that are used for a short time, 5 years maximum.
- Medium-term preservation – solutions that are used during a system's lifetime, 10 years maximum.
- Long-term preservation – solutions that are used after the originating system's lifetime, number of years unspecified.

That goes well in hand with an opinion expressed by many of the cultural heritage institutions that digitised objects and "born-digital" objects are to be considered differently, mostly because of the link that exists between a digitised object and its "original" physical form (i.e. a digital object may be just considered a digital copy of a certain physical object).

It seems to be a common opinion that preservations solutions proposed by DCH-RP should be tailored towards domain specific requirements, but on the other hand many cultural heritage institutions say, when asked, that the projects shall propose “hybrid” solutions including both generic elements and others specific to the cultural heritage domain.

There is a general concern in the cultural heritage domain about the continually increased amounts of digital heritage content, which will induce higher costs both for managing it (including storage) and for preservation. For those cultural heritage institutions that already have digital objects, but no process and specific systems in place for treating them long-term, the situation is some time close to be desperate and not manageable. Therefore, the answer if DCH-RP project shall offer any advice on how digital materials are selected for preservation is in most cases a clear YES (in big letters).

The collections and holdings maintained by cultural heritage institutions are some times described in a (so called) ”traditional” way, and international standards for metadata are in these cases seldom implemented fully. Thus, metadata structures differ much between the institutions. However, within the library community, stable and widely accepted standards are frequently used, both for metadata and raw data, and there are often systems for long-term preservation available.

[21] The main inputs used for this section are from deliverable D4.3 *Report of the Concertation Work shop. With e-infrastructures and DCH organisation*, the r*eport Results from the* on-line questionnaire based on deliverable D3.4 *The Intermediate Roadmap*, and the report *Digital Preservation at Cultural Heritage Institutions in Sweden: Analysis of the Current Situation and of Future Needs and Requirements.*

Archive information (both metadata and raw data) is described and preserved according to established principles and international standards. As for libraries, there are often systems for long-term preservation in place.

Inside the institutions a clarification of internal roles in digital preservation (which responsibilities the administrators and managers of information, and their counterparts for systems, should have) is often asked for.

Cultural heritage institutions seem fairly often to be in favour of centralised solutions for storage/preservation and centralised support functions like handbooks and training - but within the cultural heritage sector!

Most of them say, when asked, that they think it is vital, very important or at least desirable that the DCH-RP project shall address outsourcing issues in the context of using shared digital preservation infrastructure and services, but only a third clearly state that they consider outsourcing of any of their preservation- related responsibilities to a shared service.

Most cultural heritage institutions appear not to have links with e-Infrastructure, and about one half of them say that they might be prepared to consider private sector solutions or partnerships as components in their digital preservation programmes. The other half do not know. Just a small number of institutions say no, when asked.

The lack of training is often an upcoming issue in the discussion about digital preservation. Most Member States seem not to have an establish organisation or focal point for professional cultural heritage training in this area. But surprisingly many institutions declare that they are prepared to act as “centre of excellence” to spread best practice on DCH preservation to other institutions.

5.2.2 Distributed digital preservation services

Different parts of the DCH domain have different needs, depending on if they are small or large, the kind of digital objects they have etc. The conditions (e.g. resources) for managing digital preservation differ also quite much. Services for distributed digital preservation, therefore, have not only to be flexible, but also easy to adapt and utilise, and address several areas.

Some basic issue for addressing distributed digital preservation services are:

- functional requirements;
- service types and objects to address;
- type of service architecture;
- level of maturity;
- licensing conditions.

Functional requirements

Getting the right set of functional requirements in place is crucial but also that they are defined in a way that makes it possible to measure how the requirements are met. In the previous Chapter 4 a number of services to address and their functional requirements are set up, following the OAIS model and/or the preservation process. It is done in a way so they can be evaluated and checked.

Service types to choice

There are two main levels of services for distributed digital preservation, which can be considered as basic for the DCH community:

- Level 1: those already available or could easily be made available by e-infrastructures to support digital preservation activities conducted by the cultural heritage institutions. This “kiosk-model” could contain supplementary services like federated authentication, audit and certification, persistent identifiers distribution, which are typical network services that would make work easier for institutions or networks of institutions that manage digital preservation "on their own".
- Level 2: those cloud or grid based "turn-key" services that can offer the entire process covering all the phases and functions of the OAIS model, with a particular focus on storage, curation services and other organisational aspects like trust.

The advantages of such two-level service architecture would be:

- It would allow a gradual approach to digital preservation services, paid or payable on the cloud or grid-based, by cultural heritage institutions that have digital objects but difficulties in managing them; an institutions can initially use the services of level 1 and later upgrade to level 2;
- The different levels of services for digital preservation would be associated with different patterns of costs and, therefore, highly flexible when it comes to decisions about what is reasonable taking into account the financial resources at hand.

Close to the “kiosk-model” is an approach called “microservices” presented just a few years ago. It represents a step away from integrated digital archive systems and is, therefore, under discussion in the DCH community. The key idea with “microservices” is that they allow flexible combinations of specialised solutions for preservation depending on the requirements of a DCH institution. “Microservices” for digital preservation are currently used in the open archival information system Archivematica.[22]

In annex 2 some general service models are presented that summarise existing combinations of offering digital preservation services as federated digital archives or as distributed services relying on cloud or grid providers.

Objects to address

As discussed earlier in this report, preservation is a complex activity. This is not only because of the increasing complexity of digital objects and their growing number; it is also because the contexts of active use needs to be re-created, which means sustaining not only the data, but also any specific software which was used to work with it, and the technological infrastructure. The gradual expansion of preservation towards various types of objects is presented in the following figure:

[22] http://archivematica.org/wiki/index.php?title=Development_roadmap

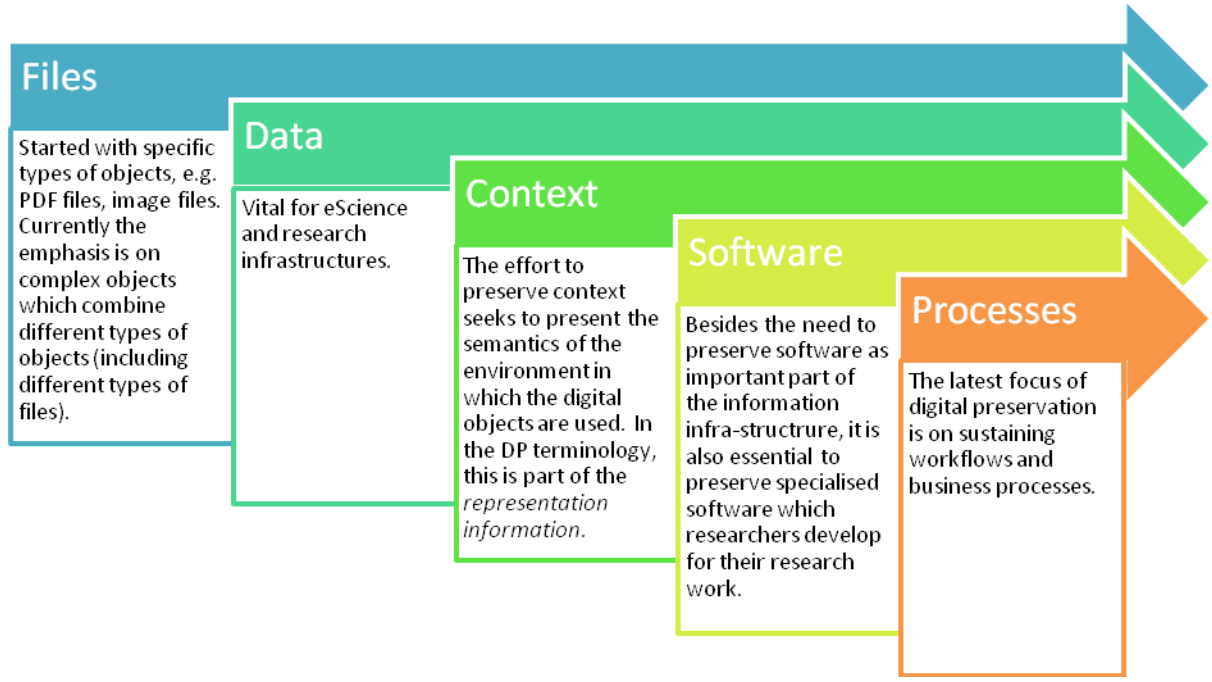


Figure 6: Evolution of digital objects addressed by digital preservation

Source: Digital Preservation Services: State of the Art Analysis by Raivo Ruusalepp and Milena Dobreva (report for the DC-NET project, available at http://www.dc-net.eu

All these different types of digital objects are relevant for digital preservation within cultural heritage institutions as well as in humanities and arts research. Although in many cases the emphasis is on the preservation of computer files, it is important to analyse the need to preserve software, the context of digital objects necessary for their future use, and any processes which also need to be preserved.

Type of service architecture

As mentioned above, the OAIS reference model provides the basic archiving workflow, but it does not articulate clearly how distributed archiving architectures can be arranged. E-Infrastructure service architectures vary significantly and do not allow for a uniform mapping of preservation tools and services to a single architectural model. Conceptualising and modelling of joint service architecture have been undertaken by only a few recent initiatives, and remain in a developmental phase.

The EUDAT project has presented the architecture of a conceptual model that integrates various infrastructures with vast amounts of research data, and adds services for curation and trust in addition to the interface to users. This architecture illustrates a process that will have to be accommodated in the future by most preservation work, where solutions for preservation and curation can be used to support multiple different infrastructures.

As it stands, this model represents basic stakeholder needs in the research area: ensure the trustworthiness of data, provide for its curation, and permit an easy interchange among the generators and users of data. These needs could also be said to be basic ones in the cultural heritage community, and the EUDAT projects conceptual model can, therefore, serve as a base for further development in the cultural heritage sector.

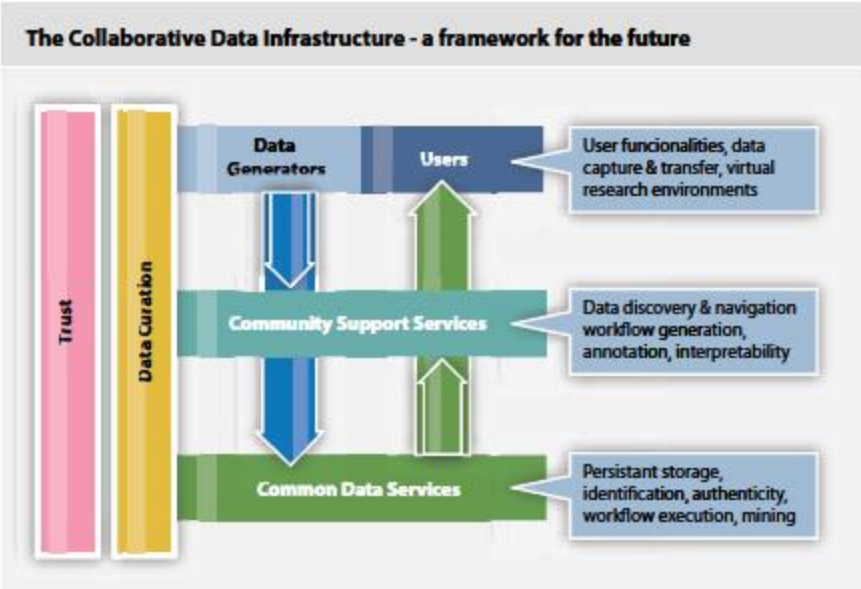


Figure 7: The collaborative data infrastructure - a framework for the future; from *Riding the Wave*, p. 31

Improvements and adjustments of the model have already been made in, for example, the area of research data. The Data Archiving and Networking Services (DANS) in the Netherlands has developed based on the EUDAT conceptual model a federated data infrastructure with three layers of roles and responsibilities for the various stakeholders (The Front office – Back office model) [23]

Since preservation is part of digital objects' lifecycle, it has implications both for the working processes within the institutions and for the professionals working there. The organisational structure of cultural heritage institutions varies and understanding their specific requirements from the distributed preservation infrastructure could be a challenge that is not so easy to handle. It is sometimes argued by the DCH institutions that the uniqueness of their digital holdings requires tailor-made approaches. A comparison of digital preservation provision across major European national libraries and the German Computer Game museum, made some years ago, showed significant differences in the type of holdings which need to be preserved, collection policies, preservation systems and standards used.[24]

It is undoubtedly true, that continuing investment in in-house preservation systems will contribute to the lack of interoperability and fragmentation of resources into "digital silos". Stand-alone solutions that are not transferrable and interchangeable lead to fragmentation and do not offer economies of scale. Instead, shared solutions for creation, storage and use of digital resources, including the e-Infrastructures, will become the major component of the future knowledge economy.

In order to move ahead from the current state into shared, decentralised solutions, it is important to define key institutional requirements in a standardised way. The use of enterprise architecture models is one possible approach because enterprise architectures seek to address system complexity while aligning technological developments with the institutional needs. There are a number of approaches for defining

[23] See www.dans.knaw.nl

[24] The National Library of France develops its in-house preservation system SPAR, OAIS-compliant and based on the use of METS and PREMIS-compliant metadata; The Royal Library of the Netherlands uses the e-Depot system which is based on the IBM DIAS and uses extended Dublin Core bibliographic metadata; The German National Library deployed a combination of tools including kopal-DIAS, koLibRI and has developed its own preservation metadata format, LMER (KEEP, 2009, 54-59;. *Preliminary document analysing and summarizing metadata standards and issues across Europe* (KEEP project deliverable D3.1). Available: http://www.keep-project.eu/ezpub2/index.php?/eng/Products-Results/Public-deliverables

enterprise architectures; one of the popular ones is the Open Group Architectural Framework (TOGAF)[25] and its eight-stage Architecture Development Method that help to manage requirements within complex systems.

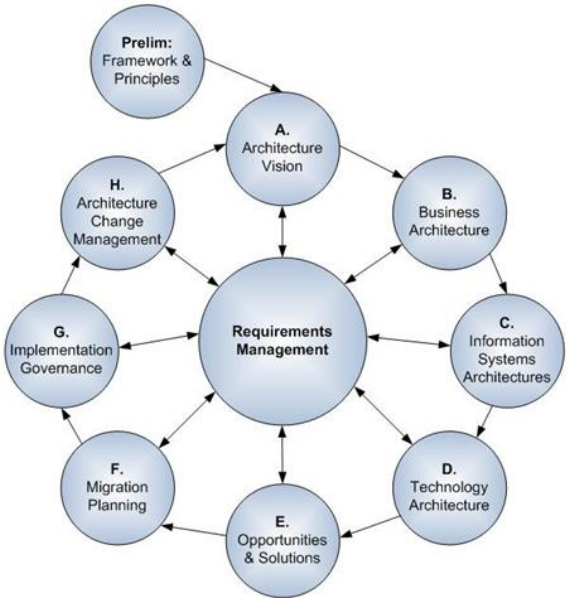


Figure 8: Architecture Development Method, TOGAF.

An earlier framework that looks at the various roles within an organisation and helps to summarise perspectives of various stakeholders on basic modalities of the organisation is the Zachman framework.[26] An adaptation of the Zachman Framework into the digital preservation domain has been done by Raivo Rusalepp and Milena Dobreva in a report conducted for the DC-NET project.[27]

Service architecture as a technical area is very close to service-oriented architecture (SOA), which is a software design and software architecture design pattern based on pieces of software that provides functionality as a service easy to combine into different kind of applications. Services mean in this case not services for the users but services in terms of written functions ready to be used by programmers, and by other applications.

SOA can be seen in a continuum: from older concepts of distributed computing and modular programming on to current practices of mashups, SaaS, and cloud computing, which some see as the offspring of SOA. In the context of the DCH-RP roadmap, aiming at the use of e-Infrastructure, SOA can clearly be regarded as a concept to get inspiration from.

Level of maturity

Tessella has described in a Maturity Model how durable storage, information management and preservation provide increased levels of sophistication aiming at a complete digital preservation strategy.[28]

[25] http://www.opengroup.org/togaf/
[26] Zachman, J. *Concise Definition of The Zachman Framework*. http://zachman.com/about-the-zachman-framework
[27] See Digital Preservation Services: State of the Art Analysis by Raivo Ruusalepp and Milena Dobreva (report for the DC-NET project, avilable at http://www.dc-net.eu)
[28] Preservica – white paper (July 2013) http://preservica.com/resource/praesent-ante-stiam-white-paper/

The term Maturity Model is used to imply layers of sophistication in processes. The first layer must be complete before graduating to the next. In digital preservation, there is no point having a good information management system if you do not have secure storage.

The Digital Archiving Maturity Model has three main parts:

Durable Storage (layers 1-3 in the Model) provides increasing levels of safety and security in the storage of the raw bits used to hold information. A level 3 compliant system implies you can be confident that your information will not be lost and that it has not been manipulated.

Information Management (layers 4-5) ensures that the preserved raw bits are organised. These layers have a hierarchy, descriptive metadata, and security, and they have a set of powerful tools to allow upload, management, search, browse and download.

Information Preservation (layer 6) is critical for information that must be retained for more than the lifetime of the application that created it. It ensures the file formats in which the information is held remain relevant to the applications available at the time the information is required, thus enabling it to be used immediately.

A simple storage archive would fulfill durable storage (layers 1-3) but no more, and a content management archive the information management parts (levels 4 – 5). A specialist digital preservation platform would fulfill all 6 layers.[29]

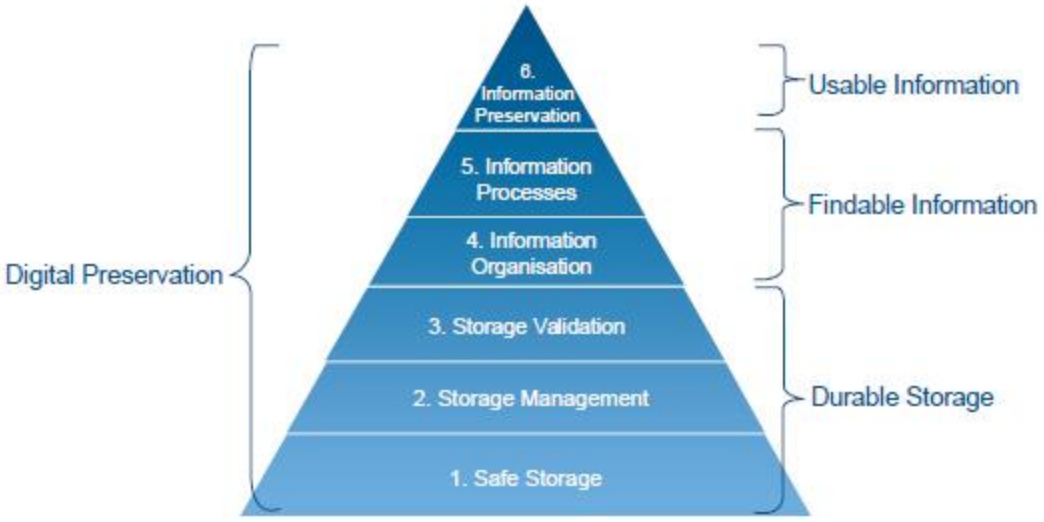


Figure 9: The Digital Archiving Maturity Model

Standard licenses and methods of license expression

The extensive use of relevant and open standards is a vital pre-requisite for the cultural heritage community when promoting interoperability, encourage widespread access and control costs in its digital preservation programmes, regardless if they are built on in-house or distributed solutions. Extensive reviews under the auspices of earlier EU financed projects like Minerva (2008), Athena (2009) and Linked Heritage (2011) have already categorized and described many of the standards that are most applicable or recommended in the area of the DCH-RP project. However, moving into the field of distributed digital preservation services makes it increasingly important to understand and communicate the license

[29] See also Safety Deposit Box (http://www.digital-preservation.com/sdb) and Preservica Preservation as a Service (http://www.preservica.com)

agreements and terms of usage associated with digital resources, whether these are 'born digital' or are digitised representations of other cultural heritage artefacts. The Linked Heritage project investigated this topic and the result is summarised in Annex 3.

5.3 THE MAIN COMPONENTS OF THE ROADMAP

5.3.1 A vision

Distributed preservation solutions are becoming more and more common, but there is an apparent lack of basic concepts that the DCH community has agreed on for implementing distributed preservation solutions, like architectural design or best practice. There is obviously no commonly agreed vision of distributed digital preservation architecture relying on e-Infrastructures. Such a vision is an important piece in the puzzle and, therefore, urgently needed.

The overall vision for the DCH-RP roadmap is to implement a federated infrastructure, dedicated to support the application of open science in the arts and the humanities, which will make digital cultural heritage accessible and usable long term. This will be done by exploiting and integrating what already exists and to creating only those parts that are not yet available. The key to success is to use existing e-Infrastructures for research and academia (including NREN, NGI and the newer data infrastructures) as an efficient channel for the delivery of advanced services also to the digital cultural heritage. Connecting these facilities to the DCH sector will also contribute to developing the research capacities of this sector. This is simplified by the fact that DCH data and scientific data have overlapping layers of information and therefore can be expected to have rather low barriers for sharing common services.

5.3.2 A timeframe

The DCH-RP roadmap should make it possible for each cultural heritage institution to define its own practical action plan with a realistic timeframe for the implementation of its stages.

- Short-term (2014-2015)

  A short-term action plan (2014) is proposed by the DCH-RP project in order to initiate the development of a preservation services infrastructure on a level that will be self-sustainable and continue to progress on its own. This further progress is defined in terms of two further proposed time spans:

- Medium-term (2016-2017), i.e. two years after the end of DCH-RP), and
- Long-term (2018 and beyond) for the logical continuation of the DCH-RP work.

5.3.3 Appraisal and selection

In a digital environment, decisions taken at creation and selection of digital objects have significant implications for their preservation. The link between access and preservation is far more explicit than it is for paper and other carrier of information, as access to a digital object can be lost within a short period of time if actions are not taken from the beginning to guarantee that it is preserved. A useful approach is, therefore, to incorporate appraisal and retention functionalities into the design of information systems applications.

While many of the general principles for appraisal and selection in the traditional preservation environment can be applied in a digital environment, policies and procedures need to be adapted to the

new technical conditions. In a paper based environment, the decisions to select for preservation, and to actually preserve, can be taken separately and within a timeframe which may span over decades. But the brief lifetime of digital objects before becoming inaccessible, makes the decisions about selection and preservation to be taken simultaneously.

For networked digital resources, where access does not necessarily require bringing these resources physically into a collection or holding, appraisal and selection is quite different from traditional procedures. Whereas in a non-networked environment acquiring a resource normally means keeping it, is it in the networked digital environment possible to provide access to a resource without undertaking any preservation commitment either short- or long-term. Access is instead provided by making copies/mirrors for access in combination with hyper links, online catalogues, and other kind of finding aids. Appraisal and selection criteria are in this case based mainly on the number of requests and cost-benefit considerations.

Employing evaluation criteria and selection procedures for all potential digital acquisitions ensures that collections development is carefully prioritised and sustainable. Such review will normally be required for digital objects acquired before institutional policies and procedures were in place. One of the first steps that an institution undertakes in implementing a digital preservation policy may therefore be to quantify its current digital holdings and assessing preservation risks.

Over time the need may also arise to review collections and collections policy to reflect changing conditions and requirements in the surrounding world. The necessity of making early decisions on selection for preservation in a digital environment may result in needs of future reviews in the preservation lifecycle. However, for digital objects selected for long-term preservation such reviews have to be conducted under strictly controlled circumstances.

5.3.4 A sustainability plan for the Roadmap

The basic part of a sustainability plan for the Roadmap is to ensure a clear direction. This means first of all to give it an endpoint. But the Roadmap also needs to have a structure that makes it possible to maintain until this endpoint has been reached. In section 5.3.4 above a timeframe has been set for the DCH-RP Roadmap.

A second part is the sustainability of the distributed services for digital preservation to which the roadmap leads. Each cultural heritage institution has to summarise steps taken in the short-term stage of the Roadmap and transform them into a plan for how to implement the Roadmap. We call such a plan “Preservation as a service” (PraaS). Depending on circumstances like internal conditions, mandates of the institutions and considerations regarding the Roadmap, this plan (Praas) can differ quite much between the cultural heritage institutions.

The Praas also has to take into consideration the sustainability (political, financial, organisational, and technological) of the preservation infrastructure services that are chosen. Identification of target groups is one of the key factors to ensure that it will not be only a mayfly. These target groups could belong to different areas. They could be policymakers (EC, Member States, and Regions), owners of institutions, or e-Infrastructure providers, and the important part is that they on some level interact or at least share the same view on digital preservation. Important questions to consider are, therefore:

- Which scenes of action (existing, or to be created) are there for the identified target groups to interact on?
- How can these target groups be motivated to participate in the implementation and continues support of the Roadmap? Which are their interests?

- How are issues on digital preservation and access considered in each of the identified target group? How are these two issues interlinked in each group?

The answers will indicate

- on what level access is needed (e.g. politically and economically possible to ask for) by the cultural heritage sector which the e-Infrastructures have to serve;
- the amount of services in long term preservation that need to be offered by the e-infrastructures, and which level of sustainability need to be requested (e.g. politically and economically possible to ask for) by the cultural heritage sector.

Another key factor is cost aspects that per se are crucial for the sustainability of the Roadmap. 'Who will pay?' is a question that always will be raised – sooner or later. In chapter 6.2.2 below is the issue of business models highlighted.

Social and cultural factors are also expected to play a driving role in the implementation of the Roadmap but also in sustaining it. The implementation as well as the sustainability of the Roadmap is a complex process that involves many different actors that may change over time, each with different vocations, knowledge, technical skills, and economical capability.

A sustainability plan for the Roadmap has to consider which practical actions are needed to avoid those social and cultural factors to become obstacles for the implementation and sustainability of both the Roadmap itself but also the outcome of the Roadmap exercise.

In deliverable D2.2 *Report on community building, stakeholders' consultation and sustainability strategy*, a sustainability plan is described for the DCH-RP as consortium, to foster after the end of the project the launch in each partner countries of national plans for the implementing use of e-Infrastructure by the DCH community. It also proposes a strategy for keeping alive the recommendations, information, services and collaborative network established during the formal lifetime of the DCH-RP project.

5.3.5 Major areas to concentrate on

The roadmap exercise as such is aiming to produce an instrument that will facilitate policy makers as well as management within cultural heritage institutions. To achieve this, the roadmap is concentrated on four areas which identify the policy domains that require intervention:

Harmonisation of data storage and preservation: would allow integrating in common environments the curation of research data with other digital objects – two domains which are currently addressed separately;

Improved interoperability: includes better integration of preservation within the overall workflows for digitisation and online access; in a way this is a set of measures to avoid building 'digital silos' within the organisation, for example when digitisation is carried out without taking into account needs for preservation, and/or accessibility online is disjointed from preservation; an area of importance that need to be integrated is the selection of what to preserve (see section 4.3.2 above)

Establishment of conditions for cross-sector integration: a key condition for maximising the efficiency of successful solutions, transferring knowledge and know-how;

Governance models for infrastructure integration: a necessary condition for successful institutional participation in larger e-Infrastructure initiatives, and aggregation and re-use of digital resources.

These four areas were selected in order to help consolidating experience gained in individual institutions and to merge it into useful knowledge for the cultural heritage sector as a whole. For each area a set of prioritised actions are suggested (see chapter 6.2).

The short-term, medium-term and long-term perspectives will then articulate, within specific further steps, the evolution of these four areas.

5.4 CONDENSED VERSIONS OF THE ROADMAP

In this section are presented condensed versions of the roadmap - in short-term, medium-term, and long-term perspectives

5.4.1 Short-term (2014 – 2015)

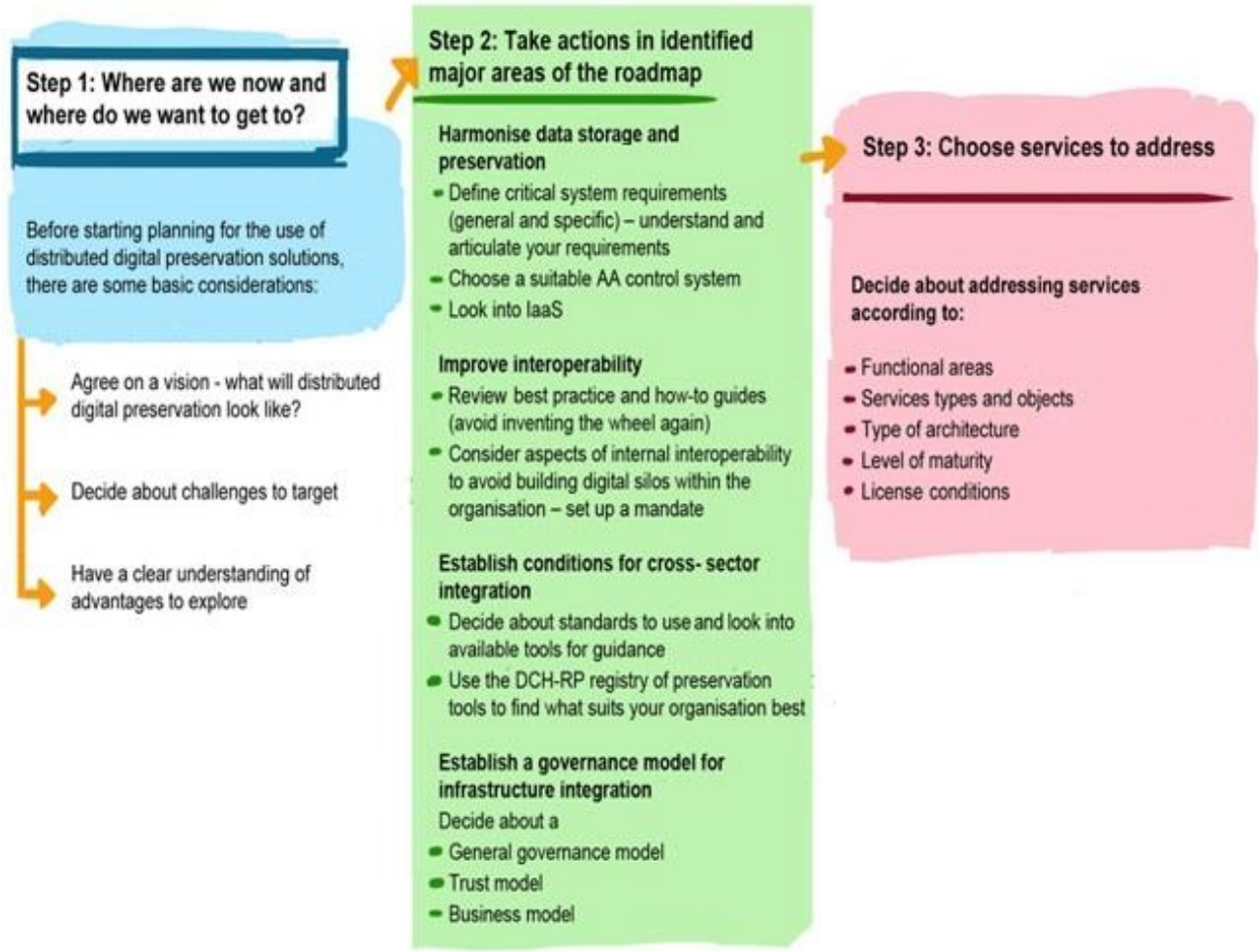


Figure 10: The Condensed version of the intermediate roadmap – short-term

5.4.2 Medium-term (20116 – 2017)

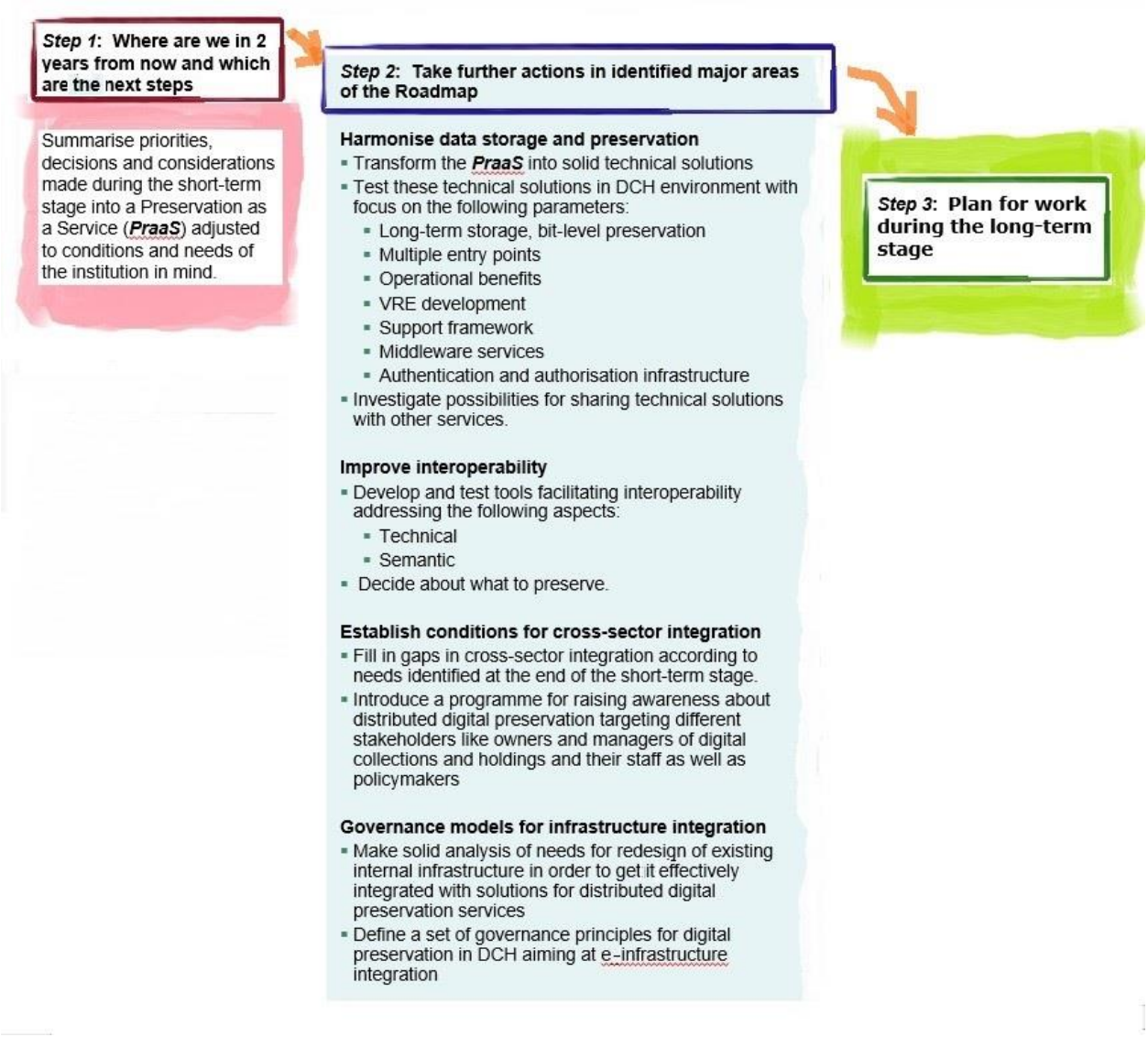


Figure 11: The Condensed version of the roadmap – medium-term

5.4.3 Long-term (2018 and beyond)

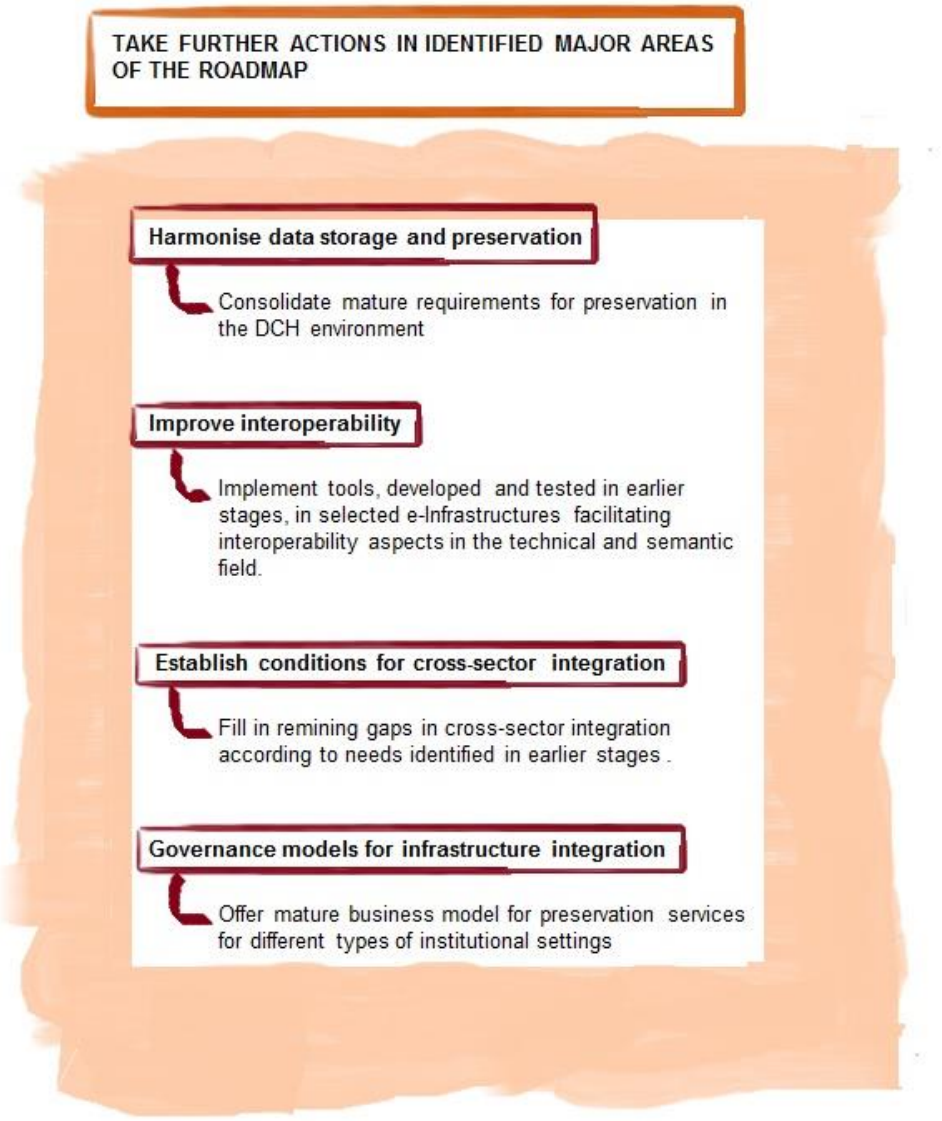


Figure 12: The Condensed version of the roadmap – long-term

6. AN ACTION PLAN

6.1 ESTABLISH A VALUE CHAIN

The DCH-RP project has looked into other domains, to see if there are experiences concerning value in distributed digital preservation that are transferrable to the DCH domain. Obviously, very little has been done so far, but in the e-journal preservation community much has been achieved in terms of evolving mechanisms and organisations to look after digital preservation. The technical, organisational and financial challenges have been proved to be solvable, given strong commitment from the communities involved. The key issue appears to be the ways in which these communities have organised themselves to bring about long-term agreements and infrastructures to make preservation happened.

Cost will clearly be a key variable when deciding whether or not to contract out digital preservation to an external service provider. But there are also other factors to consider, and the advantages and disadvantages of each of them need to be balanced against the overall mission of the institution. For example, legal provisions due to privacy or confidentiality may influence whether outsourcing is appropriate or not. The extent to which the potential advantages of using distributed preservation services can be maximised and the potential disadvantages minimised is also dependent on the possibilities for dedicating staff resources to preservation activities. The costs for these staff resources need to be added to the overall contract costs when calculating the cost benefit of using distributed services for digital preservation. However, one have to be aware of that most of these costs will be or should be received even when digital preservation is not outsourced.

Digital Preservation Coalition has listed a number of issues and potential advantages and disadvantages of using distributed services in digital preservation activities.[30]

[30] See Preservation Management of Digital Materials: The Handbook, p. http://www.dpconline.org/advice/preservationhandbook /

Issue	Potential advantage of using 3rd party services	Potential disadvantage of using 3rd party services
Limited practical experience in preserving complex digital objects over time	Avoids the need to develop costly infrastructure (particularly important for small institutions) Allows the institution to focus on other aspects of service provision. Provides specialist skills and experience which may not be available within the institution If there are economies of scale, outsourcing may well be cost effective. Allows action to be taken in the short to medium term, pending development of infrastructure.	Without some practical experience and expertise, it will be difficult to develop and monitor effective contracts. Without practical experience it will also be difficult effectively to communicate the requirements of the organisation (or to assess whether they are technically feasible or not) Danger of either not developing or losing skills base. There is no established bench marking. It is still too new an area. Risk of business failure Until the market increases there may be an overdependence on one contractor Unless there are adequate exit strategies, may be locked into an outsourcing contract longer than intended.
Access considerations	Monitoring usage may be more efficient (assuming the contractor has a demonstrated ability to deliver meaningful usage statistics). There may be synergies and cost savings in outsourcing access and preservation together.	Difficult to control response times which may be unacceptably low and/or more costly, especially for high-use items.
Rights Management	Avoids what is often a resourceintensive activity for the institution.	May significantly increase the cost of the contract and/or complicate negotiations with rights holders

Security	Contract can guarantee security arrangements required by the institution	Lack of control, especially for sensitive material
Quality control	A watertight contract will build in stringent quality control requirements	Risk of loss or distortion may still be unacceptably high for highly significant and/or sensitive material

Major advantages, specific for the DCH sector when using distributed services offered by e-Infrastructures, could for example be the following:

- Long-term preservation (i.e., bit-level preservation) and access to digital objects of different kind, also so called "live" content (e.g., streaming audio and video collections);
- Multiple entry-points that suit a variety of user interfaces (e.g. APIs, protocols). New cloud based search engines are under development, based on multilevel nodes that can combine different data sources (documents, images, books etc) from multiple content providers;
- The DCH-community can focus on its own areas of specialisation by deploying new services for monitoring and management tools that ensure smooth and secure running of distributed operations;
- Forming a community of practice or a Virtual Research Community that transcends discipline and national boundaries while achieving economies of scale by bringing together international communities;
- Benefitting from integration within the research and educational e-Infrastructures support framework;
- Central hosting and monitoring of middleware services;
- Simple authentication and authorisation infrastructures for large (and potentially unbounded) user groups;
- Connections to shared services in other countries and sectors. (e.g. research data centres, commercial businesses, etc.).

To summarise: it is important for cultural heritage institutions to have a clear understanding of what to exploit, before taking a decision about the use of distributed digital preservation services.

Research and development on the use of digital preservation services built on distributed facilities instead of ones performed in-house has just started.[31] Some identified drivers that probably will underpin an enhanced the use of distributed digital preservation services are:

- increased flexibility in digital preservation architectures based on granular or layered structures (e.g. SaaS, PaaS, IaaS) that are easy to adapt to a variety of preservation scenarios;
- clearly defined sets of metrics or benchmarks for comparing preservation tools and services and their performance;
- terminology and standards that no longer converge along professional community borderlines but instead are agreed cross-sectorial.

6.2 ACTIONS TO TAKE

6.2.1 Harmonise data storage and preservation

SHORT TERM PRIORITIES

Today, an ever-broadening range of preservation software tools is available, and institutions can combine and tailor digital preservation components according to their specific needs and context. The typical digital preservation workflow incorporates generic tools, e.g. virus checking, metadata generators or format identifiers, specific preservation services, as well as services that relate to storage management in distributed preservation environments. The aim here is to establish the necessary conditions for various services to coexist and to be orchestrated into a suitable digital preservation "eco-system", regardless of whether the services are targeted on research data or other digital objects.

Tests of existing technical solutions in a DCH environment have being carried out by the DCH-RP project. The results achieved are reported in Annex 4 *Best practice*.

In Annex 2 are listed prioritised short term actions to take in this area of the Roadmap.

MEDIUM TERM PRIORITIES

The Praas has to be transformed into solid technical solutions aimed at the DCH environment. These solutions must, then, be tested more specifically addressing aspects like:

- Long-term storage (bit-level preservation)
- Multiple entry points
- Operational benefits
- VRE development
- Support framework
- Middleware services
- Authentication and authorisation infrastructure

Possibilities for sharing technical solutions with other services have also to be investigated.

LONG TERM PRIORITIES

[31] See as an example the InterPARES Trust (ITrust 2013-2018), a multi-national, interdisciplinary research project exploring issues concerning digital records and data entrusted to the Internet (http://interparestrust.ord)

The main priority in this stage is to consolidate mature requirements for distributed digital preservation in the DCH environment.

6.2 2 Improve interoperability

SHORT-TERM PRIORITIES

In Annex 2 are listed prioritised short term actions to take in this area of the Roadmap.

MEDIUM TERM PRIORITIES

Improved interoperability is an area of action that focuses mainly on DCH institutions internal conditions (see above under Short- term priorities). It is important during this stage to develop and test tools that facilitate interoperability addressing both technical and semantic aspects.

During this stage the question of what to preserve has to be raised. An important issue, called for by several stakeholders, is that analogue data carrier converted into digital ones by digitisation and "born-digital" objects have to be treated differently; mostly because digitised objects may be needed during a shorter period and sometimes are considered just as digital copies of physical objects.

One way of starting the process of appraisal is to try to get an overview of the basic conditions for preservation of different media types in digital form. A first attempt is made in figure 13 below.

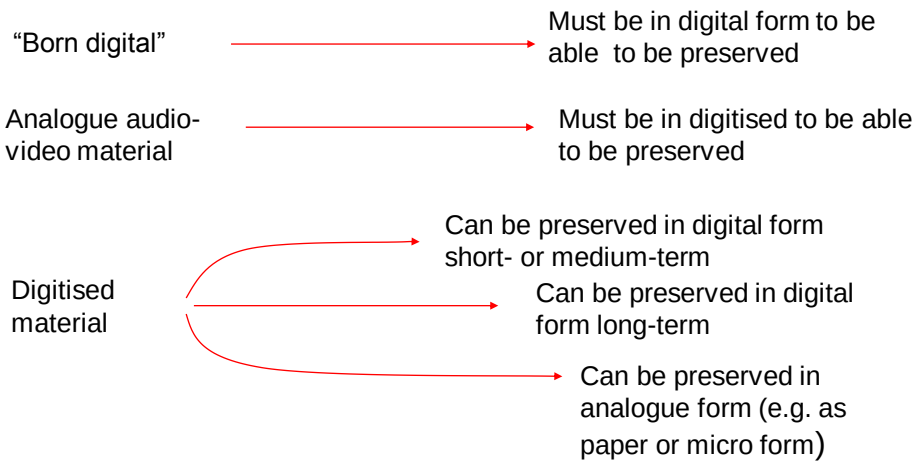


Figure 13: Basic conditions for preservation of different media types in digital form

Applied on the actual contents in the collections and holdings kept by a cultural heritage institution, the results can be used in a matrix where the other axis shows for example the key factors noted in the sustainability plan (interest and intentions of different target groups, cost aspects and social and cultural factors, see chapter 5.3.4 above).

LONG TERM PRIORITIES

By now, some e-Infrastructures should have been identified as designated for distributed digital preservation The main priority in this stage is, therefore, to implement selected e-Infrastructures tools that have been developed and tested in earlier stages to facilitate interoperability aspects in the technical as well as in the semantic field.

6.2.3 Establish conditions for cross-sector integration

SHORT TERM PRIORITIES

Analyse what impact emerging and established standards have on grid and cloud preservation architectures

The DCH-RP projects deliverable D3.2 Standards and interoperability best practice report is about existing projects and initiatives as well as standards, guides, and tools, which are useful for the DCH and e-Infrastructures communities when approaching the digital preservation issues. This deliverable is public and can be downloaded from the projects homepage www.dch-rp.eu.

One of the challenges for the DCH community is to choose among the vast number of standards that are already available. This may be problematic, especially for small DCH institutions with limited knowledge in and/or resources in this field. There are also non-technical issues that have to be resolved. One is differences in the legal system between countries, especially when data is covered by copyright or classified.

The conclusion is that much work has already been done, but more efforts are still needed before these standards (including guides and tools etc.) can give substantial help to the DCH community. For example, many of them need to be more user-friendly in order to be understandable for non-technical personnel. Furthermore, practical tests made within the DCH-RP project have shown that already developed e-Infrastructure services must be modified and/or improved in order to provide a “pan-European” solution for the DCH community.

In this deliverable we are not bringing forward arguments for adopting or recommending specific standards, but information about standards are reported in Annex 3.

Registry of preservation tools and services

The development of the DCH-RP preservation services registry is a key step in the construction of the Roadmap. In this regard, it should be noted that the collection and summarisation of information on services is quite an onerous task, because over the last decade the number of tools and services produced within the community has been quite impressive; however, more work needs to be done on the characterisation of services in order to make them usable in a distributed e-Infrastructure and currently there are no testing tools which would help to run systematic evaluation on the behaviour of tools – either singly or in combination

There are a few hundred software tools on offer to support automation of preservation tasks, yet their support status, interoperability status, level of documentation, quality, and reliability are poorly documented. There continues to be inadequate support for decision-making, selecting, testing and benchmarking tools for preservation. While a number of digital preservation tools registries/collections are already in place, there is no such collection addressing grid and cloud services. The DCH-RP projects deliverable D3.3 Registry of services fills this gap by presenting a registry of the services available to support preservation activities, with particular regard to the services that can better fit the requirements of the DCH sector. This deliverable is public and available on the projects homepage www.dch-rp.eu. Through integration of the work of the EU project APARSEN in the DCH-RP Registry of Services and

discussions with other initiative (e.g. COPTR recommended by the Digital Preservation Coalition[32]) sustainability of the registry will be accomplished.

Analyse which PaaS composition of services best matches' digital preservation requirements

To find a suitable mixture of distributed services that matches the individual institutions requirements on distributed digital preservation can be a challenge. The DCH-RP project has conducted a number of practical experiments in the framework its Proofs of Concept that are meant to be used as best practice. The results from these experiments are together with some general recommendations reported in annex 4.

Identify gaps in provision and establish a plan for medium- and long-term developments to address the gaps

A plan for medium- and long-term work to address identified gaps needs to be made in the end of the short-term stage.

MEDIUM TERM PRIORITIES

The main challenge during this stage will be to fill in gaps in cross-sector integration according to a plan made in the end of the short-term stage.

To make the introduction of the Roadmap successful, raising awareness in one of the main challenges identified in chapter 3.3 above. What is needed is introduction programme for raising awareness about distributed digital preservation. Targeted stakeholder should be policymakers, owners and managers of digital collections and holdings and their staff.

As a premise to the actual implementation of the Roadmap by cultural heritage institutions, it will be necessary also to plan for an initial training/awareness phase, where the cultural managers and their staff need to familiarise with the concept of a Roadmap. The management staff should improve its awareness about the need to conceive their institution and its data as a living body, which should move along a 'road', to arrive to new destinations. The terminology used in the Roadmap presents some elements of discontinuity with the traditional terminology, which also need to be explained in order to get the most of the benefits from the adoption of the Roadmap.

LONG TERM PRIORITIES

The main challenge during this stage will be to fill in gaps in cross-sector integration according to a plan made already by the end of the short-term stage.

Establish a governance model for infrastructure integration

SHORT TERM PRIORITIES

Analyse major information governance patterns and windows of opportunities

The model for governance to use must be tailored to the concept of distributed digital preservation. The following framework can be seen as a role model for how to achieve good governance:

[32] About the recommendation of COPTR, see http://www.dpconline.org/advice/tools-coptr

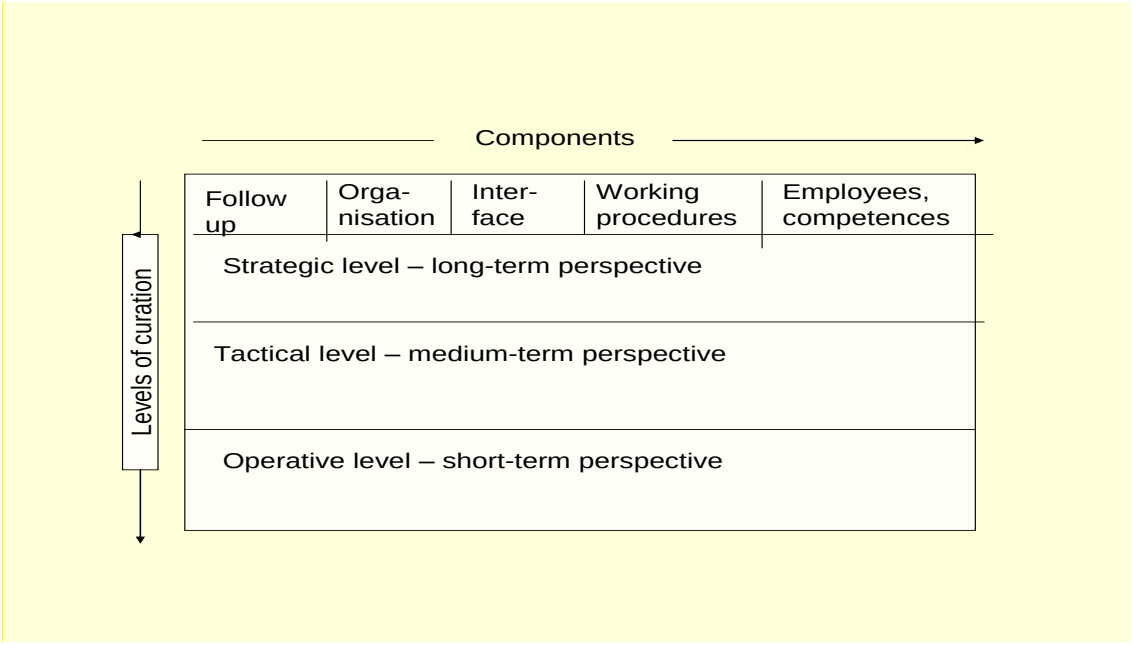


Figure 14: A framework for the governance of distributed digital preservation services

This framework consists of five components that highlight different dimensions of governance focusing on three different levels (strategic, tactical and operative). The components are:

- Follow up (including how to manage distributed digital preservation services)
- Organisation (including definitions of roles and responsibilities)
- Interface (including forum for clients and service providers to meet)
- Working procedures
- Employees and competences.

The levels of governance each have different focus and perspectives:

- Strategic level: aiming at securing the long-term perspective; this is done from both an internal and an external perspective through, firstly, follow up and managing a consolidated service provider portfolio, and, secondly, establishing a forward-looking relation between the client and the service-provider;
- Tactical level: has a time middle-term perspective with focus on securing services and agreements at hand and that they are up to date;
- Operative level: focus is here on securing the follow up of the daily work and that problems and incidents that arise are handled in proper way.

Depending on which type of service is involved (see section 5.3.3), the service providers can be classified as being strategic/non-strategic and providing services that are easily accessible/not easily accessible. For the cultural heritage institutions the results of such a classification will inform their approach to managing the situation.

Explore the issue of trust-building

There is no trust model of a distributed repository system in place today in the DCH domain. The only similar example in existence is the "circular chain trust model" of the LOCKSS system where all partners using the software also share a trust network. The CESSDA is working on one.[33]

Trust work is also going on in the APARSEN project, but this is about the repository level of trust and is predominantly occupied with auditing of digital repositories. The underlying concept there is that trustworthiness of a repository can be established through an audit. This is derived from the 2002 RLG/OCLC report *Trusted digital repositories: attributes and responsibilities*. In the APARSEN sense there are three levels of trust that can be established through audits:

1) Self-assessment, using the Data Seal of Approval (a toolkit developed by DANS for research data archives) or DRAMBORA;
2) Self-audit using ISO 16363 or DIN 31466 (both are originally based on the TRAC checklist that was developed by RLG and NARA);
3) Formal audit using ISO 16363 or DIN 31466 using external auditors that leads to certification.

In parallel with this initiative there is the Center for Research Libraries (CRL) in the US still conducts TRAC audits and issues certificates to repositories and their cooperatives[34].

Neither of these approaches are not directly relevant to DCH-RP purposes, because NRENs are likely not interested in undergoing a full digital repository audit. NRENs are for understandable reasons not that keen to become full-scale digital preservation repositories for DCH alone, because this is not really their sole core business. What is needed is a more flexible method of auditing of a distributed digital preservation service where a repository is outsourcing some of its services to an NREN. And this does not readily exist yet. In chapter 5.2.2 this approach to distributed preservation services is called the „kiosk" model.

DCH preservation has sometimes a tendency to be project-based. Therefore, there is also an urgent need for national and international programmes that assure long-term sustainability of e-Infrastructures.

There is one very new development that is more relevant for the DCH-RP project. This is called the Distributed Digital Preservation reference model (DDP) that is trying to enhance the original OAIS model that suits best a single repository.[35] As part of the DDP model there are plans to develop a distributed trust model, but this work has not proceeded very far yet.

In deliverable D4.1 *Trust building report* the DCH-RP project has outlined the design of a new trust model suitable for the use of e-Infrastructures, including recommendations for user authentication and access control system(s). It is important to strengthen the capability of cultural heritage institution to articulate their trust requirements.

This new trust model is attached in Annex 1.

Establish a possible business model

A business model describes the rationale of how an organisation creates, delivers, and captures economic, social, cultural, or other forms of value. In both theory and practice, the term business model is used for a broad range of informal and formal descriptions to represent core aspects of a business, including purpose, target customers, offerings, strategies, infrastructure, organisational structures, trading

[33] See DCH-RP deliverable D3.1
[34] See http://www.crl.edu/archiving-preservation/digital-archives/certification-and-assessment-digital-repositories
[35] See a guide: http://www.metaarchive.org/GDDP

practices, and operational processes and policies. There is also a clear connection between the business model used and trust-building.

It is obvious that a business model based on passive preservation is not an option. While there is understandable concern that the costs of preserving digital materials will be high, it is equally important to consider the costs and implications of not preserving them. The costs of recreating a digital resource may be much higher than those for preserving it; further, the opportunity to do so may no longer exist when the digital resources concerned is needed. An increasing dependence on both digitally produced and accessed information means that there is a rapidly growing body of digital material for which there are legal, ethical, economic and/or cultural imperatives to retain the material, at least for a defined period of time and, in some cases, forever. If active steps are not taken to protect these digital materials, they will inevitably become inaccessible and unusable within a relatively brief timeframe. Innovative funding models must, therefore, be investigated, for example:

- models where the public and private sectors enters into new partnerships (e.g. the re-use of digital cultural content by creative industries, non-IPR based models for the exploitation of digital cultural resources in applications for educational and research, commodification of cultural heritage and cultural tourism exploitation, etc.);
- models where the cultural heritage integrates with new e-government processes (e.g. a new role for the archives in the digital world).

Digital preservation built on a distributed model needs also a business model suitable for the integration between the cultural heritage community and the e-Infrastructures. ITC managements have today started to implement new concepts for outsourcing, whether cloud-based or not. One of them is Vested Outsourcing. This is a hybrid business model, based on research conducted by the University of Tennessee Center for Executive Education and funded by the U.S. Air Force, In this model both clients and service providers in an outsourcing or business relationship focus on shared values and goals to create an arrangement that is mutually beneficial to each, in contrast to traditional outsourcing and businesses relationships that, according to Vested Outsourcing, focus on win-lose arrangements.[36]

The basic philosophy in the Vested model is “What´s in it for We”, and it consists of five rules that have to be implemented in a relation-based contract, in this case for distributed digital preservation:

Focus on results and not on transactions: conform to a business model that will give both parties unanimous interest with focus both on valuable results and on a joint vision for the partnership.

Focus on what to do instead of how to do it: this approach means to concentrate on what to achieve instead of how it shall be done. Traditional outsourcing contracts often have detailed texts on how a service provider shall provide a service. This, sometimes called the “outsourcing-paradox”, can ends up in a situation where the client outsource a service to an expert organisation, but at the same time describe in detail how this expert organisation shall provide its expertise. The Vested model instead points out the need for both a definition of functions and a roadmap with strategic goals for how the service provider shall support the client in achieving his or hers objectives.

Agree on clearly defined and measurable goals and deliverables: traditional contracts on outsourcing often contain agreements about measuring different levels of services and how to compensate the client if the agreed levels are not reached. However, this is not the same as the client being satisfied with the results. In a result based business model, focusing on what to do, the goals and achievements must be clearly defined from the beginning.

[36] http://en.wikipedia.org/wiki/Vested_outsourcing

Establish a pricing model with optimal incentives for the agreed partnership: the traditional price list is not used in the Vested model. Instead, the service provider shall be economically compensated depending on how the strategic goals are achieved. But the conditions for every pricing model are constantly changing, and both partners must, therefore, have a high degree of transparency regarding their actual costs and economical situation. Otherwise fruitful negotiations about changes of prices will not be possible.

Establish a governance model that gives both parties both overview and insight: the important part in good governance is - according to the Vested model - to focus on the partnership as such and not on the partners. The partners work with a stratified structure, usually found in governance models (see above), but instead of just one interface for communication, with one responsible person per partner, several interfaces are used, one for each specific field in the contract.

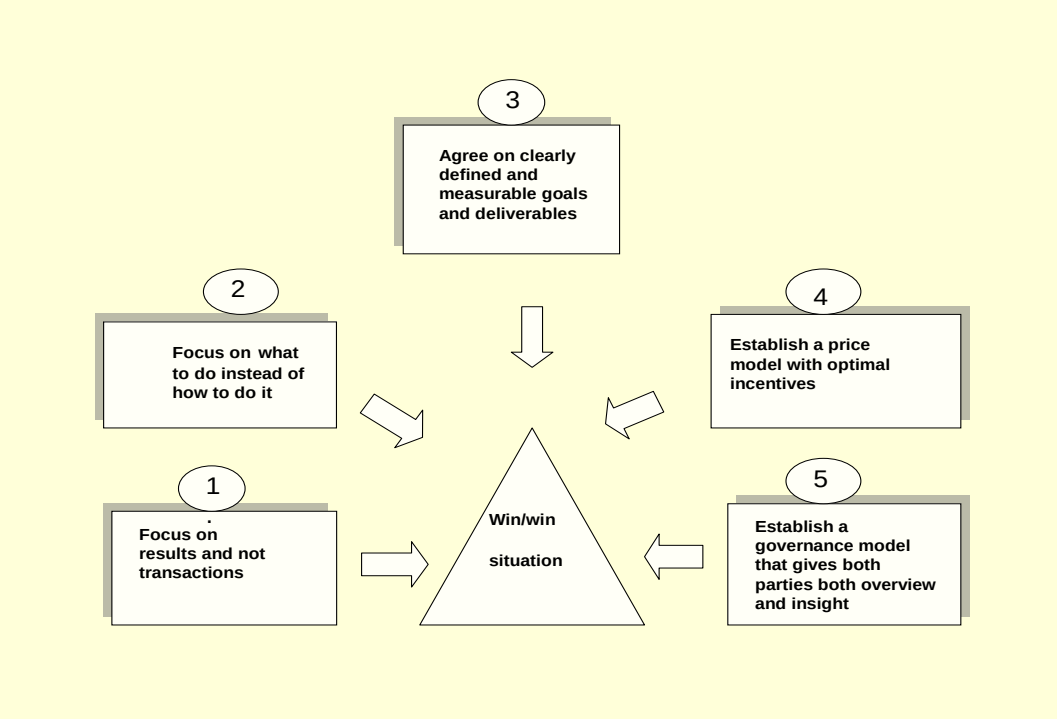


Figure 16: The Vested model

MEDIUM TERM PRIORITIES

In this stage the highest priority actions to take are:

- To make solid analyses of needs for redesign of the cultural heritage institutions existing internal infrastructure, in order to get it effectively integrated with distributed digital preservation services;
- To define a set of governance principles for digital preservation in DCH aiming at e-Infrastructure integration.

LONG TERM PRIORITIES

Most important in this stage, is the possibility to offer mature business models for distributed digital preservation services for different types of institutional settings (context and environment).

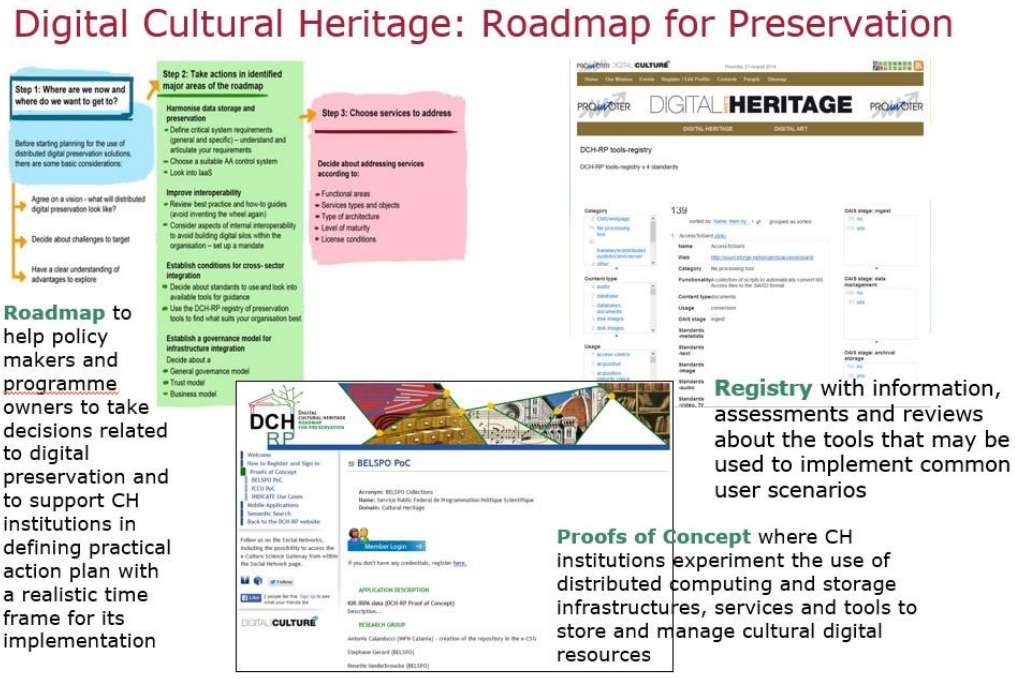
7. A WEB-SPACE FOR THE ROADMAP

The Roadmap for the implementation of a preservation e-infrastructure for Digital Cultural Content represents the main outcome of the DCH-RP project.

By definition, a Roadmap is not useful if it is not widely disseminated, validated and endorsed by the user groups that it aims to target. DCH-RP contributed substantially to the creation of a wide community of people coming from different sectors (policy makers, cultural institutions, e-infrastructure providers, etc.) who demonstrated interest in the work done for the development of the Roadmap. Now it is important to keep alive and continue to nurture this community, creating awareness about the final version of the Roadmap and fostering its diffusion and implementation in Europe and worldwide.

Furthermore, a Roadmap cannot be considered as a final step. It has on the contrary to be considered as a living document that needs to be continuously maintained, updated and improved as time passes, technology changes, new requirements have to be taken into account, and so on.

It is for these reasons that we decided to create a dedicated web-space where it is possible to download the last version of the Roadmap, but also where it is possible for everyone to provide feedback and comments, a kind of Forum dedicated to the use of e-infrastructure services and facilities for the long-term preservation of digital cultural content.



Apart from presenting and discussing the Roadmap, this web-space will link also to other relevant material, information and services that are linked to the Roadmap itself and that contribute to supplement it.

In particular, a section will be dedicated to the Registry of Services and Tools that was developed in DCH-RP as a practical instrument to help decision makers, DCH communities, institutions and projects to plan the implementation of their digital preservation processes.

Another section will be dedicated to present the results a range of proof of concepts conducted during the project, where cultural institutions and e-infrastructure providers worked together on concrete

experiments targeted at demonstrating how e-Infrastructures can be of benefit for the DCH community, in particular for the preservation of digital cultural content.

By the end of the project, the web-space is hosted as a section of the DCH-RP showcase in Digital meets Culture (http://www.digitalmeetsculture.net/heritage-showcases/dch-rp/dch-rp-roadmap-for-preservation). The partners are committed to continue the work on the Roadmap even after the end of the project period and in this framework they are discussing about creating a URL dedicated to the Roadmap to be maintained on a longer period.

8. CONCLUSIONS

In Greek mythology, the god Proteus is the keeper of knowledge of the past, present and future. Anyone who wants this knowledge must catch Proteus, who will change into many forms to escape. Once someone is persistent enough to hold Proteus through all his changes, he or she will reveal his knowledge. Therefore, when Proteus is mentioned, it normally refers to anyone or anything that is flexible, able to change and adapt or having many forms. Preserving digital objects is in many ways like trying to catch Proteus.

The cultural heritage sector is faced with a number of challenges in making current and future digital information accessible and usable over time. In short: solutions for preservation must have a high level of automation and self-reliance to be able handle the rapidly growing amount of DCH information; the tremendous rapidness in the development of new technology requires preservation solutions adaptable and flexible enough to really solve permanence and longevity issues; the infrastructure and organisational models must be highly scalable and adaptable to the various levels of input, storage and access.

What is needed is in other words a readiness for handling perpetual change, and using existing e-Infrastructures for research and academia for distributed digital preservation can be one way of meeting these demands. Keywords are distinct functional and technical requirements, solid models for handling business issues, governance and trust, and a service architecture that altogether can guarantee the authenticity of the digital resources over time, physically and technically preserve them over time, and verify that they are accessible and usable over time. Despite the fact that NRENs and NGIs are national entities, sometimes with different policies and procedures for access and usage, it is in most cases possible to establish common policies, processes and protocols to allow digital DCH organisations to access e-Infrastructures. Obstacles to overcome are often to find in areas like politics and economy or based on social and cultural factors.

A ground breaking part of the concept is the possibilities to customise the services provided by e-Infrastructure, i.e. tailoring the service portfolio and characteristics to the actual preservation tasks and requirements. However, even if the e-Infrastructure resources seems to be allocated in ways that could support preservation functions and sub-functions quite well, the general conclusion must be that the market for distributed digital preservation services is still in its infancy, even if this market is developing quiet rapidly with a focus on the reach domain..

Another important issue is the level of maturity in the DCH sector to handle distributed digital preservation solutions. E-Infrastructures can reach their maximum potential in serving the DCH preservation practice only if the DCH sector is prepared to exploit the opportunities of the e-Infrastructure. This is obviously not the case today. Both e-Infrastructure and DCH institutions express feelings of dissatisfaction, the latter also reporting about difficulties in utilising the offered facilities and tools. The DCH-RP projects aim is have a practical approach with a strong focus on what to do, and this approach has become even more important than expected in the beginning of the project.

Future developments will also underpin an enhanced use of distributed digital preservation services, like

- increased flexibility in digital preservation architectures based on granular or layered structures (e.g. SaaS, PaaS, IaaS) that are easy to adapt to a variety of preservation scenarios
- clearly defined sets of metrics or benchmarks for comparing preservation tools and services and their performance
- terminology and standards that no longer converge along professional community borderlines but instead are agreed cross-sectorial.

9. ANNEX 1 A TRUST MODEL SUITABLE FOR THE USE OF E-INFRASTRUCTURES

1. TRUST AND DIGITAL PRESERVATION – STATE OF THE ART

Digital curation is a complex field that requires competence in preservation, technology, metadata, risk management and so on, as well as availability of infrastructure and tools for carrying out both passive and active digital preservation. Not all digital repositories can be expected to deliver all digital preservation services to the same level of quality. Hence, questions over why should one trust a given repository to preserve digital content appear justified. When the repository involves third party service providers to preserve its clients' data, the trustworthiness and reliability of the third parties also becomes a demonstrable need.

1.1 The concept of a trusted Digital Archive

Claims of trustworthiness of digital archives are easy to make but are difficult to justify or objectively prove. A trusted digital repository is one whose mission is to provide reliable, long - term access to managed digital resources to its designated community, now and in the future.[37] Trust in a digital repository is related not only to trusting the preservation methods applied by the repository, but also to broad organisational issues like funding base, policy framework, staff training, existence of transferable skills, and so on. A trustworthy digital repository will understand threats to and risks within its systems and organisation.

In 1996, the Commission on Preservation and Access (CPA) and the Research Libraries Group (RLG) joint Task Force on Archiving of Digital Information called the existence of a sufficient number of trusted organizations capable of storing, migrating, and providing access to digital collections "a critical component of the digital archiving infrastructure". The Task Force report proposed that a "process for certification of digital archives is needed to create an overall climate of trust about the prospects of preserving digital information".[38]

An understanding of what are digital archive components and how is the preservation function embedded into the overall archive workflow is presented in the OAIS reference model. OAIS is a pivotal standard in the digital preservation domain, ISO 14721 *Space data and information transfer systems – Open archival information system – Reference model*. It is a functional framework that presents main components and basic data flows within a digital preservation system. As a reference model, the OAIS standard does not imply a specific design or formal method of implementation. Instead, it is left to the users to develop their own implementation by analysing existing business processes and matching them to OAIS functions.

Among the first to explore the characteristics of a trusted digital repository was the RLG and Online Computer Library Centre (OCLC) Working Group on Digital Archive Attributes. It released its report *Trusted Digital Repositories: Attributes and Responsibilities* in 2002.[39] RLG and OCLC sought to define the characteristics of "sustainable digital archives that could serve large-scale, heterogeneous digital collections held by national libraries, university libraries, special collections, archives, and museums". One of the qualities of the trusted digital repository (TDR) was set as: "compliance with the Reference Model for an Open Archival Information System (OAIS)". The OAIS Reference Model supplies a common

[37] RLG/OCLC (2002). *Trusted Digital Repositories: Attributes and Responsibilities, p. 5* RLG. http://www.oclc.org/research/activities/past/rlg/trustedrep/repositories.pdf.
[38] CPA/RLG (1996). *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information.* RLG
[39] See foot note n:o 37

framework, including terminology and concepts, for describing architectures and operations of digital archives.

Through this conjecture the concept of a 'trustworthy digital repository' became linked with a standard workflow model that a digital archive has to follow. Although the OAIS reference model does not prescribe any specific technologies or technology architectures, the 'trusted digital repository' also came to be understood as a centralised, single organisation-based preservation service model where the institution that provides the preservation service is also the owner of the digital repository system that houses digital objects. The practice of applying the TDR criteria over the next decade has demonstrated that the word 'trusted' in this concept should more appropriately have been 'quality' because essentially the TDR is about ensuring quality at the operational level of repository work. Being trusted to deliver quality service requires a step further – making the compliance with quality criteria transparent and verifiable by external stakeholders. Thanks to a strong striving towards increased reputation among repositories, the digital preservation community has accepted the term 'trusted' as a replacement for 'quality' and has gone on to develop audit methods that instead of establishing compliance with quality standards are claiming to establish the trustworthiness of a repository.

1.2 The trusted digital repository audit methods

To begin answering questions on trustworthiness of digital preservation repositories a number of approaches have been proposed that rely on different methods of audit (see deliverable D4.1 Trust building Report).

In 2012 an ISO standard in support of the OAIS reference mode – ISO 16363:2012 *Audit and certification of trustworthy digital repositories* was accepted. The scope of the checklist is explicitly the entire range of digital repositories; its criteria are empirically derived and consistent measures of effectiveness have been ascertained. A team of experts as also conducted a series of pilot audits as part of the APARSEN project to test the methodology of the ISO 16363 standard. [40] (APARSEN 2012).

The same working group is working on an adjunct standard *Requirements for Bodies Providing Certification of Candidate Trustworthy Digital Repositories* (ISO/DIS 16919). Once completed, this standard will provide normative rules against which an organization providing audit and certification of digital repositories may be judged, and it describes the auditing process.

The certification process based on these standards is guided by a Memorandum of Understanding that was signed to define a European Framework for Audit and Certification of Digital Repositories.[41] It names three certification levels:

- Basic Certification (based on DSA)
- Extended Certification (self-assessment based on DSA plus self-audit based on ISO 16363 or DIN 31644)
- Formal Certification (self-assessment based on DSA plus full external audit of ISO 16363 or DIN 31644).

In summary, despite that the criteria and checklists for assessing trustworthiness of digital repositories having been around for over a decade, the practice of applying them has been limited to self-assessment and only a handful of repositories have been formally audited and certified. Self-assessment can improve

[40] D33.1A *Report on Peer Review of Digital Repositories*. APARSEN project http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/04/APARSEN-REP-D33_1B-01-1_1.pdf

[41] http://www.trusteddigitalrepository.eu/Site/Trusted%20Digital%20Repository.html

the quality of repository work, but if the results of this assessment are not made public, its positive impact on trust towards the repository is indirect and only emerges over a long period of time. If the results of self-assessment are made public, this may increase trust towards (and eventually reputation of) the repository among some stakeholders. However, there is no objective benchmark available for conferment of 'trust' or for measuring how much the trust will increase, because the assessment criteria deal instead with quality of operations within the digital repository. The quality and trustworthiness of digital repository services to external stakeholders is not explicitly part of the existing five assessment methods because the assessment does not involve external parties, their expectations or satisfaction with the services delivered. The trust models in digital archives require, therefore, further analysis and expansion to include architectures that involve third party service providers. The tools to evaluate and establish trustworthiness of a digital repository that have been developed do not cater easily to a situation where some services are contracted out to third parties or shared between institutions. The current thinking is that the third party service provider would have to meet the exact same requirements as the digital repository does in an audit. However, most e-Infrastructure service providers have no ambition to become certified as trusted digital repositories or even to act as repositories solely for the DCH sector. Instead their core business is to provide services to many customer segments. Models and assessment criteria for trusting distributed digital preservation services are, as yet, not there.

1.3 Trust in distributed preservation services

SERVICE MODELS

The DCH-RP deliverable D3.1 *Study on a Roadmap for Preservation* analysed digital preservation service models and concluded that although the basic archiving workflow is provided by the OAIS reference model, it does not articulate clearly how it can cater for distributed archiving architectures. Cloud, grid and e-Infrastructure service architectures vary significantly and do not allow for a uniform mapping of preservation services to a single architectural model. Conceptualising and modelling the joint service architecture is only in developing phases.

At present there are not formal (reference) models that describe distributed digital preservation services because the practice of using distributed service architectures is only emerging. An early description of the distributed digital preservation (DDP) model was described in the Educopia Institute and MetaArchive report *A Guide to Distributed Digital Preservation*.[42] The report describes the principles and advantages of a federated repository architecture based on the MetaArchive Cooperative experience with a Private LOCKSS Network (PLN). The report sets requirements for the number of copies of each archived object and their storage conditions that focus on best practice disaster preparedness:

- Content should be replicated at least three times;
- Sites preserving the same content should not be within a 75-125-mile radius of one another;
- Preservation sites should be distributed beyond the typical pathways of natural disasters, such as hurricanes, typhoons, and tornadoes;
- Preservation sites should be distributed across different power grids and under the control of different systems administrators;
- Content preserved in disparate sites should be on live media and should be checked on a regular basis for bit-rot and other issues.

[42] Skinner, Katherine, Schultz, Matt (2010). *A Guide to Distributed Digital Preservation*. Educopia Institute http://metaarchive.org/sites/metaarchive.org/files/GDDP_Educopia.pdf

The following generalised service models summarize the existing permutations of offering digital preservation services as federated digital archives or as distributed services relying on cloud or grid providers:

1. A *cooperative* file sharing model where each participating archive is a node in a network that hosts some other node's data. The best-known example of such a network is LOCKSS (Lots of Copies Keeps Stuff Safe).16. In the case of LOCKSS networks, digital preservation is limited to bit-level preservation.

The LOCKSS service model is based on a secure, closed-access network of servers set up between the network members. Each institution in the network runs a server that is linked securely to the network but maintained by different systems administrators. A new ingested object is replicated to other nodes in the network for preservation. The servers also check in with each other to make sure that all copies of the objects are identical. If a mismatch is detected, the servers come to a consensus regarding which copies are correct and which do not match, and then the network repairs the "bad" files.

2. A *centralised* archive that acts as a service provider for a number of institutions participating in a network. The central archive uses external cloud or grid service for its storage layer or as an extra off-site storage Examples of such centralised service provision models are numerous since the replication of storage adds to the security of preservation services

3. A *network of repositories* that share a cloud or grid-based storage that is replicated between multiple sites to achieve more secure replication of stored data. A shared storage service layer based on grid or cloud infrastructure is analogous to the previous service model (see 2) above) but there are extra services agreed between several storage providers to ensure data redundancy and fixity. This service model has recently been deployed by the EUDAT project that offers data repositories a data replication service as part of its B2SAFE service.[43]

4. *Repository outsources* one or several functions other than storage to a cloud or grid service provider. The outsourced functions could be computationally intensive, like quality assurance and conversion at the ingest stage of archive workflow; metadata management combined with user access and authentication. An example of such a solution is the e-Culture Science Gateway that was developed as part of the Indicate project, one of the predecessors of the DCH-RP project and has now been updated to become the DCH-RP e-Culture Science Gateway (eCSG).[44] The Indicate project developed eCSG to host the catalogue and user access system to Italian libraries.

5. *A cloud or grid service provider offers* all repository services and effectively becomes a digital preservation repository. Although no examples of full-scale digital repositories being supported on cloud or grid platforms are available, these solutions are being discussed and it is likely that an institution somewhere may have implemented a private cloud technology to host its repository system. The DuraCloud service in the US is at present the closest known example to this scenario.[45]

MODELING TRUST

Similar to the lack of a reference model for distributed digital preservation services, no trust model for a distributed preservation repository system yet exists. The need for a trust model for distributed digital preservation solutions has been discussed through a number of research papers (see deliverable D4.1)

[43] http://eudat.bsc.es/b2safe
[44] http://ecsg.dch-rp.eu/
[45] http://duracloud.org/

The recent EUDAT report on *Trust Establishment* describes the attributes of data objects that contribute towards trustable data and discusses organisational components that engender trust in a networked service:[46]

- Agreements, legal framework, governing structure
- Years of collaboration
- Relations within communities
- Configurable solutions.

The trust network for EUDAT's replication service between two data centres and their partners relies strongly on a shared user authentication and identification solution (ID Provider) and a service providing persistent identifiers to the objects in the digital repository (Handle System).

The need for a transitive trust model for distributed digital preservation solutions is, thus, accepted in the research literature, but as yet none of them have been implemented or could be relied on as working best practice.

2. RISK ASSESSMENT AS A FORM OF ESTABLISHING TRUST

In the absence of a universally accepted trust model for distributed digital preservation architectures, the search for alternatives has led to risk assessment as a method of establishing and communicating trustworthiness of a preservation service. The Digital Repository Assessment Method Based on Risk Assessment (DRAMBORA)[47] has been in active use since 2007 and has proved that risk registries are an effective means of engaging stakeholders and managers of repositories in discussion of trust and sustainability of services. Indeed, risk is viewed by many of these stakeholders as the "other side of the coin" of trust.

DRAMBORA describes a formalised process that assists repositories in establishing a comprehensive self-awareness of their objectives, activities and assets before identifying, assessing and managing the risks implicit within their organisation. The assessment report is essentially a risk register, presented in ten categories that helps communicating the problem areas to the repository staff and management, but also supports the trust decision-making for partners and external stakeholders who need to be able to estimate the risks they are taking when entrusting the repository. Examples of the development of risk profiles are given in deliverable D4.1 *Trust building Report*. The results have also led to certification frameworks like the Cloud Security Alliance Security, the Trust & Assurance Registry (STAR)[48] and systematic studies of risks around outsourcing digital preservation services to the cloud.[49]

The key concerns with outsourcing preservation services to third parties like cloud or e-Infrastructures have roots in different jurisdictions that govern cultural heritage institutions and e-Infrastructure providers, as well as with the general nature of distributed computing. The main areas of risk are related to:

- legal and governance – incompatibility of regulatory frameworks, legal liabilities;

[46] EUDAT (2013). D4.3.1: *Trust Establishment Report* http://www.eudat.eu/system/files/EUDAT-DEL-WP4-D4%203-Trust%20Establishment%20Report.pdf

[47] http://www.repositoryaudit.eu/

[48] https://cloudsecurityalliance.org/star/

[49] Aitken, B., McCann, P., McHugh, A., Miller, K. (2012). *Digital Curation and the Cloud*. Final Report. Produced by the Digital Curation Centre for JISC's Curation in the Cloud Workshop, Hallam Conference Centre http://www.jisc.ac.uk/media/D/0/1/%7BD01C1CDB-AF99-4A20-A9BC-12E73DB224DD%7DCuration-in-the-Cloud.pdf

- security – loss of data or service;
- data transfer – bottlenecks due to bandwidth restrictions, entrenchment due to vendor "lock-in".

Addressing these groups of risks is vital for the digital repository for both maintaining its level of service as well as the level of trust it enjoys with its stakeholders.

Appendix 1:1 includes a risk analysis tool that repositories can use to assess the policy, legal and organisational level risks when negotiating a service contract with an e-Infrastructure or, indeed a cloud service provider.

Security risks that relate primarily to fixity of information, information loss and security, multi-tenancy and shared technology issues in distributed infrastructures, but also to insecure or incomplete data deletion, are well documented in literature and standards (e.g. ISO 27001). Complete risk registers developed for outsourcing preservation service can be consulted (see for example Cloud Sweden 2011[50] and Cloud Security Alliance https://cloudsecurityalliance.org/[51]).

Data transfer issues and exit strategies can be mitigated in service level agreements between the digital repository and the service provider.

Risks specific to preservation activities can be identified with the help of the DRAMBORA toolkit.[52] Since in a majority of cases the repository will not be outsourcing core digital preservation decision-making to an e-Infrastructure, these risks are not part of the trust-forming issues. Nevertheless, transparency and communicated accountability for digital preservation activities would contribute towards increased trustworthiness of the repository. The intended use of the risk analysis tool in Appendix 1:1 is described below as a use case scenario that cultural heritage institutions (CHI) can modify according to their own specific needs of services that they are outsourcing to e-Infrastructures.

1) CHI conducts a risk analysis of its own operations or the particular service that it is looking to outsource, using the DRAMBORA toolkit, the risk analysis tool presented in this report or a risk profile tailored specifically for the CHI or its service.

2) CHI drafts requirements for the service(s) it plans to outsource and highlights the specific vulnerabilities / risk areas that it considers vital components for its services to continue to be trusted.

3) The e-Infrastructure and CHI jointly analyse risks related to the listed service requirements, agree on risk mitigation measures and how these can be made public (without disclosing technical or business details that may jeopardise the competitive advantage of either or both parties).

4) The resulting risk register is published, reviewed and updated at regular intervals.

5) CHI can additionally conduct a self-assessment using one of the repository assessment methods (DSA, TRAC, DIN 31644, ISO 16363) and include the risk register of outsourced services in the assessment results. Once auditing and formal certification service of digital repositories becomes available, the CHI may consider applying for certification based on the results of the self-assessment.

[50] Cloud Sweden (2011). Areas and problems to consider within information security and digital preservation during procurement and use of cloud services.Guidelines.
http://cloudsweden.files.wordpress.com/2011/11/cloud_sweden_security-digitalpreservation_v1-1-1_english_final.pdf

[51] https://cloudsecurityalliance.org/

[52] http://www.repositoryaudit.eu/

3. FEDERATED ACCESS

Methods of accessing services have evolved dramatically in the last decade. One of most important changes relates to the way in which users access applications and how applications manage users.

In the past user access was managed centrally by each application, which meant that users had to register and get application-specific credentials. This model did not work efficiently with the proliferation of applications and with the need from institutions to offer services beyond their organisational borders.

The current best practice is that authentication and authorisation are decoupled from the application:

- Authentication of the users is done by their user **Identity Providers** (i.e. the user's organisation), while
- Authorisation is done by the services (**Service Providers or Relying Parties**) based on the information (identity information) received by the Identity Providers and on the characteristics of the services.

Access to resources that follows this model is known as **Federated Access.** Identity Federations are the infrastructures deployed to enable federated access: these encompass a number of institutions that agree to inter-operate and offer services under a set of well-defined rules.

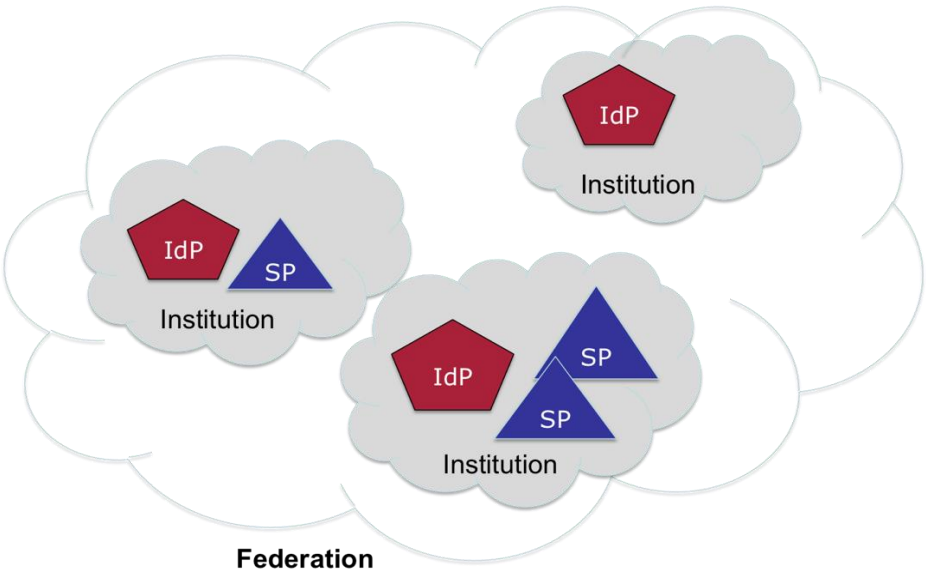


Figure 17: Identity Federation Model

Federated Access has brought several advantages both for users, who can benefit from a better user experience (fewer credentials to remember, log in once and access multiple applications, lower risk of forgetting their credentials) and for the service operators, who in practice outsource the user management life-cycle and can focus on authorisation. Federated access also increases security, by using a trusted connection between the identity provider (IdP) and the service provider; this trust connection is built by using standard protocols, legal framework and policies that are shared by the participating entities.

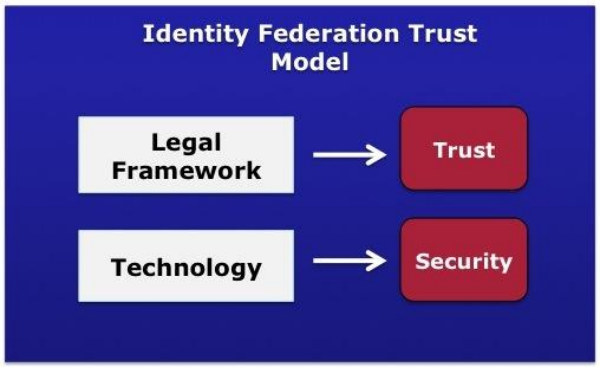


Figure 28: Trust Model in Federated Access

Typically R&E federations are operated nationally by the National Research and Education Networks (NRENs) for their community in the country or region concerned. In a federation resources are offered to the participating partners of that federation

Federated access requires considerable technical expertise to set up the technical infrastructure, whether this is about creating an IdP or an SP. In the context of digital cultural heritage this is potentially a barrier as emerged from the survey. Sadly at the moment there is no software that can be easily installed.

Some federations offer greater support to their users, for instance by installing/operating the IdPs and by offering technical support for non-commercial services. However the cost-recovery model and the availability of manpower make it hard to follow this model for all federations.

- *Recommendation:* Use a managed service to operate your IdP, whether a commercial offering (such as OpenAthens, Gluu, Ping Identity and equivalent) or one offered by the NRENs (such as GARR's IdP in the Cloud).

Although institutions and services are free to implement authentication processes as they wish, especially for users with limited technical know-how, authentication based on digital certificates should be avoided. As the grid world has demonstrated there are a number of usability issues related to digital certificates.

- *Recommendation:* Avoid the usage of digital certificates; if services require a digital certificate (i.e. grid facilities), use solutions like the e-CSG to hide the complexity.
- *Recommendation:* The usage of social network identities should not be discarded; there may be applications for which a social network account is sufficient.

Service providers should design their interface to be easy to use; particularly in the case of federated access it is important to follow accepted best practices to implement federated login in ways that improve user satisfaction and increase successful logins. The REFEDS group has produced guidelines to help login for federated access.[53]

- *Recommendation: Applications should use simple graphic interfaces, rather than command line, to encourage wider usage.*

[53] REFEDS stands for Research and Education Federations. REFEDS is a membership group of identity federations from around the world. See https://refeds.org

- *Recommendation: Especially if federated access is provided, best practice guidelines should be followed to improve user satisfaction*

The deployment of an e-Infrastructure for the digital cultural heritage domain requires significant investments, even if existing infrastructures are reused. Therefore, it is important to engage national and international decision makers to secure funding. Federated access is recommended although there may be cases (for instance if there are no plans to offer the service widely) where local access can be a better option. Federated access works well for web-based applications. The technology used to date to support federated access for applications that do not run in a browser is still immature; this should be considered when deciding to provide federated access mechanisms.

10. ANNEX 1.1 POLICY, ORGANISATIONAL AND LEGAL RISKS IN A DISTRIBUTED DIGITAL PRESERVATION SERVICE

Policy risks

Risk Identifier:	R01
Risk Name:	Vendor lock-in
Risk Description:	The organization becomes dependent on the services offered by the service provider, or is unable to change to another service provider without high switching-costs or losing assets.
Example Risk Manifestation(s):	Lack of standard technologies among the service providers to allow data portability (APIs, formats, procedures…). Difficulties on migration from one provider to another or to in-house services (portability and interoperability issues). Increase of data lock-in at the same rate as the amount of data stored in the e-Infrastructure if portability is not provided.
Mitigation strategies:	Negotiation of exit-strategy with the service provider. Selection of provider with the most suitable options regarding interoperability with the organization. Use of open standards, whenever applicable.
Risk Identifier:	R02
Risk Name:	Loss of governance
Risk Description:	The organization cedes to the e-Infrastructure provider governing responsibilities over a number of issues concerning the assets stored in the e-Infrastructure.
Example Risk Manifestation(s):	The organization transfers the responsibility for issues affecting security to the e-Infrastructure provider and no reports/logs are shared with them, making it impossible to audit or control the assets. Security procedures of provider are unknown, not agreed upon or are not aligned with the organizational ones. Compliance challenges with regulatory or legal environment due to the lack of guarantees on the authenticity, integrity and reliability of information stored.
Mitigation strategies:	Ensure that Service Level Agreements (SLA), contracts or any other agreements are complete and clarify roles and responsibilities of each of the parties. Reservation of rights by the e-Infrastructure provider should be analysed in detail and restrained when necessary. Clarify the potential provision of services by third-parties and their compliance with the guarantees provided by the service provider.
Risk Identifier:	R03
Risk Name:	Loss of ownership
Risk Description:	Organization cedes ownership of digital assets or related information to the e-Infrastructure provider.
Example Risk Manifestation(s):	Service provider takes control of assets due to a lack of transparency on the agreements. Use of transactional and relationship information collected by the e-Infrastructure provider that might be revealing or commercially valuable.
Mitigation strategies:	Clear terms of contract and service, including statements on the ownership of the assets. Clear roles and responsibilities in the contract.

Risk Identifier:	R04
Risk Name:	Non-compliance with certification and accreditation requirements
Risk Description:	Stakeholders are not able to meet confirmation of the characteristics required to fulfil such certification and accreditation frameworks.
Example Risk Manifestation(s):	Standards not adapted to the use and characteristics of the e-Infrastructure infrastructures, thus there might be an impact on compliance or certification to them. Control on the location of the data could be mandatory to fulfil the standards' requirements for accreditation, and in some cases it is unknown by the organization. E-Infrastructure provider is not certified by standards that could increase the assurances on information security to the organization. E-Infrastructure provider does not allow the organization to audit their processes.
Mitigation strategies:	Selection of e-Infrastructure providers accredited by relevant certification schemes. Negotiation with e-Infrastructure provider on the requirements on communication, reporting and audit to ensure trustworthiness on their procedures and compliance with the SLA. Selection of e-Infrastructure providers that allow selection of the jurisdictional areas to allocate the organizational assets.
Risk Identifier:	R05
Risk Name:	Loss of service level or availability
Risk Description:	The e-Infrastructure provider fails in providing availability of the service or its quality levels are compromised.
Example Risk Manifestation(s):	E-Infrastructure provider does not reach levels of availability specified on SLA. Service credit or insurance does not compensate for loss of service, income and reputation. Planned downtimes are not included in terms of lack of availability. Organization cannot control / measure availability levels and communication procedures / reporting not established. Data loss and inaccessibility. Business continuity and data recovery plans are not ensured. Scalability expectations (either up or down) not met in a timely fashion by service provider. Failure on backups leading to data loss.
Mitigation strategies:	Make sure that SLA is detailed enough on the availability of the service and possible downtimes. Assess whether the compensation measures for downtimes adequately compensates for impact for the potential loss of service. Establish procedures to get timely communication and reporting from the service provider. Ensure that adequate plans for business continuity, data and disaster recovery or incident response are in place. Understand resource provisioning policies and procedures of e-Infrastructure provider (resource overload, scalability, etc.) and assess whether it fits with organizational needs. Agree upon a protocol for system updates and other planned maintenance activities to minimize impact on service.

Risk Identifier:	R06
Risk Name:	Non-compliance with existing information management and preservation policies and procedures
Risk Description:	Organization policies and procedures to manage their digital assets throughout their life-cycles are not aligned with the requirements of using e-Infrastructure technologies.
Example Risk Manifestation(s):	Non-alignment with the OAIS reference model causes difficulties on transferring responsibility to an external party of some processes. Preservation tasks not offered by the e-Infrastructure provider. Removal actions are not transparent or appropriately carried out. There are severe technical difficulties around the destruction of records in the e-Infrastructure and its verification. Disposal of public records not achieved as specified by the organization's retention and disposal schedule (multiplicity of locations, backups…). Difficulties to put in practice retention policies for the assets stored in the e-Infrastructure. Metadata mismanagement results in portability becoming possible. Diminished level of metadata quality, losing usability for preservation. Loss of control on provenance of the data. Loss of integrity and authenticity of the data. Lack of transparency on data migration and transformation actions.
Mitigation strategies:	Keep track of backup copies or any action that the service provider takes on the stored assets. Detail policies and procedures including all aspects that might be affected by transferring parts or the whole workflow to the e-Infrastructure. Seek compliance from the e-Infrastructure provider with policies and procedures to ensure bit and logical preservation. Ensure complete record and control over the processes on the chain of preservation, to assure authenticity and reliability.
Risk Identifier:	R07
Risk Name:	Difficulties in monitoring, auditing and reporting e-Infrastructure services
Risk Description:	E-Infrastructure provider does not allow the organization to monitor the service to check compliance with SLA in aspects such as information security and performance measurements. There are no guarantees that SLA is being fully accomplished or that the quality of service is adequate.
Example Risk Manifestation(s):	No access to logs provided to the organization. E-Infrastructure provider does not provide standard audit documentation and reports. E-Infrastructure provider is not audited internally or by external bodies. E-Infrastructure provider does not provide the appropriate tools to the organization to allow auditing of performance. SLA does not reflect in which ways performance can be measured and controlled.
Mitigation strategies:	Establish requirements on audit and reporting, and how these should be carried out by the service provider. Ensure the correct level of readiness in the organization to fulfil these new tasks.

Risk Identifier:	R08
Risk Name:	Non-compliance with organization's security policy
Risk Description:	Security policies and procedures of the e-Infrastructure provider are not aligned with the organization's own policies and fail to fulfil its needs.
Example Risk Manifestation(s):	Organization's security policy has to be adapted to be aligned to those of the e-Infrastructure provider. Access restrictions are not under the organization's control, and there is no assurance on unauthorised access.
Mitigation strategies:	Reach agreements with the service provider on particular conditions in SLA, contracts, terms of service. Clarify roles and responsibilities of each party to avoid security areas not being covered. Establish adequate communication and reporting protocols that the provider must comply with. Understand and agree upon authentication and access management policies to be carried out by the e-Infrastructure provider.
Risk Identifier:	R09
Risk Name:	Limitation of liabilities on Service Level Agreements
Risk Description:	The definition of the levels of service includes limitations on the responsibility the service provider holds.
Example Risk Manifestation(s):	Unclear definition of roles and responsibilities in the agreements between organization and e-Infrastructure provider. Low level of transferability of liability to e-Infrastructure provider. Legal and reputational implications faced by the organization.
Mitigation strategies:	Clarify roles and responsibilities of both parties on agreements. Identify cases with no responsibility or obligation to compensation and assess whether the model fits with the organization's purposes.
Risk Identifier:	R10
Risk Name:	Organization fails to revise its own policies and procedures
Risk Description:	Rationale and/or business activities and processes are not adapted to the new architecture of the service, leading to inefficiencies or contradictions.
Example Risk Manifestation(s):	New workflows not included in the organization's procedures. Organization's security policy not updated or aligned with that of the e-Infrastructure provider.
Mitigation strategies:	Assess own policies and procedures and revise them according to the agreements reached with the service provider. Adjust roles and responsibilities in the organization.

Organizational risks

Risk Identifier:	R11
Risk Name:	Lack of sustainability related to financial resources
Risk Description:	The organization dismisses cost implications of e-Infrastructure services in the long-term.
Example Risk Manifestation(s):	Organization does not own the resource, which implies on-going payment for the e-Infrastructure provider services due to usage-based pricing. Cost implications of regular accesses or processes in the e-Infrastructure not considered. Increments in the needs for bandwidth or storage significantly vary the costs. Increase of computational expenses due to new operations (e.g. data/text mining) not affordable for the organization. Additional costs might arise: hidden-costs, extraction process related costs, licensing costs, metadata updates, etc.
Mitigation strategies:	Clarify all possible additional costs and likeliness of increases. Seek guarantees on the ability to switch between vendors, avoiding lock-in. Ensure the level of budget. Request additional funding or revise objectives when this is not possible. Maintain contingency fund. Review funding strategy.
Risk Identifier:	R12
Risk Name:	Loss of business or service reputation
Risk Description:	Organization's stakeholders change their opinion about and lose confidence and trust in the service provided by the organization.
Example Risk Manifestation(s):	Lack of reputational isolation leads to a contagious effect due to negative activities on the part of co-tenants. Lack of resource isolation in physical resources shared by multiple customers allowing unauthorized access or manipulation. Negative stakeholders' perceptions towards the use of the e-Infrastructure to store data with privacy implications. Lack of transparency on the use of cross-organizational authentication systems and perception of privacy infringing on end-users.
Mitigation strategies:	Comply with all relevant certification schemes. Increase transparency towards end-users on the use of their personal data. Ensure that possible vulnerabilities (e.g. hypervisor security model) are under control by the e-Infrastructure provider.
Risk Identifier:	R13
Risk Name:	Role changes of organization's staff
Risk Description:	The use of e-Infrastructure computing requires different capabilities and modifications in the roles played by the organization's staff.
Example Risk Manifestation(s):	Management and maintenance tasks might differ or increase, if there is a need to manage and secure the operating system, applications and virtual instances. Organization has to monitor e-Infrastructure services to check performance of SLA. Difficulties for the staff to transition to an e-Infrastructure service.
Mitigation strategies:	Define new roles and profiles according to the new workflow/tasks. Monitor performance and adapt plans after assessment. Implement a training plan for the staff to improve competences and raise awareness on issues concerning the new system.

Risk Identifier:	R14
Risk Name:	Staff skills become obsolete
Risk Description:	The introduction of new roles brings up the need for a whole new set of skills.
Example Risk Manifestation(s):	No training plans have been established before/after the introduction of the new systems.
Mitigation strategies:	Determine the organizational needs to address the new tasks and assess whether the staff members need additional training or there is a need for new members of staff. Review performance regularly and implement training plans accordingly.
Risk Identifier:	R15
Risk Name:	Resistance to change in the organization
Risk Description:	Perceptions of organization's staff towards the use of e-Infrastructure technologies do not contribute to the acceptance of the new model.
Example Risk Manifestation(s):	The change process is not well understood or followed within the organization. Difficulties in implementation and failures in usability of the new systems.
Mitigation strategies:	Assess new organizational needs and identify staff expectations and experiences. Establish change management plan.
Risk Identifier:	R16
Risk Name:	Management failure
Risk Description:	Organizational management shortcomings produce a failure on the achievement of its objectives.
Example Risk Manifestation(s):	Insufficient allocation of resources considering the cost-models used by e-Infrastructure providers. Organization does not have a business continuity plan to mitigate effects of a crisis involving critical processes or assets.
Mitigation strategies:	Design and adequate the policies and procedures according to the changes in the organization and establish review mechanisms. Establish business continuity plans or any other mechanisms to mitigate and overcome the failure.
Risk Identifier:	R17
Risk Name:	Business objectives not met
Risk Description:	Organization fails totally or partially to achieve the foreseen outcomes.
Example Risk Manifestation(s):	Preservation of the assets is not adequately achieved due to poor performance of operations in the e-Infrastructure. Difficulties to prove authenticity and integrity of information preserved. Personal data leakage/disclosure to third parties.
Mitigation strategies:	Ensure compliance with organizational policies and procedures. Monitor and review service provider performance. Review preservation policies and procedures. Monitor business objectives and redefine them whenever necessary.

Risk Identifier:	R18
Risk Name:	Enforced cessation of organization's operations
Risk Description:	Impossible to continue organization's activities.
Example Risk Manifestation(s):	Bankruptcy of service provider without an adequate strategy leads to the loss of the assets stored in the e-Infrastructure. Technical failure affects the stored data causing an irreparable loss of the main digital assets. Failure in outsourced critical business process. Organization lacks succession plan for its digital assets.
Mitigation strategies:	Establish succession plans. Establish exit strategy. Establish policies and procedures ensuring security of assets.
Risk Identifier:	R19
Risk Name:	Inability to evaluate organization's success
Risk Description:	Organization is not able to determine whether its objectives have been achieved or not and to what extent.
Example Risk Manifestation(s):	E-Infrastructure provider lacks transparency and does not provide the organization with sufficient information through audit reports about the state of the stored assets. Organization has no mechanisms in place to monitor the performance of the e-Infrastructure provider. Organization does not engage with stakeholders to determine satisfaction levels.
Mitigation strategies:	Establish means of assessment of both internal and external actions. Use external certification to determine the degree of competence.
Risk Identifier:	R20
Risk Name:	Difficulties in negotiating contracts and terms of service
Risk Description:	Organization lacks the ability to negotiate agreements with e-Infrastructure provider.
Example Risk Manifestation(s):	Organization does not have access to a legal counsellor able to determine the most suitable conditions for the organization. E-Infrastructure provider has standard contract and additional agreements that cannot be negotiated. Organization does not conduct due diligence assessment before entering into a contract.
Mitigation strategies:	Seek legal advice to give support on the negotiation of contracts and other agreements.

Legal risks

Risk Identifier:	R21
Risk Name:	Location and jurisdictional implications
Risk Description:	Location of the e-Infrastructure resource unknown, established in a different jurisdictional area to that where the organization is located.
Example Risk Manifestation(s):	Legal practices and regulations differ from those in the organization's jurisdictional area. E-Infrastructure provider does not give the organization choices on the location of information. E-Infrastructure provider does not provide timely information about changes of location of the e-Infrastructure resource.
Mitigation strategies:	Establish agreement with the e-Infrastructure provider about the jurisdictional areas where organization's assets can be stored. Request notification on proposed changes of location. Have control over the regulations of the jurisdictional areas that can affect organization's assets. If possible, reach contractual agreement on the court and applicable law in case of eventual legal dispute.
Risk Identifier:	R22
Risk Name:	Non-compliance with data protection laws
Risk Description:	Breach of regulatory requirements of protected data such as those containing personal or sensitive information.
Example Risk Manifestation(s):	Sharing protected information with e-Infrastructure providers might be non-compliant with privacy laws. Records management and disposal laws may introduce limitations on the ability of government agencies to share information with e-Infrastructure providers. Data stored in the e-Infrastructure is accessed by unauthorised people, intercepted or leaked to the public. Breach of the limits on privacy set up by regulations on using personal information in a cross-organisational setting for the purpose of identity management. Mismanagement of encryption keys leads to the loss of confidentiality of the information stored.
Mitigation strategies:	Ensure compliance through formal agreements with the e-Infrastructure provider and get assurance of its levels of liability for unlawful actions. Select, on the role of controller of personal data, a processor with adequate guarantees on security measures. Request e-Infrastructure provider's assurance on reporting on any data processing that they carry out Organization should be informed on data security activities and the data controls e-Infrastructure provider has in place. Get guarantees of a robust system for authentication, authorization and accounting. Establish a hybrid model with highly sensitive data stored in a private e-Infrastructure. Carry out an appropriate Privacy Impact Assessment (PIA) before entering into agreements with e-Infrastructure provider.

Risk Identifier:	R23
Risk Name:	Non-compliance with IPR regulations
Risk Description:	Breach of regulatory requirements of copyright, patent infringement or other IPR-related misdemeanour.
Example Risk Manifestation(s):	The organization fails on fulfilling IPR laws requirements by using e-Infrastructure technologies to store protected materials. The organization does not properly manage rights and restrictions of protected materials stored in the e-Infrastructure. Information is not properly classified according to rights and restrictions. Metadata mismanagement causes the loss of rights metadata and thus the lack of sufficient contextual information to identify the level of protection. Actions taken for digital preservation of assets are not-compliant with IPR regulations.
Mitigation strategies:	Assess whether the assets stored in the e-Infrastructure are subject of IPR restrictions. Seek legal advice to determine legality of the activity. Establish the conditions and ensure e-Infrastructure provider's compliance with organization's requirements, without diminishing the quality of the service. Establish and monitor agreements with the rights-holders, when necessary.
Risk Identifier:	R24
Risk Name:	E-Infrastructure Provider disclosure obligations implications
Risk Description:	Legal requests enforce the e-Infrastructure provider to give access to the information under their supervision.
Example Risk Manifestation(s):	E-Infrastructure provider might be obliged to examine user records to find evidence of irregular activities.
Mitigation strategies:	Require from the e-Infrastructure provider information about procedures and conditions for disclosure and timely notification for any requested disclosure. Demand guarantees on security of organization's data when co-tenants are subject of disclosure.
Risk Identifier:	R25
Risk Name:	Unintentional disclosure in multi-tenant environments
Risk Description:	E-Infrastructure resource is shared by multiple tenants and isolation failure might allow third parties to access to organization's data.
Example Risk Manifestation(s):	Physical drives are shared with other tenants that are involved in a legal case and whose information disclosure is enforced.
Mitigation strategies:	Request guarantees on the full isolation of resource, even not sharing physical machines in the case of critical data. Ensure sufficient levels of encryption and reliable key management.
Risk Identifier:	R26
Risk Name:	Inadequacy of regulations and legislation to e-Infrastructure
Risk Description:	Regulations affecting the organization’s assets in the e-Infrastructure do not contemplate the challenges imposed by the use of e-Infrastructure technologies.
Example Risk Manifestation(s):	Contractual relationships are the only ones filling gaps within the regulation framework. Impossibility of compliance and possible liability for infringement of regulations.
Mitigation strategies:	Seek legal advice to avoid non-compliance with legal framework.

Risk Identifier:	R27
Risk Name:	Liability for infringement of legal requirements and regulations
Risk Description:	Organization is legally accountable for not fulfilling responsibilities or acting beyond the scope of what is allowed on the basis of legal and regulatory instruments.
Example Risk Manifestation(s):	Organization has to face legal consequences of the infringement of the laws protecting information even in the event of e-Infrastructure provider actions. No clear delineation of liability has been set up between parties.
Mitigation strategies:	Monitor legal framework to ensure compliance of the actions, procedures, policies, agreements, etc. Seek legal advice to determine legality of activities with respect to legislation. Establish policies and procedures to follow in the event of legal challenge.
Risk Identifier:	R28
Risk Name:	Exit-strategy deficient or not defined by the organization
Risk Description:	Lack of assurance that contractual, technological or planning resources are in place to move out or replace e-Infrastructure computing services.
Example Risk Manifestation(s):	The e-Infrastructure provider does not offer a standardised export procedure for information and the organisation needs to develop its own programme to extract its information. Information retrieval requires a change in format, with possible consequences for authenticity, reliability or legal admissibility.
Mitigation strategies:	Analyse and document all the procedures and properties that are critical for the assets stored in the e-Infrastructure. Define and establish an exit-strategy according to them.
Risk Identifier:	R29
Risk Name:	Acquisition of e-Infrastructure provider
Risk Description:	The ownership of the service provider changes and operations and assets are transferred.
Example Risk Manifestation(s):	Policies, procedures and terms of service might change. Unknown accountability or affiliations of new e-Infrastructure provider.
Mitigation strategies:	Include guarantees in the contract to keep the conditions agreed upon in the event of changes in provider ownership. Establish exit-strategy.
Risk Identifier:	R30
Risk Name:	E-Infrastructure provider ceases business
Risk Description:	The service provider goes out of business and ceases operations.
Example Risk Manifestation(s):	Limitations on retrieving data in the event that the provider ends business operations with little or no warning. Difficulties in information and metadata portability. There is no business-continuity strategy established by the cloud provider.
Mitigation strategies:	Conduct due diligence to get assurance on cloud provider accountability, maturity, viability, etc. before entering into a contract. Establish exit-strategy.

Risk Identifier:	R31
Risk Name:	Subcontract to third-parties by e-Infrastructure provider
Risk Description:	Third-party subcontractors provide the e-Infrastructure provider with parts of the service or infrastructure for the deployment of the service to the customer.
Example Risk Manifestation(s):	Third-parties subcontractors have different policies and procedures. Subcontractor does not offer the same guarantees of service availability. Subcontractor established in different jurisdictional area.
Mitigation strategies:	Acknowledge which services are subcontracted by third parties and establish the necessary procedures. Get assurance that SLA with third parties does not diminish levels of service. Ensure that their third party performance levels and security compliance are monitored by the e-Infrastructure provider. Ensure that all organizational requirements are met in case of contracting services through an e-Infrastructure services broker.
Risk Identifier:	R32
Risk Name:	E-Infrastructure provider's reservation of rights
Risk Description:	E-Infrastructure provider reserves certain rights on the use of customer's assets under their supervision.
Example Risk Manifestation(s):	E-Infrastructure provider changes its terms and policies unilaterally. Secondary use of customer information by the e-Infrastructure provider.
Mitigation strategies:	Ensure transparency on agreements. Establish the conditions that the e-Infrastructure provider should comply with to avoid unlawful actions.
Risk Identifier:	R33
Risk Name:	Evidential value of information diminished
Risk Description:	It is not possible to prove authenticity and integrity of records stored in outsourced e-Infrastructure facilities.
Example Risk Manifestation(s):	No reliability in e-Infrastructure provider procedures on migration processes, backups, etc. Information on security policies of e-Infrastructure provider is not documented or accessible. E-Infrastructure provider does not update organization about issues concerning corruption, loss or data changes.
Mitigation strategies:	Monitor actions taken by the e-Infrastructure provider on the data stored in the e-Infrastructure. Require transparency in e-Infrastructure provider's policies and procedures.
Risk Identifier:	R34
Risk Name:	Liability for breach of contractual or licensing relationships
Risk Description:	Organization is legally accountable for not fulfilling responsibilities or acting beyond the scope of what is allowed in contractual relationships with stakeholders.
Example Risk Manifestation(s):	Protected materials are stored in a third-party storage facility without the consent of the right holders.
Mitigation strategies:	Monitor contractual relationships to ensure their terms are corresponded. Seek legal advice to ensure no breaches of contractual relationships. Establish policies and procedures to follow in the event of contractual challenge.

11. ANNEX 2 PRIORITISED ACTIONS SHORT TERM IN SELECTED AREAS OF THE ROADMAP

1. HARMONISE DATA STORAGE AND PRESERVATION

1.1 Define an initial set of critical system requirements

General needs and requirements in a digital preservation context.

Examples (listed regardless of priority):

Miscellaneous issues

- Reliability and robustness
- Assurance of valid licensing procedures, commercial conditions, and transactions
- Open, scalable, and flexible solutions (built on open industry standards like J2EE and XML)
- Ease of use (for example, user-friendly interfaces)
- OAIS compliance
- Multilingualism

Content/information issues and metadata issues

- Mechanisms for integration and automation of appraisal and ingestion of digital material
- Automatic metadata capture and extraction
- Separation of content (information) and metadata
- Various content formats (from print-based documents to digitized images)
- Ontologies for both visual and textual concepts
- Annotation services

Performance issues

- Scalability (up to hundred terabytes or more)
- Performance for hundreds of thousands of electronic documents

Trust issues and security issues

- Authenticity and integrity of data
- Continuity (which means the handling of information, both data and metadata, for at least the next 100 years)
- Identification of digital objects which are in danger of becoming inaccessible due to changes in technology
- Security during transmissions of files between countries
- Validation (certification) of software and hardware environments required to render the digital objects

Infrastructure-related issues

- Distributed systems
- Virtualisation

Hardware-related issues

- Support of many storage media and devices
- Backup and restore

Specific requirements

Need for simplicity

Integrating preservation workflows with e-Infrastructures normally requires significant levels of computing and IT expertise, not always available in cultural heritage institutions. The solutions developed need, therefore, to be tested for their simplicity of installation, management and use.

Metadata

The metadata connected to a digital object is crucial for the possibilities to preserve it for future use. It has to include basic descriptive information about the file as well as information about the file format of the object. The metadata collected about a digital object helps to place it in context, as well as give specific information, which is essential for making sure the object in mind is authentic (hasn't been added to or modified in any way). This is especially important for digital files, which in contrast to print media can be easily changed in ways that may not be easily apparent. Metadata can be linked to the digital object or encapsulated with the digital object itself. Encapsulating the metadata with the object ensures that the information stays with the file, no matter where it goes. Linking the metadata but storing it separately ensures that the information about the file can be recovered even if the object itself is lost. Depending on the actual situation, a decision about metadata has to taken before a cultural heritage institution enters into distributed digital preservation.

Storage in different locations

Archival data (master files) can often be stored offline, since they are infrequently accessed. It is best practice in many cultural heritage institutions to write digital archival data to more than one type of media and then store these in different locations.

Digital resources in continual use (surrogate delivery files) will typically be stored online. Online storage is often mirrored across multiple disks using redundant disk arrays (RAID).

Today clustered (data center) and distributed storage systems are normally used for distributed storage. A storage cluster consists of at least two independent storage nodes, running under the control of relevant software. When one of the nodes fails, the other immediately takes over all of its duties.

A data center is a facility housing computer systems and associated components like telecommunications and storage systems. It generally includes services such as redundant or backup power supplies, redundant data communications connections, environmental controls (e.g., air conditioning, fire suppression) and security devices. The concept Dynamic Infrastructure is a design of data centers making it possible for the underlying hardware and software to respond dynamically to changing levels of demand in more fundamental and efficient ways. This concept is also known as *Infrastructure 2.0* and *Next Generation Data Center*.

Cloud storage is often implemented with complex, multi-layered distributed systems built on top of clusters of servers and disk drives. Sophisticated management, load balancing and recovery techniques are needed to achieve high performance and availability. While there is a relative wealth of failure studies

of individual components of storage systems, such as disk drives, relatively little can be found reported, so far, on the overall availability behavior of large cloud-based storage services. Special care has therefore to be devoted to this issue before entering into a solution based on distributed preservation.

Migration of data and metadata

A routine error-checking schedule should be implemented and a strategy drawn up for migrating data and metadata to suitable formats as necessary. If a file format is becoming obsolete and a migration is planned, archival master files should be migrated to new formats that are non-proprietary. Quality control checks should follow any migration or refreshment so that any loss of data integrity can be identified and quickly addressed.

1.2 Needs and conditions for infrastructure federation

The needs to access networked applications and remote/distributed data is evolving dramatically. Authentication and authorisation are often separated from the application and the data themselves: authentication of the users is done by the users Identity Providers while the authorisation is done by the services based on the information received by the Identity Providers.

Access that follows this model is known as federated access and has advantages for both users and application developers. However, the usage of federated access requires that some technical and trust issues have to be solved.

For the DCH-RP project federated access is a key element, both in terms of using federated storage to handle preservation of cultural heritage data distributed all over Europe and in terms of user management. Federated access is in fact particularly desirable in a situation where services are offered across institutions to users that do not belong to the same institution that offers the service. In line with the objectives of the DCH-RP project, the ambition is not to establish a separate authentication and authorisation (AA) infrastructure for the DCH service and user community, but to use the most suitable AA services available in the research and education community.

Federated access provides the technical and policy framework to allow for services to be shared in a trustworthy fashion across borders. How authentication is carried out by the institutions and how rights management is carried out by the service provider is left up to the respective parties.

When deciding whether to offer federated access, e-Infrastructures offering services should assess their potential user-base: whether they expect many local users or many users coming from different institutions. Federated access caters for the latter use-case and brings the following benefits:

- *Users* will be able to log in once (single sign-in) using their institutional credentials and access multiple services (sign on), Single Sign-On, whilst having the assurance that their personal data will not be disclosed to third parties.
- *Digital cultural curators and cultural institutions participating* will be free of the burden of user name and password administration, and will have access to more tools for managing data. On a large scale of users this means reduced administration and service provisioning costs; and it avoids duplications of identity stores.
- *Collaboration* among different parties becomes easier.

The eCulture Science Gateway of INFN (Instituto Nazionale di Fisica Nucleare), is based on federation identities. eCulture Science Gateway was developed within the framework of the earlier INDICATE project. It will be upgraded with new functions by the Italian DCH-RP partner INFN and used for the DCH-RP projects Proofs of Concept.

When joining a federation, the following entities are expected to show up:

1. *Identity Providers* (IdPs) – typically organisations that hold information about users and manage user credentials, used to access to resources
2. *Service Providers* (SPs) – publishers, storage services, data management services, blogs, wikis – in fact anyone who wants to provide a 'sign-in' to resources without the hassle of managing user information.
3. *A policy or agreement* – that IdPs and SPs sign up to agree how to interact with each other. These are typically implemented at a national level.
4. *Registration* – a place to sign up and give to a federation information about your IdP or SP - also called your 'entity'.
5. *Metadata* – the collected information about entities, brought together in one place and typically digitally signed by a federation and published to its members.
6. *Discovery service* – a tool used by Service Providers to allow users to select their own Identity Provider.

Institutions in a federated context can act both as IdPs and SPs, or they can only act as either IdPs or SPs.

The first step to join a federation is to talk to the federation operator in a specific country. The list of existing federations is available online at: https://refeds.org/resources/resources_list.html

More information about federated access is available in annex 1.

1.3 Ongoing experience with grids and cloud solutions applied in cultural heritage institutions

One of the basic assumptions for the DCH-RP project is that grid and clouds approaches can offer a stable and reliable storage and computing platform to the digital cultural heritage community. In general it seems that this community's first priority, when it comes to digital preservation activities, is storage. Other identified priorities are computer capacity for integrity checks and access to advanced virtualisation services. One conclusion is, therefore, that at least two main approaches to preservation services must be in place for distributed solutions. In chapter 5.2.2 they are referred to as the "kiosk" model and the "turn-key" model respectively. What in the same section is called "micro services" could also be a fruitful approach to look into. However, if various micro services are to be used, they must be orchestrated in a way that assures that requirements for authenticity and integrity of preserved digital objects are not compromised.

When reviewing the limited experiences of distributed preservation of digital cultural heritage to date, the most striking observations are a feeling on the part of the e-Infrastructure developers and the operator's of frequent dissatisfaction on the users' behalf, and of users regularly reporting about difficulties in utilising the facilities and tools offered. Therefore, a roadmap establishing future approaches and methods of preservation definitely has to put special emphasis on how to bring the e-Infrastructure closer to the users, how to make the e-Infrastructure providers more sensitive to user demands and, on the other hand, how users can better exploit the opportunities offered by the e-Infrastructure.

1.4 Examples of use of platform as a service (PaaS) and of benefits offered by virtualisation

Although a number of preservation tools are available, their uptake and use in practice is very hard to measure, and so is the whole market for digital preservation services. The models for evaluating market

maturity are too general to fit easily a niche area like digital preservation. The Planets project conducted interviews with leading IT companies to explore the emerging market-place for digital preservation tools and services. Results of this study confirm that engagement is being led by memory institutions and driven primarily by legislation. There is perceived high demand for technology to support automation of digital preservation processes and for consultancy, training, awareness-raising and exchange of best practice, but the overall description of the services market was as a "market in its infancy".[54]

In recent years some new distributed services in digital preservation has been introduced. One example is the Data Archiving and Networked Services (DANS). In the Netherlands a federated data infrastructure is developing with DANS as a trusted digital repository, in the first place for research data, performing back-office functions like expertise in data governance and long term storage and accessibility.[55] Another example is Preservica, a cloud-based service to safeguard digital information. Preservica conforms to the OAIS model (ISO 14721:2003) and marketing themselves as providing all the tools required for building a long term digital preservation solution.[56]

2. IMPROVE INTEROPERABILITY

2.1 Identify and promote best practices

Annex 4 in the roadmap is dedicated to best practices, presenting an overview of the most important practical guidelines and lessons learned connected with the integration between the cultural heritage community and the e-infrastructure providers.

The second round of Proofs of Concept has been an important instrument for capturing best practices WP4 will support the validation of the results of WP3 through a range of tasks interacting with the DCH community.

2.2 Analyse interoperability issues

To avoid building 'digital silos' within the organisation, the following aspects need to be considered:

1. *Technical aspects*: a storage solution should be decided upon before producing any digital output, as it is of prime importance for the following steps in an organisations digital preservation programme; strategies for both online and offline storage should be considered for the digital resources to be stored, otherwise storage of digitised resources runs the risk of competing with limited resources for maintaining the administration platform; due to the large size of master files, an entire digital collection can be very substantial in size, possibly requiring a mixed architecture for data storage; the size of both master files and any surrogate files have implications for the amount of storage space required and should be calculated at the outset of the project.

2. *Semantic aspects*: there are many vocabulary sources already available and it makes sense to check these out before inventing a new one. Depending on its needs an organisation might:

- Use an existing controlled vocabulary;
- Adapt or customise a vocabulary in use;

---

54. *An Emerging Market: Establishing Demand for Digital Preservation Tools and Services*. Available: http://www.planets-project.eu/docs/reports/Planets-VENDOR-White-Paperv4.pdf (PLANETS 2010)

55 See www.dans.knaw.nl

56 Preservica Preservation as a Service (http://www.preservica.com)

- Developing its own vocabulary (not recommended though sometimes unavoidable);
- Use an "uncontrolled" vocabulary - i.e. keywords entered by the organisations cataloguers or its users – should not be done under any circumstances as it makes interoperability impossible or very hard to achieve.

Of course, it can be quite reasonable to use a combination of these approaches, for example a formal controlled vocabulary plus additional keywords to assist in retrieval.

In choosing a vocabulary, it is important to have in mind:

- The end users - are the terms used going to be meaningful to them?
- The community - it makes good sense to use vocabularies that similar collections are using.
- The nature and extent of the collection - if the collection is small, it will probably not need a detailed vocabulary.
- Copyright issues - it will maybe be necessary to check whether permission or a license is required to use the vocabulary in the way the organisation wish to.

3. Organisational and inter-community issues: while it is clear that a technical strategy is necessary to ensure digital preservation, it is also important that digital preservation receives an organisational commitment.

4. *Legal issues*: the transfer of personal data has to be in line with European directives on data protection and their implementation in national legislation; harmonisation of legal frameworks in general have also to be addressed, for example concerning the issue of cross boarder storage and differences in legal positions regarding preservation of master files

5. *Political/human aspects*: digital preservation is an active task, and it is imperative that the responsibility for all digital resources is firmly assigned and known to all stakeholders - digitisation projects should have, as part of their project specifications, a policy which covers:

- Who the digital resource or collection belongs to in the organisation and who is responsible for its upkeep;
- What the process is for deciding when and how refreshment/migration takes place and who makes the decision;
- Where the budget is coming from for this ongoing digital preservation investment.

12. ANNEX 3 LICENSE AGREEMENTS AND TERMS OF USAGE

As mentioned in chapter 5.2.2, the use distributed digital preservation services makes it important to understand and communicate the license agreements and terms of usage that are associated with digital resources, "born digital" ones as well as digitised representations of other cultural heritage artefacts. The Linked Heritage project investigated this topic and reported seven overall license types relevant here and broke these out further, for example describing at least four variations of the Creative Commons (CC) licenses in routine use.

The following table briefly summarises the licenses mentioned.[57] The table also mentions a highly structured method for license expression, namely ONIX-PL; this is not a license in itself but rather a machine-readable framework for conveying licensing and usage terms, conditions and prohibitions.

License	Description/purpose	More information
BSD Berkeley Software Distribution	One of a group of permissive software licenses, imposing minimal restrictions on the redistribution of the software covered by the license	http://en.wikipedia.org/wiki/BSD_licenses
CC Creative Commons	A series of public copyright licenses. Currently seven such license types exist	http://creativecommons.org/licenses/ See the website for more information on each license type: CC BY, CC BY-SA, CC BY-NC, CC BY-ND, CC BY-NC-SA , CC BY-NC-ND and CC0
GNU FDL GNU Free Documentation License	A "copyleft" licence designed for the free documentation of software, but which can be used for other text works	http://www.gnu.org/copyleft/fdl.html
GNU GPL GNU General Public License	A free software licence granting the licensee the right to change and redistribute the software free of the prohibitions of copyright law	http://www.gnu.org/copyleft/gpl.html
ODbL Open Database License	A license covering data in databases and allowing licensees, under certain conditions, to share create or adapt the database or its content	http://opendatacommons.org/licenses/odbl/

[57] More details can be found in Linked Heritage deliverables.

License	Description/purpose	More information
ODC PDDL Open Data Commons Public Domain Dedication and Licence	A license covering data in databases and allowing licensees, without attribution, to share create or adapt the database or its content	http://opendatacommons.org/licenses/pddl/1-0/
ONIX-PL ONIX for Publication Licenses	An XML format for the communication of license terms for digital publications in a structured and substantially encoded form	http://www.editeur.org/21/ONIX-PL/

13. ANNEX 4 BEST PRACTICES

Collecting best practices

The DCH-RP project has conducted practical experiments (proofs of concept) in the project partners' countries. The results are reported in two deliverables:

- D5.3 *Report on the First Proof of Concept*, handling seven out of 14 identified scenarios covering some fundamental concerns of digital preservation: (1) Operational challenges, (2) End user concerns and (3) New services and infrastructure integration
- D5.4 *Report on Second Proof of Concept* focused on integrated solutions and services tested in five experiments

The outcomes of the Second Proof of Concept have the form of best practice and are, therefore, highlighted in this annex. Both Proofs of Concept underlined the importance of assessing the software for the two most paramount requirements regarding the targeted users:

1. Ease of use of the tool or service for the *end user*
2. Ease of installation/provisioning for small IT departments or IT-experienced individuals.

Experiments

Experiment 1*:* explores a tool ("Matchbox") developed by the SCAPE project that allows automating the task of finding duplicate images in a set of files. "Data hygiene" activity is a necessary filter for diligently preparing a dataset for archiving, and for regular quality assurance and repository certification for preservation.

Conclusion

The basic Tool seems to work. The code is clearly written and it is stable enough to handle broken files in the test. It looks promising and there can be a demand for such a tool.

However several drawbacks exist that need to be addressed before MATCHBOX can be usable in a production environment. To name a few of these drawbacks:

- Output was hard to interpret certainly in an end-users perspective as was the a remark from the photographer when the tool was demonstrated;
- work-flow and practical use by the end-users must be taken into consideration:
- although FE was quick, cross-comparison of the 6 (six) images' features took several minutes.

Note: KIK-IRPA (one of the testing institutions) has ~1 million images in their database, and images and scans are added daily. Both the tool's performance, and above all speed and user-acceptance (ie. clear output, even for trivial cases like an indication of resolution reduction or added meta-data) are important to get this tool 'sold'.

Experiment 2: looked at the HAPPI (Handling Authenticity Provenance and Persistent Identifiers) service developed by the SCIDIP-ES project. Cultural heritage data is often included in various projects over a long period of time, which raises a number of needs and requirements as follows:

- Digital asset authenticity – establishing and maintaining the originality of the asset
- Data provenance – Keeping a trail of data usage events for audits and data usage indication
- Data reference persistence and validity – Idempotent data reference/identifier resolution over time and space to the correct storage location

Conclusion

For the first experimentation, i.e. deployment and setup, HAPPI toolkit has been successfully packaged into a virtual appliance that is deployed on the EGI federated Cloud infrastructure. This means that the DCH-RP community has available a running instance of SCIDIP-ES HAPPI toolkit 1.5.0 available at

http://90.147.102.191:8080/happi-server-1.5.0/.

This experimentation has demonstrated the extremely ease of installation/provisioning for small IT departments or IT-experienced individuals. Indeed, the software artifacts are provided with predefined configurations. So, the toolkit could be deployed and few easy steps (i.e. download packages, unzip DB server, run DB server and deploy HAPPI on tomcat).

Since its deployment and setup, HAPPI toolkit 1.5.0 is continuously running without having experienced issues and interruption of operation. This allows to assess its good level of maturity, as well as the underlying Cloud Infrastructures.

Moreover, the HAPPI toolkit instance does not integrate with the EGI authentication framework, demonstrating effective separation of infrastructure management authentication and infrastructure user authentication.

Even if the second experimentation is still ongoing, it is reasonable to assert that “HAPPI is a sample service for data provenance, facilitating repeatable science”, as well as it could be applied to DCH-RP community too, for its generic provenance model based on OPM and PREMIS.

For the above reasons, it has been decided to keep running the SCIDIP-ES HAPPI toolkit 1.5.0 over the end of DCH-RP project, in order to make it available for further experimentations and assessment.

Experiment 3: assesses a combination of services provided by the EUDAT project (B2SHARE and B2SAFE) in combination with a service (Platon) provided by PSNC to its national digital libraries and archives. The aim is to evaluate EUDAT’s services for curating and publishing a research community’s digital assets, in DCH-RP’s case the preservation of digitised and born-digital cultural heritage.

Conclusion

Preliminary results of the evaluation show that B2SHARE service provides data sharing and publication solution suitable for the needs of small cultural institutions and “citizen” “publishers” or “curators”.

However there are bugs and limitations that prohibit from using this service in a production environment. More thorough testing needs to be done to detect more major and minor bugs and users should be consulted to upgrade a number of functions.

It can be said that:

- Mass scale uploads and sharing may require more domain-optimised and specialised approach. Ensuring long-term data availability is not part of B2SHARE’s service scope and intent; therefore B2SHARE should be orchestrated with additional layers such as EUDAT B2SAFE and PLATON’s Archival Services.
- An effort should be made to provide a service that works more correctly;
- The search engine needs to be further developed taking into account user requirements;
- Metadata functions, tagging, etc. should be added.

Experiment 4: is investigating some of the results of the experiments in the first Proofs of Concept phase More specifically, this experiment revisited the use case of uploading digital assets to a remote Grid/Cloud infrastructure in conjunction with the e-Cultural Science Gateway (eCSG) developed by INFN-Catania. Including federated identity management and AAI into this experiment, this experiment is addressing two of the main outcomes of the previous experiment in the first PoC phase.

Conclusion

It was successfully demonstrated that customised uploaders can allow DCH institutions to make use of eCSG for the storing of their digital assets in automatic way. Moreover, ICCU has now a concrete example of the benefits of using federated credentials to access Service Providers belonging to the IDEM federation.

While doing this, INFN Catania have learned how to build an uploader portlet that can be customised in an easy and quick way for different metadata schemas and formats and this will allow further adaptations to other kind of repositories straightforward.

Experiment 5: concludes the second PoC phase with the aim of assembling a general-purpose digital preservation platform implementing a Service oriented Architecture (SOA). The focus of this experiment lies on reducing the total cost of ownership (TCO) of such a preservation platform through integrating as many generic services as possible, implementing as many preservation-specific standards as necessary, and addressing the needs of as many user communities as is feasible. In collaboration with the APA (through the APARSEN project) this experiment will also explore how an external, independent service provider might offer services around such a platform to the target market while integrating underpinning services delivered by, for example, EGI or EUDAT, or other suitable infrastructure providers.

Conclusion

The first use case was applied with the use case partner OpenAire. OpenAire is a metadata repository service and provides search and access to a variety of resources. The following steps were investigated:

- Harvest a collection of data objects including meta-data and supplementary data consisting of PDF documents via OAI-PMH.
- Focus on OAIS compliant metadata packaging.

A collection of 214 documents and their metadata was harvested. The records are then packaged into an OAIS compliant package for long-term preservation. The packages were tested within our storage system and can be searched and downloaded.

Following this experiment we will investigate with other use case partners more complex supplementary data objects like 3D visualisations, which need a different ingest processing than ordinary PDF documents

National experiments

During the project lifetime, some national side-line projects were undertaken, that ran along or affiliated with experiments coordinated through DCH-RP.

Identity federation experiment: to ease uptake and support locally at ICCU for conducting experiment 4, a side-project was conducted to establish and configure an Identity provider (IdP) service at ICCU with the help of INFN Catania and GARR. This IdP service provides authentication services so that institutional users can keep using their institutional authentication credentials for using remote services.

Specifically, this would allow curators working at ICCU to upload digital assets (i.e., data and metadata) through the eCSG using their institutional credentials. Authentication and authorization are then decoupled: the former is done by the user's organisation (ICCU in this specific case), while the latter is done by the Service Provider (the eCSG in this specific case).

Conclusion

Using “IdP in the cloud”, GARR provided an IdP as a service to ICCU, populated it with a subset of ICCU staff information, and linked to ICCU’s backend credential management system. Also, this IdP been registered in the Italian Identity Federation IDEM (www.idem.garr.it), which is also managed and operated by GARR. Through IDEM this ICCU IdP has also been registered in the eduGAIN (www.edugain.org) inter-identity provider federation.

Experimenting with national e-Infrastructures: this experiment involved Polish DCH institution: Silesian Digital Library (http://www.sbc.org.pl) and Polish e-Infrastructure services namely Archival Services of the PLATON - Science Services Platform (http://storage.pionier.net.pl/en). Silesian Digital Library (SDL) is the second largest regional digital library in Poland (100 000 items). The content creators include public libraries, academic and educational institutions, cultural institutions, publishers and archives, museums and Protestant commune. The assets include institutional collections: regional heritage, rare materials, educational materials, scientific and research publications, doctoral theses, periodicals and special collections as well as private collections. While the capability of the SDL infrastructure addresses today’s needs, it is predicted than in several years, the volume of the digitized content exceeds current capacity of the SDL infrastructure. The level of the data protection on the physical level must be improved in future, in order to preserve data even from local disasters. Collecting the data from distributed locations is still partially manual. While most institutions upload the data to SDL servers already, quite a few contributors still provides the data by sending the storage media (disks, DVDs, Blue-Rays) using a surface mail or courier service.

Conclusion

Experiment involving Silesian Digital Library and Archival Services of the PLATON project in Poland prove that proper application of the e-Infrastructure services to implementing digital preservation processes may be effective and have limited negative impact on the user experience. Usage of data replication functionality of the Archival Services enabled improving the data durability, safety and availability. Virtually unlimited storage capacity of PLATON infrastructure enabled extending the storage space available to SDL and its partnering institutions. NDS2 tools enabled users to keep their methods and habits related to storing and accessing data, while performing data acquisition, processing and preparation for archival. Performance offered by the remote storage system was acceptable for the use case.

Important observations related to the role of the Silesian Digital Library in the e-Infrastructure services take up. SDL is the example of the very open, collaborating however demanding community side-partner of the e-Infrastructure providers. Thanks to its technical competence as well as awareness of the opportunities brought by the e-Infrastructures it is an early adopter of the services provided by e-Infrastructures. It also helps defining high-level and real-life requirements of the solutions to be provided in order to address DP processes.

General recommendations

Recommendation 1: Tools designed for installation on end user IT equipment, and intended for installation by end users, should be *as easy as possible to install* – ideally by a single action. It should be as easy as copying a number of files into one directory, followed by double-clicking an icon. Exemplar applications are the Eclipse Foundation's IDE "drops", or Firefox releases that literally require little more than copying a number of files into a directory of choice, or on a platform level, the Mac OS X application installation process comprising of one simple dragging the application icon to drop it over the system's Applications folder.

Recommendation 2: Tools integrating with typical Linux package management systems such as apt-get for Debian based Linux distributions or yum for Red-Hat based systems must provide an appropriate package for all supported hardware architectures (32bit and 64bit), including a well-defined and well-managed dependency manifest, so that, after downloading the package, a single command to install that package automatically installs any missing dependency without further unnecessary interaction.

Recommendation 3: Ideally, tools identified as suitable for inclusion in the DP roadmap should have active maintainers for the used/desired target platforms who ensure that recommendations 1 and 2 are adequately met, so that installing an application, tool, or service requires little more than issuing a command similar to "`sudo apt-get install scape-matchbox`".

Recommendation 4: If some software does not entirely match DCH requirements, investigate whether it has a modular design, preferably including well-documented extension interfaces (c.f. "plug-in" and "connector" design), for which DCH-specific extensions might be developed at greatly reduced cost. Aim to find partners and communities in the same market segment that might join in the maintenance effort for either the entire tool, or specific plugins.

Recommendation 5: Aim to avoid vendor lock-in by developing a service-oriented architecture for the DCH digital preservation landscape (or a desired "Preservation-as-a-Service" platform) including strategically placed and mandated publicly defined standards governing the interfaces between the services within the platform. Aim to avoid or reduce to an absolute minimum second-level dependencies such as one service directly depending on one or more specific instances of other services – operational maintenance and reliable rollout is next to impossible in an entangled network of dependencies. Ideally, an SOA with the right abstraction level and service scoping allows upgrading one service entirely independently from any other service.

Recommendation 6: Before defining the technical architecture of the preservation services and platforms, define and agree on the business process(es) you wish to implement in the technical architecture. Good business process modelling results not only in a business process architecture satisfying the requirements, but allows changes in its orchestration and sequences without redefining or altering the defined activities.

Recommendation 7: In the process of further developing the roadmap, describe each service that is required, and which capabilities it is expected to implement. For example, describing a storage service the roadmap might attach the following capabilities to it:

- Bit-level preservation of each digital object stored in and managed through this service;
- Data access and modification policies: Read-only, copy-on-write, transactional, or version-controlled;
- Self-service configuration of object replicas
- Self-service configuration of geospatial distribution of replicas
- Central or distributed data access points
- Transparent storage medium obsolescence management

These recommendations are arguably very technical in their nature. However, describing the overall results from both Proof of Concepts phases in the project one observation is key: Those experiments that were conducted with help and support of technical domain experts (Experiment 3, 4, 5) thrived well, installation was done smoothly and provided more crisp results. This is not to negate the value of the other experiments. The point is to reinforce the observation made already in the first Proofs of Concept phase in the project: *CH users are neither IT experts (or savvy with IT management and operation) – nor are they supposed to be.*

When engaging with e-Infrastructures, user communities and especially the CH community needs to be aware of different mandates hence different objectives of e-Infrastructures and customers, which will inevitably result in a gap analysis of "services needed vs services provided". While it is clear that e-Infrastructure are supposed to support research and scientific communities in Europe it is neither clearly stated nor mandated *how exactly* this has to happen. While e-Infrastructures such as PRACE, EUDAT have a clear mission and mandate bestowed upon them by their members, these are targeting specific communities in Europe hence able to provide more focused services towards these communities. EGI, on the other hand, has a clear mission to scale out its support from High-energy Physics towards essentially *any* research community in Europe.

Regardless, there is a clear gap emerging from the experiments conducted in the entire project, which we wish to convey as the last two but not least recommendation:

Recommendation 8: The DCH community relies very heavily on appropriate ICT support geared towards real end users. This is an observation, not a judgment, which needs to be appropriately taken into account. When engaging with e-Infrastructures, a third stakeholder must be considered for inclusion: The first stakeholder is clearly the DCH community as the consumer of any ICT services provided to them. The second stakeholders are the e-Infrastructures in Europe (and potentially worldwide) that provide a certain set of underpinning ICT services. The third, possibly new, stakeholders are service integrators and platform providers offering services tailored to the DCH community. The business relationships and value chain up to the memory institutes most likely will look like this:

1. Service consumer – Memory institutes, digital libraries, etc.
2. Service provider – ICT experts who are domain experts in the CH field
3. E-Infrastructure suppliers – Providing general-purpose infrastructure services on-demand and at scale to service providers.

Recommendation 9: Regardless of who is taking up the task of doing so, the strategy of sketching, developing, refining and eventually executing a strategy of providing a preservation as a service Cloud platform to the DCH community, the involved stakeholders need to be very clear in who their target audience is, and which institutes among these are suitable for early adoption and serve as multipliers into the "market" of DCH and DP.