# SURVEY ON THE FIRST PROOF OF CONCEPTS

## Introduction to DCH-RP project

DCH-RP (www.dch-rp.eu), started on the 1st of October 2012, is a coordination action co-funded by the European Union under its Seventh Framework Programme whose outcome is a roadmap for the implementation of a preservation infrastructure for DCH which will be the first instance of an Open Science Infrastructure for DCH in 2020[1].

Its main goals are:

- Prepare a Roadmap for Preservation (RP)
- Cover Digital Cultural Heritage (DCH) aspects
- Conduct a coordination action within EU FP7
- Think about implementing a federated e-Infrastructure

DCH-RP is articulated in five work packages (WP)

- WP1 - Project Management
- WP2 - Dissemination and sustainability
- WP3 - Preservation Roadmap
- WP4 - Case Studies and Best Practice
- WP5 - Proofs of concept

and it involves 13 partners:

- ICCU Italy
- RIKSARKIVET Sweden
- BELSPO Belgium
- EVK Estonia
- COLLECTIONS TRUST United Kingdom
- PROMOTER Italy
- EGI.eu the Netherlands
- INFN Italy
- PSNC Poland
- NIIFI Hungary
- EDItEUR United Kingdom
- TERENA the Netherlands
- MICHAEL CULTURE Belgium

## The Proofs of Concepts

A special task of the project is devoted in the execution of Proofs of Concept (PoCs). These PoCs are experiments where e-infrastructure providers and cultural institution work together to test the actual use of distributed computing and storage infrastructures to store and manage cultural digital resources, starting from real scenarios and use cases. The aim of the group involved in the Proofs of Concepts (PoC) is to validate in concrete experiments assumptions and concepts expressed in the DCH roadmap to preservation, that is the goal of the Preservation Roadmap working group and of the project in general.
The team members have been charged with coordinating and conducting Proofs of Concept (PoC) that will inform the results achieved by the members of the Preservation Roadmap working group with the outcomes of these PoC. The questions in need of answers were not known a priori, however, these questions can broadly fall into the following categories:

[1] For more details see http://www.dch-rp.eu/

- Functionality of the preservation infrastructure components
- Non-functional capabilities (e.g. scalability, reliability) of the examined e-Infrastructure
- Acceptance and usability of services
- Sustainability and technology insertion options into local and international DCH infrastructures

From a practical point of view, by these experiments team members tried to put in practice some scenarios developed by the Working Group in charge for designing the requirements for a preservation Roadmap for DCH (WP3). In short, the scenarios can be grouped as follows:

1. **Theme 1 – "Organisational challenges"**
    1. Use specialised DP tools on in-house data
    2. Integrating a new tool into existing infrastructure
    3. Select an existing DP solution at an institute with best effort IT support
    4. Preservation from a consortium of collections on the cloud
    5. Preserving a 3D visualisation
    6. Retrieve archived data
2. **Theme 2 – "End user concerns"**
    1. Researcher discovers a historical database
    2. Research and select a tool serving a specific purpose
    3. Accessing digitised content from schools
    4. Gain access to archived websites
3. **Theme 3 – "New services & infrastructure integration"**
    1. Proof of authenticity in distributed archiving
    2. Defining new services
    3. Integrating new services into existing infrastructure[2]

In this context, the e-Cultural Science Gateway, a standard-based web 2.0 demonstrative platform, developed by the INFN of Catania has been used to conduct some tests[3]. Other e-infrastructure services and tools are currently under evaluation to be possibly tested in the second run of the Proofs of Concept. Examples are: EUDAT B2SAFE and B2SHARE services, SCIDIP-ES tools for managing provenance and authenticity, SCAPE scalability platform, APARSEN prototypes and services, etc.

## The first Proof of Concept

The first PoC has been carried out from March 2013 to September 2013.

At the end of these activities important conclusions have been achieved and they can be summarized as follows:
7 scenarios (out of a total of 14) were explored in six concrete experiments conducted by partners. A cross-partner collaboration happened between Belgium and Italy covering Scenario 1.1 (Using specialised research tools), Scenario 1.2 (Integrating a new tool into existing institutional infrastructure) and Scenario 1.4 (Preservation from a consortium of collections on the cloud).. While Belgium was exclusively concerned with organisational challenges indicating a concern on the sustainability and adequateness of local or national solutions for potentially being used on a European or even global level, the Swedish partners put slightly more focus on the end user's experience on practical preservation activities.
To sum up we can say that:

- Cultural data are managed by many different persons: data management and administration + user access control are very important.
- Security of the data is very important for cultural institutions: trust building is a key factor when it is not determined where data are stored
- Functionalities and services offered by e-infrastructures should not impact on the outgoing traffic of the institution
- Access to the e-infrastructure services should be simple without requiring IT specialist knowledge

## Using existing infrastructures to store and retrieve digital assets

The Proof of Concept covering scenarios 1.1, 1.2 and 1.4 involved the use of the EGI Grid as a backend storage system and the e-Cultural Science Gateway hosted by INFN Catania as the frontend service.

[2] For a closer look to the scenarios http://www.dch-rp.eu/index.php?en/61/deliverables D3.1 Study on a Roadmap for preservation.
[3] http://www.dch-rp.eu/index.php?en/98/e-culture-science-gateway

Concerning the activities regarding the Scenario 1.1 (Using specialised research tools), the activities carried out can be summarized as follows:
BELSPO (the Belgian partner) uploaded 1012 files, for a total of about 128 GB, were transferred by its staff directly to a grid-storage running the DPM version with the HTTP/S interface. About 75% of files (for a total of 92 GB) were transferred on the 19th of August while the remaining 25% (for a total of 36 GB) were transferred on the 20th of August.
ICCU (the Italian partner) uploaded 32520 file jpg (web version 100 dpi resolution) and 85 file xml encoded with the MAG standard, the ICCU metadata schema, describing 85 ancient books (XVI-XVII centuries) the 12th of September, for a total of 11.85 GB.The uploading has been conduct by people from the CH sector, even if with the collaboration of ICT experts, and with some impact on the other outgoing traffic of the institution.

As expected by the project, this first Proof of Concept demonstrated the existence of some gaps between the knowledge of Cultural Institutions and the one of e-infrastructures providers. The reasons for this gap are to some extent also cultural/social discrepancies but even more technical incompatibilities in the understanding of authentication & authorisation infrastructures, and the mechanics of infrastructure-wide support (or non-support) of a certain required feature at a particular point in time: when taking on the decision to integrate with any e-Infrastructure of choice, Any Research Community must take into account that infrastructure providers do have feature roadmaps of their technology in place that cannot be changed ad-hoc. Rather, reliable e-Infrastructure providers will have mature change management processes in place to implement the roadmap in the communicated timeframe.
Another fundamental lesson achieved is the need to develop a digital cultural heritage vision which is expressive yet concise enough to drive which tools and services needed to be tested to support a roadmap towards achieving this vision. At the same time, thanks to this experience DCH institutions and e-infrastructures providers have developed a common workflow, useful to share their knowledge, in particular regarding the problems that a CH institute can find during the use of these kind of tools. On their side e-infrastructures provider should try to be more focused on the semantic aspects of the interoperability with the outcome of digital cultural preservation programmes[4].

## Planned activities

1. **Conduct a Proof of Concept for the preservation of 3D data**
   In collaboration with Collection Trust, who has access to large collections of 3D visualisations, assess the preservation of 3d data on remote storage locations.
2. **Re-execute the PoC on scenarios 1.1 and 1.4 with an improved e-CSG**
   Several improvements have been implemented in the e-CSG during and after the initial joint Belgian-Italian PoC. Since the underlying framework is due to support Cloud Storage, the PoC may consider switching from Grid to Cloud storage (provided by EGI and Belnet) or support both.
3. **Integrate EUDAT storage services**
   The Polish partner PSNC is an active partner in both EUDAT and DCH-RP. Some EUDAT services are of interest to DCH-RP for future uptake. Depending of the exact definition of this PoC it may cover different scenarios
4. **Providing access to preserved data for scientific publishers**
   Together with Elsevier, Editeur and INFN Catania, this Proof of Concept will explore the use case of giving a publisher access to scientific data stored in the SCH e-Infrastructure. A possible architecture foresees using the EGI Cloud storage, the e-CSG as the main user facing access, and Elsevier and Editeur as publishers. A partner for the preserved data is yet required.
5. **Value-added metadata analysis services**
   This Proof of Concept will explore the use case of advanced metadata analysis services. By harvesting large amounts of available and accessible metadata collections, new metadata may be generated for consumption of other services. This PoC is interested in exploiting the generic metadata capabilities of CDMI-based Cloud storage services, usability and resolvability of PIDs pointing into datasets (e.g. study supplements) in the Cloud, and the feasibility of cross-referencing studies stored in different repositories (e.g. open access/OpenAIRE and closed access/Elsevier/Springer).

[4] For a complete overview of the results achieved see http://www.dch-rp.eu/index.php?en/61/deliverables

## This survey

In order to know the potential and the market of the services provided by the e-infrastructures in the context of Digital Cultural Heritage, the project now carry out this survey. Here we aim to find out who can be the users, what services will be attractive by the potential users, what kind of services the user are adopting or are interested in using, and if users would like to join the activities carried out by the working group involved in the Proofs of Concept.

**PRELIMINARY QUESTIONS**

- Person who fills the survey (Name, Surname, E-mail, Telephone, Skype account)
- Organization (Full name, Acronym, Address)
- What is the type of your organization?
  (Academic, Research, Cultural, National, International, Public, Commercial, Other)
- How many staff members does your organization have?
  (1-50, 51-500, 501-2000, more than 2000)
- What is your role in your organization?
  (Management, Research staff, Technologist staff, Operational staff, Other)

**QUESTIONS FOR MANAGERS**

1. Which kind of data does your institution manage?
2. Are you aware of the e-infrastructures?
3. Is your institution using any service provided by e-infrastructures providers? If yes which one?
4. How much are you paying for these services?
5. Does your institution planning a cost reducing programme by using e-infrastructures services?
6. Do you think that your institution can be interested in integrating a new tool into existing infrastructure?
7. Is your institution member of a consortium of e-infrastructures providers/consumers? If yes, which one?
8. Do you think that your institution can be interested in joining a consortium of e-infrastructures providers/consumers ?
9. Does your institution run a programme for the long term preservation?
10. Does your institution run a programme for retrieval of data?
11. Are you aware of the existence of the DCH-RP project?
12. Have you ever been in contact with other projects regarding the digital preservation? If yes, which ones?
13. Do you think that the tests conducted by the WP5 can fit with your experience and needs?
14. Do you think that your institution can be interested in conducting some tests in the context of DCH-RP project? If yes in which one?
15. Are you interested in joining the next activities of DCH-RP WP5? If yes which one?

**QUESTIONS FOR SYSTEM ADMINISTRATORS / DCH CURATORS**

1. What is the total size in Gigabytes, of data that your institute is managing/curating?
2. How many individual data items does your data comprise of?
3. What are the smallest, median and largest sizes of your managed data?
4. Do you manage data collections, or individual data items only?
5. How many individual items, on average, do your data collections comprise of?
6. If you are managing data collections, do their members appear in more than one collection at any point in time?
7. Does your overall preserved data collection grow, or is it stable in size, or may it also shrink over time?
8. In case it changes size (whether growing, shrinking, or both), is showing steady, predictable size changes, or does it at times change erratically (i.e. unpredictably)?
9. In case if predictable growth, which are the most contributing factors to that growth?
10. When managing remote data preservation infrastructure services, which are your preferred management interfaces (web interface, mobile app, desktop client)?
11. When accessing (i.e uploading, downloading, etc.) data items within your managed data, which are the preferred, required, desired protocols (e.g. HTTP, FTP, WebDAV)? Please briefly indicate your rationales for each of them.
12. What are your data retention requirements? Do they differ from data preservation requirements?
13. What are your requirements for data access service availability and reliability? Availability measures the actual time a service was accessible over a defined period of time. Reliability is similar, but filters out any planned service downtime.
14. What are your data access bandwidth requirements?
15. Which are your availability and reliability requirements for the management parts of the preservation services?
16. Which are your requirements on Authentication and Authorisation? Do they differ for access to management interfaces and data access interfaces, or are they the same?
17. Do you accept or require social identities (e.g. Facebook account, LinkedIn account, etc) for granting access to preservation services?
18. Do you accept or require vetted institutional identities (i.e. a user's existing account with her home institute) for granting access to services?
19. Do you accept or require vetted external user identities (e.g. X.509 certificates)?
20. Do you manage preserved data that is also available publically for anonymous (read-only) access?