**Proofs of Concept**

**Executive report for Roadmap consideration**

**Scenario 1.1**

**Scenario 1.4**

**Revision: draft 1.1**

**Authors:**

**Rosette Vandenbroucke (Belspo)**

**Patrizia Martini (ICCU)**

**Giovanni Ciccaglioni (ICCU)**

|  |  |  |
| --- | --- | --- |
| **Project co-funded by the European Commission within the ICT Policy Support Programme** | | |
| **Dissemination Level** | | |
| **P** | **Public** |  |
| **C** | **Confidential, only for members of the consortium and the Commission Services** |  |

**Revision History**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rev.** | **Date** | **Author** | **Organisation** | **Description** |
|  | 22 Sep 2013 | Rosette Vandenbroucke | Belspo | First description of the PoC introduced |
| 1 | 24 Sep 2013 | Rosette vandenbroucke | Belspo | Added information from the Italian partners, refined existing text |
|  | 27 Sep 2013 | Giovanni Ciccaglioni Patrizia Martini | ICCU | Information and data from the ICCU PoC |
|  |  |  |  |  |

Table of Contents

1 EXECUTIVE SUMMARY 4

1.1 Grading 4

1.2 Recommendations 4

2 TITLE 1 Errore. Il segnalibro non è definito.

2.1 TITLE 2 **Errore. Il segnalibro non è definito.**

2.1.1 Title 3 7

3 Conclusion 9

ANNEx 1 10

# EXECUTIVE SUMMARY

The scenarios 1.1 and 1.4 were used to create a life scenario for the first run of the DCH-RP PoCs. This life scenario can be described as follows: The Belgian and Italian partners want to look at preserving their data on an external e-infrastructure in order to find preservation solutions beyond the use of local storage. Several options were available (“e-infrastructures for research” or commercial e-infrastructures, grid storage or cloud storage).Control of the location where the data is stored could be a necessity. The efficient and easy access of the data is also a must.

A basic choice for this PoC was to use grid storage available on the European Grid Infrastrucure to store data and to use the e-Culture Science Gateway (eCSG) as the tool to copy data from the local store to the grid store and to access the data afterwards. The Belgian partner Belspo and the Italian partner ICCU took part in this PoC.

A series of steps were defined. Their order of execution is important as each step depends on its predecessor.

The different steps are:

For ICCU: use the eCSG to copy a collection to the grid storage, to make the metadata available on the grid storage and to test the access performance to the grid storage.

For Belspo: use the eCSG to copy a collection from the KIK-IRPA to the Belgian grid storage, to the Italian grid storage, to make the metadata available on the grid storage and to test the access performance to the grid storage, Install an eCSG in Belgium and repeat the data copy.

Only few of the steps could be executed due to problems with the eCSG. These problems were related to the storage software used on the grid storage, to the copying of a collection (= more than a few files), with retrieving automatically the metadata from the original data to the eCSG metadata format. The support from the eCSG was not sufficient to solve the problems in an acceptable time.

## Grading

For an explanation and the definitions of the various aspects see Annex 1.

|  |  |
| --- | --- |
| **Aspect** | **Score** |
| * Using grid storage via eCSG | **1** |
| * Copy data to grid storage via eCSG | **0** |
| * Automatically copy metadata to the eCSG metadata format | **0** |

## Recommendations

The eCSG in its current form is not fit to be used for realizing the preservation of collections of a DCH organization. Its usability is limited to manually copy files to an external storage (grid, cloud, …) and to fill out the metadata manually.

**An adequate portal will be needed to realize the transfer of data to external storage for preservation and to solve the metadata problem. Regarding this last point, this means, from the point of view of the CH institutions, that they have to be involved in the mapping between the native metadata and the e-CSG metadata, always before the beginning of the uploading activities.**

# Scenario overview

<https://wiki.egi.eu/wiki/DCH-RP:PoC_1_Belgium>

The test scenario was built up as follows:

1. Use the eCSG installed in Catania with storage in Catania

(this solution has the advantage of a working eCSG with possibility to upload data) (ICCU) (Belspi if 2. Is not possible)

1. Use the eCSG installed in Catania with storage on BEgrid

(for this scenario there is a need to adapt dcache, the storage management system in BEgrid, to eCSG) (Belspo)

1. Install the eCSG on BEgrid and use with BEgrid storage (Belspo)

### Execution

* E1: Exploit the possibility to have eCSG working with dcache
* E2: Dependent on the outcome choose scenario 1 or 2
* E3: Sort out the metadata aspect (metadata on the gateway, connection between metadata and data
* E4: Put data on the grid storage
* E5: Import metadata in the eCSG portal
* E6: Define the access measurement tools
* E7: Define the userfriendliness measurement tool
* E8: Do the measurements
* E8: Exploit the technical requirements to install eCSG on BEgrid
* E9: Depending on the outcome of E6, install the eCSG
* E10: Repeat E4-E5 on the BEgrid eCSG

## document structure

After the basic description of the tool testing environment we will describe in Chapter 3 each of the steps that could be executed.

## Scenario / Tool testing environment

An eCSG is installed in Catania and can make use of the Catanian grid storage. It could make use of other grid storages as long as those storage use a storage management software that is compatible with eCSG. The partners ICCU and Belspo who will use the eCSG have access to a collection of a DCH organization in their country.

# Tests To be executed

## Check if BEgrid storage can be used with ECSG

BEgrid uses dcache as the data management software. It has to be checked if this version includes the characteristics that are needed by eCSG.

## Test description

A grid expert checked the characteristics.

**Results**

The most recent version of dcache that is installed on the BEgrid storage does not have the characteristics needed by eCSG. Dcache developers have been approached to see if the required modifications could be made to the software. Even if these modifications will be done they will not be available in the timeframe of the DCH-RP PoC 1. Hence the Belgian grid storage cannot be used for this proof of concept.

## Copy data to the Italian grid storage

The eCSG includes a user interface to copy data. Use this interface to copy a collection that is available on the local storage of a DCH organisation to the grid storage.

**Test description**

Access had to be gained to the eCSG in Catania. Also a webinar on the use of the eCSG had to be attended. The explications given during the webinar revealed that it was not possible to use the eCSG user interface to copy data that was on a remote server (relative to the user who wants to do the transfer). In fact the eCSG can only transfer files from the PC of the user to the grid storage.

Eventually the collection data from Belgium was stored to the Italian grid storage by a grid expert who even had to solve another number of problems.

ICCU managed to copy the collection via SFTP because, as already mentioned, INFN is still developing a direct GUI for the direct uploading and retrieval on the e-CSG. During the uploading ICCU and INFN found different problems regarding the management of the SFTP that it is possible to sum up this way:

* During a first try (2013-09-11, 14.00 CEST) the port number 22 (INFN side) was closed. INFN tried to insert the ICCU IP in a white list of users of its port 22.
* During a second try (15.15 CEST) ICCU found a problem with another port, number 4422. The port (ICCU side) was closed, because isn’t a standard port and is managed by the ICCU data center under specific rules. ICCU asked to INFN to open the standard port 22, as already asked.
* With the third and last try (2013-09-12, 11.50 CEST) the problems were solved, and at the end of the uploading, the e-CSG people managed the transfer of the ICCU digital assets from the SFTP storage to the grid storage.

The total amount of the ICCU digital assets consists of 11,85 GB for:

* 32520 file jpg (web wersion 100 dpi resolution).
* 85 file xml encoded with the MAG standard, ICCU metadata schema, describing 85 ancient books (XVI-XVII centuries).

The uploading has been conduct by people from the CH sector, even if with the collaboration of ICT experts, and this is a problem that it’s necessary to solve, because CH people should be able to conduct the uploading by themselves. A second aspect to consider, even if is totally out of the control of the E-infrastructure providers, is the traffic overloading, that means that during the uploading can happen some decreases in the velocity of transfer, that involve the other outgoing traffic of a CH institution.

Now the ICCU digital objects reside on the e-CSG, with their metadata, and people can search and retrieve them, but, as already noted, the solution adopted to achieve this goal can’t be considered the standard one.

**Results**

The eCSG proved to be unusable for copying directly data from the local store of a DCH organisation to the grid storage.

Anyway, thanks to this experience ICCU and INFN have developed a common workflow, useful to share their knowledge, in particular regarding the problems that a CH institute can find during this kind of activities, when a FTP is adopted or, in a future perspective, when the direct access to the grid will be available.

## Making metadata available on the eCSG

In order to use data on the grid storage via the eCSG, metadata has to be made available on the eCSG. We had to test an automatic way to translate the original metadata to the metadata format of the eCSG.

Regarding this activity, from the ICCU point of view, it is mandatory the development of a tool to make available directly the metadata on the e-CSG. As far as the Italian experience, the ICCU team members were aware of the lack of such a tool since the beginning of the PoC. According to the decisions assumed during a webinar organized by the INFN (2013-07-29), the ICCU members sent to the INFN (2013-07-31) the MAG metadata schema, in order to let the e-CSG managers to set up a first version of the mapping between the MAG standard metadata and the e-CSG metadata.

Technical meetings should be scheduled on the metadata mapping and on the selection of metadata to be made available on e-CSG, useful to permit the search and retrieval of the data. The metadata mapping must be considered the first step, looking at the semantic interoperability, to achieve the automatic way to translate the original metadata to the metadata format of the e-CSG.

**Test description**

We find no possibility to create the metadata via the user interface of the eCSG. Intervention of the eCSG developers was needed to create the metadata. This was the case for both Belspo and ICCU.

**Results**

The test to use the eCSG to make available directly the metadata on the eCSG failed.

# Conclusion

Executing this PoC was a lengthy process and revealed some problems with the tested software. This experience learned that a good initial description of the software tool is essential in order to have a realistic idea on what the tool can do and what it cannot do. Another lesson learned is that support of a tool needs to be organized and help should be available in an acceptable time.

A further lesson learned is that a tighter interaction between the needs of CH organizations and the e-infrastructure providers is fundamental for the development of a tool that can satisfy the CH community requirements. At the same time, CH institutions should be more involved in order to share their main knowledge in this field, that is the management and development of metadata, both for search and retrieval and for preservation. In addition, an important goal have been achieved and it concerns the awareness that the integration between E-Infrastructure and CH sector goes through the understanding of problems existing in adopting the E-infrastrucutre technologies, the solutions agreed and the existing good practices and standards

The eCSG as is technically not ready for the preservation of data of DCH organisations. New features have to be added and the use of the gateway should be made possible for DCH people that are not ICT experts.

# ANNEx 1

This annex provides an extensive list of aspects that were assessed during the Proof of Concept reported in this document. For each discussed aspect a definition and the respective grading scale is provided.

|  |  |
| --- | --- |
| Using grid storage via eCSG | |
| The eCSG provides functionality to copy files to several types of storage including grid storage. | |
| Grade | Description |
| 1 | Due to several factors, including the functioning of the grid environment, only one type of grid storage can be accesses by the eCSG. This leads to a limited number of places that can be used for storage. |
| 2 |  |
| 3 |  |
| 4 |  |
| 5 |  |

|  |  |
| --- | --- |
| Copy data to grid storage via eCSG | |
| The eCSG supports a function to copy files to the grid storage. | |
| Grade | Description |
| 0 | It proved to be impossible to copy files from an external storage to the grid storage via the eCSG. The current eCSG copy function only permits to copy files from the user PC that has accessed the eCSG. This solution is not fit to copy a collection of thousands of files. |
| 1 |  |
| 2 |  |
| 3 |  |
| 4 |  |
| 5 |  |

|  |  |
| --- | --- |
| **Automatically copy metadata to the eCSG metadata format** | |
| The eCSG includes the metadata to enable an easy search and access to the data that has been copied to the storage. However in the current eCSG the metadata information has to be entered | |
| Grade | Description |
| 0 | In the current eCSG the metadata information has to be entered manually for each file that is uploaded. This approach is totally impossible when copying whole collections that contain thousands of files: it is important to highlight that the last one is the ordinary situation for the CH sector, within which the production of large amount of digital objects, with the relative metadata, is the everyday business activity |
| 1 |  |
| 2 |  |
| 3 |  |
| 4 |  |