DCH-RP and PREFORMA: Two Case Studies on the Digital Preservation of Cultural Heritage

Antonella Fresa, Claudio Prandoni

Promoter Srl, via Boccioni, 2, 56037, Peccioli (PI), Italy
{fresa, prandoni}@promoter.it

Abstract. The huge amount of digital cultural heritage data is growing every day thanks to the digitisation programmes carried out by cultural institutions. These data should continue to be available in the future to anyone who wishes or needs to access them. This is a very important challenge memory institutions have to face nowadays, and several questions need to be answered to this regard: How will we ensure the long-term preservation and access to our digital information? Who should select which digital data should be saved? How will we successfully store and migrate data as technology evolves? Understanding of digital preservation requires consideration of two main aspects: organisation and technology. Several projects have been funded in the last years to work on these topics by the European Commission. This paper presents the work that is being carried out in two of these projects: DCH-RP and PREFORMA.

Keywords: digital cultural heritage, digital preservation, roadmap, e-infrastructures, standard file format, conformance checking.

1 Introduction: Digital Cultural Heritage and Digital Preservation

Since early 2000s, a wide range of activities was carried out by the European Member States in order to accept the challenge of driving the European cultural heritage through the digital age.

On one side, memory institutions feel the unavoidable need of digitising their content, both for preserving it in a digital format and for granting and enlarging the access to them by researchers, students and citizens. It is esteemed that only a very small part of the European cultural heritage has been digitised until now [1].

On the other side, our society is like never before accumulating a huge amount of digital-born material (result data from the research, materials' analysis, digital art, bibliographies and so on). The digital-born heritage is therefore adding data and content to the digitization process output.

As a consequence of the above, the volume of digital cultural heritage data is incredibly growing year after year, so that it became necessary to reflect upon the tools which permit to manage such a huge amount of data in an efficient and selective way,

in order to make the data available to the researchers and the citizens in a European dimension, and towards a global dimension too [2].

In addition, the digital cultural heritage (DCH) sector has the challenge of the complexity of the information itself, due to the relationships that each cultural object has with its collection, with the memory institutions where it is held, with the other objects of the same nature and/or culturally interconnected, etc.

The importance of long-term preservation and its complementarities to digitisation efforts was highlighted in 'The New Renaissance', the report of the Comité des Sages appointed by the European Commission in 2010 that clearly stated the digital preservation as a mandate of the memory institutions that are engaging with the digitisation of cultural heritage in Europe [3].

Understanding of digital preservation requires consideration of two main aspects: organisation and technology.

On the organisational side, the problem of preserving digital cultural heritage content is common to all the cultural institutions in Europe and beyond but it is at the moment, addressed by individual projects, without really any shared approaches. Often the same problems are studied repeatedly and successful solutions are unknown by others working on the same issues. The preservation of DCH is an ongoing action, to be periodically revised, in order to update data sets and metadata formats. These are time consuming activities, in particular if carried out independently by each cultural institution. 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 (namely DCH) data which is stated as the priority for the global knowledge society. To this regard, the report of the Comité des Sages indicates the urgency of a Digital Preservation infrastructure to contribute to cost reduction.

On the technological side, memory institutions working with digital preservation currently use several strategies: techno-centric, incremental, analytical, durable digital objects, etc. [4]. Of these strategies, the migration has been the dominant one. To avoid technical obsolescence, the digital objects are converted to new standardised file formats as technology changes. These conversions are expected to be done without information losses. For this reason, it is of vital importance to fully control the file formats at ingestion time. Although many software tools exist to carry out preservation tasks, their support status, quality, reliability, etc. are not under the control of the content owner and are too uncertain and unproven to be fully trusted and implemented in digital preservation workflows. Memory institutions need to be sure that what is produced according to a standard, tested for conformity, and (if needed) re-processed for corrections, happens within an iteration that is under full control of the institutions themselves or of the curators that are in charge of managing and preserving the electronic documents and other media content in the long term. This in particularly valid for digitisation projects and institutional archives, where archivists may have more input into record creation/management.

This paper presents the work that is being carried out in two EU-funded projects: DCH-RP (Digital Cultural Heritage – Roadmap for Preservation) and PREFORMA

(PREservation FORMAts for culture information/e-archives). DCH-RP [1] is mostly focused on the planning and organisational aspects. Its main outcome is the definition of a roadmap for the implementation of a federated infrastructure for the digital preservation of DCH and more in general dedicated to support the application of open science in the arts and the humanities. PREFORMA[2] focuses instead on a technological challenge: the implementation of an open source conformance checker that verifies whether the file to be ingested in the digital archives have been produced according to the specifications of a standard file format and that this process is under the direct control of the memory institutions.

2 Roadmaps, Digital Preservation and E-Infrastructures

Roadmaps are useful instruments 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[3] which presents environmental, policy, preservation, presentation, standards, skills, and communication developments over time. The APARSEN project roadmap[4] presents research topics and larger themes; preservation services are a research topic under the theme of sustainability. Some projects use roadmaps to present various formats, e.g. the PrestoSpace[5] 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.[6]

The Long Term Preservation Case Study conducted by the International Network for a Digital Cultural Heritage e-Infrastructure project (INDICATE)[7] highlighted that cultural institutions implement their own methods to address the digital preservation issues and that there is a strong need for common policies and strategies; that institutions and their staff are mostly unaware of the e-infrastructure possibilities, in some cases they are not even connected to the e-infrastructure providers [5]. The continuing investment in in-house preservation systems is contributing to the lack of interoperability and fragmentation of resources into 'digital silos'. This stand-alone service architecture cannot be sustained over the longer term as further fragmentation does 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 this context, DCH-RP project continued the work

1 http://www.dch-rp.eu/

2 http://www.preforma-project.eu/

3 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.

4 http://www.alliancepermanentaccess.org/index.php/aparsen/

5 http://www.preservationguide.co.uk/RDWiki/pmwiki.php

6 See for example the Danish roadmap for RI http://en.fi.dk/publications/2011/danish-roadmap-for-research-infrastructure-2011/uk-roadmap.pdf

7 http://www.indicate-project.eu/

started with Digital Cultural heritage NETwork (DC-NET)[8] and INDICATE towards the establishment of a roadmap for the digital preservation of cultural heritage data making use of e-infrastructure facilities.

A shared implementation of common e-Infrastructure layers will be beneficial and cost effective for the DCH sector, as it can:

- Be an efficient and cost-effective channel for the delivery of advanced services in the DCH sector (e.g. advanced visualisation services of 3D objects, implementation of virtual environments, exploring data resources by the use of semantic links automatically generated, etc.).
- Realise data infrastructure more quickly, at international scale.
- Contribute to data standardisation, which is a pre-requisite for interoperability and preservation.
- Allow for cost reduction. The researchers can avoid duplication of efforts by sharing results of similar activities that require processing high volume of data and using high performance computing. Also digitisation workflow, that is currently mostly based on human intervention, can be substituted by cheaper machine processes, using, for example data mining on textual and visual resources for automatic extraction of metadata.
- Embed scalability of the resources which is a difficult task for cultural institutions that are not used and not skilled to do it.

2.1 The DCH-RP Roadmap

The DCH-RP project, supported by the e-Infrastructure Capacities programme of the 7th Framework Programme for Research and Technologies Development of the European Commission, has lasted for two years, from October 2012 until September 2014, with the participation and factual contributions of thirteen partners from eight European countries.

Aim of the project was to develop a roadmap to implement a preservation infrastructure for digital cultural heritage. The roadmap has been supplemented by practical tools for decision makers, including a registry of services and tools and commented list of best practices, and it has been validated through a range of proofs of concept, where content owners and e-Infrastructure providers worked together on concrete experiments.

The DCH-RP roadmap exercise aimed to produce an instrument to facilitate the work of policy makers as well as management activities within cultural heritage institutions. To achieve this, the roadmap focused on four areas which identify the policy domains that require intervention [6]:

1. *Harmonisation of data storage and preservation*: allows integrating in common environments the curation of research data with other digital objects – two domains which are currently addressed separately.

[8] http://www.dc-net.org/

2. *Improved interoperability*: includes better integration of preservation practices within the overall workflow for digitisation and online access.
3. *Establishment of conditions for cross-sector integration*: is a key condition for maximising the efficiency of successful solutions, transferring knowledge and know-how.
4. *Governance models for infrastructure integration*: is a necessary condition for successful institutional participation in larger e-Infrastructure initiatives, for the aggregation and re-use of digital resources.

The DCH-RP roadmap makes 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* is proposed 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. Beyond the duration of the DCH-RP project, the roadmap identifies two further planning periods: *medium-term* (approximately two years after the end of DCH-RP) and *long-term* for the logical continuation of the DCH-RP work into the future.

The following pictures present a condensed version of the Roadmap, highlighting the main steps and actions to be taken in the short, medium and long term.

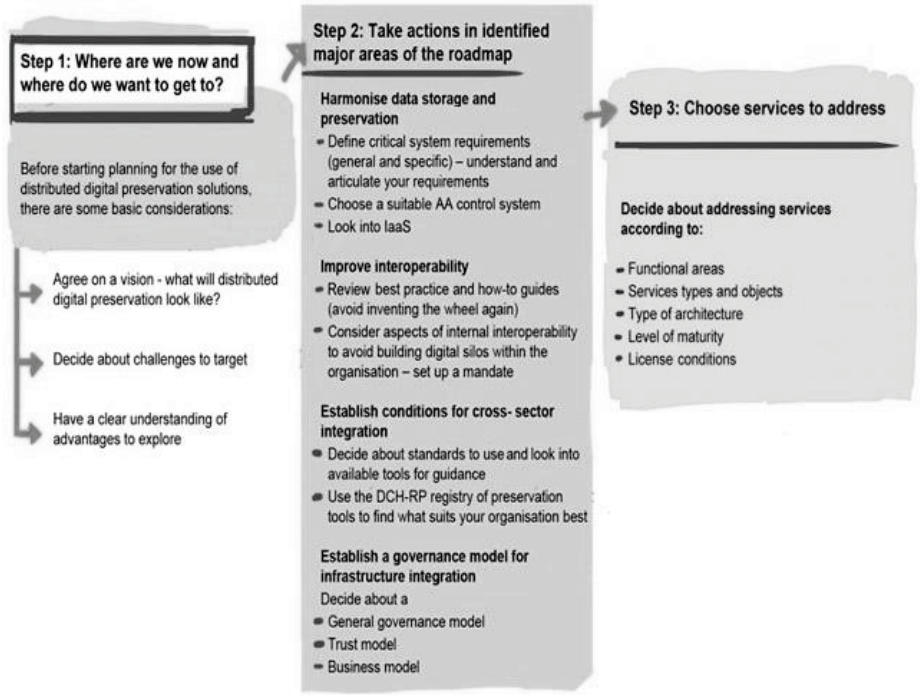

Fig. 1. The Condensed version of the intermediate roadmap – short-term

When the short-term steps are completed, the following will be achieved:

- A clear view of both the actual situation and the main goals for using distributed digital preservation services.
- An initial set of functional as well as technical and administrative requirements.
- An idea about the services to be addressed (types, objects, architecture, etc.).

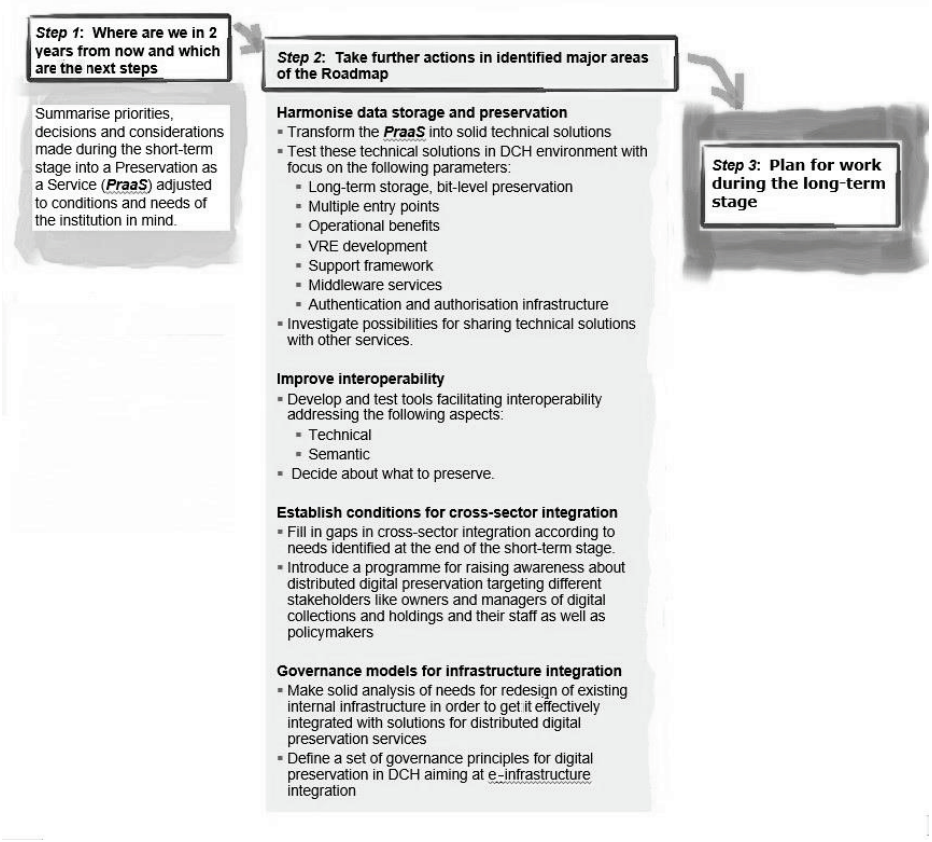

Fig. 2. The Condensed version of the roadmap – medium-term

When the medium-term steps are completed, the following will be achieved:

- A Preservation as a Service (PraaS) plan and tests of technically feasible solutions; possible e-Infrastructures should also have been identified as appropriate for distributed digital preservation.
- Decisions about the digital resources on which to apply distributed digital preservation services.
- Preparation of the internal organisation through awareness/training programmes, adaption of internal infrastructure, and decisions on governance principles.

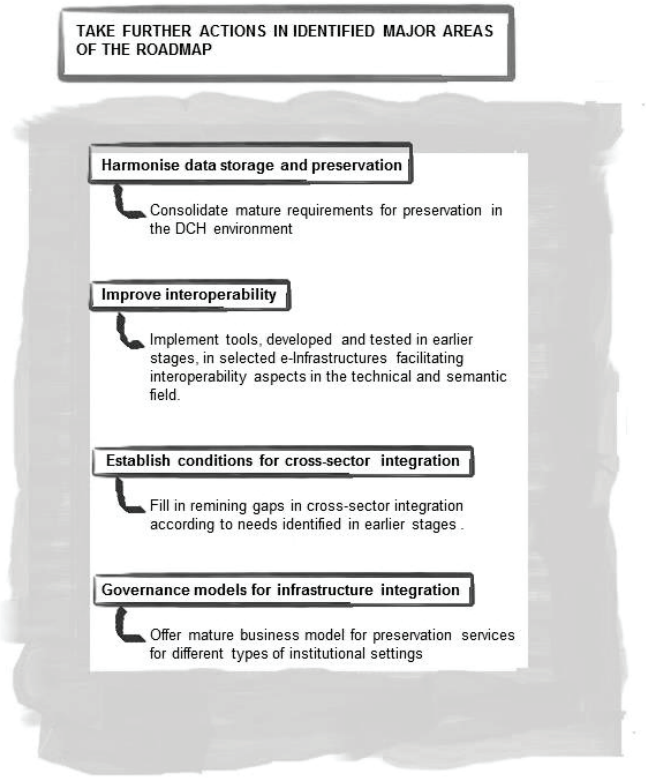

Fig. 3. The Condensed version of the roadmap – long-term

At the end of the last step, the selected e-Infrastructures tools and services that have been developed and tested in the earlier stages should be implemented.

Results from the work conducted in DCH-RP shows that the two basic assumptions on which the roadmap is built are achievable:

- Firstly, existing e-Infrastructures for research and academia are also efficient channels for the digital cultural heritage sector to use them for distributed digital preservation.
- Secondly, 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.

However, a ground breaking part of the concept is the possibility of customising the services provided by e-Infrastructures, i.e. tailoring the service portfolio and charac-

teristics to the actual preservation tasks and requirements. Even if the e-Infrastructure resources seem to be already able today to 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.

In order to maintain the roadmap as a living document, to be updated and improved as time passes, technology changes, new requirements emerge, etc., the DCH-RP project created a dedicated web-space where it is possible to download the latest version of the roadmap as a handbook and to provide feedback and comments. This web-space is hosted as a section of the DCH-RP showcase in Digital Meets Culture[9].

3 File Format Conformance Checking

Memory institutions are facing increasing transfers of electronic documents and other media content for long-term preservation. Preservation models are often inspired by standards, such as the Open Archival Information System (OAIS) reference model[10], where transfers and preservation are built on information packages containing both data and metadata. Metadata are normally stored or exported in XML and specified in different schemas. XML is a stable and easily accessible standard, and the schema specifications, like METS, PREMIS, EAD, are controlled by the community of professional curators in digital preservation through different international boards and committees. The data content, on the other hand, is normally stored in specific file formats for documents, images, sound, video, etc., depending on the originating system. These files are usually produced by software from different vendors. However, even if the transferred files with data content are in standard formats, the implementation of these standards cannot be guaranteed. The main reason is that software used for implementation of standards for producing electronic files is not controlled neither by the institution that produces them, nor by the memory institution that holds the archive. As a result, memory institutions have to make conformance tests before accepting transfers of electronic collections, to verify that they have been produced according to the specifications of a standard file format, and hence that they match the acceptance criteria for long-term preservation established by the memory institution. However, the software used to perform these tests are, in turn, not controlled by the institution [7].

This situation, when conformance to standards is not guaranteed, may result in increasing costs. Furthermore, this poses problems for the long-term preservation, since data objects passing through an uncontrolled degenerative process can jeopardise the whole preservation exercise. Migration of data files can for example be more or less impossible to carry out with the authenticity and integrity of the files still in place. This is particularly true in the case of born-digital archives and is thus the reason why preservation of these objects is so difficult.

---

[9] http://www.digitalmeetsculture.net/heritage-showcases/dch-rp/dch-rp-roadmap-for-preservation

[10] http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=57284

3.1 The PREFORMA Solution

PREFORMA is a Pre-Commercial Procurement (PCP) project co-funded by the European Commission, under its FP7-ICT Programme. The project started on January 1st, 2014 and it will last 4 years, until the end of December 2017.

The consortium of PREFORMA is composed by 14 partners, from 9 European countries, ranging from the North to the South of the EU and including memory institutions, research centres, universities and SMEs.

A pre-commercial-procurement has been launched by the project for the development and deployment of an open source software licensed reference implementation for file format standards, aimed at any memory institution (or other organisation with a preservation task) wishing to check conformance with a specific standard. Six suppliers have been selected (two for each media type) to carry on the first design phase that lasted 4 months and completed with the ranking of the delivered designs. The suppliers of the best three designs (one for each domain) were then requested to proceed to the prototyping phase, which includes two releases and the re-design phase and which lasts until December 2016.

The research objective underlying the work of PREFORMA is to explore critical factors in the quality of standard implementation [8]. This involves acquiring knowledge about:

- How to establish a methodology or an objective frame of reference to interpret and implement the standard specifications against the background of the current variations of interpretations and implementations by software vendors.
- How to determine whether a file is what it claims to be, i.e., in this context, what makes a file a valid file, i.e., conform to the "standard".
- How to develop and sustain the open source project in the short and long run.

PREFORMA research and development activities aim to empower memory institutions to gain full control over the technical properties of preservation files.

The first activity is to develop an open-source toolset for conformance checking of digital files, intended for long-term preservation in memory institutions. The conformance checker has to:

- Verify whether a file has been produced according to the specifications of a standard file format.
- Verify whether a file matches the acceptance criteria for long-term preservation by the memory institution.
- Report in human and machine readable format which properties deviate from the standard specification and acceptance criteria.
- Perform automated fixes for simple deviations in the metadata of the preservation file, leaving the original bitstream untouched and created a correct copy of the object to be preserved.

The development of the conformance checker focuses on four use cases that facilitate the interaction between the supplier, academic research and memory institution. They

are compliant with the OAIS Reference Model and represent conformance checking procedures at different moments in the life cycle of a preservation file:

- Conformance Checking at *Creation Time*: Producers pro-actively check if technical properties of a file meet the acceptance criteria of an OAIS Archive, e.g. government agencies checking conformance of text documents to be deposited at public archives when the document is made available.
- Conformance Checking at *Transfer time*: Archives check the technical properties of files ingested in the OAIS Archive, assessing whether they meet the acceptance criteria for ingest and conformance to the relevant preservation file formats, e.g. libraries monitor the preservation status of digital publications deposited in their digital repository.
- Conformance Checking at *Digitization time*: Archives check the technical properties of digital representations of collection items, internally or externally produced, assessing whether they meet the requirements specified in the digitization tender, e.g. museums doing quality control on the digital representations and documentation, produced by photographers.
- Conformance Checking at *Migration time*: Archives check the technical properties of files that are repackaged or transcoded, following the rules defined in the preservation strategy of the OAIS Archive, e.g. libraries performing quality control when transcoding audiovisual files from a 'transmission' to a 'preservation' format.

The second activity is to establish a network of common interest in order to gain control over the technical properties of preservation files. This involves the adoption of a 'reference implementation' by other software applications, and continuous improvement of the 'standard' specification through engagement in the standardization process. The network gathers all stakeholders that control different stages in the lifecycle of a preservation file, providing a sustainable and viable ecosystem for the deployment of tools developed by PREFORMA as well as tools adopting the reference implementation. These stakeholders include:

- *Developers*, controlling the production of preservation files, e.g. by file editors or transcoders, thus aiming at improving the effectiveness and interoperability of their software.
- *Digital preservationists*, controlling the acceptance and management of preservation files in digital repositories, thus aiming at improving the preservation status of the digital collection they maintain and the effectiveness of the ingest procedures.
- *Standardization bodies*, maintaining the formal specifications of file formats in standards, thus aiming to improve the specification of the standard.

Fundamental concepts for achieving these objectives are open file formats, open source software, and an open approach. Memory institutions heavily rely on open source components for building preservation workflows, since only software under an open license allows them to build tailor-made solutions and to avoid lock-in situations. PREFORMA aims for strengthening this approach by procuring an open source

validation platform that enables memory institutions to make a better assessment of the digital objects they acquire and facilitates the improvement of the preservation infrastructure they already have in place.

An Open Source Portal[11] has been implemented to provide an overview and references to each open source project that is currently working in the prototyping phase. It acts as an entry point for all interested suppliers and memory institutions allowing easy navigation to all externally hosted resources.

4 Conclusions

In the last decade, cultural institutions started to move their cultural content on the digital world. This implied to establish new instruments, new rules of access, new standards to exhibit digital heritage, new ways of communication.

Consequently, the issue of preserving this content is becoming very urgent: preserving DCH is a priority as it is for the tangible heritage. The cultural institutions should urgently migrate from existing practices to appropriate preservation practices and the use of policy-based approaches can make this possible. National policies about cultural preservation need to re-use best practices, to share solutions and to avoid duplication of efforts, in particular in the current context of economic crisis where the resources available are often smaller.

PREFORMA aims to progress in this direction, providing the necessary middleware and application services through the development of new technologies and the integration of these technologies into innovative tools.

Our vision for the future is that, after all these tools have been developed, it is necessary to make them easily accessible and usable by the community. The deployment of these services in an e-Infrastructure dedicated to the preservation of DCH represents a challenging and effective area of future investigation and DCH-RP provided a roadmap for that. Deployment of this infrastructure will secure future access to the data, fostering mainstream usage of these digital assets; i.e. promoting a significant expansion of the number of potential users of preserved data across research fields, beyond the research territory and by future generations of researchers, economic players and society as a whole.

In addition to the technological area, two other domains are of extreme importance to establish an efficient preservation practice: (1) innovations around the internal workflows of the organisations operating in the DCH sector and (2) policy support for the implementation of federated infrastructures.

Internal workflows currently encountered among DCH players imply that a number of actions need to be taken by many memory institutions that are engaged in digital archiving, in order to make their digital resources accessible and usable in the long-term. Firstly, roles inside the organisation have to be re-organised to guarantee that preservation is a mandatory step in the process of handling their digital resources. Secondly, in order to create new skills and competences, cultural heritage practition-

[11] http://www.preforma-project.eu/open-source-portal.html

ers have to be trained in both understanding and the handling of the new conditions associated with digital preservation, i.e. the changing forms of artefacts and metadata, the changing methods of work, and the rapid changes in technology itself. Furthermore, decisions have to be taken about the procurement of services related to storage and computing resources, not just for storing the 'raw' data, but also to take into account preservation requirements. All these actions require time to be performed and financially resourced. Advocacy of the need for digital preservation is, therefore, another important action in order to create the conditions required for understanding, acceptance, and endorsement by decision makers.

Policy support is another fundamental component for the successful implementation of the federated infrastructure. Policy makers should be encouraged to endorse the DCH-RP roadmap in order to define policy and financial support to the federated infrastructure. The varying interests of different stakeholders should be taken into account and harmonised. On one hand, e-Infrastructure providers should convincingly acknowledge the potential represented by the emerging DCH sector as an area for investment, in the same way as for other scientific domains. On the other hand, the cultural heritage sector must find methods for establishing and communicating trust in distributed digital preservation services provided by the e-Infrastructures so they can be commonly regarded as safe, secure and trustworthy. All this requires changes in professional approaches that need to be supported by awareness raising activities, and eventually also reflected in operating procedures, regulations and legislation.

References

1. Stroeker N., Panteia R. V.: ENUMERATE survey report on digitisation in European cultural heritage institutions (2012)
2. Fresa A.: Digital Cultural Heritage Roadmap for Preservation. In: International Journal of Humanities and Arts Computing 8 (2014)
3. European Union, Comité des Sages: The New Renaissance, Report of the 'Comité des Sages' Reflection group on bringing Europe's Cultural Heritage online (2010)
4. Dobreva, M., Ruusalepp, R.: Digital preservation: interoperability ad modum. Chowdhury, G.G.; Foo, Schubert (Toim.). In: Digital libraries and information access. Research perspectives, pp. 193 – 215 (2012)
5. Aydinonat B., Barbera R., Fernandez S., Fresa A., Khoury A., Kollias S., Ozluk H.K., López M., Messina M., Piccininno M., Tomassini S., Zakrajsek F.: INDICATE case study report – long term preservation (2011)
6. Justrell B., Fresa A.: Roadmap for the preservation of digital cultural heritage (2014)
7. Fresa A., Justrell B., Prandoni C.: Digital curation and quality standards for memory institutions: PREFORMA research project. In: Archival Science - Special Issue Digital curation (2015)
8. Lemmens B., Elfner P., Lundell B., Prandoni C.: PREFORMA tender specifications (2014)