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 D3.4 Intermediate version of the Roadmap

Revision: final

Authors:

Borje Justrell (RA)
Lajos Balint (NIIFI)
Eva Toller (RA)
Raivo Ruusalepp (EVKM)

Reviewers:

Maciej Brzezniak (PSNC)
Maurizio Messina (ICCU/Venice National Library)
Tim Devenport (EDItEUR)

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	10/10/2013	Borje Justrell	RA	Initial structure (outline)
0.2	31/10/2013	Claudio Prandoni	PROMOTER	Formatting and editing the outline
0.3	17/11/20213	Borje Justrell	RA	Reorganised the outline text into a deliverable text
0.4	13/12/2013	Borje Justrell	RA	Reorganised the text in all sections except in section 1
0.5	17/12/2013	Borje Justrell Eva Toller	RA RA	Added text in sections 2.2, 3.1, and 5.2 and made language polishing in all sections.
0.6	26/12/2013	Borje Justrell Tim Davenport Sara Di Giorgio Lajos Balint Maciej Brzeźniak	RA Editeur ICCU NIIFI	Added text in almost all sections
0.7	02/01/2014	Borje Justrell	RA	Added text and/or editing in Glossary, Abbreviations and sections 2.1, 2.2, 3.2, 4.1, 4.2, 4.3.3, 5.2, 5.3.2,5.3.3, 5.3.4, 6, and 7; new annex 5 added
0.8	04/01/2014	Borje Justrell	RA	Added text and/ or editing in Glossary, Abbreviations and sections 1, 2.1, 2.2, 3.1, 3.2, 3.3, 4.1, 4.2.2, 4.2.3, 4.3.1, 4.3.3, 5.1.1, 5.2.1, 5.2.2, 5.2.3, 5.2.4, 5.3.1, 5.3.2, 5.3.3, 5.3.5, 6, 7 and in annexes 1 – 4
0.9	06/01/2014	Borje Justrell, Rolf Källman, Raivo Ruusalepp	RA RA EVKM	Made adjustment in language, add new sections 6 and 7, and taken down the number of annexes from 5 to 3.
1.0	07/01/2014	Borje Justrell	RA	Pre-final version.
1.1	09/01/2014	Borje Justrell	RA	Integrate last comments from reviewers.
1.2	10/01/2014	Claudio Prandoni Antonella Fresa	PROMOTER PROMOTER	Formal check and final adjustments.

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This deliverable contains original unpublished work except where clearly indicated otherwise.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

> ''The last decade and a half has produced more records than any previous similar period of human activity. The fact, that the majority of these records is less reliable, retrievable or accessible than ever before, is one of the ironies of the modern information age''
>
> Prof Luciana Duranti, University of Brittish Columbia (1999)

D3.4 GLOSSARY

Specific terms and the definitions used in this deliverable:

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, and can 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 which is required to manage, search, understand, use, and preserve it.

Mushup - 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 Instituion
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
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
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
TB	TeraBytes
VPN	Virtual Private Network
VRC	Virtual research Community
VRE	Virtual Research Environment
VRO	Virtual Research Organization

1. EXECUTIVE SUMMARY

This deliverable presents an intermediate version of a roadmap for digital preservation that the DCH-RP project is tasked to design. The main objective is that the roadmap shall be an evolution of the deliverable D3.1 (Study on a Roadmap for Preservation), providing a first description of what the roadmap of preservation will look like. It is a work in progress, which targets primarily two main communities: cultural heritage institutions and e-Infrastructures that already include digital archiving functions in their preservation programmes. The final version of the roadmap will be submitted in a future deliverable (D3.5) by the end of the DCH-RP project.

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 describes a working model for the implementation of distributed digital preservation services for the DCH community, including an action plan for (a) concrete steps to take and (b) which services to address. It also gives a condensed version of the intermediate roadmap with focus on what to do and when in a short term perspective.

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

A total of three annexes are also included:

- An outline of a trust model suitable for the use of e-Infrastructures;
- Analyses of IaaS and future DCH preservation opportunities;
- Examples of the current use of distributed digital preservation services by the cultural heritage community.

2. INTRODUCTION

2.1 STRUCTURE OF THE DOCUMENT

This deliverable is an intermediate version of the roadmap for digital preservation that the DCH-RP project will design, in other words a work in progress. The intermediate version targets primarily two main communities: cultural heritage institutions and e-Infrastructures that already include digital archiving functions in their preservation programmes. It may also serve as a consultation document for other stakeholders that would like to give feedback.

The final version of the DCH- RP digital preservation roadmap, to be developed over the course of the project and reported in a coming deliverable (D3.5), will mainly target policy-makers on different levels and owners of digital preservation programmes at cultural heritage institutions.

This deliverable is organised as follows:

Section 2 - sets out the structure of the document and the objectives of the deliverable;

Section 3 - offers an overview of the general context for the deliverable;

Section 4 - describes the intermediate roadmap, starting with a working model for the implementation of distributed digital preservation which also serves as a framework for this deliverable;

Section 5 - presents a proposed action plan focusing on concrete steps to take and services to address, when using distributed digital preservation services;

Section 6 - gives a condensed version of the road map short-term with a focus on what to do and when;

Section 7 - summarizes on a general level the results in previous sections;

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

Annex 2 - analyses IaaS and future DCH preservation opportunities;

Annex 3 - contain some examples of the current use of distributed digital preservation services in the project partners home countries.

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 should be coherent and realistic in order to help policy makers and programme owners to plan ahead and also assist managerial teams of cultural heritage institutions in taking decisions related to digital preservation. The design of the roadmap will be 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 annaual budgets for digitisation at European cultural heritage institutions was 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.

The main objective of this deliverable (D3.4) is to be an evolution of deliverable D3.1 and to provide a first description of what the roadmap of preservation will look like. It will take into account the feedback of all other activities in the DCH-RP project during the first year of its life-time.

The main input has, so far, been provided by DCH-RPs work packages:

- WP3 (Preservation Roadmap) that, besides the above mentioned deliverable D3.1 (Study on Roadmap for Preservation), has produced the deliverables D3.2 (Standards and Interoperability

[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

Best Practices Report) and D3.3 (Registry of Services), and also an analysis of IaaS and future DCH preservation opportunities, presented in Annex 2.

- WP5 (Proofs of Concept) that has conducted the first proofs of concept using the SCRUM methodology and with WP3 acting in the role of the product owner. The main results are reported in deliverable D5.3 (Report on the First Proof of Concept).
- WP4 (Case Studies and Best Practice) that in a coming deliverable D4.1 (Trust Building Report) will report on trust and trust building, which has been identified as a key issue for the DCH-RP roadmap. Deliverable D4.1 will be submitted after the present deliverable (3.4), but WP4 has already produced an outline of that discusses some preliminary thoughts and also documents the results of a WP4 survey of access and authentication services from e-Infrastructures to support trustable services of memory institutions. Most of the text in this outline has been used in Annex 1.

3. SETTING THE SCENE

3.1 PRESERVING DIGITAL OBJECTS: 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 (see section 5.1.1)

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:

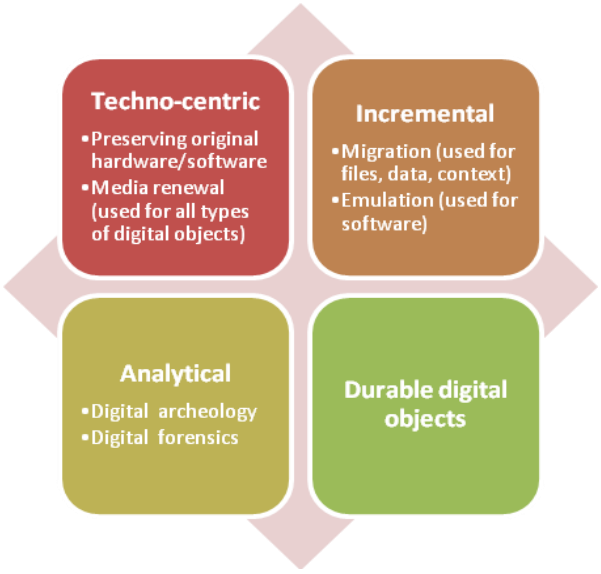

Figure 1: Strategies for sustaining the use of digital objects

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.

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

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

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 foreseable 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 which strategy or combination of strategies that is chosen, 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.

3.2 THE DIGITAL ARCHIVE LIFE-CYCLE

The diversity of 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.

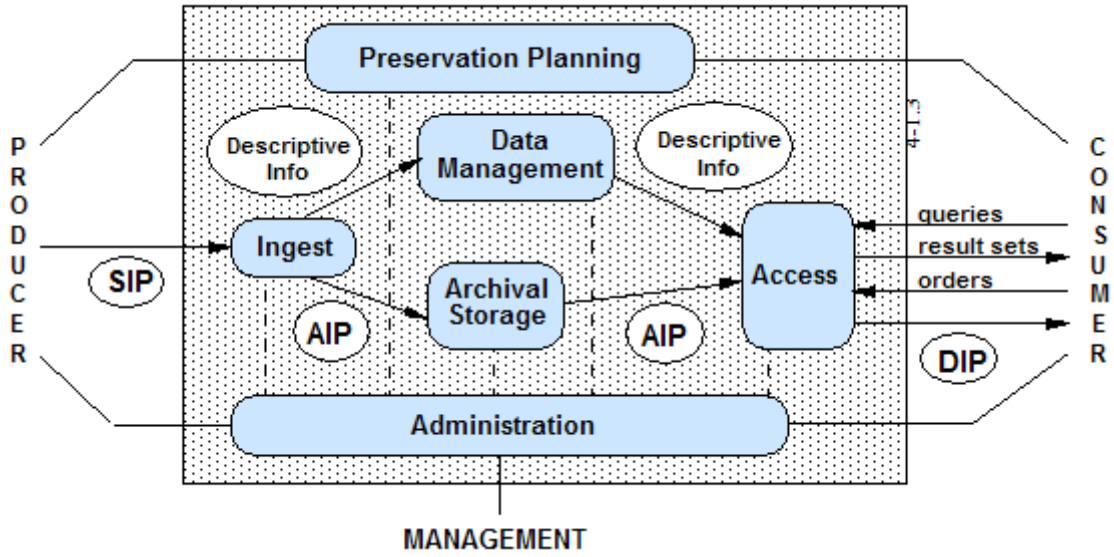

Figure 2: The OAIS functional model

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,[8] Tessella,[9] ExLibris[10] 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.

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

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.

3.3 DIGITAL PRESERVATION AND ROADMAPS IN A 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 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.[11] Also important is the EC Recommendation on digitisation and online accessibility of cultural material and digital preservation[12] 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 Commission´s expert workshop The Future of the Past, held in Luxembourg in May 2011.[13] 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. One of the speakers, Dr Ross King, Senior Scientist at the Austrian Institute of Technology (AIT), foresaw that in the future, issues around security and trust would increase. He also proposed a number of topics worthy of further research work, 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)[14] which presents environmental, policy, preservation, presentation, standards, skills, and communication developments over time. The Open Planets Foundation developed a *Tools and Services Roadmap*[15] to outline their software development plans. The APARSEN project roadmap[16] presents research topics and larger themes; preservation services are a research topic under the theme of sustainability. Some

[11] The New Renaissance, 2011: 6

[12] 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

[13] 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

[14] 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.

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

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

projects use roadmaps to present various formats, e.g. the PrestoSpace[17] project 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.[18]

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

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

[18] 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/

4. AN INTERMEDIATE ROADMAP FOR DIGITAL PRESERVATION

4.1 A WORKING MODEL FOR THE ROADMAP

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.

When bringing together two different worlds – centralised in-house digital preservation and 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 will have a strong focus on what to do. The first review of the DC-RP project the need to focus on the usability of services and technologies and on working solutions in the roadmap was underlined.

The conclusion is that the DCH-RP roadmap needs to be both multidimensional and to contain several layers. In order to achieve this, the project has adapted the following working model consisting of:

Firstly, the **roadmap** itself with its main components and detailed descriptions of each different part;

Secondly, an **action plan** with challenges and advantages to target, practical actions to take up, and services to address;

Thirdly, a **condensed version** of the intermediate roadmap focusing on what to do and when to do it**.**

ROADMAP Main components	Vision	Major areas to concentrate on	Timeframe	Infrastructure model for distributed digital preservation	
Detailed descriptions	Formulating a vision	Connecting the major areas of the road map to a time-line Short term (2014) Medium term (2016) Long term (2018)		Service architecture	Infrastructure framework
ACTION PLAN	Challenges and advantages to target on	Actions to take		Services to address	*To be specified in the final version of the roadmap*
CONDENSED ROADMAP	What to do and when - short term				

Figure 3: Working model for the DCH-RP roadmap

4.2 THE MAIN COMPONENTS OF THE ROADMAP

4.2.1 The 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. The main reason is that there is no commonly agreed vision of distributed digital preservation architecture relying on e-Infrastructures. For the DCH-RP project such a vision is an important piece in the puzzle and, therefore, urgently needed, which is underlined both in deliverable D3.1 (Study on road map for preservation) and in deliverable D5.3 (Report on the first proofs of concept). The latter says that "… *the DCH roadmap expressed in D3.4* ***must*** *improve on the community's vision, and give a clear guidance towards the aspired architecture of a DCH e-Infrastructure."*

4.2.2 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 should concentrate on at least 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[19]: 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;

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 section 5.2).

4.2.3 The 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.

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, 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.

For these later time spans we present only a brief description of the main actions points to be addressed. Further details will be elaborated in the course of the project.

[19] Called "Progress of inter-organisational communication" in Deliverable D3.1

4.2.4 The infrastructure model

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 a joint service architecture have been undertaken by only a few recent initiatives, and remain in a developmental phase. The DCH-RP project needs to advance the current state of the art and to develop its own model for an e-Infrastructure based preservation services architecture for the cultural heritage sector.

4.3 DESCRIPTION OF THE MAIN COMPONENTS

4.3.1 Formulating a vision

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.

4.3.2 Connecting the major areas of the roadmap to a timeline

Deliverable D3.1 proposed a matrix structure as a method for connecting the major areas of a roadmap to a chosen timeline. In figures 3 and 4 this has been done for DCH-RP: the first of these deals with relatively short-term priorities (defined here as those necessary through 2014) whilst the second focuses on medium-term (through 2016) and long-term priorities (through 2018 and beyond).

The matrices in figures 3 and 4 are also populated with action points that are described more in detail in the action plan (see section 5).

	Harmonisation of data storage and preservation	Improved interoperability	Establishment of conditions for cross-sector integration	Governance models for infrastructure integration
Short term (2014)	Test existing technical solutions in DCH environment: • Define an initial set of critical system requirements • Analyse the needs and conditions for infrastructure federation – e.g. NGIs, NRENs, EGI, EUDAT, CLARIN, DARIAH, DASISH, PLATON and commercial infrastructures • Summarise ongoing experience with grids and cloud solutions applied in cultural institutions • Identify examples of use of PaaS – and promote the benefits offered by virtualisation	Identify and promote best practices Analyse interoperability issues including the following aspects: • Technical • Semantic • Organisational and inter-community • Legal • Political/human • Cross-border	Analyse what impact do emerging and established standards have on grid and cloud preservation architectures Establish and update a registry of preservation tools and services Analyse which PaaS composition of services best matches digital preservation requirements Identify gaps in provision and establish a plan for medium- and long-term developments to address the gaps	Analyse major information governance patterns and windows of opportunities Explore the issues of trust-building through pilot systems Suggest possible business models for typical scenarios

Figure 4a: Major areas of the roadmap in the short-term (through 2014)

	Harmonisation of data storage and preservation	Improved interoperability	Establishment of conditions for cross-sector integration	Governance models for infrastructure integration
Medium term (2016)	Test technical solutions in DCH environment • Long-term storage, bit-level preservation • Multiple entry points • Operational benefits • VRE development • Support framework • Middleware services • Authentication and authorisation infrastructure Sharing of other services	Develop and test tools facilitating interoperability addressing the following aspects: • Technical • Semantic	Fill in gaps in provision [plan for medium-term work needs to be made in the end of the short-term stage]	Analyse needs for redesign of existing local (institutional) infrastructures Define a set of governance principles for digital preservation in DCH
Long term (2018 and beyond)	Consolidate mature requirements for preservation in the DCH environment	Implement tools in selected e-Infrastructures facilitating interoperability aspects: • Technical • Semantic	Fill in gaps in provision [plan for long-term work needs to be made in the end of the short-term stage]	Offer mature business model for preservation services for different types of institutional settings

Figure 4b: Major areas of the roadmap in the middle term (2016) and long-term (through 2014 and beyond)

4.3.3 An infrastructure model for distributed digital preservation

Data infrastructure framework

The EUDAT project 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 could also be said to be basic needs 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.

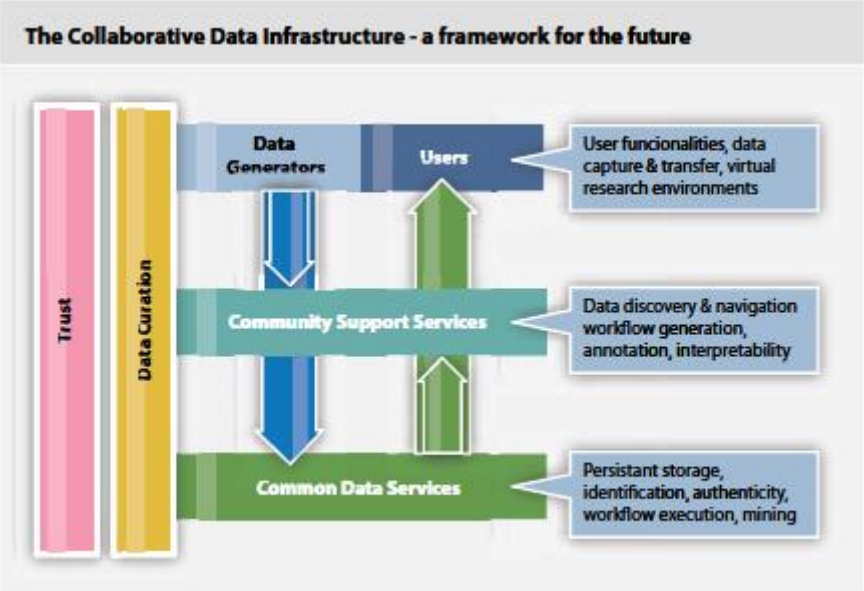

Figure 5: The collaborative data infrastructure - a framework for the future; from Riding the Wave, p. 31

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 from 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) [20]

The DCH-RP project aims to improve the EUDAT model in order to adapt it to distributed digital preservation. This will be done during the second half of the project.

[20] See www.dans.knaw.nl

Service architecture

Service architecture for distributed digital preservation has to address several areas of service. The basic approach is to address services according to their:

- Functional areas, following the OAIS model and/or the preservation process (pre-ingest, ingest, archival storage, preservation planning, data management, access and use);
- Their type (micro services, services) and objects addressed (files, bitstreams);
- Type of architecture (addressing a simple and well defined task, grid- and cloud oriented);
- Level of maturity;
- Licensing conditions.

In section 5.3 these service areas to address are turned into practical actions as a part of an action plan.

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 DCH-RP roadmap, aiming at the use of e-Infrastructure, SOA is regarded as a concept to get inspiration from rather than one to implement.

The DCH-RP project has, at its present stage, not yet defined a final vision of a service architecture to target. In the intermediate roadmap, therefore, several approaches are presented.

5. AN ACTION PLAN

5.1 CHALLENGES AND POTENTIAL ADVANTAGES

5.1.1 Challenges to meet

The challenges that cultural heritage institutions are facing today in maintaining accessibility and usability to digital resources over time are in many ways related to notable differences between digital and paper-based material.

The Digital Preservation Coalition has pointed out six such differences: [21]

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.

The issues above are all interconnected, and together they clearly indicate that a radically different approach is required in managing digital objects compared with paper-based materials, an approach in which action needs to be taken, and planned for, at regular intervals. 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.

By using existing e-Infrastructures for research and academia (including NREN, NGI and the newer data infrastructures) as a channel for the delivery of advanced services in digital preservation to cultural heritage institutions, these differences or challenges can be met. An important step before entering distributed digital preservation is, though, to decide which of them are to be targeted (all of them or just a few).

5.1.2 Potential advantages to develop

The major advantages for the DCH sector when using e-Infrastructures could include these:

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

- 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.).

Also when evaluating potential advantages, it is important for cultural institutions to have a clear understanding of what to exploit, before taking a decision about the use of distributed digital preservation services.

5.2 ACTIONS TO TAKE

5.2.1 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 wether the services are targeted on research data or other digital objects.

Tests of existing technical solutions in a DCH environment are being carried out by the DCH-RP project with the aims of

- Defining an initial set of critical system requirements;
- Analysing the needs and conditions for infrastructure federation;
- Summarising ongoing experience with grids and cloud solutions applied in cultural institutions;
- Identifying examples of use of Preservation as a Service (Praas) and promote the benefits offered by virtualisation.

The results achieved so far are as follows. Some of them are outcomes of the First Proofs of Concepts, some of other deliverables and studies in the DCH-RP project (see section 2.2 above).

Define an initial set of critical system requirements

1. *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

2. *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. During the DCH-RP project, tests are being made in two rounds of Proofs of Concept (see deliverable D.5.1 *Technical Plan for DCH-RP proofs of concept* and D5.3 *Report on First Proof of Concept*).

In section 5.3.1 below some functional requirements are specified.

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.

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 user´s 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, aiming at storage and preservation of cultural heritage data distributed all over Europe, which is in fact a wider concept including also federated storage systems, namespaces, metadata hubs etc. WP4 has a focus on federated issues and how they can benefit users of memory institutions. During the coming year WP4 will analyse the various organisations offering services and/or contents in the context of the project and produce a set of recommendations for user authentication and access control system(s) that would be most suited for the cultural heritage institutions. 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.

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 project´s Proofs of Concept.

Ongoing experience with grids and cloud solutions applied in cultural 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 section 5.3.2 they are refered to as the "kiosk" model and the "turn-key" model respectively. What in the same section is called "microservices" could also be a fruitful approach to look into. However, if various microservices 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.

If we review 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 dissatisfation 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. In section 5.3.3 below is made an attempt is made to change the perspective and look upon the service architecture from the e-Infrastructures point of view.

In Annex 2, the DCH preservation process has been analysed and the impact of possible forms of Infrastructure as a Service (IaaS) evaluated. Integrating the IaaS aspects into the roadmap will follow in the second phase of the project. The findings will appear in the final version of the roadmap (deliverable D3.5), including the outcome of studying other versions of cloud offerings such as PaaS and SaaS.

Examples of use of preservation as a service 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".[22]

In Annex 3 some examples are presented taken from partner countries in the DCH-RP project, where cultural heritage institutions are using distributed digital preservation services.

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.[23] 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.[24]

MEDIUM TERM PRIORITIES

Recommendations, best practice and lessons learned from tests of existing technical solutions - executed during the phase of short-term priorities - have to be transformed into solid technical solutions aimed at the DCH environment. These solutions must, then, to be tested more specifically addressing aspects like these:

- Long-term storage (bit-level preservation)
- Multiple entry points
- Operational benefits

---

[22]. *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)

[23] See www.dans.knaw.nl

[24] Preservica Preservation as a Service (http://www.preservica.com)

- VRE development
- Support framework
- Middleware services
- Authentication and authorisation infrastructure

However, Ii is not considered to be within the scope of the DCH-RP project to conduct these tests.

LONG TERM PRIORITIES

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

5.2.2 Improve interoperability

SHORT-TERM PRIORITIES

Identify and promote best practices

In the final version of the roadmap (deliverable D3.5) the aim is to have a section 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 will be an important instrument for capturing best practices (some useful information is already in place as outcomes from the first one). The discussion forum and the events organised by the project will also be used as channels to fulfil this task.

WP4 will support the validation of the results of WP3 through a range of tasks interacting with the DCH community.

Analyse interoperability issues

Fostering better integration of digital preservation within the overall internal workflows for digitisation and online access in cultural heritage institutions is very much a matter of providing best practices, role models etc. 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 organisation´s 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;

- Developing its own vocabulary (not recommended though sometimes unavoidable);
- Use an "uncontrolled" vocabulary - i.e. keywords entered by the organisation´s 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.

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.

LONG TERM PRIORITIES

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

5.2.3 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 available on 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 (see also section 5.2.2 above)

The conclusion is that much work has already been done, but that more efforts are still needed before these standards, guides, and tools, etc can be of essential help for 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-infrastructures 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. This will be an issue for the final version of the road map (Deliverable D3.5). One approach can be to recommend standards aiming at building up the actual trust it requires for cultural heritage institutions to relinquish immediate control of DCH information by using digital preservation services outside their own institutions.

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.

Analyse which PaaS composition of services best matches digital preservation requirements

One planned outcome of the Proofs of Concept conducted by the DCH-RP project is a suitable mixture of distributed services that matches the DCH sector´s preservation requirements. The results will be reported in the final version of the roadmap.

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. An outline how it can be done will be presented in the final roadmap.

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.

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.

5.2.4 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:

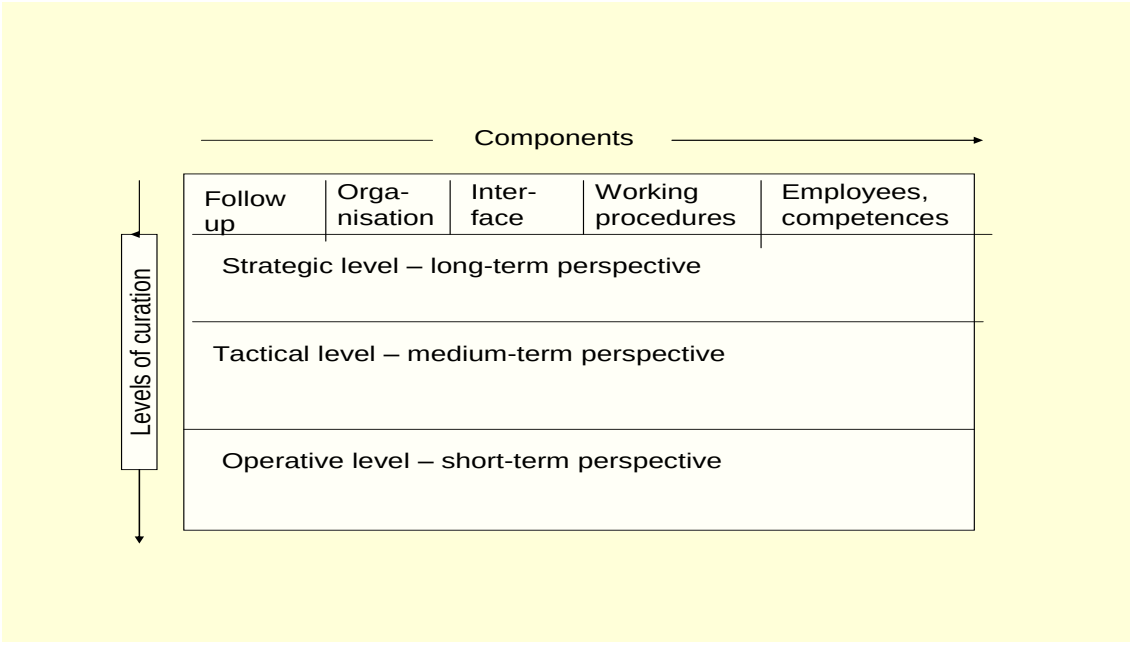

Figure 6: A framework for the governance of distributed digital preservation services

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.[25]

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[26].

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

[25] See DCH-RP deliverable D3.1

[26] See http://www.crl.edu/archiving-preservation/digital-archives/certification-and-assessment-digital-repositories

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 section 5.3.2 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.[27] 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 will outline 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 will be reported in the final roadmap (Deliverable D3.5). In Annex 1 is attached a text based on an outline to deliverable D4.1.

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 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.

Digital preservation built on a distributed model needs a business model 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.[28]

[27] See a guide: http://www.metaarchive.org/GDDP

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

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.

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.

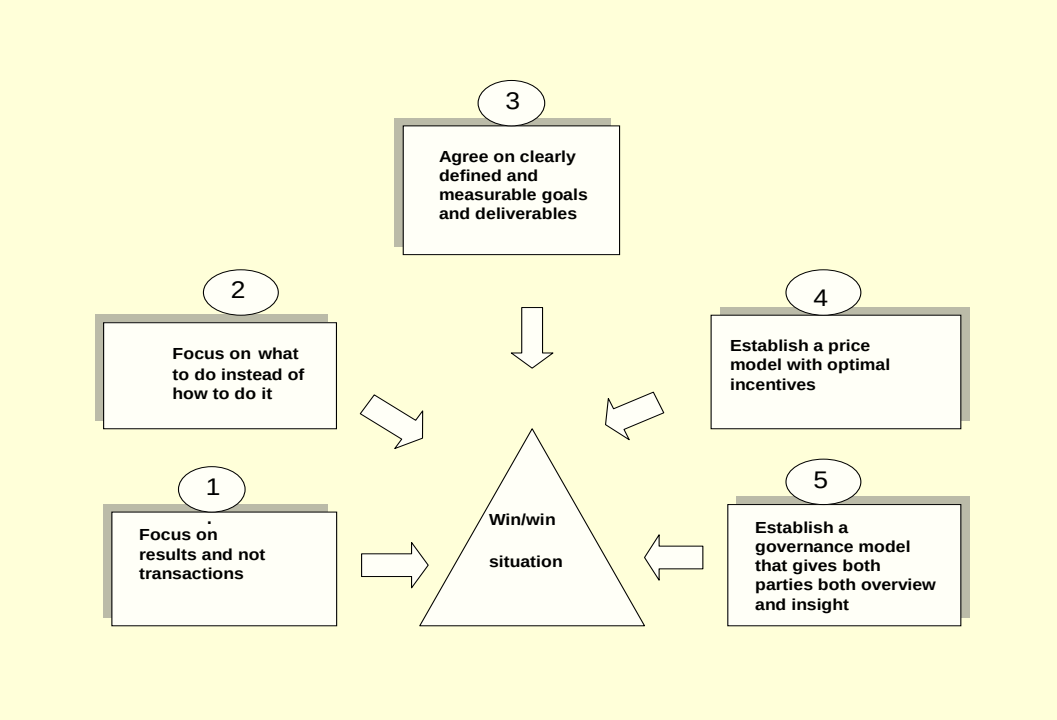

Figure 7: The Vested model

MEDIUM TERM PRIORITIES

In this stage the highest priority actions to take are:

- To do solid analyses of needs for redesign of the cultural heritage institution´s existing internal infrastructure, in able to get it effectively integrated with distributed digital preservation services;
- To define a set of governance principles for digital preservation in DCH aiming at 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).

5.3 SERVICES TO ADDRESS

5.3.1 Functional areas

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 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.

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.

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 actally 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.

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.

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.

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

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.

5.3.2 Service types and objects to be addressed

Service types to be addressed

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

Firstly, 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, that is typical network services that would make work easier for institutions or networks of institutions that manage digital preservation "on their own".

Secondly, 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 a 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 an DCH institution. “Microservices” for digital preservation are currently used in the open archival information system Archivematica.[29]

Objects to be addressed

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 context 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:

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

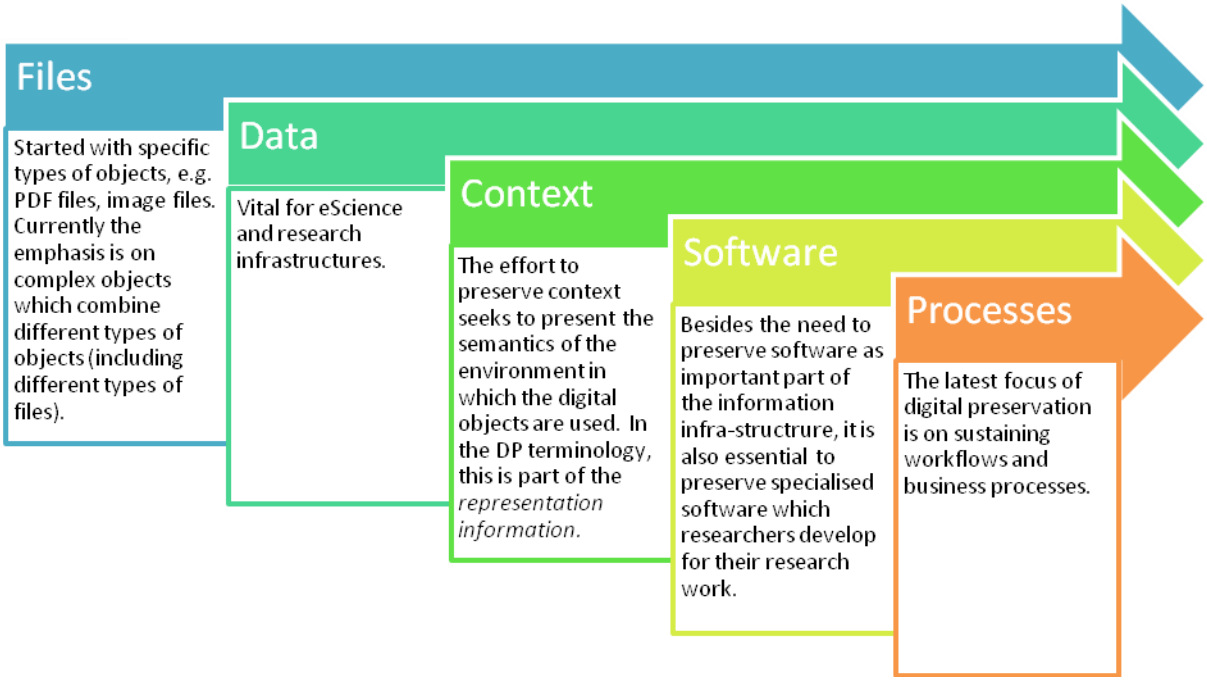

Figure 8: Evolution of digital objects addressed by digital preservation

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.

5.3.3 Type of architecture

An important issue is whether the service architecture will address a simple and well defined task, grid- and cloud oriented, or more complex service patterns. Most sections of this report are addressing the nature of preservation and major functional entities that preservation systems should support, but it could be useful to analyse preservation also from an enterprise perspective.

Since preservation is part of digital objects' lifecycle, it has implications for the processes and professionals within the institution. 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.[30]

[30] 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

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 enterprise architectures; one of the popular ones is the Open Group Architectural Framework (TOGAF)[31] and its eight-stage Architecture Development Method that help to manage requirements within complex systems.

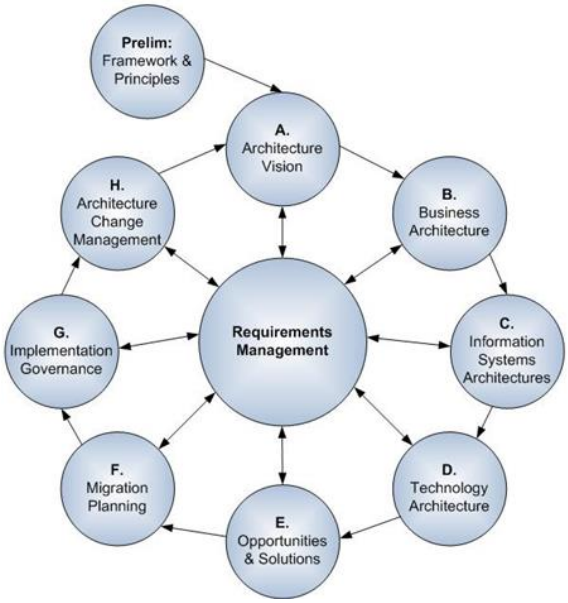

Figure 9: Architecture Development Method, TOGAF.

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.[32] 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.[33]

5.3.4 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.[34]

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.

[31] http://www.opengroup.org/togaf/
[32] Zachman, J. *Concise Definition of The Zachman Framework*. http://zachman.com/about-the-zachman-framework
[33] 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)
[34] Preservica – white paper (July 2013) http://preservica.com/resource/praesent-ante-stiam-white-paper/

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 fulfil 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 fulfil all 6 layers.[35]

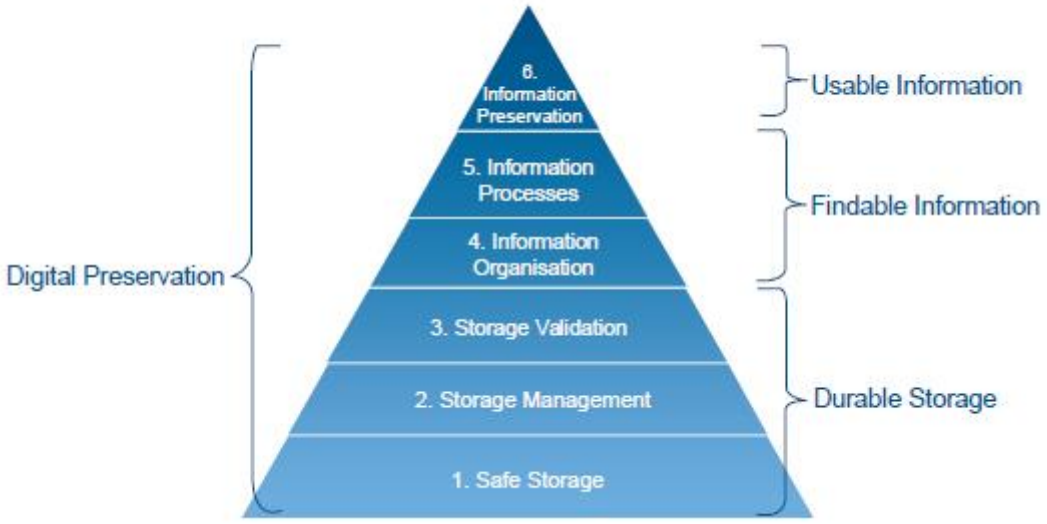

Figure 10: The Digital Archiving Maturity Model

5.3.5 Licensing conditions

Standard licenses and methods of license expression

It is becoming increasingly important to understand and communicate the license 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 reported seven overall license types relevant here and broke these out further, for example describing at least four variants of the Creative Commons (CC) licenses in routine use.

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

The following table briefly summarises the licenses mentioned.[36] 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/
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/

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

6. CONDENSED VERSION OF THE INTERMEDIATE ROADMAP – SHORT-TERM

In this section is presented a condensed version of the intermediate roadmap, but only in a short-term perspective. A stated in section 4.2.3 above, medium-term and long-term actions will not be presented in details. During each step references are made to relevant parts in sections 4 and 5 above.

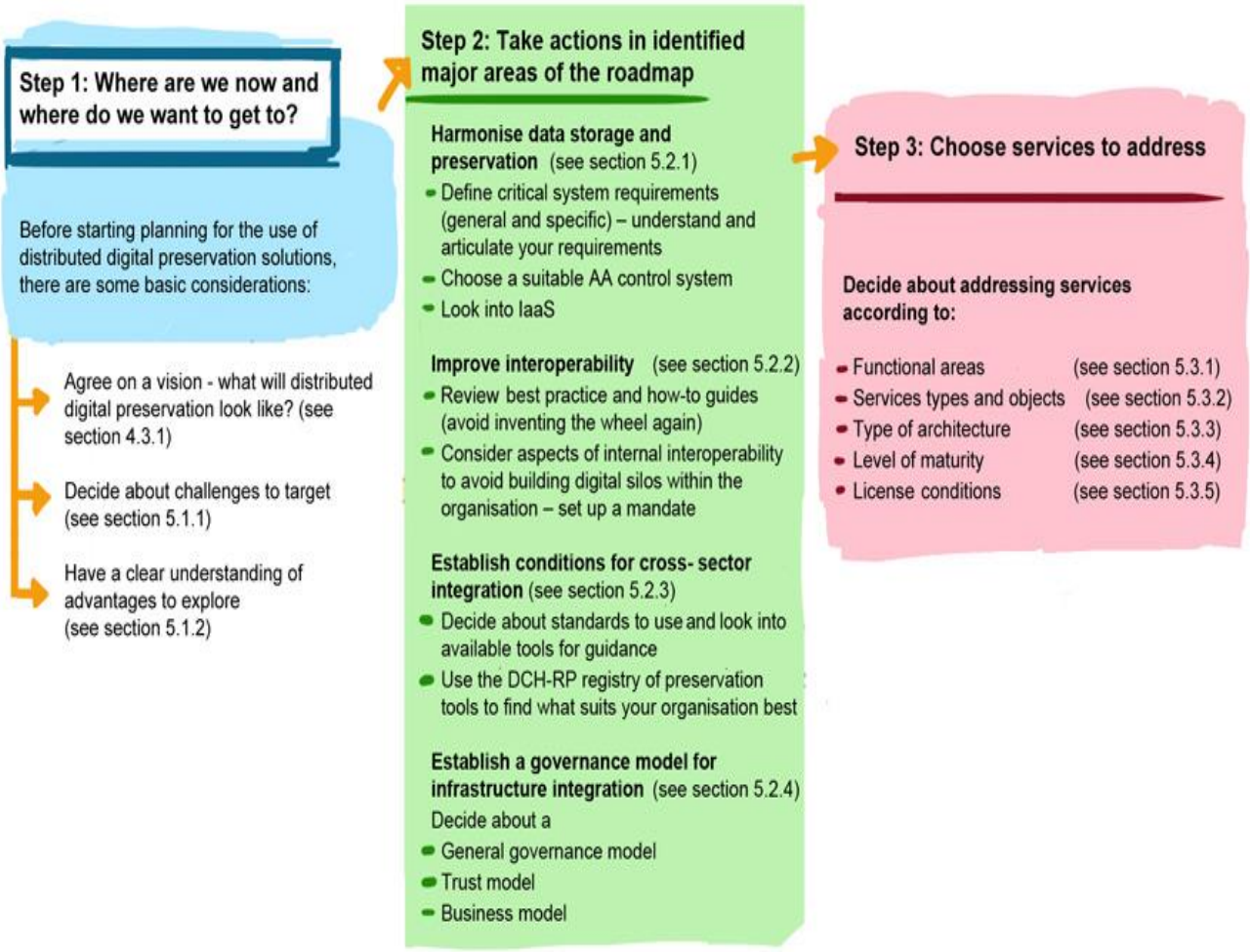

Figure 11:The Condensed version of the intermediate roadmap – short-term

7. 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. 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.

When summarising the work on the DCH-RP projects road map, so far, the use of e-Infrastructure in meeting these demands looks promising. The two basic assumptions that the DCH-RP roadmap is built on are achievable:

- existing e-Infrastructures for research and academia are efficient channels also for digital cultural heritage sector to be used for distributed digital preservation
- it is possible to establish common policies, processes and protocols to 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.

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.

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. To find ways to bridge this gap is a main challenge for the final version of the roadmap. The DCH-RP projects initial aim was have a practical approach with a strong focus on what to do, and it 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

Proteus is far from being caught, he is just being pushed a little bit into a corner!

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

This annex presents an outline - produced by WP4 - of the coming deliverable D4.1 (Trust Building Report)

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. To begin answering questions on trustworthiness of digital preservation repositories a number of approaches have been proposed that rely on different methods of audit.

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” (CPA/RLG 1996).

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 (RLG 2002). 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. This report provided a comprehensive look at the organisational context for a digital preservation program and made a direct call for the development of a digital certification program.

2. THE TRUSTED DIGITAL REPOSITORY AUDIT METHODS

The following year RLG and NARA established the joint Digital Repository Certification Task Force with membership from the U.S., U.K., France, and the Netherlands representing multiple domains including archives, libraries, research laboratories, and data centres from government, academic, non-profit, e-science, and professional organizations (Ambacher 2007, p. 3). The task force worked on developing an audit checklist that was released as a draft for public comment in August 2005 (RLG/NARA 2005) with the comment period extending through mid-January 2006.

The draft audit checklist aimed to develop criteria to “identify digital repositories capable of reliably storing, migrating, and providing access to digital collections” (RLG/NARA 2005). Certification of repositories was foreseen to instil confidence in data creators, resource allocators, and users that the repository – a certified repository – meets recognised standards and can fulfil its preservation and access provision mission.

The final version of the audit checklist was published in March 2007 as the Trustworthy Repository Audit and Certification (TRAC) Criteria and Checklist (TRAC 2007). The checklist presents almost 90 organisational, technological and digital object management criteria for digital repositories. Many are based heavily on the principles, terminology and functional characteristics outlined in the OAIS Reference Model (ISO 14271:2003).

In 2004 the German Network of Expertise in Long-term Storage of Digital Resources (nestor) established a working group on the certification of trustworthy archives.[37] Building on the RLG/NARA draft version of TRAC checklist, the nestor group focused on identifying features and values that are relevant for evaluating both existing and planned digital object repositories. The first version of the nestor criteria for auditing digital preservation repositories was released in 2006 (nestor, 2006) with an update in 2008 (nestor, 2008). This checklist covers the technical, organizational, and financial characteristics of a digital repository. It is structured similarly to the TRAC checklist, but additionally provides examples and perspectives that are of particular relevance to the legal and economic contexts and operational situation in Germany. On the conclusion of the nestor project, work on the trustworthiness criteria was transferred to the German national standards body7 and a new version of the criteria was published as a national standard DIN 31644:2010. Certification based on this standard is expected to commence in late 2013.

In early 2007 the DigitalPreservationEurope project (DPE) and the UK Digital Curation Centre (DCC) published their joint work as the Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) (Hofman et al. 2007). This tool presents a methodology for repository self-assessment and characterises digital curation as a risk-management activity; the job of digital curator is to rationalise the uncertainties and threats that inhibit efforts to maintain digital object authenticity and understandability, transforming them into manageable risks. An online assessment tool was released in 2008 to guide and document the repository assessment.[38]

The Data Archiving and Networked Services (DANS) in the Netherlands published 16 guidelines to help a data archiving institution striving to become a trusted digital repository in 2008. The guidelines are called the Data Seal of Approval (DSA 2009).[39] The audit is a two-stage process where a repository carries out its own assessment, publishes the results and then applies for an external audit that is carried out by a member of the international DSA assessment group on the basis of the available assessment document. The board determines whether the guidelines have been complied with and whether the DSA logo can be awarded to the data repository (Harmsen 2008, p. 1).

The Center for Research Libraries (CRL) – the maintenance agency of the original TRAC checklist – has established a Certification Advisory Panel and is conducting audits using the TRAC checklist. It awards certificates of a trustworthy digital repository based on successful audits.[40]

An international joint effort undertaken to develop a set of criteria on which full audit and certification of digital repositories can be based resulted in a 2011 ISO standard in support of the OAIS reference mode. The ISO 16363:2011 Audit and certification of trustworthy digital repositories is based on the previous TRAC checklist, but with more detailed specification of criteria by which digital repositories are to be audited. 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. (Ruusalepp et al. 2012, p. 124) A team of experts conducted a series of pilot audits in 2011 as part of the APARSEN project, to test the methodology promoted by the ISO 16363 standard. (see APARSEN 2012)

The same working group has also developed a standard with *Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories* (ISO 16919:2011). This standard is meant primarily for those setting up and managing organizations that perform the auditing and certification of digital repositories. The standard provides normative rules against which an organization providing audit and certification of digital repositories may be judged, and it describes the auditing process.

[38] http://www.repositoryaudit.eu/
[39] http://www.datasealofapproval.org/
[40] http://www.crl.edu/archiving-preservation/digital-archives/certification-and-assessment-digital-repositories

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).

It is important to note that the scope and level of detail of individual criteria of the DSA against the DIN and the ISO standards means they are not directly comparable (nor are they intended to be) though it is to be expected that as organisations progress through the European Framework it will become clear how evidence for specific DIN or ISO metrics maps to evidence for the 16 DSA guidelines. Providing guidance on these relationships in the future is expected to drive adoption of both the DIN and ISO standards through Extended and Formal certification. (APARSEN 2012, p. 11)

Conclusions

Ever since the publication of the RLG/OCLC *Trusted digital repository: attributes and responsibilities* report in 2002, the trust model of digital preservation function has become repository-centric, i.e. it applies (mostly) to a single organisation in charge of a digital repository system that it owns. 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.

3. TRUST IN DISTRIBUTED PRESERVATION SERVICES

Similar to the lack of a distributed digital preservation service models (see DCH-RP deliverable 3.1), no appropriate trust model yet exists for a distributed preservation repository system does not yet exist. The need for a trust model for distributed digital preservation solutions has been discussed through a number of research papers.

Berman et al. (2007) describe the Chronoplis cooperative as a virtual organisation (federation) that exhibits trust both from dispositional (the natural tendency of an individual to trust other people) and situational (dispositional trust combined with structural and situational factors) perspectives. They conclude that “in formalizing the trust relationships between preservation providers, partners, and users, many issues are left unresolved”.

Day (2008) discusses how trust comes to the fore in many areas of digital preservation where collaboration is necessary; this includes participation in strategic alliances and research initiatives, and in the provision of shared services like registries. He suggests that self-assessment tools like DRAMBORA

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

could be used to help develop shared organisational cultures that are focused on solving long-term preservation challenges in an incremental and managed way.

Walters and McDonald (2008) use the example of the US Federal Reserve Bank regional governance (trust federation) model as an exemplar for centralized authority while providing for distributed independent organizational governance.

Schultz and Gore (2010) stress that "distributed digital preservation solutions must communicate trust to their Designated Communities as they continue to mature". Applying the TRAC checklist to the MetaArchive Cooperative distributed digital preservation solution revealed that the current metrics for gauging trust in digital preservation could be readily applied to distributed solutions but because these metrics often presume a more centralized approach to preservation, there is a pressing need to "apply them carefully and with great thought".

The EU High Level Experts Group on Scientific Data advise in their final report (Riding the Wave 2010, p. 17) that "if science is to advance, […] questions of trust must be answered by the infrastructure, itself" because data-intensive science operates at a distance and in a distributed way, often among people who have never met, never spoken, and, sometimes, never communicated directly in any form whatsoever. They must share results, opinions and data, but in truth, they have no real way of knowing for sure if, on the other end of the line, they will find man or machine, collaborator or competitor, reliable partner or con-artist, careful archivist or data slob. How will we judge the reliability and authenticity of data that moves from a personal archive into a common scientific repository?

The need for a transitory trust model for distributed digital preservation solutions is, thus, accepted in the research literature, but as yet no working model has been proposed.

Some cooperative efforts have already emerged, mostly in the United States, that outsource or share some functions and services of a digital archive:

- LOCKSS[42] - the LOCKSS Program is an open-source, library-led digital preservation system built on the principle that "lots of copies keep stuff safe".
- Data Preservation Alliance for the Social Sciences (Data-PASS)[43] – is a partnership of five major U.S. institutions with a strong focus on archiving social science research.
- Chronopolis[44] - is a digital preservation data grid framework developed by the San Diego Supercomputer Center (SDSC). It provides cross-domain collection sharing for long-term preservation.
- MetaArchive Cooperative[45] - is a cooperative membership organization where each member runs a server for the MetaArchive network and prepares its own content for ingest.
- University of California Curation Center (UC3)[46] - is a creative partnership that services University of California campuses for their digital curation needs.

The trust model used by these cooperatives is usually described as "circular chain trust model", like the one used for example, by the LOCKSS network where all partners using the software also share a trust network.

[42] http://www.lockss.org/
[43] http://thedata.harvard.edu/dvn/dv/datapass/
[44] http://chronopolis.sdsc.edu/
[45] http://metaarchive.org/
[46] http://www.cdlib.org/uc3

4. 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.

First results of developing domain-specific repository risk profiles have recently started to appear (Ross et al. 2008; OCLC 2010) and have the potential of evolution into an ontology of repository attributes. As individual classes of repository are increasingly identified and described, their common services and characteristics can be understood and ultimately linked with objective measures of success.

At the same time, risk profiles for third-party services have emerged for cloud services and specifically for outsourcing digital archives services. Early groundwork on this was done by collaborative working groups like the European Union Agency for Network and Information Security (ENISA)[48] on risk management (ENISA 2009); the UK and Ireland Archives and Records Association (ARA) study on storing information in the cloud (ARA 2010a and 2010b); Cloud Sweden's recommendations on outsourcing preservation services to cloud providers (Cloud Sweden 2011); the US National Institute of Standards and Technology (NISO 2011). The results of these teams have led to certification frameworks like the Cloud Security Alliance Security, Trust & Assurance Registry (STAR),[49] and systematic studies of risks around outsourcing digital preservation services to the cloud (Aitken et al. 2012).

These works demonstrate that risk has proved itself as a useful and universally understood concept when communicating trustworthiness factors and that we are closer to a risk profile for distributed services architecture than a formal trust model.

5. PROPOSED WORK FOR THE COMING DELIVERABLE D4.1

Since it is conceptually sound and useful for the digital preservation community, it is proposed that D4.1 develop a risk profile or a risk-based trust model for distributed digital preservation architecture that relies on e-Infrastructures.

It will be tested with the e-Infrastructure partners in the project and associated with the project via proofs of concept in WP5.

It can be further verified with e-Infrastructure and DCH partners in the project who are already or are intending to use third party services.

The outcome of this exercise can be defined as an embryonic maturity model for memory institutions to assess:

- Whether they themselves are ready to outsource some of their services to a third party;
- The typical risks associated with third party services as a risk profile/registry or a checklist.

[47] http://www.repositoryaudit.eu/
[48] http://www.enisa.europa.eu/activities/risk-management
[49] https://cloudsecurityalliance.org/star/

A separate section will survey and discuss authentication services as components of a trust model. This is a separate exercise undertaken by the partner TERENA.

The deliverable report will also offer recommendations for the final roadmap (Deliverable D3.5) on the EC, trans-national and national level.

ANNEX 2 IAAS AND FUTURE DCH PRESERVATION OPPORTUNITIES

1. EVOLUTIONARY AND REVOLUTIONARY DEVELOPMENT – A BRIEF INTRODUCTION

This section is devoted to thinking ahead over a longer range when developing a roadmap for various development scenarios with respect to DCH preservation. IaaS ("Infrastructure as a Service") is an assumed correct answer to most of the questions arising today with respect to future e-Infrastructure developments, and this assumption about the perspectives is also mirrored by the title of this section.

Moreover, there is an intention here also to look beyond the directly foreseeable IaaS applications, by trying to forecast how the best exploitation of continuously (or sometimes abrupt) development of the e-Infrastructure resources (frameworks, facilities, tools) can be exploited by improved use of the principle and practice of virtualization. In other words, both evolutionary (tradition-based) and revolutionary (completely new) development are taken into account in our investigations.

Needless to say that the above intention assumes a wise combination of realistic extrapolation and imaginary forecasting – as is always a natural way of thinking about the future. This also means that elaborating on future scenarios certainly necessitates the thinking about an appropriate set of components for a really comprehensive and multifaceted roadmap, including appropriate details on the assumed philosophies, paradigms, architectures, technologies, services, environments, support, management, influence, etc. Indeed, in the case of IaaS, attention is to be paid to the questions of both the realization of the IaaS framework (technological aspects) and the provision of the service (organizational aspects), in accordance with the similarly double faced paradigm behind IaaS.

As far as the intended coverage of this annex 2 is concerned, in the followings we will go into details mostly with respect to the latter (organizational) side of the problem (i.e.the provision of the e-Infrastructure services) and will only touch on the technological side of the complex IaaS topic.

2. BACKGROUND

Advanced computer (networking) infrastructures allow users to access and share distributed HPC and storage systems. These solutions offer cost-effective shared exploitation of the available technology and services for practically all potential users. Accessibility is already especially excellent for the academic and research user community.

Utilization of the resulting advanced service portfolio available today allows the users to search for data in distributed sites, to provide AAI (user authentication and authorization), to grant agreed level of security, to enable real-time monitoring of activities, etc.

Thanks to the above progress, present day e-Infrastructures (frameworks, tools, services) are obviously well mirroring the considerable development achieved during the last several years both in ICT advancement and in the progress within Research Infrastructures. Moreover, present day e-Infrastructure features in many senses demonstrate the characteristics of IaaS mentioned above.

One of the reasons for this pleasing situation is that e-Infrastructure service providers do widely use cloud technology by exploiting the capabilities of cloud technique in enabling efficient, integrated, user-centric, friendly usability of distributed resources. This also means that practical experiences stemming from the application of the presently available e-Infrastructure services for solving DCH preservation tasks can provide useful information for an extrapolation-based preparation of a valid roadmap as an output of the DCH-RP project.

However, it should always be kept in mind that the (well approachable but never completely achievable) basic goal in developing and operating e-Infrastructures should be the provision of individually tailored (or custom compiled) service built on a wide set of standard (interoperable) building blocks, by exploiting a classification of individual demands and requirements on the basis of recognizing the similarities among them. Such a set of building blocks, together with the underlying (integrating) operational framework, if organized by IaaS principles, leads to capable overall solutions (sometimes called turnkey systems) built up to cover practically all regular tasks/phases of the related application area – in our case DCH preservation.

Practically all the above mentioned aspects are of overall validity but nevertheless, there are some specific aspects to be taken into consideration in the case of DCH preservation. Perhaps the most important ones are the capability of handling metadata information, enabling curation, and granting long term availability/accessibility to the preserved DCH content. These aspects are to be duly taken into consideration when thinking about roadmap scenarios for e-Infrastructure development for the application area of DCH preservation, together with some other (also in other e-Infrastructure applications indispensable) aspects, like identification (federated AAI), reliability (dependability, security) and trust (another aspect of outstanding importance in the case of DCH preservation).

A further important aspect is that although this annex is dealing with e-Infrastructure matters, a valid roadmap for DCH preservation should put similarly weighted emphasis on the capabilities, preparedness, and willingness of the DCH community (as users of the e-Infrastructure) to properly exploit the potential benefits stemming from professional use of the e-Infrastructure services. Such proficiency requires conscious efforts from both communities (DCH experts and e-Infrastructure developers/operators alike) to co-operate in solving their own challenges but also taking part in supporting each other's activities.

3. IAAS BASICS

The popularity of service models (SaaS, PaaS, IaaS) is rapidly increasing and Infrastructure as a Service is the probably the best known, most widely applied, and most extensivly exploitated of them all. Here the components of the construct providing services are of high level, high complexity and capability eg. data transmission, computing, or storage equipment. Depending on the role, function or purpose of these types of equipment, IaaS may have various functions, among others those of e-Infrastructures as a subclass of Research Infrastructures. Moreover, in accordance with the diverse application areas of e-Infrastructures, they may also serve as environments and supporting frameworks for specific tasks such as DCH preservation.

In all IaaS based e-Infrastructure realizations there is a move from traditional technology- or product-orientation towards service-orientation, on the basis of sophisticated virtualization technologies that create virtual resources from physical ones. Indeed, virtualization allows an e-infrastructure or part of it to be virtually separated and dedicated to the delivery of specific functions to specific users.

IaaS (namely the applied virtualization) enables e-Infrastructure service providers to offer "on demand" provision of custom tailored services for specific user groups. Services offered by this way include not only virtual computing and storage resources but also e-Infrastructure services, eg. web portals, databases, user and application management, etc.

The main elements of these innovations are increased interoperability (as a result of standardization), easy on-demand access (based on integrating available functions and hierarchical layers), improved resource utilization (thanks to virtualization), and elevated flexibility and adaptivity (due to service orientation).

The potential emergence of IaaS solutions in e-Infrastructures for DCH preservation will be handled in the followings sections, with the goal of presenting a roadmap. However, before entering into the details of IaaS for DCH preservation, a summary of the likely demands and requirements is provided below in order to illustrate the usefulness of introducing IaaS realizations of e-Infrastructures in that specific area of preserving DCH content and, as the final goal of the DCH-RP project is concerned, why IaaS is to be involved in the development scenarios appearing in the roadmap.

4. SPECIFIC IAAS ASPECTS OF DCH PRESERVATION

In accordance with the aims of the DCH-RP project, future developments in the field of preservation methods, as well as in the area of e-Infrastructures exploited by preservation practice, should aim at harmonizing the preservation policies in the (global) DCH sector, strengthening the co-operation between DCH institutions and e-Infrastructure organizations (public and private ones), establishing appropriate conditions for these sectors to integrate their efforts into a common work, developing suitable models for the governance and maintenance of the related preservation efforts, and offering sustainability for the in DCH and infrastructure activities/organizations involved. Forthcoming developments in the e-Infrastructure area are to be matched to these requirements and IaaS is probably the best tool for that.

The DCH-RP project is concerned, it intends to present a roadmap for implementing a federated e-infrastructure dedicated to support best preservation practices and to enable the preservation community to optimally exploit the infrastructural services in the arts and the humanities. Coordination among the DCH and e-infrastructure organizations is therefore a crucial task where IaaS based e-Infrastructure developments can have a considerable stimulating effect. It is important therefore not only to define a roadmap and the related practical tools for preservation but also to enable the monitoring of the involved activities and support the shared implementation of the common e-Infrastructure services.

Although preservation on one hand and the e-Infrastructures on the other hand are both complex, multifaceted systems, the above requirements should be taken into consideration when working on the roadmap by focusing on the storage phase, which includes both long-term preservation (including dark archives) and short-term preservation (storage for a relatively short period of access). In any case either exploiting present services or using novel services to be introduced later, one assumes in harmony with the roadmap scenarios, or their combination are all to be taken into account.

This means that, on one hand, the experiences gained from proofs of concept (PoCs) by applying present methods and tools for solving present practical preservation examples are to be taken into consideration when extrapolating future developments from today's achievements. However, this PoC-based approach should be complemented by a visionary approach which isn't closely connected to present tools and practices but tries to think ahead with respect to the roadmap. On top of that, community building and establishment of a network of common interest and knowledge are to be involved in the investigations as another way of improving preservation efficiency and reliability.

Additional aspects to consider within the DCH-RP project are, among others (as it has emerged from earlier investigations):

- Joint inclusion of multiple so called PEST (political, economic, societal, technological) viewpoints in the investigations,
- Involvement of the real boundary conditions into the discussions,
- Exploitation of a digital preservation services registry,
- Use of short-, mid-, and long-term measures in the e-Infrastructure supported preservation practice,
- Alignment of the roadmap characteristics to the EC and national research agendas,
- Establishment of key partnerships with relevant e-Infrastructures,
- Overall treatment of the related innovation effects by advanced preservation practices, and also

- Introduction of innovative elements into the preservation process.

5. IAAS AND THE ROADMAP FOR DCH PRESERVATION

Taking into consideration the above aspects and requirements when preparing the roadmap is a must. Moreover, in the second phase of the DCH-RP project the specification of the digital preservation service, as well as the definition of the stakeholders, needs, scenarios, and business models, is unavoidable. Setting critical system requirements (and mapping them into technology drivers as well as key performance indicators) will be major elements of the investigations, together with analysing the current technological offerings and also the gaps that are not yet covered. Helping in this way the process of developing the registry of tools and services will probably also be the one of the main tasks of DCH-RP activities in 2014. Of course the results will also influence the definition of the suggested preservation framework and the drivers for making a shift in institutional practices – all depending on the scenarios appearing in the roadmap.

All the above aspects require a well established, visionary but at the same time realistic handling of the issues to be covered by the roadmap for preservation. These issues include services, architectures, standards, interoperability, registry of services and tools, and, last but not least, a methodology that optimally combines all these elements into a working system, for the benefit of the users, in this case for the DCH preservation community. Realism and due consideration of all the conditions enabling well operable (well exploitable) e-Infrastructure services are of outstanding significance if sustainability aspects are to be appropriately treated in the roadmap. (Here sustainability is not restricted to sustainable financing but also other, political, legal, technical, organizational, etc. sustainability factors to be taken into account.) They together characterize the overall sustainability of the e-Infrastructure. Obviously financial sustainability is of outstanding importance here, and makes the applied business model a crucial element in operating the infrastructure.

As it has been mentioned earlier, such a methodology is based, from the preservation point of view, on the IaaS paradigm, but also has to meet, as far as possible, the needs of the specific user community interested in DCH preservation.

Although IaaS is a joint result of the increasing user demands (see above) and, on the other hand, a technological breakthrough (virtualization, allowing flexibility and easy modification of the virtual constructs stemming from an interconnected and thus integrated complex of physical components comprising the infrastructure), here, and also in the followings, text aspects of user-oriented service provision rather than the technological realization details are emphasised and investigated from among the two major facets of the IaaS solutions.

6. MORE ABOUT IAAS

IaaS is both a product of technological development, first of all by the advent of virtualization, and a response to user demands, with respect to efficiency, reliability, and friendliness of the e-Infrastructure services – not only in the case of DCH preservation.

We can examine IaaS in some more detail, by referring to the e-IRG (the e-Infrastructure Reflection Group). Some recent recommendations of the e-IRG about IaaS are first cited here:

> "The adoption of an Infrastructure as a Service (IaaS) model should be strongly stimulated and supported with the aim of increasing the sustainability of e-Infrastructure as well as identifying and providing innovative solutions which could find a wider use in society. Use virtualization and SOA when developing and introducing new e-Infrastructure services wherever this is efficient. Apply simplified access, transparent service offerings, customised support, standardization,

improved governance models and sustainable business models in the definition and deployment of e-Infrastructure services.

- Helps infrastructure developers, resource and content providers, and also user communities in order to build and exploit reliable and robust data services suitable to real needs.
- Helps in definition and exploitation of e-Infrastructure services, content-related curation, preservation and data exploitation, interoperability, data access, federation, and openness.
- Utilizes emerging use of virtualization in ICT service provision, whereby physical resources are shared by users in a manner which appears to support each user independently, optimizing resource utilization, reliability, energy efficiency and maintenance costs."

This summary of the trends and this brief provision of a forecast on behalf of the high level expert body of e-IRG well illustrate the direction and significance of the progress which results in IaaS e-Infrastructures optimally combining the well known and useful cloud and grid features. The intended combination of features is illustrated below by bold printing of those features intrduced by the IaaS approach into e-Infrastructures from cloud and grid practice:

	Cloud	Grid
Type of Applications	**provision + use**	**collaboration**
Major users	commerce/industry	**academia/research**
User community	**open**	closed
Ownership	**single**	multiple
Financing	charge per use	**cost sharing**
Funding	users	**owners (projects)**
User management	**decentralized**	centralized
Character of the applications	**interactive**	batch
Major application goal	**storage**	**HPC**
Access means	**web**	grid middleware
Major system components	physical	**virtual**
Security means	user separation	**federated AAI**
Application control	**centralized**	decentralized
Resource control	**high**	medium
Standardization	weak	**medium**
Interoperability of elements	medium	**high**
Ease of use	**medium**	low

Concerning the application of such IaaS based e-Infrastructures in DCH preservation, it is to be kept in mind that indeed the demands are high, and is increasing. They are well demonstrated by the project goals expressing the general needs and requirements in DCH preservation. A selected (non-prioritised)

list from an earlier summary of requirements is provided below, illustrating where the IaaS approach can considerably contribute to improvements in DCH preservation:

- Reliability and robustness
- Openness, scalability, and flexibility (based on open industry standards)
- Ease of use (user-friendly interactions and interfaces, etc.)
- OAIS (Open Archival Information System) compliance
- Mechanisms for integration and automation of ingesting digital material
- Automatic metadata capture and extraction
- Separation of content (information) and metadata
- Various content formats (from print-based documents to digitized images)
- Annotation services
- Scalability (up to the TB-PB level and more)
- Performance for millions of electronic documents
- Authenticity and integrity of data
- Continuity (long range – up to 100s of years - handling of information)
- Identification of digital objects in danger (of inaccessibility, etc.)
- Security of long distance (also international) transmissions of information
- Validation-certification of HW-SW environments vs. rendering digital objects
- Distributed systems operation enabled for preservation
- Virtualization possibility hidden for the user
- Support of many storage media and devices
- Backup and restore possibilities

7. PRELIMINARY ASSUMPTIONS ABOUT IAAS BASED ROADMAP SCENARIOS

It should be noted that present day e-Infrastructures (including the ones having been exploited by the PoCs performed and evaluated within the frameworks of the DCH-RP project), are approaching the principles of IaaS and that the grid and cloud solutions they apply are utilising virtualization techniques in order to allow distributed operation of computational and storage components. At the same time the GÉANT network (the pan-European backbone of the NRENs) operates as the “glue” by serving as the provider of the interconnections between the components of the grid and cloud components.

This also means that IaaS itself won’t appear in the roadmap as a completely new way of innovating within the infrastructure, but rather as an opportunity to further develop the available e-Infrastructure facilities and services while taking into consideration the special, more demanding requirements of the DCH domain and of course also introducing completely new solutions and tools into the future infrastructure development process. In this way the coming generations of e-Infrastructures will get step by step closer to truly evolutionary IaaS realizations while also enabling also revolutionary new solutions.

IaaS architectures in e-Infrastructure systems over the next several years are assumed to cover the entire spectrum of ICT functions as elements of a service portfolio for DCH preservation. This means that all the functions of input/output, transferring, processing, and storing of related information will be built into the IaaS realizations. Therefore IaaS will comprises high capability elements for networking (interconnections by communication equipment and SW), computing (processing by HPC equipment and SW), and storing (offering high reliability, controlled accessibility storage equipment and SW).

Other important features to be taken into account when developing the roadmap are listed here:

- Concerning the geographic location of the involved elements of the infrastructure, IaaS organization and operation can be either centralised (single location) or, as the more general class, distributed (integrating distant, remote infrastructure components).
- The lifetime of e-infrastructures is either medium/long (with some more or less fixed configuration) or short, where temporary constructs are established by using VPNs (Virtual Private Networks) for interconnecting the tools needed for specific e-Infrastructure functions in the case of specific applications, like DCH preservation.
- Another aspect is related to the ownership of the components and their subsets (or the entire OAIS based or similar structure): one of the options here is to allow outside (commercial) components (eg. HPC or storage equipment) to be built into the infrastructure, another option is to build just on proprietary equipment.
- Although in most cases hybrid solutions are applied, it should be understood that the more complex is the infrastructure (distributed, dynamically reconfigurable, proprietary realizations), the greater the cost of development and operation will be, because of the higher costs eg. for SW (middleware) as the specific software for making IaaS operation possible by HW virtualization, SOA, on-demand resource allocation, distributed computing and storage, etc.

One of the most important features of the IaaS approach in case to supporting DCH preservation is related to customization of the services, i.e. tailoring the service portfolio and characteristics to the actual preservation task. Here the e-Infrastructure resources (especially virtual machines and middleware components) are specifically allocated to the preservation functions and sub-functions in accordance with their specific requirements. The DCH experts themselves are assumed to control that allocation process. In this way (sometimes called configurable turnkey solution)the greatest efficiency of the preservation process can be achieved in every sense.

Nevertheless, from the application's point of view the basic benefit of using IaaS is that, thanks to virtualization and to the possibility of on-demand reconfiguration, the e-Infrastructure is available for the user as a set of easily accessible and, as much as possible interoperable, services (collected and presented in registries of services and tools) rather than just an assembly of facilities offering much less friendly access possibilities. By taking into account the features of the tools and services such as popularity and acceptance, published qualification, maturity, support level, accessibility, interoperability, portability between environments, scalability, openness and modularity of the architecture, etc., the best possible combination of the required tools and services can be selected so that the IaaS approach really proves its outstanding capabilities.

8. FIRST CONCLUSIONS WITH RESPECT TO IAAS IN DCH PRESERVATION

Let it be stressed again: the attractive features of the IaaS approach are due to the opportunities provided by virtualization, distributed resources (remote access), and redundant architecture/topology. The result can be summarised in terms of higher speed, lower power, cheaper access, ease of use, and elevated efficiency. It is worth adding here that in practice the well established NREN services, together with the availability of several well developed NGI offers, including their joint pan-European services, and the easily reachable commercial offers, do help building really flexible and dependable IaaS realizations, also enabling further development, either evolutionary or revolutionary.

Needless to say, IaaS (or any other, traditional or later generations of e-Infrastructures) can reach their maximum potential and best serve the DCH preservation requirements if and only if the users (the DCH community) are well prepared to best exploit the opportunities stemming from those e-Infrastructure realizations. Therefore the e-Infrastructure community and the DCH community should work together on

how they can best prepare themselves for the advent of true IaaS provision, and beyond, by the e-Infrastructure communities.

As a final note, it should be stressed that nowadays the IaaS concept and the flexible use of virtualised resources (service oriented infrastructure provision) are currently most visible almost everywhere just on the processing and storage level rather than at higher lavels (application, among other preservation layers) or in terms of extension to VRE/VRO for VRCs and for the wider public. This means that in the coming years considerable development is foreseen, perhaps drastically changing the environment (for example establishing the concept of Preservation as a Service) and practice (for example involving eg. wide social participation) in DCH preservation. These factors need to be seriously taken into consideration in the roadmap to be prepared by the DCH-RP project consortium.

ANNEX 3 COUNTRY EXAMPLES ON THE USE OF DISTRIBUTED DIGITAL PRESERVATION SERVICES

This annex presents some example from partner countries in the DCH-RP project where cultural heritage institutions are using distributed digital preservation services. The text is taken from the report on Digital Preservation Services: State of the Art Analysis by Raivo Ruusalepp and Milena Dobreva for the DC-NET project.

1. ITALY

The interest in digital preservation in Italy is prominent and long-standing. In 2003, Italian professionals made an international survey[50] on legislation, rules and policies for the preservation of digital resources. In 2008, a round table on "Digital preservation in Italy: experiences face to face" was hosted by the National Library in Florence and was attended by 90 participants.[51] Rome has hosted workshops, among others, by the PLANETS project in 2008, and the KEEP project in 2011, as well as workshops of LIBER on digital preservation and DELOS summer schools in the same domain. The choice of major EC-funded projects to organise events in this country seems to respond to a well identified interest in this area and a professional community that is eager to follow the most recent developments.

The Digital Stacks project[52] is developing a long term digital preservation system for electronic publications that fall under the legal deposit law. The architecture of the solution is distributed between two sites of the National Library (in Florence and Rome) that allow deposits into the archive[53] and a dark archive for preservation only in Venice. The service is operated by Fondazione Rinascimento Digitale from Florence. The multiple sites offer redundancy of content (altogether six copies of each object are being kept) and the open source platforms used for managing the collections ensure few vendor dependencies.

The Consorzio COMETA[54] – the Sicilian Grid service provider – has developed the gLibrary platform[55] that provides a simple yet powerful system to store, organize, search and retrieve digital assets in repositories built on e-Infrastructures. This effectively hides the underlying technical details of the service from the end users and provides a user-friendly way to archive digital objects.

2. ESTONIA

The Estonian Ministry of Culture has been co-ordinating the preservation of digital cultural heritage under the auspices of a national strategy since 2003.[56] A network of competence centres has been established to support both digitisation and digital preservation in memory institutions.[57]

One of the national competence centres – the Estonian Public Broadcasting Company[58] is offering secure back-up storage to other memory institutions as a secondary storage site. The terms of this service have been negotiated by the Ministry of Culture, which also funds the bulk of the cost.

[50] http://eprints.erpanet.org/65/01/Dossier1_English_version_Full.pdf
[51] http://www.digitalpreservationeurope.eu/publications/reports/Report_roundtable_event.pdf
[52] See: http://www.indicate-project.eu/getFile.php?id=233
[53] See: http://www.depositolegale.it/
[54] http://www.consorzio-cometa.it/en/home
[55] http://www.consorzio-cometa.it/en/descrizione
[56] http://www.kul.ee/index.php?path=0x838
[57] http://digiveeb.kul.ee/index.php?id=10429
[58] http://arhiiv.err.ee/

The Restoration Centre “Kanut” is a competence centre for museums, offering specialised digitised services and also digital storage on behalf of museums.[59]

There is a common gateway to digital collections in memory institutions that harvests metadata from a variety of in-house catalogues.[60]

3. HUNGARY

NIIF,[61] the developer and operator of the e-Infrastructure in Hungary, together with Hungarnet, the association and representative of the e-Infrastructure users, have jointly been putting emphasis on co-operation with the academic, research and digital cultural heritage communities. The requirements of these communities have been taken into consideration, as much as possible, when developing the service portfolio. However, approaching a truly service-oriented e-Infrastructure has turned out to be possible only more recently, with the advent of virtualization, enabling IaaS, an attractive solution for the majority of e-Infrastructure applications – practically all of them, except the most demanding, most complex ones where the need for special joint treatment of the high complexity research problem and the extraordinary application aspects does not allow utilizing the standard and widely exploitable IaaS solutions.

As far as digital cultural heritage preservation and related activities are concerned, the IaaS is proving to be a promising way of using e-Infrastructures in most cases. NIIFI is well prepared to introduce these new approaches and offer them to the cultural heritage sector, including networking, grid, cloud, HPC, and storage. (see also Annex 4)

4. POLAND

Many national e-Infrastructures are already offering basic digital archiving and storage services, including the Poznan Supercomputing and Networking Centre. The PSNC runs the PLATON project[62] that provides archiving and data storage service for research data from scientific and academic community. The services offered by PLATON include:

- Automated and transparent data replication, ensuring the durability of the stored data; file system-level metadata are replicated and protected against disasters, to ensure logical namespace consistency and accessibility.
- Data objects are persistently identified within the logical filesystem namespace; their physical location is masked by the virtual filesystem layer and the underlying logic.
- Abstract, universal data access interface: the virtual filesystem is accessible through WebDAV (over HTTPs), SFTP and GridFTP protocols.

Standard data access methods make it possible to:

- Use the long-term storage service as a networked filesystem or (with additional tools) as the networked drive;
- Integrate the service with the Content Management Systems (e.g. dLibra[63]), by using the standard libraries for C, C++, Java and others in order to build customized service clients.

The PLATON approach lets users focus on their core business, by outsourcing the following long-term data management issues:

- Data storage technology migration (e.g. due to the technology development) is performed on the service provider side, transparently to users;

[59] http://kanut.ee/index.php/digiteerimine
[60] http://e-kultuur.ee/?locale=en
[61] http://www.niif.hu/en
[62] http://www.man.poznan.pl/online/en/projects/50/PLATON.html
[63] http://www.man.poznan.pl/online/en/projects/20/dLibra.html

- Infrastructure development – equipment and software procurement, deployment and testing;
- Infrastructure maintenance and operation – monitoring, failure tracking, handling and resolving.

These are some examples of early adopters of digital preservation as e-Infrastructure in different EU countries. They demonstrate that the traditional models of preservation management are beginning to evolve towards more distributed preservation architectures.