DELIVERABLE

Project Acronym:	**DCH-RP**
Grant Agreement number:	**312274**
Project Title:	**Digital Cultural Heritage Roadmap for Preservation - Open Science Infrastructure for DCH in 2020**

D3.1 Study on a Roadmap for Preservation

Revision: version 3.1

Authors:

R. Ruusalepp (EVKM)
M Dobreva (EVKM)

Reviewers:

Roberto Barbera (INFN)
Michel Drescher (EGI.eu)

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 November 2012	R. Ruusalepp	EVKM	Extended structure of the report
0.9	13 January 2013	M. Dobreva	EVKM	Draft chapters 3, 6
1.0	03 February 2013	R. Ruusalepp	EVKM	Editing
2.0	08, 14 March 2013	M. Dobreva	EVKM	Integrating comments into deliverable text
2.1.	31 March 2013	R. Barbera, M. Drescher	EVKM	Review comments
3.0.	4 April 2013	M. Dobreva	EVKM	Final version
3.1	15 April 2013	C. Prandoni	PROMOTER	Formal check

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise.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

Table of Contents

List of figures

Executive summary

Improving digital preservation practices in cultural heritage institutions is an onerous and complex task. 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 on offer. Current solutions always require adaptation to the specific mandate of the institution, its technological infrastructure and the competences of its staff.

The need to address this situation and to offer more support to cultural heritage institutions was identified in the former INDICATE project and an initial survey of existing digital preservation tools and services was commissioned by the DC-Net project. DCH-RP seeks to offer a coherent and realistic roadmap that would help policy makers and programme owners to plan ahead, and will at the same time assist managerial teams of cultural heritage institutions to take decisions related to digital preservation.

This deliverable presents a snapshot made at the beginning of this endeavour. It looks at current offers of services in the environments suited for use by e-Infrastructures (including policy considerations and available guidance) according to the main focus of the DCH-RP, focusing on grid and cloud services, and suggests areas that need to be explored further in order to achieve a vision for short-, medium and long-term roadmap.

This deliverable is integrating components of an action plan to develop the DCH-RP roadmap which will guide further work in the project and beyond, and is linked with sustainability of the project outcomes.

This report targets first of all the partner institutions in DCH-RP project. As a public deliverable, it can also serve as a consultation document that would benefit from feedback from two main communities: cultural heritage and e-Infrastructures that already include digital archiving functions.

The final preservation roadmap of DCH-RP (to be developed over the course of the project) will mainly target policy-makers on different levels. Hence, policy makers who are following current developments in digital preservation are also welcome to consult this deliverable and offer their feedback.

The deliverable is organised as follows:

Section 1 provides a short introduction to the background and to the objectives of the document.

Section 2 offers first a summary overview of the DC-NET report[1] that this deliverable builds on. The overview is followed by an update of the analysis of preservation tools and services in Section 3, with a special focus on services in grid and cloud environments.

[1] http://www.dc-net.org/getFile.php?id=467

Section 4 looks at high level service architectures applied and summarises some emerging architectures. It is concluded with a gap analysis between the preservation tools and services and grid architectures.

The gap analysis is also intended to help synthesize ideas for the final roadmap development. A matrix is proposed in Section 5 with possible areas of intervention/co-ordination that can be used as a basis for further discussions within the DCH-RP consortium.

Section 6 begins the discussion of the DHC-RP digital preservation roadmap by looking at types of analysis that are required, possible timeline of the roadmap and offers an action plan for the short-term stage of the roadmap.

1. Introduction

This document represents a first stage in developing a Roadmap for Preservation by the work of the DCH-RP project. It provides a summary of work done under Task 3.1 Preservation services architecture within the project. The main aim of this study is to analyse key characteristics and requirements of digital preservation (DP) in cultural heritage institutions and how they could be linked with e-Infrastructure services. The use of existing e-Infrastructures for research and academia (including NREN, NGI and other data infrastructures) are seen as efficient channels for the delivery of advanced services to the digital cultural heritage sector.

1.1 Background

The recent DC-NET project[2] conducted research and analysis on how the digital cultural heritage sector can benefit from the use of e-Infrastructures. The outcome of this study was a report[3] that identified a list of digital cultural heritage research priorities and how they can be addressed within e-Infrastructures context. An order of priority for the services was produced and validated through consultation with stakeholders. Long-term preservation was identified as having the highest priority among other current needs.

The subsequent INDICATE project[4] identified digital preservation as a key area in need of a coherent and coordinated intervention in cultural heritage institutions (CHI).

A DC-NET report on digital preservation tools and services[5] surveyed the currently available software tools and services for digital archiving from the perspective of cultural heritage institutions. It concluded that there are significant unresolved issues with sustainability of services, benchmarking of tools and fragmentation of tools supporting individual tasks in the archiving workflow. A summary of the main findings of the DC-NET report is presented in Chapter 2 below.

1.2 Objectives of the study

The aim of the DCH-RP project is to develop a roadmap to implement a preservation infrastructure for digital cultural heritage. The design of the roadmap will be supported by practical, proof of concept level, experiments in project partners' countries. The role of this report is to set the stage for developing a roadmap document within the project and to identify gaps in the preservation services market that the proofs of concept could address.

This report and the subsequent roadmap will outline main concepts of long-term preservation and its key elements. However, it does not aim to resolve the substantial on-going debate on the nature and

[2] http://www.dc-net.org/
[3] See DC-NET D3.1 "Digital Cultural Heritage Services Priorities Report"; also "DC-NET Service priorities and best practices for digital cultural heritage" http://www.dc-net.org/getFile.php?id=450
[4] See, for example, D6.2 Strategy and Future Plans workshop proceedings & harmonised policy elements of the INDICATE project on http://www.indicate-project.eu/index.php?en/176/documents-and-deliverables
[5] Ruusalepp, R., Dobreva, M. (2012) Digital Preservation Services: State of the Art Analysis. http://www.dc-net.org/getFile.php?id=467

goals of digital preservation. Instead, it is tasked to capture and analyse relevant current practice that pertains to the specific use case of providing digital preservation services for the cultural heritage sector using e-Infrastructures.

2. DC-NET Report on Preservation Tools and Services

The DC-NET report[6] looked at the basic functional entities a digital archiving and preservation system needs to implement, and analysed the current offering of software services and tools for automating these tasks. The digital archiving workflow entities featured: pre-ingest (including transfer), ingest, storage, digital object analysis, preservation planning, access, and re-use, which represent a life-cycle process-oriented approach in preservation.

Based on a desktop study and a rapid analysis of some 190 currently available software tools, the report provided a high-level view on the range of instruments available for the preservation lifecycle. The report found that digital preservation services are by and large still an experimental area.

The report identified several previous and on-going efforts that systematically gather information on digital preservation tools and services. The study used these as a source for building its own registry of preservation tools: the CAIRO project[7] (54 tools under 15 categories),[8] the National Digital Infrastructure Preservation Program (NDIIPP)[9] in USA (38 tools and services), the Library of Congress[10] (10 tools for preservation metadata implementation), the AQuA project[11] (44 tools), the blogs of the OpenPlanetsFoundation (OPF),[12] the DigiBIC project,[13] as well as the SourceForge[14] platform for publishing, searching and downloading open source software.

The analysis of the distribution of the types of digital preservation tools in this report shows that:

- There is no coherently applied business model and implementation architecture of a digital preservation system. Most systems claim being compliant with the OAIS reference model[15] but OAIS compliance is referred to in vague terms and is difficult to prove.
- There is an abundance of software tools addressing specific tasks, e.g. format identification, which potentially could be transformed into services; however, the work on coherent service architectures for digital preservation is still ongoing.
- Granularity of tools in the digital preservation domain is a complex issue. The range of available tools covers the whole spectrum from well-defined atomic tasks (such as file format identification) to sets of tasks that form whole functional entities (e.g., ingest). Additionally, there is little support documentation and guidance for an easy “mix and match” approach, for example for smaller institutions with little IT expertise.

---

[6] http://www.dc-net.org/
[7] http://cairo.paradigm.ac.uk/projectdocs/index.html
[8] A caveat should be stated that while the number of tools analysed by projects which had been completed would not have changed since the time of the DC-NET 2012 report, in the cases of initiatives which continued to gather information on preservation tools, and in the case of Sourceforge there could be an increase in tools
[9] NDIIPP Partner Tools and Services Inventory, http://www.digitalpreservation.gov/partners/resources/tools/index.html
[10] http://www.loc.gov/standards/premis/tools.html
[11] http://wiki.opf-labs.org/display/AQuA/AQuA+Mashup+Tool+List
[12] http://www.openplanetsfoundation.org/
[13] http://www.digibic.eu
[14] http://sourceforge.net/
[15] ISO 14721:2003 Space data and information transfer systems - Open archival information system - Reference model

- The interoperability between existing tools – technical, semantic, policy, inter-community, legal – is not yet systematically addressed and with clearly defined aims.

The identified tools were grouped into a taxonomy based on stages of the digital archiving workflow. This allowed demonstrating the areas of preservation work that have most software tools to choose from – metadata extraction, file characterisation and file format identification. Detailed comparison of features of all these tools and testing them in practice was outside the scope of that study. However, the report pointed out a significant lack of benchmarks and metrics for comparing preservation tools for both professionals and beginners in the digital preservation business. It also recommended creating business scenarios that would ensure sustainability for the tools or their development into maintained e-services. The DCH-RP project will take steps to progress work along these recommendations, the current study being the first in a line of several deliverables.

One of the main conclusions from the DC-Net report to the DCH-RP roadmap development work is that orchestration of digital preservation tools and services from a number of disparate sources into one coherent solution is, under current circumstances, next to impossible. What is needed is a consolidated view or vision of implementation architecture of digital preservation services on the e-Infrastructures. This did not appear to exist in early 2012 and still does not seem to be agreed on. This study is going to report on a number on initiatives going in this direction but that are, as yet, hard to compare.

3. Updated Overview of Digital Preservation Tools and Services

The DC-NET report on preservation services and tools looked at currently available preservation tools and services. Since it was written, the offering of tools and services in the area of digital preservation continued to grow[16] , following some of the patterns identified in the DC-NET report:

- There are hundreds of tools mostly developed for the needs of memory institutions, as commercial products and as pilot implementations of research projects. The tools originating from the cultural heritage sector are developed by larger institutions and address their specific requirements. Those developed as pilot implementations of research projects often are proofs of concepts related to research and innovation initiatives; they rarely grow beyond a few early adopters.
- The differing size of cultural institutions means that commercial software solutions may not be accessible (affordable) to everyone. At the same time, tools addressing specific requirements of large institutions are not always compatible with those of smaller institutions. The plentiful pilot implementations are hard to match to the diversity of real-life needs of various cultural heritage institutions. There continues to be inadequate support for decision-making, selecting, testing and benchmarking tools for preservation (i.e., the process from analysing local needs, picking the best combination of available tools and to implementing a robust solution).
- The availability of digital preservation services is still limited. The software products that are offered in this domain are either addressing very specific tasks, or complex solutions combining a number of tools to suit particular institutional needs.

3.1 Merging preservation with the Grid – Some examples

Since our report was published, a roadmap was developed for the needs of the creative industries within the DigiBIC project (Teruggi, Ranzuglia 2012)[17] . This roadmap looked at the priorities for preservation and identified four of them:

- Preservation of complex objects;
- Preservation of artistic objects;[18]
- Preserving environments; and
- Self-preserving objects.[19]

The roadmap did not look into the issues of tools and services, although the DigiBIC project addressed the issue of re-use of tools developed within research projects.

The DigiBIC roadmap illustrates areas of preservation where creative industries need further work to be done. Other studies have been published that look more specifically at digital preservation services. For example, Faria et al. (2012) address the digital preservation watch component. They proposed a

[16] For a frequently updated overview of resources including services and tools, see *Digital Curation Resource Guide* by Charles W. Bailey, Jr. (at the time of writing this report the last update was from 8/12/2012), http://digital-scholarship.org/dcrg/dcrg.htm

[17] http://www.digibic.eu/

[18] Artistic objects can in most cases be considered complex objects; for example software art was one of the topics addressed within the POCOS project (Preservation of Complex Objects), see Anderson et al. (2012)

[19] Objects that have in-built capacity to evoke preservation actions

three-tier architecture (see Figure 1) with an interface tier communicating with the external world, a business logic tier and a knowledge base. The architecture includes data enrichment service, monitor service and assessment service which are all typical for the implemented approach and need to be specifically developed to communicate with the other two architectural tiers.

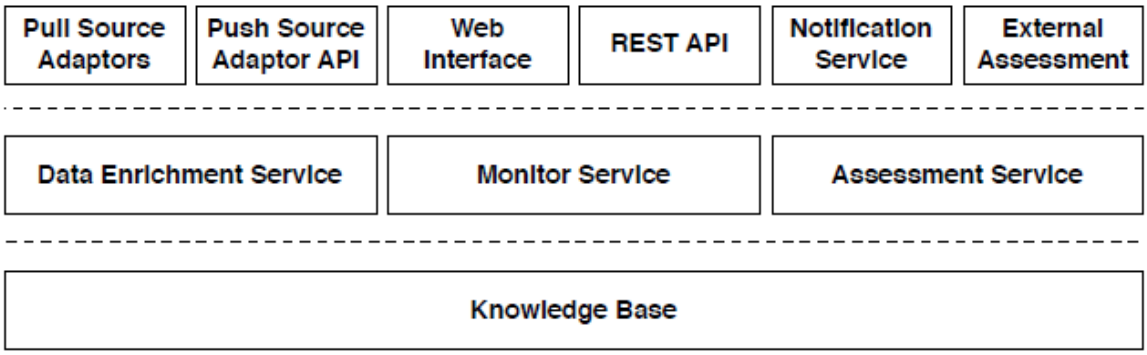

Figure 1. High level architecture of the Preservation Watch component [from: Faria et al. 2012]

Michael Peterson suggested in 2011 a Digital Preservation Services Taxonomy[20] . It has added cloud services to the concept map but is still struggling to integrate it with the rest of digital archiving services. This appears to be the overall difficulty at this stage that the role of cloud and grid services[21] is not uniformly defined within the archiving workflow.

[20] http://www.ltdprm.org/discussion-wiki/digitalpreservationtaxonomy

[21] Grid and cloud architectures are similar in using distributed resources. Grid architectures are based on shared resources while cloud computing is based on leasing resources. There are also divergences – the grids are mostly based at universities and academic institutions while cloud services mostly come from the commercial sector. Two popular types of grids are data grids and computing grids; the idea of shared storage was naturally appealing to the digital preservation community, given the scale of preservation tasks. With the introduction of cloud services, the concept of what such shared resources could offer evolved further and now includes offering of software, infrastructure and platforms as services (SaaS, IaaS, and PaaS). All these are relevant to preservation.

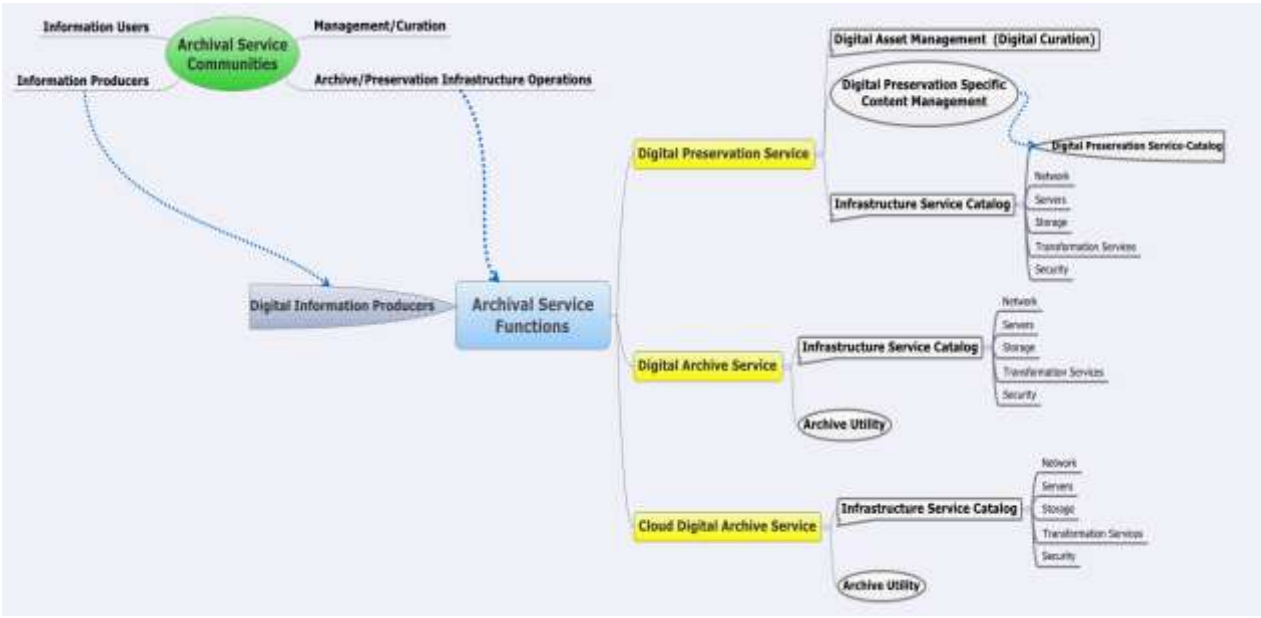

Figure 2. Suggested preservation services taxonomy by M. Peterson (2011)

A further development, relevant to the application of service architectures for digital preservation, was the release of TIMBUS project architecture in 2012 (see Galushka (2012)). The TIMBUS project looks beyond the issues of preservation of specific file formats, aiming to address business processes in their variety. Figure 3 shows the high-level view of their suggested architecture, which is concisely captured by the following description:

> "The proposed architecture consists of five modules, which cover various stages of the preservation process. The initial data is collected by DP Agents installed in the source environment. This data is combined and processed by the DP Acquisition Module. The intelligent Enterprise Risk Management Module in combination with the Legal Life-cycle Management Module utilises the processed data and generates a preservation recommendation report. This report is analysed by the DP Engine. Depending on the specified scenario, the DP Engine executes either the preservation, exhumation or both stages, where BPs are always handled together with relevant dependencies and contexts. A result from any of these operations can be verified and reported to an expert in case of any inconsistencies. All stages of the DP process are controlled via user interfaces." [Galushka (2012) p. 8]

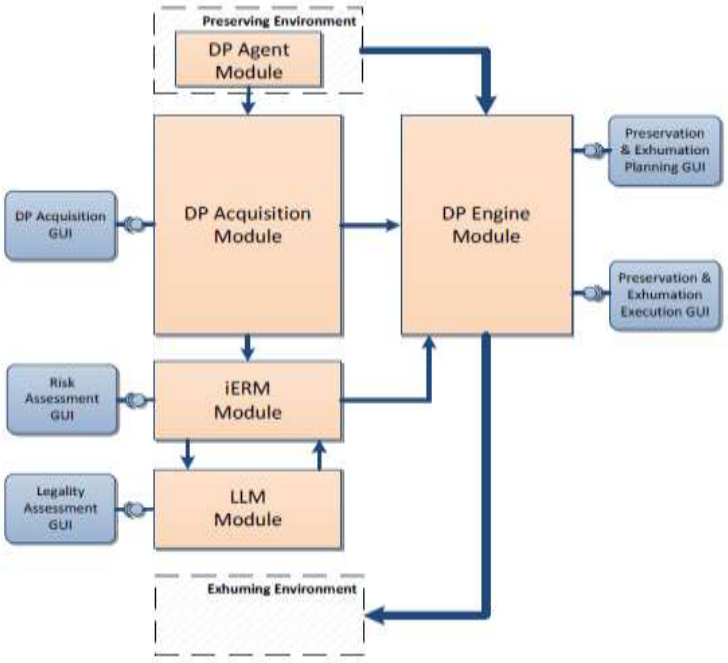

Figure 3. TIMBUS preservation architecture

Another recent development is the release of the architecture of APTrust, the system developed by a consortium of institutions in the USA under the remit to aggregate and preserve academic content. One distinguished feature of APTrust is that it is based on open source technologies with storage on the cloud. Figure 4 below summarises the architecture of APTrust.[22]

[22] https://wiki.duraspace.org/display/aptrust/Architecture

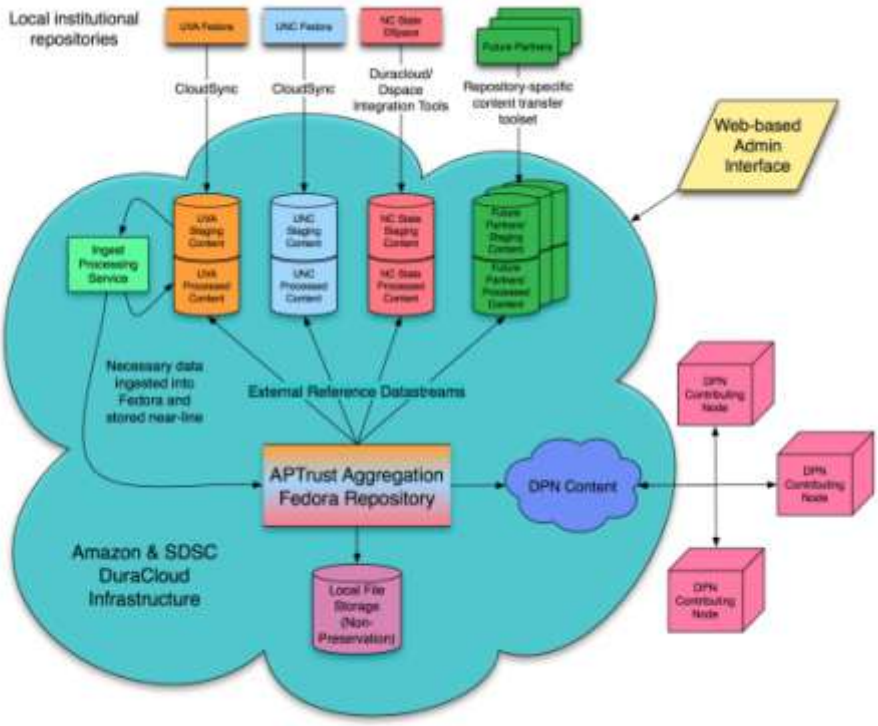

Figure 4. APTrust architecture diagram

Cloud solutions are being looked at by several projects co-funded by the European Commission, especially in the area of research data preservation, such as EU, for example, DAT[23] and SCIDIP-ES[24]. These will be considered in more detail in the next section of the report. A specialised seminar was held in late 2012 on preservation and cloud services in Schloss Dagstuhl; its findings will be published shortly[25].

These examples lead to an observation that there is still lack of coherence between digital preservation services and cloud architectures. The interoperability between digital preservation tools remains largely undefined and each initiative or project comes with its own understanding of the composition of modules and services. Work in this area continues and concepts are at the stage of "work under development" until best practice emerges for orchestrating digital preservation tools into workflows that involve grid and cloud infrastructures.

At this stage, developing preservation solutions for the cloud requires one to have a clear concept on the layers of services. In order to update the previous study and bring it to the focus of the DCH-RP project, this deliverable looks at relevant European digital preservation services, more specifically in current initiatives to develop grid and cloud environments for preservation.

[23] EUDAT: European Data Infrastructure, http://www.eudat.eu/
[24] SCIDIP-ES: Science Data Infrastructure for Preservation-Earth Science, http://www.scidip-es.eu/
[25] http://www.dagstuhl.de/en/program/calendar/semhp/?semnr=12472

3.2 Integrating Grid services and Preservation – Recent Examples

Examples of early implementations of Grid services include the work done at the SDSC in the USA on the use of iRODS for preservation, as well as pilot applications developed within the SHAMAN project for three memory institution scenarios, and the TextGrid project in Germany.

3.2.1 Use of iRODS in the SHAMAN project

The SHAMAN project[26] looked at the use of Grid technologies for preservation developing further the ideas on the use of iRODS for preservation developed earlier by Reagan Moore and MacKenzie Smith (2007) presented on the diagram below.

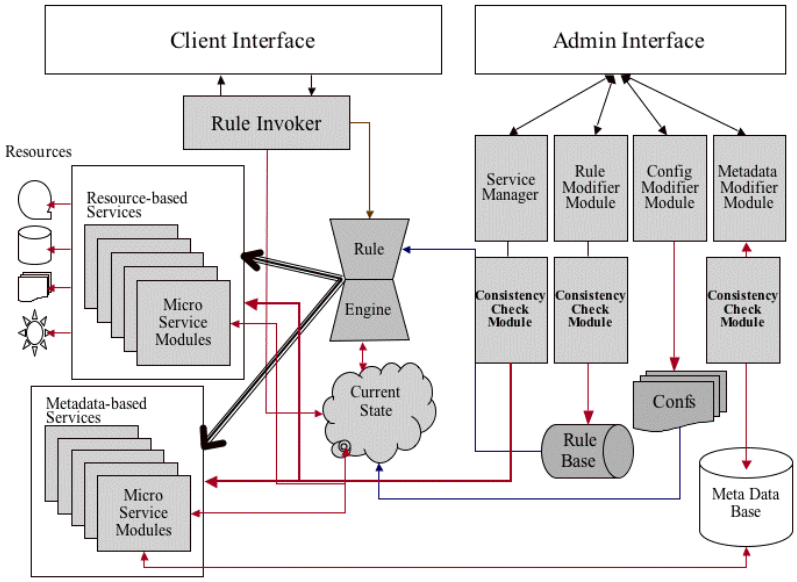

Figure 5. iRODS system architecture

SHAMAN project suggested an enterprise-driven reference architecture for digital preservation in 2009 and refined it later (see Antunes et al. (2012)):

[26] http://shaman-ip.eu/

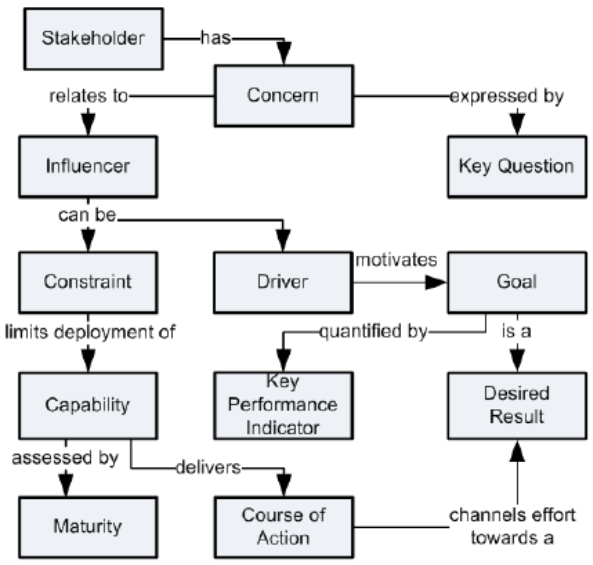

Figure 6. Main concepts of the architecture vision, SHAMAN project

SHAMAN is particularly relevant to DCH-RP because it created a demonstration case of memory institutions. The SHAMAN project did not develop a holistic tool but offered a combination of several instruments that were applied in the German National Library, Göttingen State and University Library in Germany, and the SME GLOBIT (see Birrel et al. (2010)). The three participating institutions implemented one scenario each:

- Scenario 1: Indexing and archiving book-like publications in libraries (German National Library);
- Scenario 2: Indexing and archiving digitised materials (Göttingen State and University Library);
- Scenario 3: Scientific publishing and archiving heterogeneous interlinked material (GLOBIT).

The following figure shows those components within the functional entities of a digital preservation system. iRODS was used as the basis of archival storage.

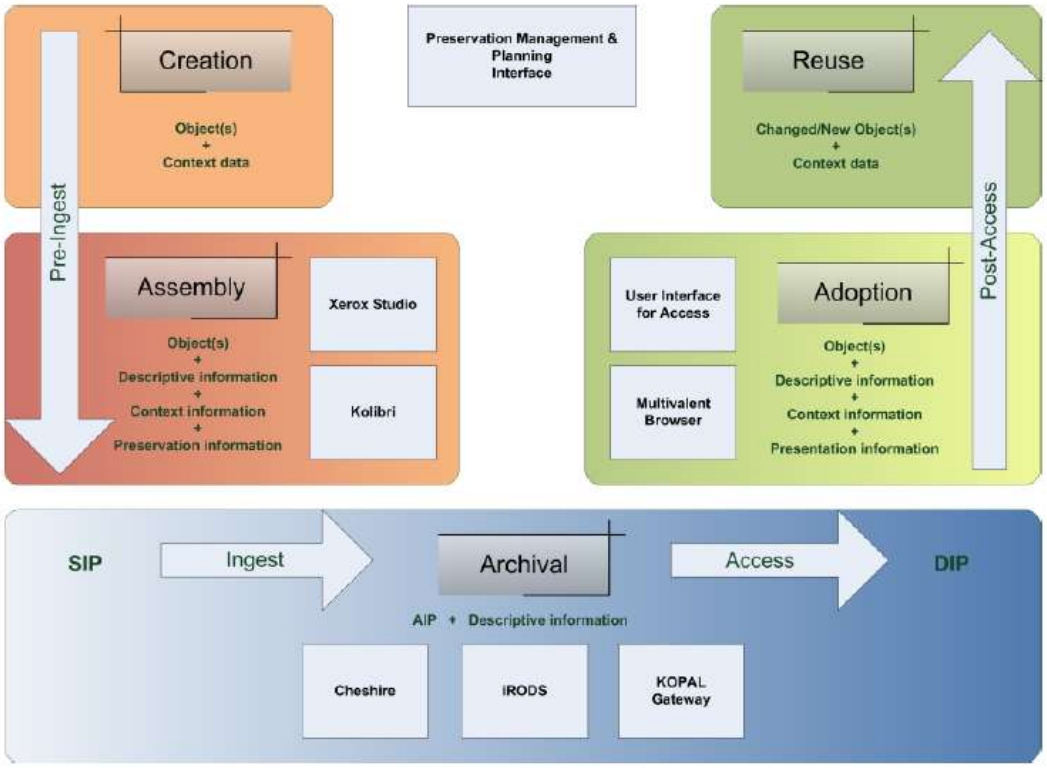

Figure 7. Tools and components included in SHAMAN demonstration for memory institutions

The above early examples did not apply their solutions on mass scale in memory institutions, yet they play an important role in developing the preservation architectures in distributed environments.

3.2.2 INDICATE project

The Indicate project[27] recently developed eCulture Science Gateway demonstration of how to implement e-Collaborative Digital Archives on e-Infrastructures and to protect them with access control and rights management. The Grid-based architecture was implemented in the COMETA centre in Catania (see Ardizzone at al. (2012) and Barbera et al. (2012)):

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

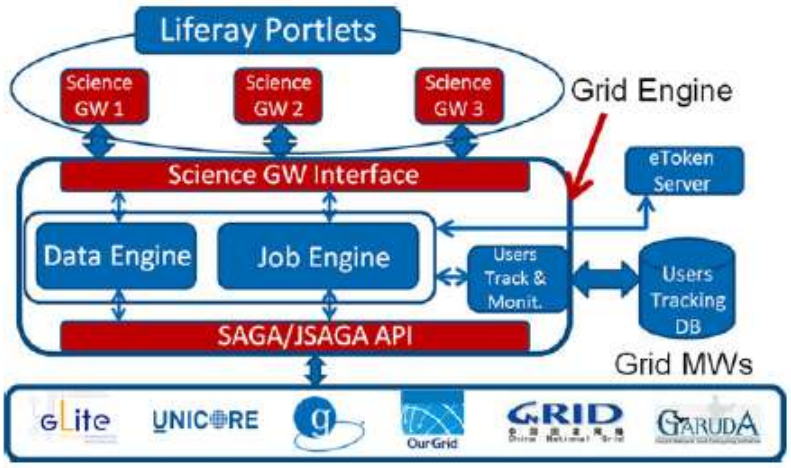

Figure 8. Grid machine architecture [Ardizzone et al. (2012)]

3.2.3 Carolina Digital Repository

One of the recent digital preservation workflows that integrates preservation repository with the grid for preservation is Carolina Digital Repository (CDR) of the University of North Carolina, Chapel Hill.[28] The Policy-Driven Repository Infrastructure project that is building the environment is especially focussed on investigating interoperability issues between repositories and grid platforms. Modelled after the OAIS reference model, the CDR employs the Fedora Commons repository as an object, model and services provider and iRODS grid as a distributed storage and preservation system (see Figure 9).[29]

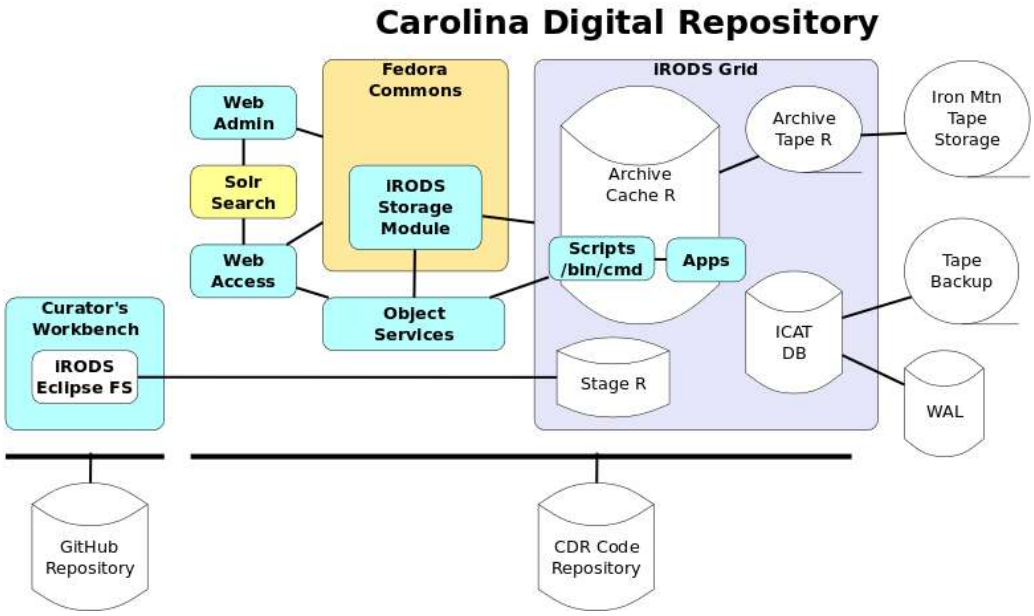

Figure 9. System overview of Carolina Digital Repository

[28] Carolina Digital Repository https://cdr.lib.unc.edu/
[29] https://cdr.lib.unc.edu/static/aboutPages/CDRSystemOverview.png

The diagram below shows the ingest workflow of CDR that has identified 'ingest service', 'staging storage', 'archival storage' and 'access storage' as separate services:[30]

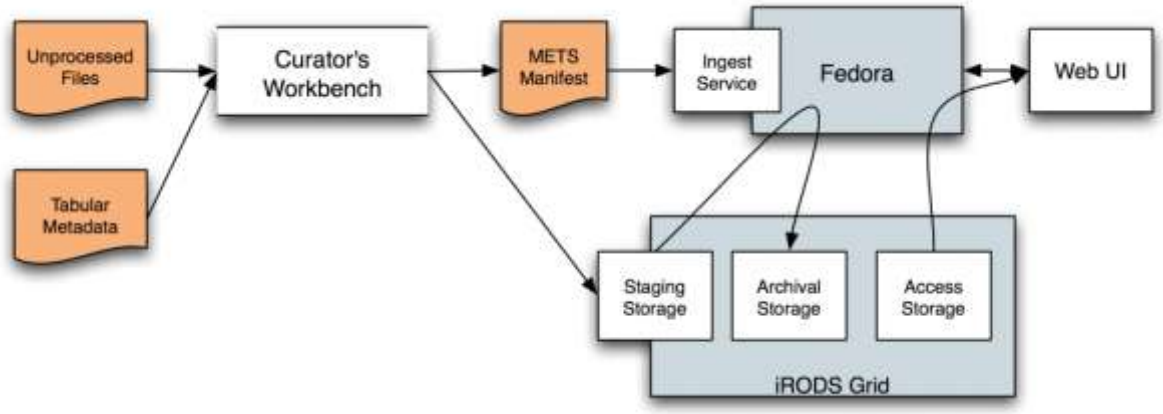

Figure 10. Ingest workflow of the Carolina Digital Repository

3.2.4 TextGrid

Another earlier example of the use of Grid in the domain of digital scholarship is the project TextGrid in Germany[31] (see the diagram below). TextGrid serves as a Virtual Research Environment based on a repository for research data in the humanities that relies on the nationwide WissGrid infrastructure. There are several architecture diagrams available for TextGrid; the one below presents interaction between preservation and Grid services.[32]

[30] http://www.lib.unc.edu/blogs/cdr/wp-content/uploads/2010/12/workbench-workflow1-1024x390.png
[31] http://textgrid.de/
[32] http://www.wissgrid.de/publikationen/deliverables/wp3/WissGrid-D3.4.3-LZA-Dienst-WDFv1.0.pdf, p. 9

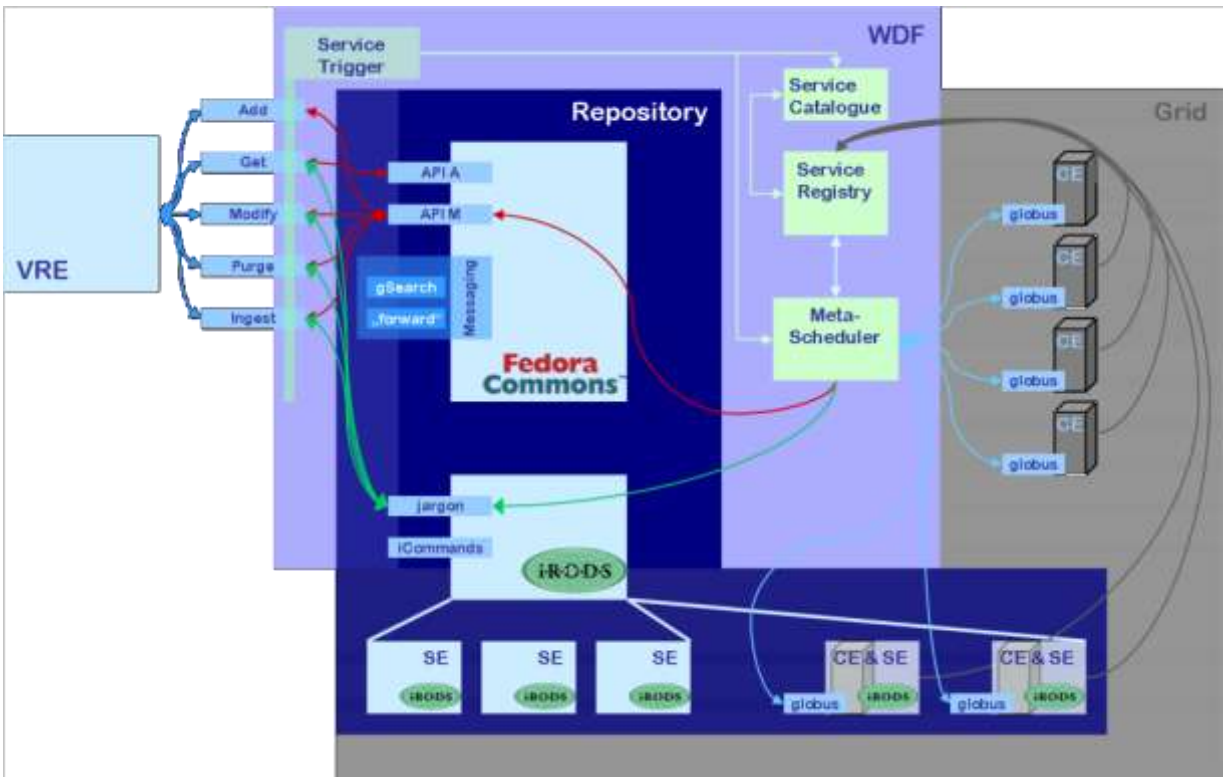

Figure 11. TextGrid architecture

3.2.5 UCL research data service

University College London (UCL) started developing a research data service in 2010[33] . It combines grid storage and preservation services, illustrated on the diagrams below. This example represents a well-developed view on the types of services that are needed to be integrated for a successful research data storage and use.

[33] http://www.ucl.ac.uk/isd/staff/research_services/research-data

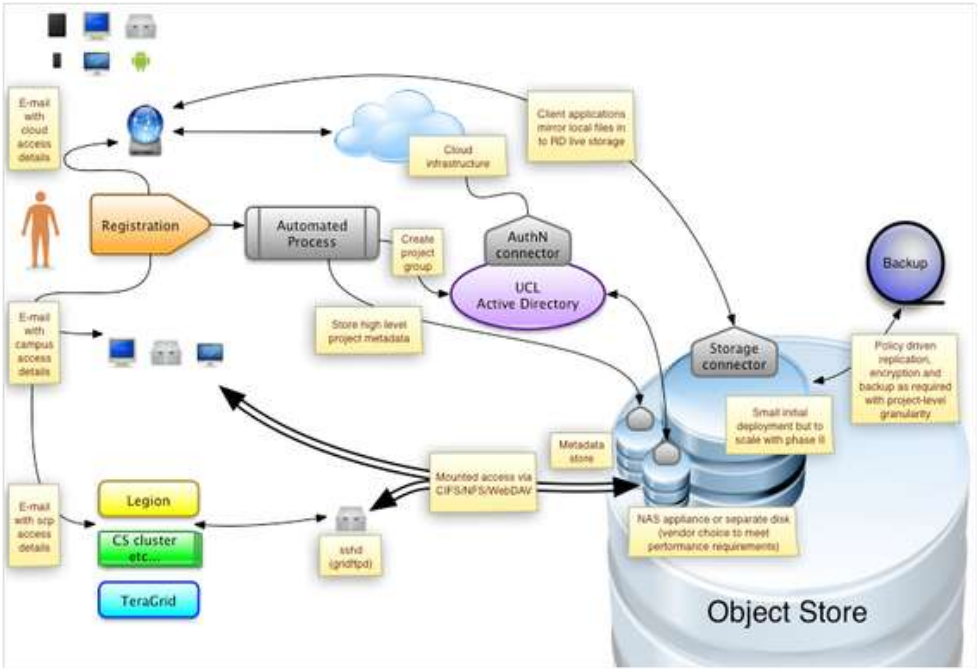

Figure 12. Logical view, UCL research data service

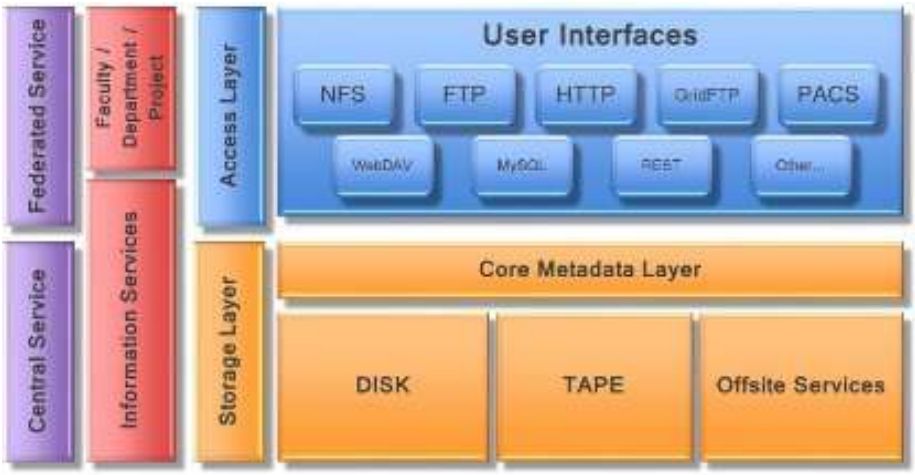

Figure 13 Service layers, UCL research data service

3.2.6 dArceo

dArceo is a specialised digital preservation service developed by the Digital Library team of the Poznań Supercomputing and Networking Center (PSNC) in Poland[34] . It is integrated into a suite which brings together digitisation workflow management, preservation and access to resources online (see diagram below).

[34] http://dlab.psnc.pl/darceo/

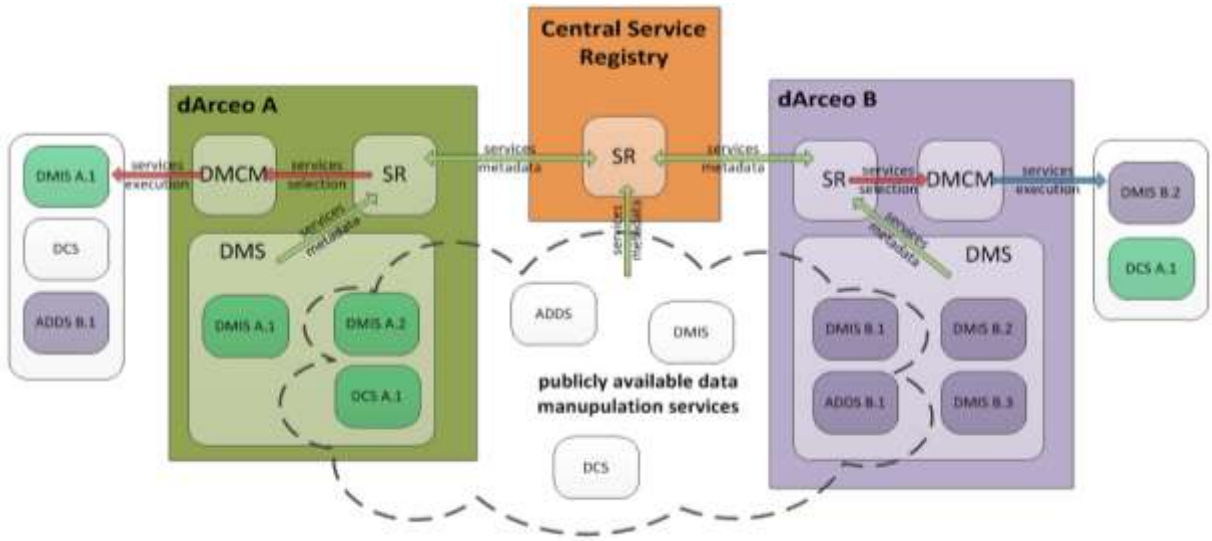

Figure 14. Suite of dArceo services from PSNC

Such examples are of particular interest to cultural heritage institutions because they combine the three major areas of their work: digitisation, online access and preservation. Identifying good solutions that illustrate how preservation services communicate with other strands of work is of importance for the DCH-RP roadmap development.

3.2.7 SCAPE project

A cloud-oriented architecture is currently being developed by the EU-funded SCAPE project.[35] In this architecture there are several layers, addressing authentication, content representation, data store, tools and resources as well as execution (see Figure 15). This architecture is a good example of clearly presented decomposition into tools and services. It is not clear whether any of the participating tools/services were used as available before, or all components are being developed anew for this environment.

[35] http://www.scape-project.eu/

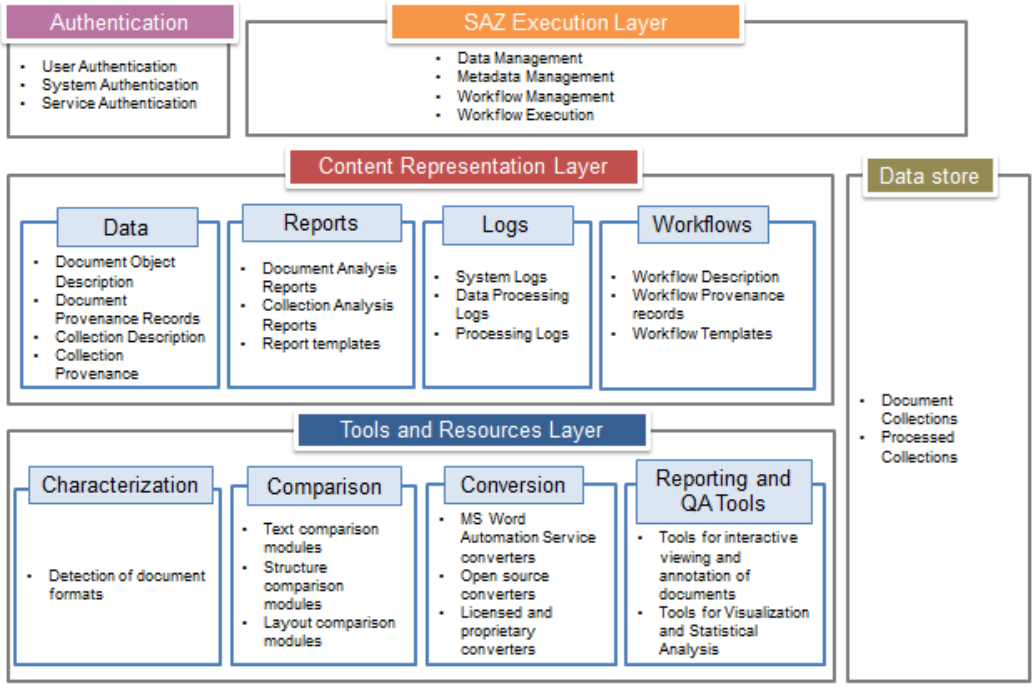

Figure 15. SCAPE project MS Azure-based preservation components

3.2.8 Conclusions

Not all recent activities in preservation services and the multiple examples of architectures integrating grids and clouds into service-oriented preservation systems can be described in their totality. More and more such examples will emerge as cloud services take hold in everyday computing. The examples provided here show that there are numerous methodological and structural differences in the interpretation of preservation. **The main challenge remains how to piece the two types of services together – whether to use the OAIS as the underlying model and map the grid/cloud services to it as "add-ons", or use the service and architecture models provided by the e-Infrastructures and embed preservation services into them.** Examples of both approaches were provided in this chapter, but best practice is yet to emerge.

Since the OAIS reference model remains too abstract for detailed implementations, preservation services are scattered and mostly atomic, it is a challenge to set up a complete digital archiving workflow based on repository software and lots of micro-services, and then add cloud/grid services.

On the other hand, when we study existing grid/cloud infrastructures (that, unlike the OAIS, do physically exist, are running, can be described in detail and tested) and attempt to "paste" the preservation services into them, different kinds of issues need to be faced – (down)scalability, adaptation of generic grid services to specific needs of DCH sector, and also how to orchestrate the whole workflow together.

This is why we need to further conceptualise the preservation architecture on grid and cloud infrastructures. The next chapter takes a look at some early attempts of doing this.

4. Conceptualising Preservation Services Architecture for the Grid

The OAIS reference model provides the basic archiving workflow. However, it does not articulate clearly how it can cater for distributed archiving architectures[36] . Cloud and grid 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 the joint service architecture has been undertaken by a few recent initiatives, but is only in developing phases.

> ***It is in the area of conceptualising a preservation services architecture that the DCH-RP project will need to advance the current state of the art and develop its own vision for a Grid-based preservation services architecture for the cultural heritage sector.***

4.1 Preservation Services Architecture for e-Infrastructure – Recent Examples

Some recent examples of preservation services architectures on e-Infrastructures come from EU-funded projects and joint brainstorming events between preservationists and cloud service providers.

4.1.1 Preservation Services on PAAS Cloud

A recent Dagstuhl seminar, "*Is the future of preservation cloudy?*"[37], held in November 2012 in Schloss Dagstuhl, Germany, held a brainstorming session which looked into the core PaaS services necessary for digital preservation on the cloud. As the necessary services it identified:

- Ingest services:
  - Add;
  - Graph exploration;
  - Object analysis (characterization of properties);
- Curate services:
  - Integrity checking;
  - Dereferencing and delete;
  - Migration and MoveOut;
  - Export;
  - Conversion and Transformation;
  - Administering retention;
  - Periodic sampling;
- Access services:
  - List items;
  - Find items;
  - Retrieve items;
  - Emulate;

[36] *Note for further elaboration: There is some discussion of it in http://www.metaarchive.org/GDDP*

[37] http://www.dagstuhl.de/en/program/calendar/semhp/?semnr=12472

– Administration of access.

The brainstorming also came with a list of orthogonal services, featuring triggered events (updates and removals), AAA (Authorisation, Authentication and Accounting), subscription plan, and reporting to users.

4.1.2 SCIDIP-ES project

An example of an e-Infrastructure that supports preservation of research is the SCIDIP-ES project[38] . Its architectural solution relies on a cloud infrastructure, but most importantly it has conceptualised a number of preservation services that are required for successful archiving of data[39]:

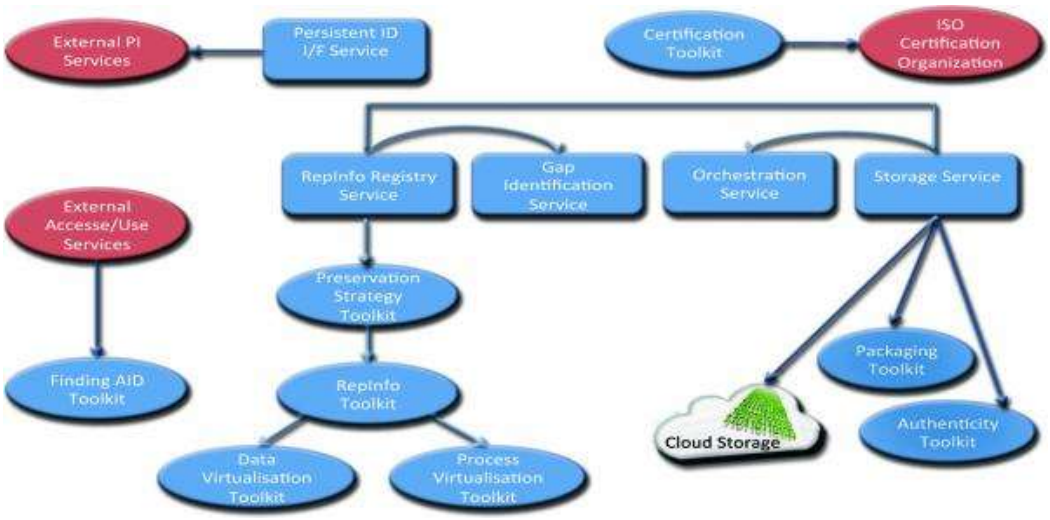

Figure 16. SCIDIP-ES general architecture

This overview of the SCIDIP-ES services and toolkits is orchestrated by the RepInfo Registry service, which helps capturing knowledge related to preserved digital objects. The Storage Service not only provides physical storage but also helps the consumers to access data they need. The Orchestration Manager is a complex notification system, allowing Data Managers – as well as users or at least other software components – to be immediately notified when something related to preserved data happens. Toolkits are software elements which are more domain-related and add specific functionalities to the system. The Service and Toolkits are not intended as a full and closed system. They have to be thought of as a set of functionalities that can work together but can also be used separately.

4.1.3 EUDAT project

The architecture of the EUDAT project integrates various infrastructures with vast amounts of research data, and adds services for curation and trust in addition to the interface to users (see Figure 17 below).

[38] http://www.scidip-es.eu/preservation/overview/
[39] http://jenkins.scidip-es.eu/joomla/index.php/documentation

This architecture illustrates a process that will have to be accommodated in future by most preservation work – solutions for preservation and curation can be used to support multiple different infrastructures. This is similar to the scaling in accessibility where aggregation of resources became a 'must' in order to offer better research discovery experience. The need in preservation is not completely identical – while in the domain of accessibility aggregation caters for better user experience, 'aggregated' preservation will facilitate not as much the end users, as the institutions.

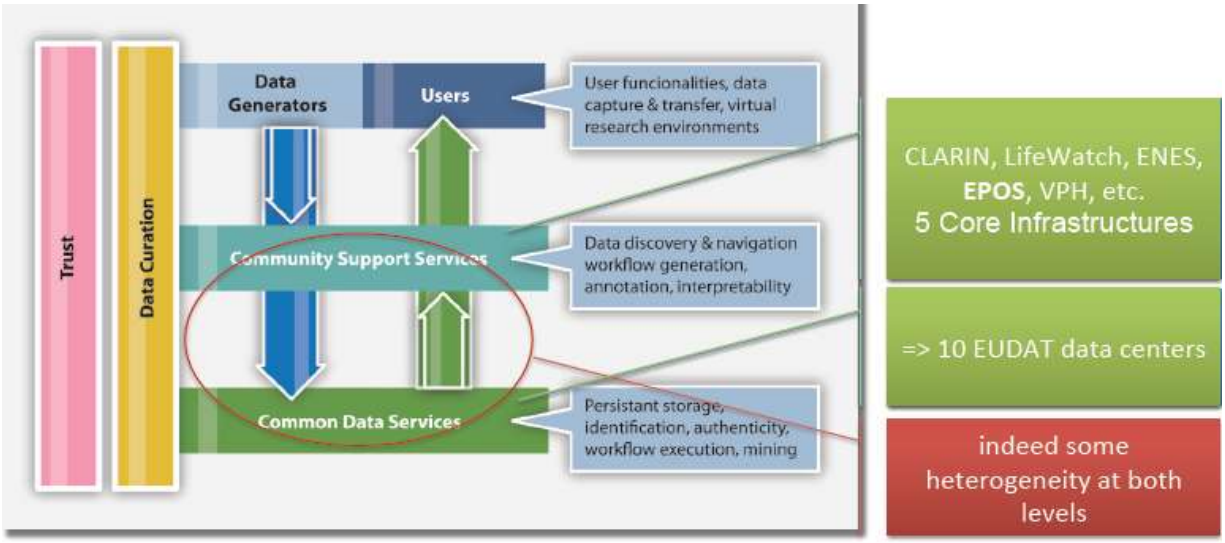

Figure 17. EUDAT high-level architecture

5. "Mind the Gap"

When bringing together two "worlds" – digital preservation software tools and e-Infrastructure services – some disjoint is inevitable such as, for example, incompatibility of purposes or scope, lack of technical or semantic interoperability, reliance on different standards, jurisdictional and legal barriers, etc. The two reports – the original DC-Net project study of digital preservation tools and services and the current baseline report for the preservation roadmap – have looked at the emerging digital preservation infrastructure from the point of view of digital preservation. The e-Infrastructures' view on developing support for preservation has been represented only through existing projects. Hence, the gap analysis is, at this stage, somewhat one-sided and only based on analysis of tools and services. Further aspects of preservation e-Infrastructure interoperability will be provided by upcoming deliverables from WP3 on standards and WP4 on trust and authentication.

The gaps here refer to areas:

- Where insufficient provision of tools and/or services exist to enable the integration of preservation and e-Infrastructures;
- Where examples have not yet been (sufficiently) developed to establish best practice or a consensus on an efficient solution;
- That have not been studied and/or piloted in sufficient detail to expose risks and point to significant shortcomings.

The gaps identified in the two reports on preservation tools can be summarised as:

- Although examples of distributed preservation solutions are now becoming more common, there is an apparent lack of a reference model, architectural design or best practice that the community has agreed on for implementing distributed preservation solutions.

  There is a need for a vision of a distributed digital preservation architecture that relies on e-Infrastructures.

- 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.

  There is a need for a registry of preservation services with clearly applied metrics, which makes tools easy to compare[40].

[40] While a number of digital preservation tools registries/collections are already in place, there is no such collection addressing grid and cloud services. Furthermore, the metrics applied to tools are not always very clearly explained. Comparing tools could only be done if the criteria for comparison are clear, hence the need in elaboration on metrics.

- With an ever-broadening range of preservation software tools available, institutions can now 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 necessary conditions need to be established under which the various services can coexist and be orchestrated into a healthy digital preservation ecosystem.

- The current examples of integrating preservation workflows with e-Infrastructures require significant levels of computing and IT expertise that both are not readily available in the majority of cultural heritage institutions.

  The solutions developed need to be tested for their simplicity of installation, management and use.

In order to facilitate further discussion on the proofs of concept to be developed in WP5, we provide a few hypothetical but typical scenarios that illustrate the range of requirements cultural institutions may have today. These scenarios or service levels could be piloted by the project in WP5 and further down the line during the roadmap implementation.

Scenario 1. Using specialised research tools from a digital humanities e-Infrastructure on material preserved in-house

A major memory institution in France, which has its own development team, is gradually implementing a solution for digital preservation. It is using local in-house storage. The institution participates in projects that aggregate content to Europeana and regularly uses social media channels to engage with the wider public. Thus, the access to its digital collections is either possible through the institutional website, or resource discovery is made via specialised portals and social media which in fact redirect the users to the institutional web server. Recently, it has happened several times that researchers ask to use specialised document analysis tools that are available through an e-Infrastructure. This raises issues of sharing content outside the institutional storage and preservation facilities on the cloud used by the e-Infrastructure, or the use of 'external' tools for processing locally stored documents. Both options raise concerns and, for the time being, there is no good solution for the end users.

Scenario 2. Integrating a new tool into an existing institutional infrastructure

A major memory institution in Germany had already developed its own preservation infrastructure. A new research project is asking for a newly developed software tool that would save time on checking file formats. However, the integration of this tool with the existing preservation solution cannot compromise any essential preservation features implemented in the local preservation system. The requirement is to analyse the difference that using the new tool will make and how to embed it with other components already in place; or how to run the new tool from a cloud-based provider and integrate this service with the existing preservation solution.

Scenario 3. Selecting a digital preservation solution in the case of an institution with only voluntary IT support

A little museum in Malta has a historical library and a digitised personal archive collection. The museum has staff of only 9 and only voluntary IT support. The director of the museum is aware of the need to organise digital preservation for the digitised documents, but is not sure how to do it. He receives periodically offers for long-term storage of digital content, but finds it difficult to select or to make a decision. He has practically no IT competence to rely on for decision-making, but is convinced that the decision should be forward-looking and accommodate the needs of the museum for the next 5 years.

Scenario 4. Preservation from a consortium of collections on the cloud

A specialised consortium of several institutions working on a complete digital repository of the works of a modern digital artist who worked and exhibited in 15 different countries has to resolve the issue of preservation of objects that are stored in different location. The works of the digital artist include a variety of digital formats as well as especially developed software tools. The curator of the collection has to identify a cost efficient solution that would also be suitable to store the complex objects in the collection. An additional difficulty is that the copyrights on the objects differ in the countries of origin of the objects.

Scenario 5. Preserving a 3D visualisation

A research lab in the UK is collaborating with an archaeological site in Italy to create a 3D visualisation of an ancient building. The visualisation is used as scientific documentation. Both institutions have to agree who will take care for the preservation in usable state of the model. There is also an issue of interoperability of the model with a free visualisation tool which can be used to show the model on a web site which is resolved producing a lower quality visualisation in an additional format. There is an on-going discussion whether it also needs to be preserved and by whom.

The next chapter will start fleshing out what the DCH-RP roadmap will need to take into account.

6. Designing a Preservation Infrastructure Roadmap for Digital Cultural Heritage

The main goal of the DCH-RP project is to design a roadmap for developing a preservation infrastructure for the cultural heritage sector. Work Package 3, that is responsible for this deliverable and the subsequent compilation of the roadmap, will need to integrate a multitude of viewpoints and aspects. This chapter provides a framework and sets out a preliminary action plan for the development of the roadmap.

6.1 Background to WP3 work

The DC-Net report on digital preservation tools and services provided a preliminary analysis of the domains that should be included in the design of a preservation roadmap (see Figure 18):

- Business change;
- Policy framework, and
- Better tools.

With the major PEST factors as the external context the roadmap has to consider. The necessary future steps are grouped in three phases: preparatory, development, and deployment & monitoring.

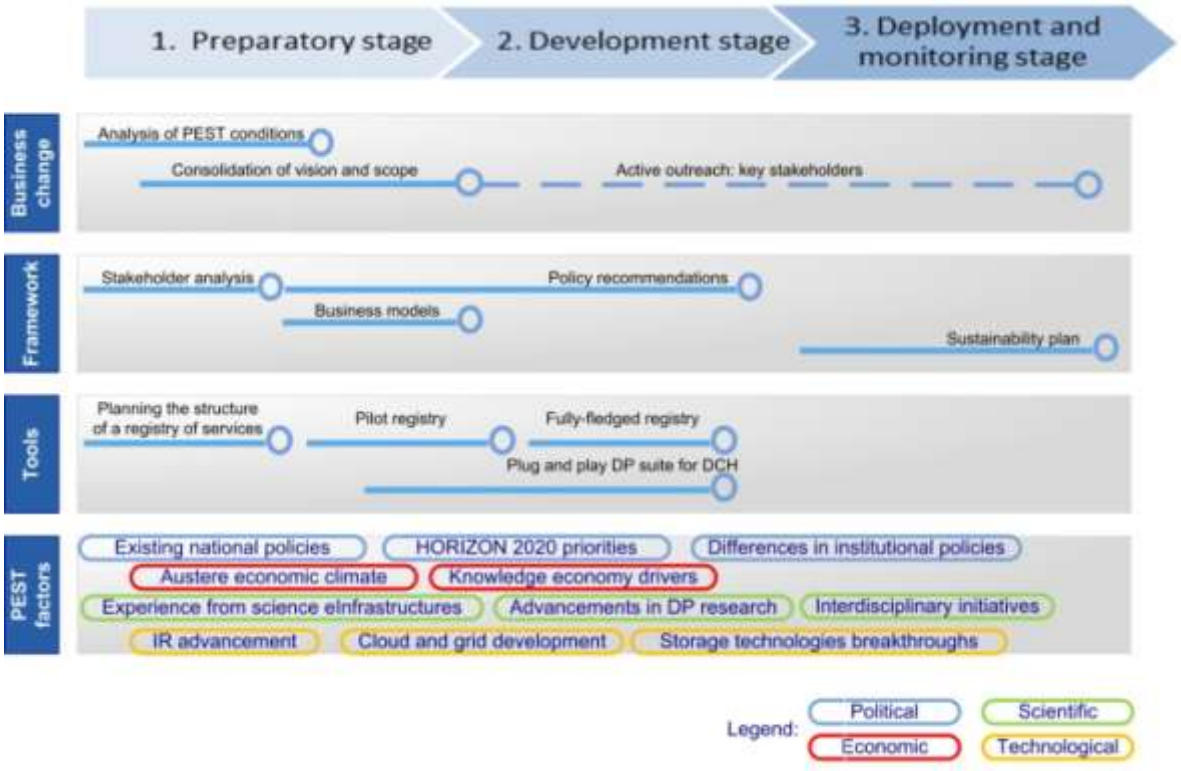

Figure 18. Roadmap - digital preservation services for CH collections from [Ruusalepp, Dobreva 2012]

The DCH-RP project will further develop the vision about what a roadmap should address taking into account the context of e-Infrastructures and the previous experience of implementing preservation in

grid and cloud in other domains. The essential first task in DCH-RP project is to select a suitable methodology, which allows setting realistic objectives and achieving credible results.

The work on the Roadmap is closely connected with Work Package 5 (Proof of Concept). The connection between WP3 and WP5 is bi-directional: WP3 is being informed by work done in WP5 but also provides feedback for its further development. Work Package 3 is also strongly connected to WP4 (Case Studies and Best Practice) and will draw examples from WP4 in the future.

The work on Proof of Concept (WP5) provides the basic framework for the look into the future within DCH-RP. It establishes short, medium, and long-term milestones in 2014, 2016 and 2018. These will also be adopted into the roadmap document.

6.2 What should be addressed in the roadmap?

In order to be practical and to fit the aims of the project, we looked into the following questions:

6.2.1 What sources should be consulted?

The effort to design a roadmap brings together several sources of information:

- Existing reference models in the domain of digital preservation;
- Existing preservation tools and services;
- Projects addressing preservation in grid and cloud environments;
- E-Infrastructure projects and services.

6.2.2 What types of analysis should be done?

The aim of the roadmap exercise is to produce an instrument that will facilitate policy makers and management within cultural heritage institutions. To achieve this, the roadmap should concentrate on at least four areas, aligned with the action plan (D5.1), which identify the policy domains that require intervention:

- Harmonisation of data storage and preservation – which would allow to integrate in common environments the curation of research data with other digital objects – two domains which are currently addressed separately;
- Progress of inter-organisational communication – including 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 where digitisation is made without taking into account needs for preservation, and accessibility online is disjointed from preservation);
- Establishment of conditions for cross-sector integration - as a key condition for maximising the efficiency of successful solutions, transferring knowledge and know-how, and
- Governance models for infrastructure integration – as a necessary condition for successful institutional participation in larger eInfrastructure initiatives, and aggregation and re-use of 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 there will be a set of actions we suggest to undertake. These actions are addressing two situations:

- Consolidating knowledge and expertise from different partner institutions (i.e., a process of synthesis of a range of on-going experiences) and
- Identifying gaps and areas that have not been properly addressed (i.e., a process of analysis of lack in provision).

6.2.3 Deciding a timeframe for actions

The roadmap should allow for the definition of a practical action plan with a realistic timeframe for the implementation of its stages. The short-term action plan is addressed by the DCH-RP project that should initiate the development of a preservation services infrastructure to a level that will be self-sustainable and continue to progress on its own.

Two further time spans should be considered: medium term (2016, i.e. two years after the end of DCH-RP), and long term (2018 and beyond) for logical continuation of the DCH-RP work.

To connect the major areas of work and the time-line, a matrix structure is proposed which can be populated during the roadmap design work of the DCH-RP (see Figure 19).

Long term (2018 and beyond)				
Medium term (2016)				
Short term (2014)				
	Harmonisation of data storage and preservation	Progress for inter-organisational communication	Establishment of conditions for cross-sector integration	Governance models for infrastructure integration

Figure 19. Structure of the roadmap matrix

When the detailed analysis is done at later stages of the project, it would also be useful to look at the possible preservation services architecture by addressing services according to their

- Functional area (pre-ingest, ingest, archival storage, preservation planning, data management, access and reuse);
- Their type (microservices, services), objects addressed (files, bitstreams);
- Type of architecture (addressing a simple well defined task, combined; grid-, and cloud-oriented);
- Level of maturity;
- Licensing conditions.

These could be documented in a registry of tools thus helping users of the project outcome with the gathered information. While various initiatives already gathered information on tools and services relevant to digital preservation (see Ruusalepp and Dobreva 2012), there is still a significant difference in the descriptions of those tools. A uniform approach aiming to offer a consistent quality of descriptions would be definitely recommended.

6.2.4 Drafting the short-term actions

A priority should be engaging relevant communities that will continue the work of DCH-RP for the cultural heritage sector beyond the project.

The major logic for the choice of actions for the short term is thus:

- What actions are already prepared/expected;
- What actions represent activities which hinder further development and will have negative effect on the economic wellbeing of CHI, if not properly supported;
- What actions are feasible to be implemented with the existing project resources (for the short term);
- What actions need to be implemented in order to guarantee sustainability of the project outcomes;
- What actions could be supported through the planned sustainability measures (for the medium and long term).

7. Draft Action Plan for WP3

The following matrix consolidates a proposal for actions across the four areas of work outlined above over their short, medium and long-term lifespan.

	Harmonisation of data storage and preservation	Progress for inter-organisational communication	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
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	Consolidate mature requirements for preservation in the DCH environment	Implement tools in selected eInfrastructures 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 20. Suggestion for a roadmap coordinated with the Proof of Concept of DCH-RP project

References

Aitken, B., McCann, P., McHugh, A., and Miller, K. (2012) Digital Curation and the Cloud Final Report. Produced by the DCC for JISC's Curation in the Cloud Workshop, Hallam Conference Centre, March 2012

Anderson, D., Delve, J., Dobreva, M., and Konstantelos, L. (Series Editors) (2012) The Preservation of Complex Objects (Volume 2): Software Art. http://www.pocos.org/index.php/publications/publications

Antunes, G., Barateiro, J., Becker, C., Borbinha, J., Proença, D., Vieira, R. (2012) SHAMAN reference architecture. http://shaman-ip.eu/sites/default/files/SHAMAN-REFERENCE%20ARCHITECTURE-Final%20Version_0.pdf

Ardizzone, V., Barbera, R., Calanducci, A., Fargetta, M., Ingrà, E. , Porro, I., La Rocca, G., Monforte, S., Ricceri, R., Rotondo R., Scardaci, D., Schenone, A. (2012) The DECIDE Science Gateway, *J Grid Computing*. Vol. 10, pp. 689–707. http://link.springer.com/content/pdf/10.1007%2Fs10723-012-9242-3

Barbera, R., Calanducci, A., Mantovani, M.-L., Tomassini, S., Tanlongo, F., Paolini, G. (2012) D 4.2. Pilot on e-Collaborative digital archives, INDICATE project. http://www.indicate-project.eu/index.php?en/176/documents-and-deliverables

Birrel, D. et al. (2010) SHAMAN deliverable D14.2 – Report on demonstration and evaluation activity in the domain of "memory institutions". http://shaman-ip.eu/sites/default/files/SHAMAN%20D14.2_Report%20on%20Demonstration%20and%20Evaluation%20activity%20in%20the%20domain%20on%20MI_0.pdf

Faria, L., Becker, C., Petrov, P., Duretec, K., Ferreira, M., and Ramalho, J. (2012) Design and architecture of a novel preservation watch system, In: Hsin-Hsi Chen, Gobinda Chowdhury (Eds.): The Outreach of Digital Libraries: A Globalized Resource Network - 14th International Conference on Asia-Pacific Digital Libraries, ICADL 2012, Taipei, Taiwan, November 12-15, 2012. Proceedings. Springer LNCS 7634, pp. 168-178

Galushka M. (2012). TIMBUS project WP 5 – Software Architecture for Digital Preservation. http://timbusproject.net/resources/publications/public-project-deliverables

Moore, R., and Smith, M. (2007) Automated Validation of Trusted Digital Repository Assessment Criteria, *Journal of Digital Information*, Vol. 8, No. 2 (2007)

Nicholson, D., Dobreva, M. (2009) Beyond OAIS: towards a reliable and consistent digital preservation implementation framework. In: 16th International Conference on Digital Signal Processing (DSP 2009), Santorini, Greece.

Ruusalepp, R., Dobreva, M. (2012) Digital Preservation Services: State of the Art Analysis. http://www.dc-net.org/getFile.php?id=467

Teruggi, D., and Ranzuglia, G. (2012), Draft Roadmap of Future Grand Challenges, Deliverable D3.4, DigiBIC project. http://www.digibic.eu/documents.asp?slevel=0z84z111&parent_id=111

Abbreviations

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
DC-NET	Digital Cultural heritage NETwork
DP	Digital preservation
OAIS	Open Archival Information System
PAAS	Platform as a service
PEST	Political, Economic, Scientific, Technological
PSNC	Poznań Supercomputing and Networking Center
SCAPE	SCAlable Preservation Environments
TIMBUS	Timeless Business Processes and Services
UCL	University College London