DCH-RP: Digital Cultural Heritage Roadmap for preservation

Advocacy Paper – Engage participation of cultural institutions in the design of an e-infrastructure for digital cultural heritage preservation

DEVELOPING AN E-INFRASTRUCTURE FOR PRESERVATION OF DIGITAL CULTURAL HERITAGE (DCH): STAKES AND GOALS

The DCH sector is producing a large volume of Digital content, which needs to be safely stored, accessed and easily re-used.
The objective of the DCH-RP project is to implement a federated infrastructure dedicated to support the application of an open science in the arts and humanities, exploiting what already exists and creating the parts that are not yet available. To achieve this goal, coordination and harmonisation are necessary at European and international level, between DCH and e-infrastructures stakeholders.

However, one of the main gaps remains the lack of vision on a common and international e-infrastructure suitable to serve the ICT needs of the DCH community in an efficient and accurate manner. This advocacy paper aims at involving the DCH community in the design of this vision, by engaging its participation to the project activities.

1ST CHALLENGE: PRESERVATION

The first step towards the e-infrastructure for DCH is to implement an infrastructure for preservation of DCH, as preservation is one of the most challenging problems.
Cultural institutions are still mostly unaware of the e-infrastructures possibilities, favouring in-house preservation systems, which contributes to the lack of interoperability and fragmentation of resources. There are commonalities among the national preservation strategies which need to be addressed in common and in a coordinated manner among the member states of the UE and internationally, to share solutions and contribute to interoperability and openness, which could be beneficial and cost effective to all.
This is a main challenge also for the e-infrastructures community, which would enable them to provide services to a wide variety of stakeholders, opening up new markets.

Preservation is here approached as the combination of preserving digitised and born-digital content, as well as the information associated to that content, the *infostructure*,

or metadata, at short, medium and long-term. The DCH-RP project will design a roadmap to be executed within future implementation projects.

USING PROOF OF CONCEPT TO VALIDATE THE ROADMAP

In that process, Proof of Concept will be used to validate in concrete experiments assumptions and concepts expressed in the DCH-RP roadmap for preservation. 3 themes were identified: organisational challenges, end user concerns, new services and infrastructure integration.
The Proofs of concept have a particular role to play in the project: indeed, they allow cultural institutions and e-infrastructure providers to work together on concrete experiments, and thus prepare future collaborations.

For each theme, specific scenarios were designed, presented in the Annex. Tools have been tested and recommendations made for their uses. The detailed results are available on the DCH-RP wiki: https://wiki.egi.eu/wiki/DCH-RP:PoC_Phase_1

RESULTS OF THE FIRST PROOF OF CONCEPT

Besides the concrete technical results detailed in the Annex, this first phase of PoC has highlighted 3 main issues:

- The DCH community is still fragmented into national and local solutions and processes
- The technical success of PoC is not enough, what is important is the applicability and meaningfulness of the results for the roadmap
- There is a need to develop a DCH vision on a common and international e-infrastructure suitable to serve the ICT needs of this community in an efficient and accurate manner, to drive which tools and services need to be tested to support a roadmap aiming at achieving this vision. Thanks to the implementation of the Proof of Concept, DCH institutions and e-infrastructures providers have started to develop a common workflow, which will be reinforced in the second phase of the project.

Indeed, to tackle these issues, it is more than necessary to involve the DCH and e-infrastructures communities over Europe, through the building of a network of common interest and concertation and consultation activities.

This first phase of proof of concepts contributed also to the Registry of tools that has been developed in the Work Package dedicated to the Roadmap (http://www.dch-

rp.eu/index.php?en/114/registry-of-services-tools), and it identified concrete metrics to be used for tools assessments:

- Easy installation
- Management
- Use of tools

In the framework of the 1st Proof of Concept, a survey has been designed, aiming at cultural institutions managers and system administrators / DCH curators. It will support the development of the DCH vision on e-infrastructure. The aim is to find out who can be the users, what services will be attractive to them, what kind of services they are adopting or are interested in using, and if they would like to join the activities carried out by the working group involved in the Proofs of Concept. The survey is accessible on line: http://www.dch-rp.eu/index.php?pageId=106&draft=0&sespre=MW_

A second phase of Proof of concept will be led by the project from January to August 2014, to improve the coverage of the scenarios. Participation and collaboration between DCH institutions and e-infrastructures providers will be key to ensure their success and relevance.

ANNEX: PRESENTATION OF THE ISSUES OF THE SCENARIOS, AND OF THE PROOF OF CONCEPT TESTED

SCENARIOS DESIGNED FOR THE PROOFS OF CONCEPT

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
    a) Extend 3.1 with repository safeguarding policies
    2. Defining new services
    3. Integrating new services into existing infrastructure

SCENARIOS TESTED FOR THE 1st PROOF OF CONCEPT

From February to September 2013, 6 PoC were led on the 2 first themes, through 7 identified prior scenario. They covered the following topics:

- Use specialised research tools
- Integrating a new tool into existing infrastructure
- Preservation from a consortium of collections on the cloud
- Select an existing DP solution in the case of an institution with only voluntary support
- Retrieve archived data
- Research and select a tool serving a specific purpose
- Gain access to archived websites

The scenarios and results are briefly presented below. More details can be found on https://wiki.egi.eu/wiki/DCH-RP:PoC_Phase_1

- **SCENARIO ON ORGANISATIONAL CHALLENGES**

Scenario 1.1: Use specialised research tools – Belgium, KIK-IRPA: Belgium DCH organisations, Italy, ICCU

- **Issues:**
  - Using external tools to manage local data
  - Using tools that require data on e-infrastructure storage
  - Access of data on remote e-infrastructure via the institute website
  - Access to the data on remote e-infrastructures via social media
- **Proof of Concept:**
  - Use existing tools and document the auditing process, on the local data of KIRK-IRPA and those of Italian partners
  - Recommendations: "scoremodel" is a useful tool to test the integrity of a collection. It can be used in the roadmap.

Scenario 1.2: Integrating a new tool into existing infrastructure

- **Issues:**
  - Embed a new tool in the existing preservation environment
  - Run the tool from a cloud based service

Scenario 1.4: Preservation from a consortium of collections on the cloud

- **Issues:**
  - Preservation of different data types
  - Upload to the preservation storage from different locations
  - Preservation of software tools developed for the preserved data
  - Copyright and/or IPR issues at national and international levels
- **Proof of concept**: common to scenario 1.2 and 1.4 (Belgium, Italy)
  - Looking at preserving their data on an external e-infrastructure to find preservation solutions beyond the use of local storage
  - Use grid storage on EGI to store data and e-Culture Science Gateway (eCSG) to copy data from local store to grid store and to access the data afterwards.
  - Recommendations : the usability of the eCSG is limited to manually copy files onto an external storage and to fill the metadata manually.

Scenario 1.3: Select an existing DP solution in the case of an institution with only

voluntary support:

- **Issues:**
  - Choosing a cost efficient solution for external preservation
  - Choosing a sustainable preservation solution

- **Proof of concept**: 4 tools for typical data preservation activities in a local institute, assessing these for potential pan-European usage were tested:
  - **ROND** (Riksarkivet OpeN Data) : used to anonymise data sets before publication or further public usage

    Recommendations : major limitation : requires a certain metadata model (ADDML), which is currently only used by Sweden and Norway. However, tools of this type could be very useful for publishing huge amounts of archival information as open data

  - **Archivist's toolkit** : Open Source archive management system
    Conclusion: installation was very complicated and beyond the expertise of a typical museum staff

  - **XENA** :file conversion tool that automatically detects the file format of a given file and converting it into digital objects suitable for digital preservation.
    Recommendations: very easy to use for batch conversion but the quality of results is not trustworthy

  - **DSPACE** : the software for academic, non-profit, and commercial organizations building open digital repositories DSpace preserves and enables easy and open access to all types of digital content including text, images, moving images, mpegs and data sets.
    Recommendations : too complicated. However, there is now also a hosted service, DspaceDirect10, that may be investigated as an alternative.

Scenario 1.6: Retrieve archived data: test of existing tools in Estonia

- **IBM Tivoli Server Manager/Client Server Version 5, Release 5, Level 2.0**: client-server licensed product that provides storage management services in a multiplatform computer environment.
  Conclusions: this tool requires advanced IT knowledge and expertise, even if the results are good.

- **SCENARIO ON END USER CONCERNS**

Scenario 2.2: Research and select a tool serving a specific purpose

- **Issue:** translation of formats
- **Proof of Concept:** testing format conversion tools:
    - **AVS document converter 2.2**: converts files of source formats into files of target formats

      Recommendations: should only be used when you want to convert small amounts of files

    - **AVS image converter 3.0**: converts files of source formats into files of target formats

      Recommendations: adequate for conversion between JPEG and PNG; no appropriate for conversion into PDF

    - **Universal Document Converter (UDC)** :printing service After installation you can choose UDC as the current printer when you want to convert a file

      Conclusion : can not be used for large amounts of images or documents, since you have to convert each image individually

    - **A-PDF DjVu to PDF14**: fast, affordable utility to allow you to batch convert DjVu (.djvu, déjà vu) into professional-quality documents in the PDF file format.

      Conclusion : tool not running on Windows 7 64 bit. this does not seem to be a good tool for the future, unless they make a 64-bit version, too.

Scenario 2.4: Gain access to archived websites

- **Issue:**
    - Persistent identifiers
- **Proof of Concept:** five tools tested to gain access to and manage the content of archived websites.
    - **HTTrack16**: free (GPL, libre/free software) and easy-to-use offline browser utility. To download a World Wide Web site from the Internet to a local directory, building recursively all directories, getting HTML, images, and other files from the server to the computer.
      Recommendations: both easy to install and use, and good quality of the result. It should be a suitable tool for the downloading of web sites when the most important aim is to give easy access to end users.

However, it remains to be investigated how good the downloaded format is for long-term preservation, and also how efficient the program is when many web sites are downloaded as a batch, simultaneously. For preservation, it would be useful to test all the different options that can be set for a download.

- **SWAT17** (Snappy Web Archiving Tool): tool designed for archiving web sites and displaying the archive in a simple way
  Conclusions: The fact of having to to download, unpack, and install, makes it doubtful if this tool can and/or should be managed by small institutions, that have only one or a few web sites to preserve and present for end-user access

- **WARC Tools**: to facilitate and promote the adoption of the WARC file format for storing web archives by the mainstream web development community by providing an open source software library, a set of command line tools, web server plug-ins and technical documentation for manipulation and management of web archive files, or WARC files.
  Conclusion: installation failed

- **Web Curator Tool** (WCT): open-source workflow management application for selective web archiving.
  Conclusion : installation failed

- **Heritrix20**: Internet Archive's open-source, extensible, web-scale, archival-quality web crawler project.
  Conclusion: It is unclear if the tool can run on other platforms than Linux. According to the FAQ, this has been tried even if it is not supported. A benefit with this tool is that a lot of third-party products don't seem to be required (besides Linux, only Java Runtime Environment is mentioned). However, the installation instructions are not sufficient for an inexperienced user.