



DELIVERABLE

Project Acronym: DCH-RP

Grant Agreement number: 312274

Project Title: Digital Cultural Heritage Roadmap for Preservation

D5.4 Report on second Proof of Concept

Revision: final updated version 2.0

Authors:

Michel Drescher, EGI.eu (Editor)
Andres Uueni, KANUT
Rosette Vandenbrouke, BELSPO
Claus-Peter Klas, FTK e.V.
Felix Engel, FTK e.V.
Maciej Brzeźniak, PSNC
Sara di Giorgio, ICCU,
Roberto Barbera, INFN-Catania
Luigi Briguglio, Engineering Italia Ltd.

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

Rev.	Date	Author	Affiliation	Description
1.1	2 Sep 2014	Michel Drescher	EGL.eu	Used D5.4 final version as starting point
1.2	18 Sep 2014	Michel Drescher	EGL.eu	Integrated contributions from BELSPO
1.3	23 Sep 2014	Michel Drescher	EGL.eu	Integrated contributions from RA, ICCU & FUH. Revised introduction & conclusions.
1.4	25 Sep 2014	Michel Drescher	EGL.eu	Added contribution from Engineering
1.5	26 Sep 2014	Michel Drescher	EGL.eu	Added Polish PoC working with SDL.
1.6	26 Sep 2014	Michel Drescher	EGL.eu	Reorganising contributions
1.7	29 Sep 2014	Michel Drescher	EGL.eu	FINAL delivery after internal review.
2.0	01 Oct 2014	Claudio Prandoni	Promoter	Final 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This deliverable contains original unpublished work except where clearly indicated otherwise.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

1 EXECUTIVE SUMMARY.....	5
2 INTRODUCTION	7
2.1 OBJECTIVES OF THE DELIVERABLE	8
2.2 STRUCTURE OF THE DOCUMENT	9
3 EXPERIMENT 1: EXPLORE THE SCAPE PROJECT’S MATCHBOX TOOL.....	10
3.1 RATIONALE	10
3.2 SETUP & DESCRIPTION OF TOOLS	10
3.3 EXPERIMENTATION	13
3.4 RESULTS & NEXT STEPS.....	15
4 EXPERIMENT 2: INVESTIGATE THE SCIDIP-ES PROJECT’S HAPPI PLATFORM.....	17
4.1 RATIONALE	17
4.2 EXPERIMENTATION	18
4.3 RESULTS & NEXT STEPS.....	19
5 EXPERIMENT 3: EVALUATE EUDAT STORAGE SERVICES.....	21
5.1 RATIONALE	21
5.2 SETUP & DESCRIPTION OF TOOLS	22
5.3 EXPERIMENTATION	24
5.4 RESULTS & NEXT STEPS.....	26
6 EXPERIMENT 4: RE-EVALUATE THE ECSG AND REMOTE GRID/ CLOUD STORAGE SERVICES FROM POC1	27
6.1 RATIONALE	27
6.2 SETUP & DESCRIPTION OF TOOLS	27
6.3 EXPERIMENTATION	27
6.4 RESULTS & NEXT STEPS.....	27
7 EXPERIMENT 5: A LONG-TERM DATA PRESERVATION PLATFORM.....	31
7.1 RATIONALE	31
7.2 SETUP & DESCRIPTION OF TOOLS	31
7.3 EXPERIMENTATION	33
7.4 RESULTS & NEXT STEPS.....	33
8 NATIONAL EXPERIMENTS	34
8.1 IDENTITY FEDERATION EXPERIMENT (ICCU, GARR).....	34
8.2 EXPERIMENTING WITH NATIONAL E-INFRASTRUCTURES (PSNC, SDL).....	35
9 CONCLUSION	39
10 REFERENCES	43
11 ANNEX 1: INSTALLING AND TESTING MATCHBOX	44
12 ANNEX 2: EXPERIMENTING WITH NATIONAL E-INFRASTRUCTURE IN POLAND.....	50



12.1	EXPERIMENT DESCRIPTION	50
12.2	PROOF OF CONCEPT SCOPE	51
12.3	APPLICATION TO ARCHIVAL SERVICES AND NATIONAL DATA STORAGE TOOLS TO DP PROCESSES	55
12.4	POC ORGANISATION.....	56
12.5	CONCLUSIONS	57

1 EXECUTIVE SUMMARY

The second Proofs of Concept phase in the DCH-RP project takes into account both the results of the first Proofs of Concept phase as reported in D5.3, and the DCH-RP Roadmap to Preservation as iteratively developed through its first study in D3.1, and the intermediate version provided in D3.4. In turn, the results of the second Proofs of Concept phase will inform and contribute to the final roadmap document D3.5 due in September 2014.

Preparations for the experiments started in mid December 2013 with a first conference call between prospective experiment leaders and participants, and the final list of five experiments was agreed and approved at the fourth project plenary meeting in Catania, Italy, in January 2014. At the same time, MoUs were agreed and signed with key FP7 projects to underpin and secure support for the experiments

Focussing more on integrated solutions and services, it is even more important to assess the software for the two most paramount requirements regarding the targeted users:

1. Ease of use of the tool or service for the *end user*
2. Ease of installation/provisioning for small IT departments or IT-experienced individuals.

The experiments cover a wide variety of solutions that have the potential to implement parts of the DCH roadmap to a satisfactory level, or with reasonable integration effort.

Experiment 1 explores a tool (“Matchbox”) developed by the SCAPE project that allows automating the task of finding duplicate images in a set of files. “Data hygiene” activity is a necessary filter for diligently preparing a dataset for archiving, and for regular quality assurance and repository certification for preservation.

Experiment 2 looked at the HAPPI (Handling Authenticity Provenance and Persistent Identifiers) service developed by the SCIDIP-ES project: Cultural data is often included in various projects over a long period of time, which raises a number of needs and requirements as follows:

- Digital asset authenticity – establishing and maintaining the originality of the asset
- Data provenance – Keeping a trail of data usage events for audits and data usage indication
- Data reference persistence and validity – Idempotent data reference/identifier resolution over time and space to the correct storage location

Experiment 3 assesses a combination of services provided by the EUDAT project (B2SHARE and B2SAFE) in combination with a service (Platon) provided by PSNC to its national digital libraries and archives. The aim is to evaluate EUDAT’s services for curating and publishing a research community’s digital assets, in DCH-RP’s case the preservation of digitised and born-digital cultural heritage.

Experiment 4 is investigating some of the results of the experiments in the first Proofs of Concept phase provided in D5.3. More specifically, this experiment revisited the use case of uploading digital assets to a remote Grid/Cloud infrastructure in conjunction with the e-Cultural Science Gateway (eCSG) developed by INFN-Catania. Including federated identity management and AAI into this experiment, this experiment is addressing two of the main outcomes of the previous experiment in the first PoC phase.

Experiment 5 concludes the second PoC phase with the aim of assembling a general-purpose digital preservation platform implementing a Service oriented Architecture (SOA). The focus of this experiment lies on reducing the total cost of ownership (TCO) of such a preservation platform through integrating as many generic services as possible, implementing as many preservation-specific standards as necessary, and addressing the needs of as many user communities as is feasible. In collaboration with the APA (through the APARSEN project) this experiment will also explore how an external, independent service provider might offer services around such a platform to the target market while integrating underpinning services delivered by, for example, EGI or EUDAT, or other suitable infrastructure providers.

Next to a number of technical recommendations one observation made during the entire duration of the project is standing out above all: The DCH community relies very heavily on appropriate ICT support geared towards real end users. This again is an observation, not a judgement that needs to be appropriately taken into account. When engaging with e-Infrastructures, a third stakeholder must be considered for inclusion: The first stakeholder is clearly the DCH community as the consumer of any ICT services provided to them. The second stakeholders are the e-Infrastructures in Europe (and potentially worldwide) that provide a certain set of underpinning ICT services. The third, possibly new, stakeholders are service integrators and platform providers offering services tailored to the DCH community.

2 INTRODUCTION

Over the course of the DCH-RP project, a number of activities have contributed to the inception and subsequent updates of the DCH sector's roadmap to digital preservation. Beginning with a study on how such a roadmap might look like and what its goals should be [R 1], the first phase of Proofs of Concepts [R 5] provided Work Package 3 with feedback on practical experimentation of a number of tools and services that are already used in daily activities similar to data preservation, or look promising to include in the Roadmap for Preservation. In December 2013 an intermediate version of the roadmap was published [R 2], which further developed the ideas and concepts described in D3.1 into a more concise roadmap for preservation for the DCH sector in Europe.

Based on the proceedings of the intermediate roadmap a number of strategic MoUs were signed with projects working in the same field or producing tools and services that could be used for preservation of digital cultural artefacts (e.g. with EUDAT, APARSEN, OpenAire, SCIDIP-ES), in order to secure general collaboration and support in conducting experiments in DCH-RP's second Proof of Concept phase in Work Package 5. These activities were closely coordinated with the planning for the second Proof of Concept phase in the project. During a phone conference on 17 December 2013¹ an initial set of 10 experiments were proposed to the consortium. Eventually, at the fourth project plenary meeting on 20-21 January in Catania Italy, the consortium took a strategic decision [R 6] to focus on five experiments, addressing actions identified in the then current intermediate roadmap [R 2] (primarily chapter 5.2), and specifically recommendation 1 given in Deliverable 4.1 [R 4]: "1) Adapt the use case scenario described in Chapter 5 to be tested and evaluated as a proof of concept in WP5".

The following vice experiments were chosen since they address at large important aspects and responsibilities of the DCH roadmap: Experiments one, two, three and four primarily focus on specific tools and services that implement key preservation capabilities (duplication detection, metadata management & data discovery, and data storing, sharing & replication) in either singular tools such as Matchbox, or tightly integrated services such as HAPPI. These experiments focus on functional capabilities of existing services and how they map into the roadmap, such as automatic metadata extraction and capture, authenticity and integrity of data, backup and restore (all D5.3 page 27.ff) Experiment 5 however, looks at organisational and political issues while providing representational preservation services integrated in a platform: The conceptualised platform addresses issues such as virtualisation, distributed systems, cross-sector integration and allows studying governance and business models for a platform that is provided as a services from a different, independent stakeholder to the DCH community at large.

1. Experiment 1: Explore the SCAPE project's Matchbox tool for detecting duplicate images
2. Experiment 2: Investigate the SCIDIP-ES project's HAPPI platform for data provenance, authenticity and identification
3. Experiment 3: Evaluate EUDAT storage services
4. Experiment 4: Re-evaluate the eCSG and remote Grid and Cloud storage services from PoC1

¹ <https://indico.eji.eu/indico/conferenceDisplay.py?confId=1980>

5. Experiment 5: A Long-term Data Preservation Platform

The overall experimentation criteria for the second phase of Proofs of Concepts are even more focussing on requirements and needs of the expected end user: The expected users of the tools and services are *digital librarians/archivists and digital preservation specialists*. These are by and large not IT savvy beyond their daily use of computers – and they neither have to be, nor are expected to have such a skill set. Their required key skills are more of scholarly nature, relying on small institute IT departments or individuals experienced in administering computers to provision the IT infrastructure they need: It is not necessary for digital archivists to know how software works; instead they need to know how to use it well.

Hence the two focal objectives of experimentation are, in order of importance:

1. Ease of use of the tool or service for the *end user*
2. Ease of installation/provisioning for small IT departments or IT-experienced individuals.

Participation in the five experiments was based on partner availability as well as the backing memory institutes interest. The following table indicates participation of partners and external institutes in these experiments. Membership of partner projects with MoUs in force is indicated where applicable.

Partner (affiliated FP7 Project)	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5
KANUT	X				
BELSPO	X		X		
Engineering Italia (SCIDIP-ES)		X			
PSNC			X		
RA			X		X
ICCU				X	
INFN-Catania				X	
GARR				X	
FTK e.V. (APARSEN)					X
Collections Trust					X
Editeur					X

2.1 OBJECTIVES OF THE DELIVERABLE

The initial planning of Work Package 5 and the timing of its deliverables envisioned a full and complete account of conducted experiments in the second Proof of Concept phase

to be published in this document. However, as often is the case reality required a change of plans and the project consortium decided to initially focus on securing external project collaboration and support to conduct the experiments since it realised that available funding within DCH-RP would not nearly cover the desired and necessary diligence in executing the experiments on its own.

As a result, a snapshot version of this deliverable was submitted in July 2014; this document constitutes the agreed update as per the project plenary meeting in Tallinn [R 7] on 24-25 April 2014 in Tallinn, Estonia. Unlike the original version submitted, this update will provide a record of all experiments from beginning to end.

2.2 STRUCTURE OF THE DOCUMENT

Following this introduction, the following five sections (3 – 7) describe the current experiments, each chapter following as closely as possible the same structuring:

- **Rationale:** Providing reasons why the experiment was conducted, the gaps and requirements it is tackling, including a brief numeration of the specific objects of the respective experiment.
- **Setup & Description of tools:** Specific references to tools and how they were set up allow for better reproduction of the experiment if the need arises. Though inspired by it this is not a scientific experiment where a meticulous description of the experiment methodology and setup is necessary.
- **Experimentation:** Provides a description of the actual (perhaps formalised) tests and trials that are planned for this experiment and how these would be conducted.
- **Results & Next steps:** Since this is a snapshot account of the experiments, some may not report any results yet. Future plans will accordingly accommodate for conducting the experimentation, or in cases where results exist (even if preliminary), the next steps will describe the future course of action as planned or altered in accordance with the results.

The document concludes with a brief analysis of the results achieved so far in section 8.

3 EXPERIMENT 1: EXPLORE THE SCAPE PROJECT'S MATCHBOX TOOL

3.1 RATIONALE

One recurring issue in archiving digital assets is unwanted duplication of content. This applies to any type of digital objects, be it audio files (e.g. recordings of fables that are passed orally from generation to generation), quantitative research data (e.g. SPSS data sources, Excel sheets, etc.), video interviews (e.g. for sociological studies) or images (e.g. a digitised copy of the Rosetta stone).

As a representative for digital object duplication detection tools, the SCAPE² project's "Matchbox" tool allows detecting duplicates in digital images.

Matchbox is tried out and tested in this first experiment. KIK-IRPA (Belgium) and KANUT (Estonia) participated in this experiment.

The single overarching objective was to test:

- Ease of installation of the tool,
- Ease of use for digital librarians and archivists,
- Tool accuracy in detecting duplicate images.

3.2 SETUP & DESCRIPTION OF TOOLS

The idea of the Matchbox is that there are numerous situations in which you may need to identify duplicate images in collections, for example:

- Ensure that a page or book has not been digitised twice
- Discover whether a master and service set of digitised images represent the same set of originals
- Confirm that all scans have gone through post-scan image processing.

Checking to identify duplicates manually is a very time-consuming and error-prone process. Matchbox aims to automate this process.

Matchbox is an open source tool which:

- Provides decision-making support for duplicate image detection in or across collections
- Identifies duplicate content, even where files are different (in format, size, rotation, cropping, colour-enhancement etc.), and if they have been scanned from different original copies of the same publication
- Applies state-of-the art image processing works where OCR will not, for example images of handwriting or music scores
- Is useful in assembling collections from multiple sources, and identifying missing files.

Matchbox provides the following benefits:

- Automated quality assurance
- Reduced manual effort and error rate
- Saved time
- Lower costs, e.g. storage, effort
- Open source, standalone tool. Also as Taverna component for easy invocation

² <http://www.scape-project.eu/>

- Invariant to format, rotation, scale, translation, illumination, resolution, cropping, warping and distortions
- May be applied to wide range of image collections, not just print images.

Documentation for Matchbox indicates that a preconfigured VM is available, but this was not suitable for this project. Also, pre-compiled binary packages were available, but only for AMD64 compatible 64bit architecture. Since the available test infrastructure supports only 32 bit, Matchbox and all its software dependencies had to be compiled and installed from scratch.

Annex 1 provides detailed instructions on how this was accomplished; this section however provides a summary of this.

3.2.1 Compiling and installing Matchbox

Matchbox is a command-line tool that uses OpenCV³ for the heavy lifting of the tasks ahead – it might be considered as a wrapper around OpenCV. OpenCV is the most popular and advanced code library for Computer Vision related applications today, spanning from many very basic tasks (capture and pre-processing of image data) to high-level algorithms (feature extraction, motion tracking, machine learning). It is free software and provides a rich API in C, C++, Java and Python. Other wrappers are available. The library itself is platform-independent and often used for real-time image processing and computer vision. OpenCV has already lot of interesting developments like face detection, similar object finder and etc. , see also screenshots below.

Naturally, the installation of Matchbox comprises of the following four phases:

1. Installing a build environment
2. Installing Python
3. Compiling and installing OpenCV
4. Compiling and installing Matchbox

While this appears to be easy enough, the actual details of these four phases require regular and frequent ICT knowledge and expertise in building software and satisfying its dependencies – skills that are certainly not present nor required for the typical digital archivist or librarian working in memory institutes. The following table provides an indication of the complexity of dependencies for each of these steps.

1. Installing a build environment		
	GCC	The GNU Compiler Collection includes front ends for C, C++, Objective-C, Fortran, Java, Ada, and Go, as well as libraries for these languages (libstdc++, libgccj,...). GCC was originally written as the compiler for the GNU operating system. The GNU system was developed to be 100% free software, free in the sense that it respects the user's freedom.
	Build-essential	This package contains an informational list of packages which are considered essential for building Debian packages. This package also

³ An open source tool for computer visualisation

		depends on the packages on that list, to make it easy to have the build-essential packages installed.
	G++	Released by the Free Software Foundation, g++ is a *nix-based C++ compiler usually operated via the command line. It often comes distributed with a *nix installation, so if you are running Unix or a Linux variant you likely have it on your system.
	CMAKE	<p>CMake is the cross-platform, open-source build system. CMake is a family of tools designed to build, test and package software. CMake is used to control the software compilation process using simple platform and compiler independent configuration files. CMake generates native makefiles and workspaces that can be used in the compiler environment of your choice.</p> <p>Important dependencies: libarchive 3.1.2, curl 7.36.0, libboost (any version)</p> <p>Optional: CMake GUI</p>
2. Install Python		
	Python 2.7	<p>Matchbox has an explicit dependency on Python 2.7 while common contemporary Linux distributions provide much more recent versions of Python. For example, Ubuntu 14.04 LTS includes Python 3.4.0.</p> <p>This forces the user to install Python 2.7 manually, which introduces further complications.</p> <p>Important dependencies: libsqlite3-dev, sqlite3, bzip2 libbz2-dev</p>
3. Install OpenCV		
	Build OpenCV	<p>OpenCV is a library that provides a wide variety of filters and detection algorithms, for which it makes extensive use of 3rd party libraries.</p> <p>Mandatory dependencies: build-essential, libgtk2.0-dev, libjpeg-dev, libtiff4-dev, libjasper-dev, libopenexr-dev, cmake, python-dev, python-numpy, python-tk, libtbb-dev, libeigen2-dev, yasm, libfaac-dev, libopencore-amrnb-dev, libopencore-amrwb-dev, libtheora-dev, libvorbis-dev, libxvidcore-dev, libx264-dev, libqt4-dev, libqt4-opengl-dev, sphinx-common, texlive-latex-extra, libv4l-dev, libdc1394-22-dev, libavcodec-dev, libavformat-dev libswscale-dev</p>
	Configuring OpenCV	Before the actual build process can start, a number of build variables need to be initialised according to the local environment – Annex 1 provides more details on this step.
4. Install Matchbox		
	CMake configuration	Building Matchbox is done using CMake (installed earlier). Annex 1 provides the important configuration options and values

	Installing Matchbox	<p>Once properly configured and built, Matchbox can be installed by issuing the final command:</p> <pre style="text-align: center;"><i>sudo make install</i></pre>
--	----------------------------	--

Table 1: Installation process overview for Matchbox

3.3 EXPERIMENTATION

The developers of Matchbox suggest a standard workflow for duplication detection as follows:

1. Extract SIFTComparison features of all images
2. Train a visual vocabulary on the extracted features
3. Extract BoWHistograms using the vocabulary and all extracted SIFTComparison features
4. Create a similarity matrix for the collection using compare on all BoWHistogram features
5. Take the top-most similar images for each image and compare their SIFTComparison features
6. Set a threshold based on the retrieved data
7. Images with an SSIM exceeding the threshold are considered to be duplicates

The Matchbox command line offers following features:

```
$ python2.7 ./FindDuplicates.py
usage: FindDuplicates.py [-h] [--threads THREADS] [--featdir FEATDIR]
                        [--precluster PRECLUSTER] [--config CONFIG]
                        [--bowski BOWSIZE] [--sdk SDK] [--clahe CLAHE]
                        [--downsample DOWNSAMPLE] [--update] [--binary]
                        [--binaryonly] [-v]
                        dir {all,extract,compare,train,bowhist,clean}
```

With all necessary information at hand, Matchbox allows mass-scanning entire folders and object lists for duplicates as illustrated below:

```
$ python2.7 ./FindDuplicates.py /home/anz/Downloads/matchbox_data all
=== extracting features from directory /home/anz/Downloads/matchbox_data ===
... extracting features of dir /home/anz/Downloads/matchbox_data
... 213 files to process
[1 of 213] PMF_1603_0072_inv.jpg done
[2 of 213] PMF_1657_0665_inv.jpg done
[3 of 213] PMF_1603_0287_inv.jpg done
[4 of 213] PMF_1603_0349_inv.jpg done
[5 of 213] PMF_1657_0195_inv.jpg done
...
[127 of 213] PMF_1603_0079_inv.jpg done
[128 of 213] PMF_0928_0027_inv.jpg done
[129 of 213] JPEG error: Corrupt Image
[130 of 213] PMF_1603_0331_inv.jpg done
...
$
```


Some screenshots illustrate the look and feel of using OpenCV and Matchbox:

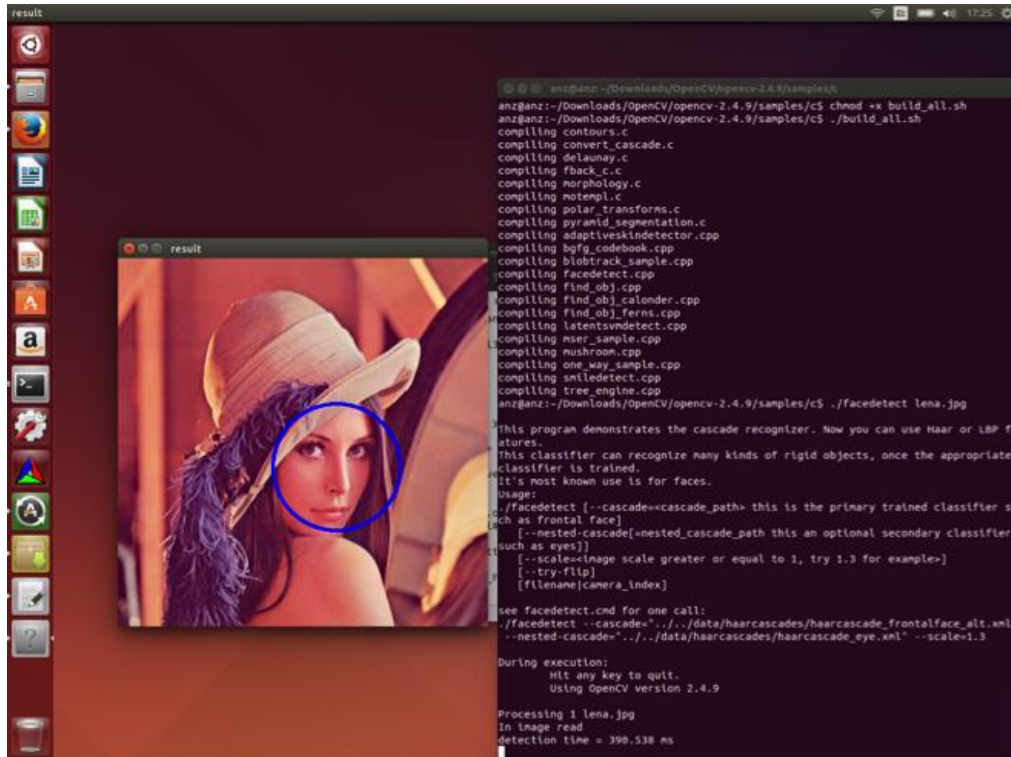


Figure 1: The OpenCV face detection feature

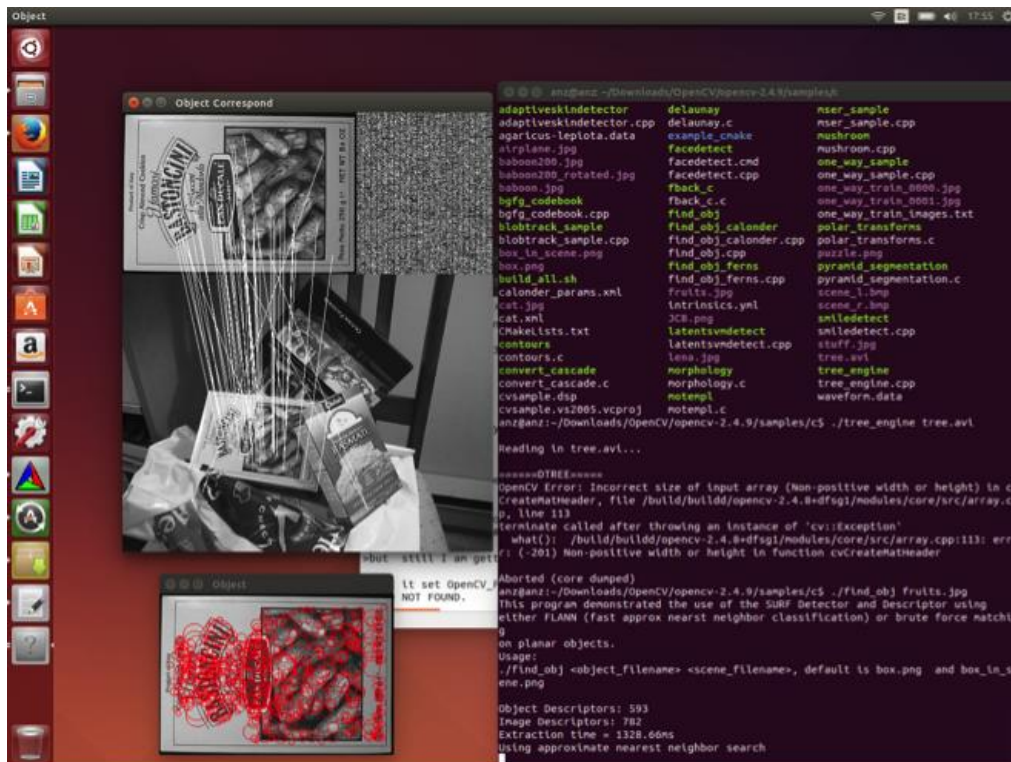


Figure 2: The OpenCV feature "detect similar objects" at work

3.4 RESULTS & NEXT STEPS

Due to the unexpected difficulties in providing Matchbox in a usable and testable environment, no results are yet available. However, in a basic test run of comparing 213 JPEG image files (or unknown file size, colour depth or image complexity) took approximately 130 minutes to finish. In a first approximation, this would translate to comparing one picture to 212 others in 36 seconds or, extrapolated, approximately 170 milliseconds to compare 2 images with each other.

Matchbox Tool works on command line, and doesn't have visual output, if necessary there are tools to parse and analyse xml files like PMF_1603_0186_inv.jpg.BOWHistogram.feats.xml.gz and PMF_1603_0186_iinv.jpg.SIFTComparison.feats.xml.gz.

At the present time Matchbox tool is available as a standalone, free of charge open source tool (Apache License Version 2.0), which needs several developer tools and tests before final implementation.

3.4.1 Test done by KIK-IRPA

KIK-IRPA owns major collections of images of paintings and statues. They are very interested in a tool like MATCHBOX to identify duplicates in a collection. Hans Opstaele has carried out the main testing.

Two tests were planned. The first was to make a limited test on the Ubuntu PC that was provided by the EVKM partner and on which the MATCHBOX tool was installed. The second was to make a full test with the data on the local storage server.

TEST 1

In the first test only a very limited amount of data could be uploaded and processed. The following files were uploaded by KIK-IRPA to the Ubuntu PC on which a cross-comparison was launched.

[1 of 6] f000510.tif F-series Glass Plate photo's, F000510 original, pattern of statues in a church, B/W

[2 of 6] f000510metadata.tif - F-series Glass Plate photo's, F000510 with added metadata

[3 of 6] f000522.tif F-series Glass Plate photo's, F000522 original, church inside, B/W

[4 of 6] f000522.png F-series Glass Plate photo's, F000510(!) renamed

[5 of 6] f000510partial.tif - Broken TIFF from F000510 (broken transfer)

[6 of 6] 14D009_Koop_1 (copy).jpg done - original image, color graffiti

[IGNORED] bis / f000510.tif - tool ignored this image in a subdirectory(?)

The MATCHBOX tool was executed and passed the tests with the following remarks:

REMARK: tool ignored an image in a subdirectory. Recursive directory traversal option not found. **FR**

REMARK: The file f000510partial is a broken, partial image; i.e. the image is not cropped, but was broken during transfer). The tool identified this correctly **PASS**

REMARK: Test on file format conversion (TIFF and png) and resolution loss. **PASS**

REMARK: Tool should keep file extension in its output to avoid confusion; Behaviour confusing if there's no file extension reported (f000522.png = f000510). **FR**

REMARK: A row of church statues in a black-white photo looks not like a coloured graffiti wall. **?BUG?**

REMARK: Difference in metadata can be very important in practise. Tool should at least warn about this. Danger to delete the wrong image (with metadata) instead of the image without its metadata or loss-thumbnail of the original. **FR**

TEST 2

The second test foreseen was making a full test with the data on the local storage server. This test would allow a more realistic test on a larger dataset and enabling to determine performance of the tool. Due to the complexity of installing the MATCHBOX tool the installation on the KIK-IRPA storage server could not be finished in time. Note that the whole installation is different between operating systems.

Conclusion

The basic Tool seems to work. The code is clearly written and it is stable enough to handle broken files in the test. It looks promising and there can be a demand for such a tool.

However several drawbacks exist that need to be addressed before MATCHBOX can be usable in a production environment. To name a few of these drawbacks:

- Output was hard to interpret certainly in an end-users perspective as was the a remark from the photographer when the tool was demonstrated;
- work-flow and practical use by the end-users must be taken into consideration;
- although FE was quick, cross-comparison of the 6 (six) images' features took several minutes.

Note: KIK-IRPA has ~1 million images in their database, and images and scans are added daily. Both the tool's performance, and above all speed and user-acceptance (ie. clear output, even for trivial cases like an indication of resolution reduction or added meta-data) are important to get this tool 'sold'.

4 EXPERIMENT 2: INVESTIGATE THE SCIDIP-ES PROJECT'S HAPPI PLATFORM

4.1 RATIONALE

For Digital preservation activities, it is important to also capture information about transformations the digital object undergoes during its life cycle. This information will form the “evidence” for later assessing the authenticity of digital object and consequently this information is called as “evidence history”⁴. It forms a crucial part of the Preservation Description Information (PDI) according to the OAIS Reference Model. Indeed, it includes provenance, reference and fixity. The SCIDIP-ES project has developed the HAPPI toolkit that supports the digital archivist in collecting this part of the PDI, the evidence history of the digital artefact that needs archiving. HAPPI is an acronym and it stands for Handling Authenticity, Provenance and Persistent Identifiers.

The collaboration between the DCH-RP and the SCIDIP-ES projects is underpinned by a mutually signed MoU (September 2014), which includes conducting this experiment. This activity started in July 2014.

The following sections describe the tests and experiments done during that period:

1. Deployment and Setup of SCIDIP-ES HAPPI in the EGI Federated Cloud Environment;
2. Evaluation of the SCIDIP-ES HAPPI Data Model in the DCH community.

The second experiment is still on-going, and the participants have agreed to keep this experiment going beyond the DCH-RP project. The SCIDIP-ES project is planning to reach out to DCH-RP partners to guide them through the HAPPI service, and run a survey afterwards.

The results will be reported in the SCIDIP-ES deliverable D24.1 “Generic services assessment and evolution definition report” (due date M40 – Dec 2014).

Setup & description of tools

Concerning the first experimentation, i.e. “Deployment and Setup of SCIDIP-ES HAPPI in the EGI Federated Cloud Environment”, the SCIDIP-ES team has provided all the documentation necessary to setup and describe the toolkit:

- SCIDIP-ES HAPPI toolkit 1.5.0 Quick Start – a brief manual (2 pages) that provides information for getting started with this toolkit;
- SCIDIP-ES HAPPI toolkit 1.5.0 Installation and User Manual – a detailed documentation including models and how to use the toolkit.

For the deployment, SCIDIP-ES HAPPI doesn't have any specific constraint, so it could be deployed mostly everywhere a JVM is available (Java 7 or greater) and an application server is running (e.g. Tomcat 7 or greater). It is enough to load the web archive package (war) of SCIDIP-ES HAPPI toolkit 1.5.0 (available at SCIDIP-ES Nexus Sonatype

⁴ Briguglio, L., Guercio, M., Salza, S.: Preserving Authenticity Evidence to Assess Provenance and Integrity of Digital Resources. In International Conference ECLAP 2013 on Information Technologies for Performing Arts, Media Access and Entertainment: Proceedings, Porto, LNCS vol. 7990, pp. 66-77 Springer, Heidelberg (2013)

repository⁵). Moreover, SCIDIP-ES HAPPI 1.5.0 persists information in a graph database: it allows using OrientDB 1.5.0 Graph Edition or Neo4j Community Edition 1.9.3.

Engineering team (from SCIDIP-ES project) and EGI team (from DCH-RP project) cooperated for this experimentation. Indeed, EGI has provided virtual machines on the EGI Federated Cloud Environment and Engineering has obtained the suitable certifications for accessing this environment and for deploying the HAPPI toolkit.

4.2 EXPERIMENTATION

For time constraints, during the preparation of this document the second experimentation, i.e. Evaluation of the SCIDIP-ES HAPPI Data Model in the DCH community, is still on-going. So this section provides a description of how this would be conducted.

Objective of this experimentation is the assessment of the software HAPPI toolkit over the requirement “Ease of use for the end user”.

For its nature, HAPPI toolkit, as any other SCIDIP-ES service and toolkit, provides its functionality through Application Programming Interface (API). This means that main end-users of HAPPI are software developers and software integrators.

In order to increase the usability of the HAPPI toolkit, code samples and tests have been published by the SCIDIP-ES team. They are available at the Sourceforge “Digital Preservation Services” Community⁶. By using the test code, a junior java developer could be able to work with the libraries and integrate the functionality in existing archives.

On top of the SCIDIP-ES HAPPI toolkit API, it has been also provided a RESTful web service. Moreover a simple, as well as effective, User Interface has been provided for demonstration purpose. This user interface is accessible via browser when invoking the HAPPI instance and it may be used by archivists and students too. This user interface may be used for assessing the completeness of HAPPI data model, i.e. confirming that provenance, reference and fixity information can satisfy needs of the DCH community.

Provenance information is based on the Open Provenance Model (OPM⁷) and it includes concepts such as Transformation (the transformation event that impacts on the digital object), Agent (who controls the transformation) and the Digital Representation (the result of the transformation). User may provide details for these concepts: the PREMIS⁸ dictionary has been adopted for ensuring ease of use to preservation experts. The picture below shows an example of information that HAPPI can manage. Further details can be provided to the “evidence” by using the so-called “significant properties”, that are

⁵ <http://nexus.scidip-es.eu/content/repositories/releases/eu/scidipes/toolkits/authenticity/happi-server/1.5.0/happi-server-1.5.0.war>

⁶ <http://sourceforge.net/p/digitalpreserve/code/HEAD/tree/SCIDIP-ES/software/toolkits/authenticity>

⁷ Open Provenance Model Core Specification v1.1 - available at <http://eprints.soton.ac.uk/271449/1/opm.pdf>

⁸ PREMIS Data Dictionary for Preservation Metadata v2.0
<http://www.loc.gov/standards/premis/v2/premis-2-2.pdf>

properties expected to be checked during the assessment, as well as by using annotations.

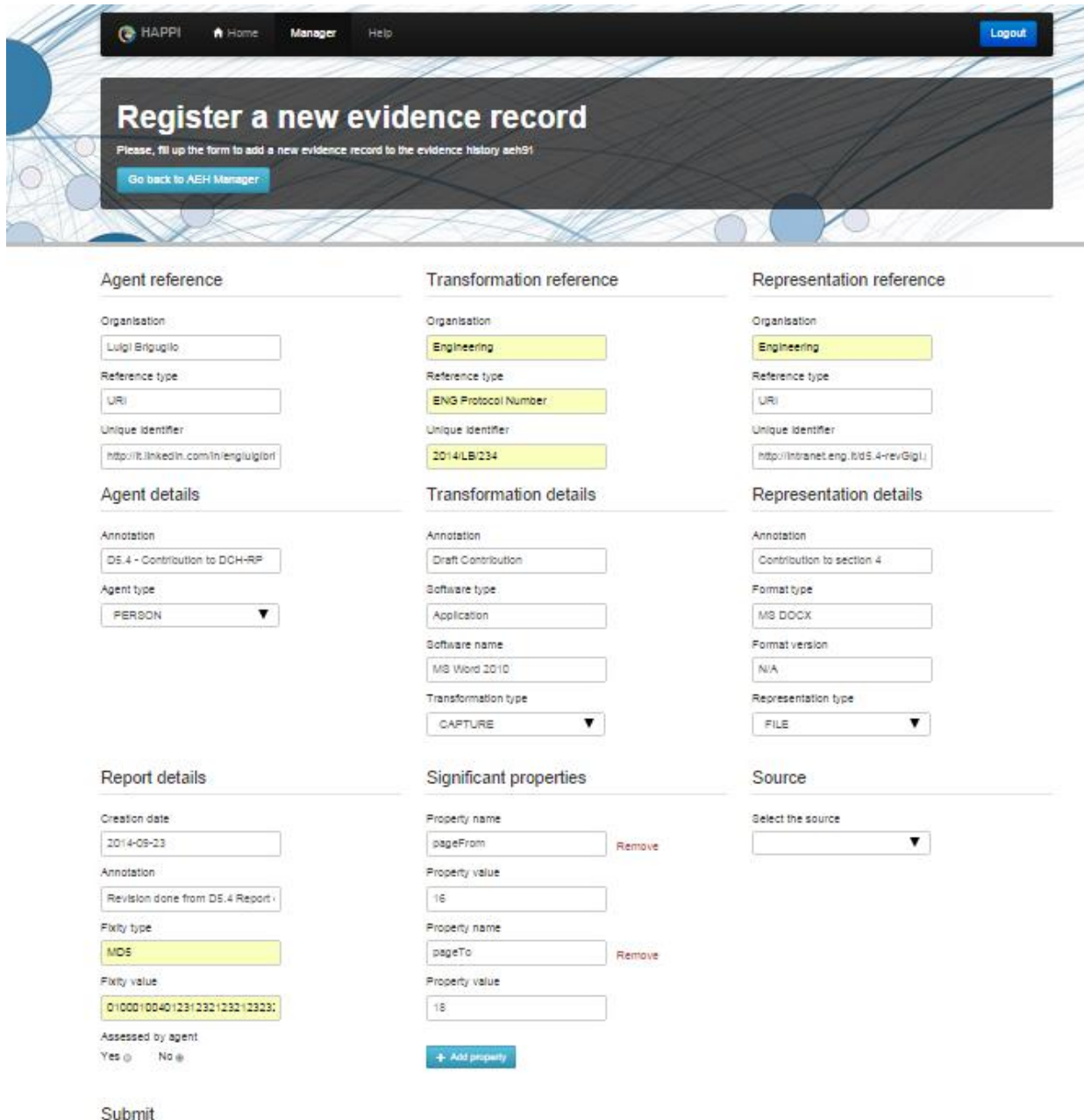


Figure 3: SCIDIP-ES HAPPI Form for capturing evidence of a transformation

4.3 RESULTS & NEXT STEPS

For the first experimentation, i.e. deployment and setup, HAPPI toolkit has been successfully packaged into a virtual appliance that is deployed on the EGI federated Cloud infrastructure. This means that the DCH-RP community has available a running instance of SCIDIP-ES HAPPI toolkit 1.5.0 available at

<http://90.147.102.191:8080/happi-server-1.5.0/>.

This experimentation has demonstrated the extremely ease of installation/provisioning for small IT departments or IT-experienced individuals. Indeed, the software artifacts are provided with predefined configurations. So, the toolkit could be deployed and few easy steps (i.e. download packages, unzip DB server, run DB server and deploy HAPPI on tomcat).

Since its deployment and setup, HAPPI toolkit 1.5.0 is continuously running without having experienced issues and interruption of operation. This allows to assess its good level of maturity, as well as the underlying Cloud Infrastructures.

Moreover, the HAPPI toolkit instance does not integrate with the EGI authentication framework, demonstrating effective separation of infrastructure management authentication and infrastructure user authentication.

Even if the second experimentation is still ongoing, it is reasonable to assert that “HAPPI is a sample service for data provenance, facilitating repeatable science”, as well as it could be applied to DCH-RP community too, for its generic provenance model based on OPM and PREMIS.

For the above reasons, it has been decided to keep running the SCIDIP-ES HAPPI toolkit 1.5.0 over the end of DCH-RP project, in order to make it available for further experimentations and assessment.

5 EXPERIMENT 3: EVALUATE EUDAT STORAGE SERVICES

5.1 RATIONALE

The DCH-RP project aims to identify suitable models and tools for the governance, maintenance and sustainability of DCH data that can be effectively used by cultural institutions across Europe.

Several projects and infrastructures aim to address long-term preservation of scientific and cultural data.

At the European level, EUDAT⁹ provides a sustainable infrastructure based on the layer of common technologies, tools and services driven by user needs. It also backs the community- and domain-specific services. EUDAT currently serves several scientific communities including CLARIN, diXa, DRIHM, ENES, EPOS, INCF, LifeWatch, VPH and further expands to new communities.

On the national level NRENs, data centres, computing centres and universities provide storage and data preservation, publication and sharing services to several communities including science and cultural sector. For instance Archiving Services¹⁰ of the PLATON project in Poland offer safely replicated storage space to academic and cultural institutions.

One of the use cases identified in DCH sector is the publication and sharing of the digital assets. Large organisations may have their own approaches, solutions and infrastructure for this purpose. However small organisations and so-called “citizen scientists” or “citizen curators” often struggle to get a reliable, adequate and affordable facilities for storing, publishing and sharing the data and metadata, as well as ensuring their long-term preservation.

5.1.1 Objectives

The aim of this experiment is to verify if and how services developed by EUDAT and nationally may address the needs of DCH community.

EUDAT’s B2SHARE¹¹ service is used as a solution for data publication and sharing. It enables users to upload their data sets, enrich them with meta-data and assign persistent identifiers. It also supports keyword-based searching, digital assets preview as well as meta-data presentation and browsing.

The scope of the experiments also includes analysis of possible orchestration of EUDAT services with other services and infrastructures such as European or nationally provided facilities for long-term data storage and preservation.

At the European level, EUDAT’s B2SAFE¹² service is considered. It ensures robust and reliable data replication and guards against data loss. It also improves data availability and locality across the continent. The service is offered by academic data centres. PSNC that represents EUDAT in DCH-RP project is one of such centres.

⁹ <http://www.eudat.eu/>

¹⁰ <http://www.platon.pionier.net.pl/online/archiwizacja.php?lang=en>

¹¹ <http://www.eudat.eu/b2share>

¹² <http://www.eudat.eu/b2safe>

At the national level, Archiving Services¹³ of the PLATON project in Poland offer 12,5PB of tape storage and 2PB of disk storage distributed in 10 locations across the country for the purposes of reliable replicated long-term storage. The service is offered to academic institutions and public sector including DCH institutions. The project is coordinated by PSNC.

The detailed objectives of this experiment are to:

1. Verify usability of EUDAT B2SHARE service for DCH communities in the terms of the following requirements:
 - simple data upload and access
 - easy and effective data sharing
 - assuring data referability of the data for long term
2. Examine the usefulness of European-wide and national solutions for long-term data and meta-data preservation. Usefulness is considered in following aspects:
 - a. reliability of the long-term storage process
 - b. transparency of the data protection mechanisms from the point of view of the service directly interfaced by end-users (such as e.g. data publication and sharing service)

5.2 SETUP & DESCRIPTION OF TOOLS

For the purposes of evaluation following tools and services setup is prepared. It consists of two layers (see Figure 4) relevant for aforementioned aspects of the experiment.

First, the upper layer provides easy interface for storing the data with simple metadata and data sharing, based on EUDAT B2SHARE service instance. Technically B2SHARE is a customised version of Invenio¹⁴ designed to offer a simple mechanism for uploading and sharing scientific data with associated metadata.

This layer is directly interfaced by the users and exposes to them several functions including data upload, meta-data editing, repository lookup, browsing, digital assets preview, meta-data presentation, search etc. This layer implements the clue of the functionality desired by the users in the considered use case.

The lower layer, which ideally should be transparent for the end users, provides an additional assurance for data sustainability as well as long-term data availability and safety.

It might be implemented using EUDAT's B2SAFE service or PSNC-coordinated Archival Services of the PLATON project. In our experiments we decided to use the Polish national service for implementing reliable long term storage and preservation. In this way we demonstrate and evaluate national facilities that complement the Europe-wide services such as EUDAT. Such approach also creates opportunity to verify the modularity and openness of the EUDAT's solutions by trying to orchestrate them with other services transparently to end-users.

¹³ <http://www.platon.pionier.net.pl/online/archiwizacja.php?lang=en>

¹⁴ <http://invenio-software.org/>

PLATON's Archival Service is a data backup/archival service with geographical replication. The service ensures long-term data availability and safety thanks to automatic data and meta-data replication as well as additional techniques such as integrity control of incoming data and periodic data scrubbing including local and remote replica integrity checks. Importantly, these mechanisms are implemented transparently to end-users. PLATON's Archival Service can be interfaced by SFTP, WebDAV and GridFTP protocols.

If enriched with NDS2-project¹⁵ provided tools the service can be accessed through convenient virtual file system interfaces from Windows and Linux clients and using portable file browser-like GUI.

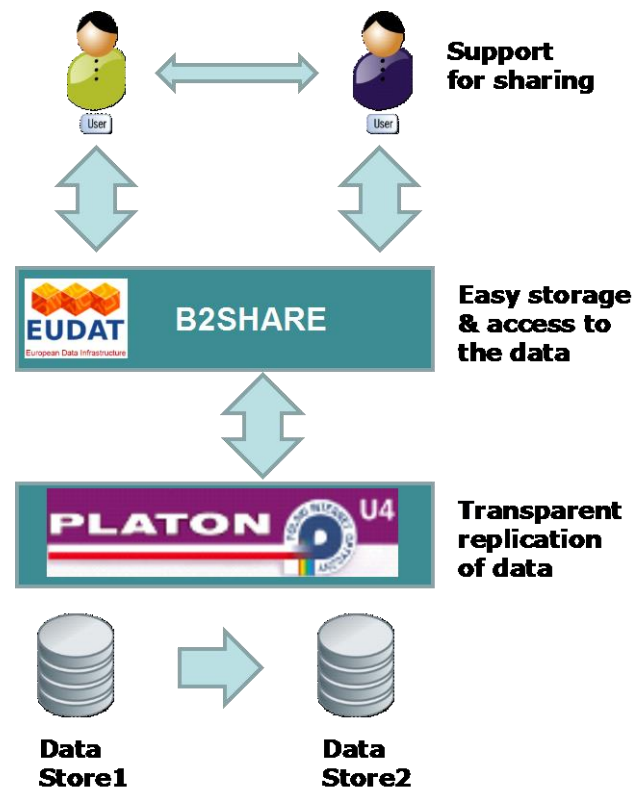


Figure 4: Setup of the EUDAT service experiment

¹⁵ <http://nds.psnc.pl>

5.3 EXPERIMENTATION

Table below summarizes actions that accomplish our experiment and provides overview of their status. Following sections discussed details of our activities and future plans.

	Action	Result / Status	Responsible
1.	Evaluations of existing B2SHARE instances	Tests performed by Swedish, Polish and Belgian partners.	Riksarkivet, PSNC, Belspo
2.	Setup dedicated B2SHARE service.	Running service.	PSNC
3.	Setup backup B2SHARE to PLATON Archival Services	In progress. Data exchange among services/layers TBD (automated backups and archival copies TBD)	PSNC
4.	Evaluation of the orchestrated services.	Tests TBD. Short report to be prepared.	Riksarkivet, PSNC, Belspo

As far our work focused on the evaluation of the EUDAT B2SHARE service using existing publicly available production and demonstration instances of the service. Several cultural institutions from Sweden and Poland were involved in these evaluations.

In addition two-level services setup is being built by PSNC involving a dedicated B2SHARE service instance integrated with PLATON's Archival Services as described in the previous section. This configuration is going to be used in the following stages of our experiment.

5.3.1 Tests done by KIK-IRPA

KIK-IRPA has huge collections of images and uploaded a very small sample on the EUDAT storage via B2SHARE and also tested out the search functions. Files were successfully stored and search queries have been made. However several problems occurred that required an intervention of the support team or even from the developers. KIK-IRPA finds the website for EUDAT-B2SHARE looking clean and intuitive to the user. However the major and minor bugs encountered make this service not yet acceptable for production work.

Those problems encountered are:

- when uploading data B2SHARE was unable to upload a directory via the user interface; the user has to do the upload on a file-by-file bases (bug#318);
- Searching, tagging and exporting metadata still has a few bugs - most problems were discussed but some seemed to 'hang' the system and required the user to restart the web browser(bug #357);

- When doing bad use testing uploading broken files was OK but very-long metadata descriptions were not all accepted. Some bad use uploads never showed up (-a reminder to the user that my upload was rejected would have been nice);
- One metadata action revealed an "internal server error" (eg. bug #346) and some metadata export formats limited the metadata descriptions (see *).

Some features that should be available in 21st century services are missing:

- **Tagging:** Tagging simply 'did not work' ("add tag" did nothing(?));
- **The search interface** was a major limitation in this release -mind you, it DID find the few objects posted, but the search functionality is too limited to be of practical use, eg. no multi-lingual support, the search options have no 'related features' (eg. through use of thesaurus or "word nets"), there's no tag cloud to refine the search, no search customization (eg. 'other people like you also liked X', like iTunes does...). An "image-content" based search would be nice too (like searching on the color histogram).

The simple search function will most probably be insufficient once the project hosts millions of objects, it will be impossible to find them back.

5.3.2 Tests performed by Riksarkivet

Overall impressions are positive. B2Share system is easy to use and it provides a lot of functions that organisations need.

Getting started/Log in

Generally, it was hard to get started because several organisations experienced problems with login, including error messages as for example "Invalid user"/ "The entered email address does not exist in the database"/internal server error, etc").

Uploading material and registration of metadata

Some of the institutions had problems with uploads. Uploads were generally very fast, but limited. It was possible to upload few objects but not to register different metadata to those objects afterwards. It was not either possible to import metadata embedded in digital photographs (EXIF-tags). Impression of being able to choose how to present metadata is very positive, however several organisations found existing descriptive metadata tags insufficient for describing all of the test collections. There is a need for metadata tags, (as for example year, district, parish, county etc.), that could be indexed in a search function.

Audio files were possible to upload but not movie files, (system hung up - too many GB?). Consequently, system requires a lot of manual work with metadata.

To summarise the impressions, it was not so easy to get started, the actual uploading of small files works well and metadata registration might work better on mass-uploads.

Search functionality

Search functionality worked very well.

There are also two current projects at the Swedish Museum of Natural History involving B2SHARE and B2STORE, testing an implementation of the archiving process for collections management system, as well as crowdsourcing functionality.

5.4 RESULTS & NEXT STEPS

5.4.1 Preliminary results

Preliminary results of the evaluation show that B2SHARE service provides data sharing and publication solution suitable for the needs of small cultural institutions and “citizen” “publishers” or “curators”.

However there are bugs and limitations that prohibit from using this service in a production environment. More thorough testing needs to be done to detect more major and minor bugs and users should be consulted to upgrade a number of functions.

It can be said that:

- Mass scale uploads and sharing may require more domain-optimised and specialised approach. Ensuring long-term data availability is not part of B2SHARE’s service scope and intent; therefore B2SHARE should be orchestrated with additional layers such as EUDAT B2SAFE and PLATON’s Archival Services.
- An effort should be made to provide a service that works more correctly;
- The search engine needs to be further developed taking into account user requirements;
- Metadata functions, tagging, etc. should be added.

5.4.2 Next steps

Activities planned for the following period include evaluation of the EUDAT’s service for data sharing and publication orchestrated with long-term storage and preservations solution. At this stage transparency and reliability of the data preservation and safety mechanisms will be evaluated.

6 EXPERIMENT 4: RE-EVALUATE THE ECSG AND REMOTE GRID/ CLOUD STORAGE SERVICES FROM POC1

6.1 RATIONALE

One of the limitations of the e-Culture Science Gateway (eCSG) identified in the first round of PoCs was that “[...] usability is limited to manually copy files to an external storage (grid, cloud, ...) and to fill out the metadata manually” (from D5.3, page 14). That limitation was indeed unavoidable because a general uploader, as the one that was developed in the first year of the project, cannot be used to seamlessly cope with a large variety of specific metadata formats and schemas.

We therefore decided in this experiment to demonstrate that specific uploaders could have made the use of the eCSG simpler and easier as well as the procedure of automatic upload of data into Grid/Cloud storage and insertion of metadata in the gLibrary-enabled repository build “underneath” or “behind” the eCSG. To ease the experiment uptake and support locally at ICCU, a side-project was conducted to establish and configure an Identity provider service at ICCU with the help of INFN Catania and GARR (see below in section 8.1).

6.2 SETUP & DESCRIPTION OF TOOLS

The tools which have been used are the eCSG, which has been already described in several other deliverables of DCH-RP, and the “IdP in the cloud” (<http://goo.gl/A3BNx6>) service of GARR that is a sub-contractor of ICCU.

For the second round of PoCs two sets of digital assets have been chosen:

- 1) A set of WARC archives of websites belonging to the .it domain and provided by the National Library of Florence (Magazzini Digitali);
- 2) A set of multi-format data belonging to SITAR (<http://sitar.archeoroma.beniculturali.it/>): the archaeological information system of the city of Rome provided by the Special Superintendence of Rome (SSBAR), also involved in the ARIADNE project.

6.3 EXPERIMENTATION

During the experiment, the following activities have been carried out as follows:

Two uploader portlets specific for the WARC and the SITAR archives have been developed and integrated in the eCSG.

6.4 RESULTS & NEXT STEPS

Figure 5 shows the uploader portlet that has been developed for the Magazzini Digitali repository of WARC files.

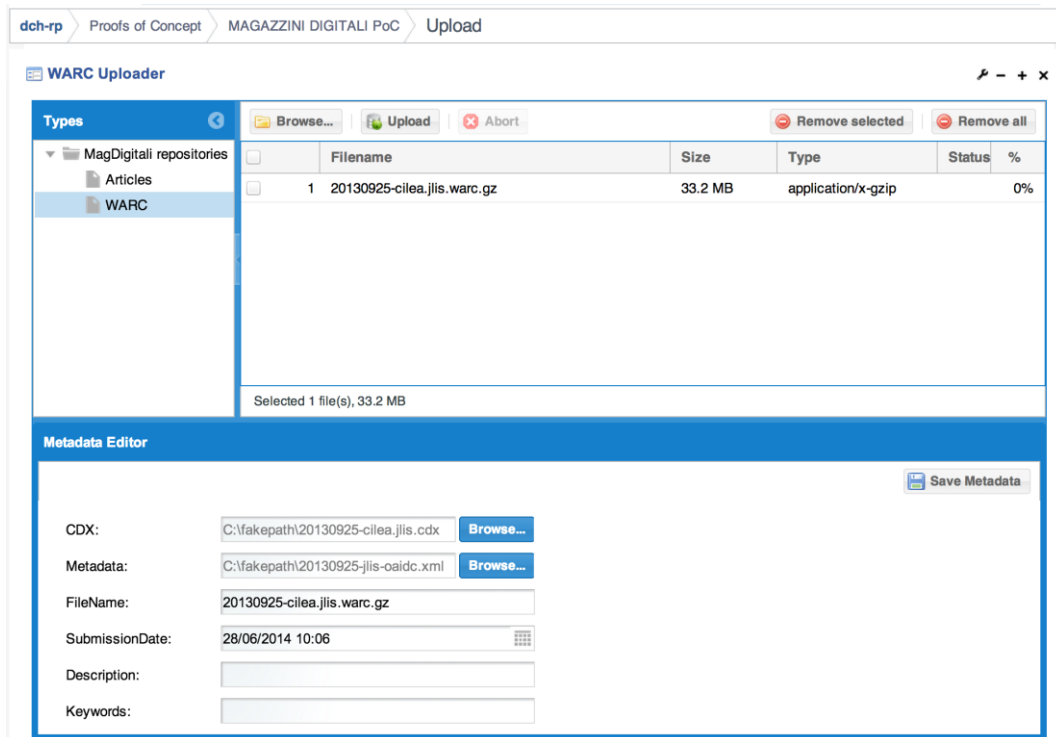


Figure 5: Uploader portlet for the Magazzini Digitali repository of WARC's

Unlike the generic uploader portlet, with the customised uploader metadata is loaded automatically from a description file that is provided by the data owners. To demonstrate fine-grained authorisation special and separate “repository uploader” roles were defined for the Magazzini Digitali and the SITAR archives and have been enforced in the portal. This forbids uploaders of one repository to upload contents on the others and vice versa. The browser portlet of the Magazzini Digitali repository of WARC's is shown in Figure 6.

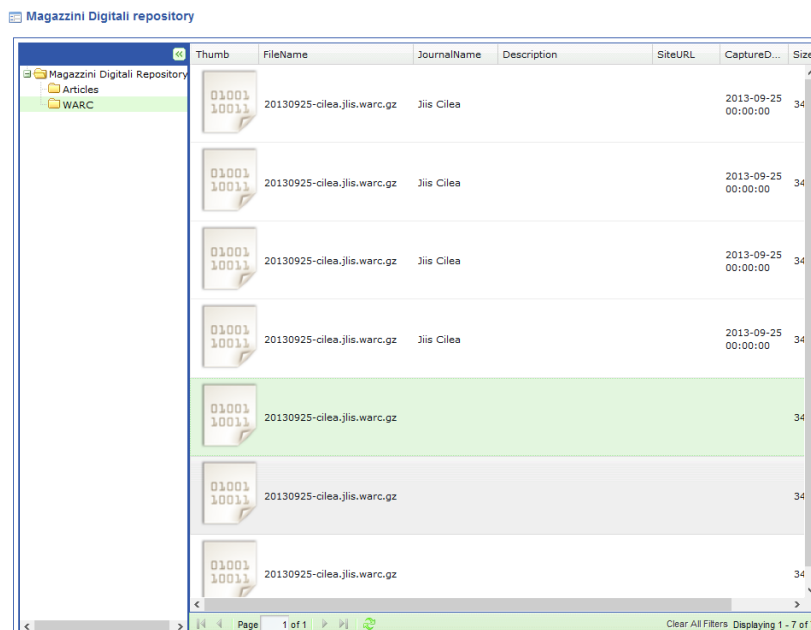


Figure 6: Browser portlet of the Magazzini Digitali repository of WARC's

The uploader and browser portlet of the SITAR repository are shown in the Figures 7 and 8 below.

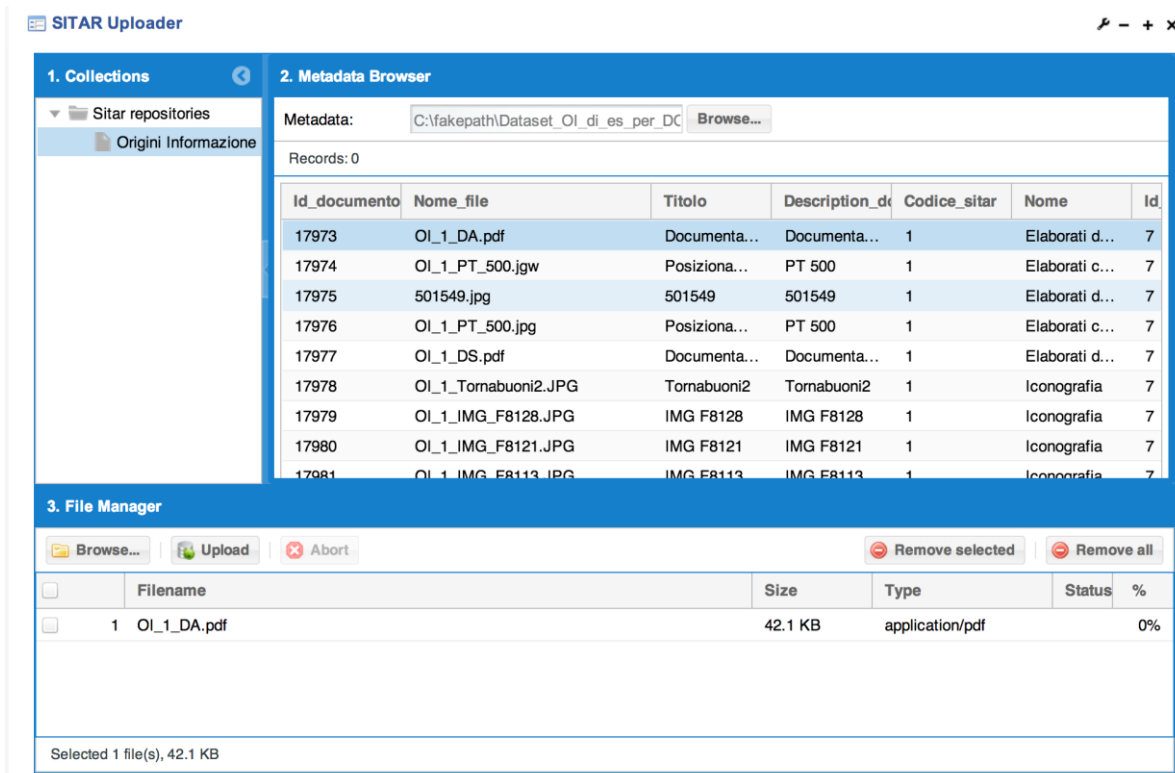


Figure 7: Uploader portlet for the SITAR repository

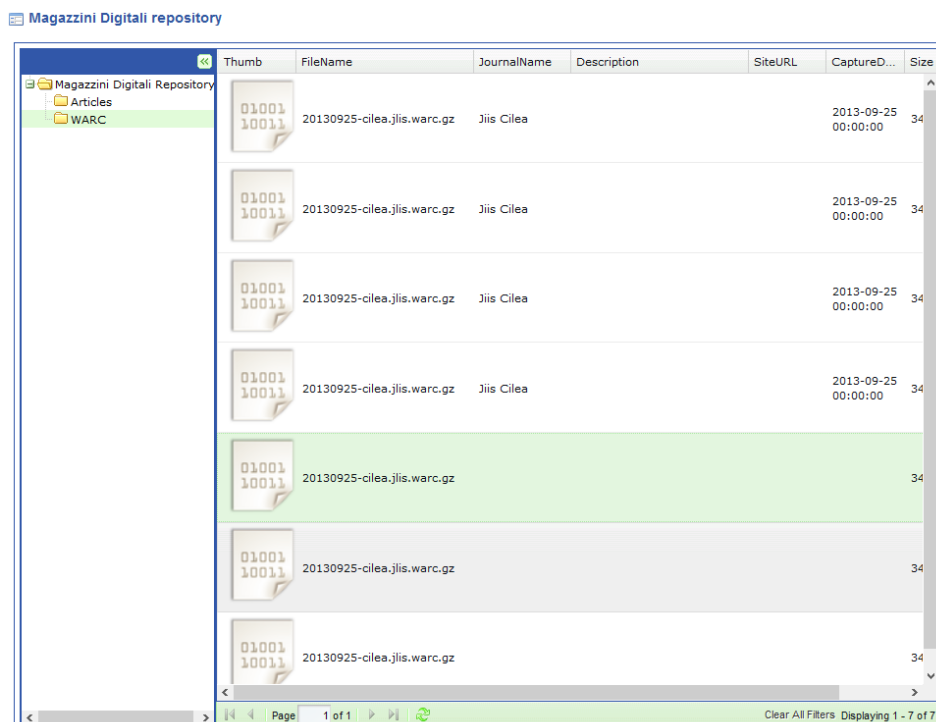


Figure 8: Browser portlet of the SITAR repository

As it emerges from the figures reported above, this second PoC has successfully demonstrated that customised uploaders can allow DCH institutions to make use of eCSG for the storing of their digital assets in automatic way. Moreover, ICCU has now a concrete example of the benefits of using federated credentials to access Service Providers belonging to the IDEM federation.

While doing this, at INFN Catania we have learned how to build an uploader portlet that can be customised in an easy and quick way for different metadata schemas and formats and this will allow further adaptations to other kind of repositories straightforward.

7 EXPERIMENT 5: A LONG-TERM DATA PRESERVATION PLATFORM

7.1 RATIONALE

The experiment on long-term data preservation explores the capabilities and service levels of data preservation with respect to standards, services and methods in the Cloud. It aims to deploy a platform that is able to create and access community specific and OAIS compliant Information Packages (IPs).

We want to gather provided data collections through the OAI-PMH protocol and build community specific IPs for further preservation.

The deployed system is based on open technologies, standards and recommendations like Tomcat, ODE, SOLR and RDF (OAI-ORE). The process of packaging is configurable through a BPEL orchestration of services. The setup and description follows in the next chapter.

For the first use case we collaborate with data from OpenAire.eu, which provides us via OAI-PMH with metadata and PDF documents to setup a running and testable system. We first focus on the metadata package generation, including community metadata.

7.2 SETUP & DESCRIPTION OF TOOLS

Our preservation system consists of a multi-layered architecture, depicted as follows:

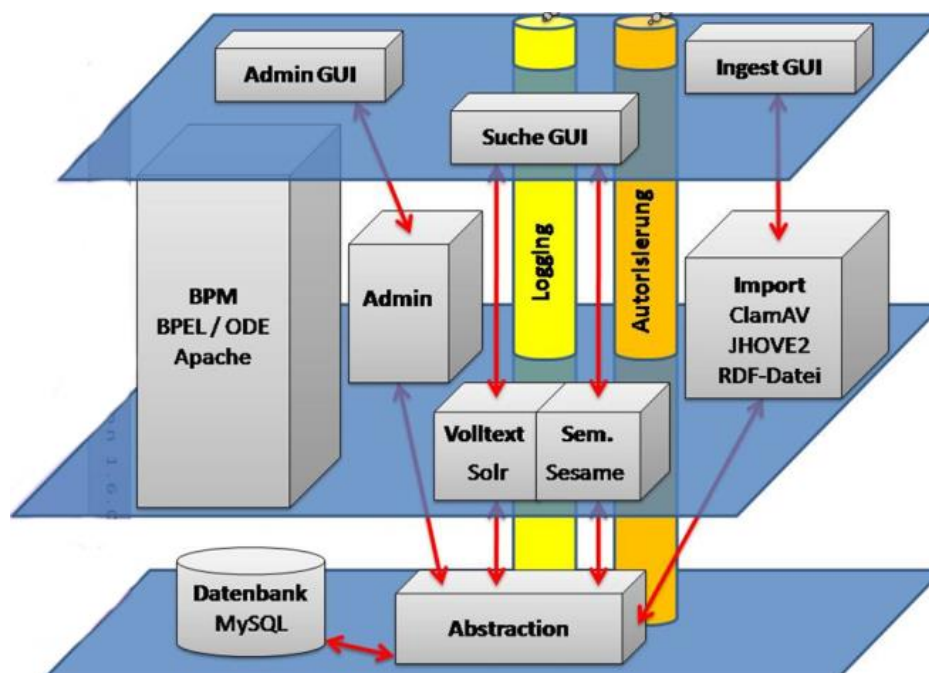


Figure 9: Architecture of the preservation system

It runs within a standard Linux operating system like Ubuntu or Debian and uses Apache as webserver and Tomcat as application server. Further main components are Apache ODE providing the BPEL-Engine to configure and run various services. MySQL

is used as database system, to provide created IPs for access (see Figure 9). SOLR run the full-text search to find the packaged information. The platform also includes an authorization mechanism and logging of provenance information.

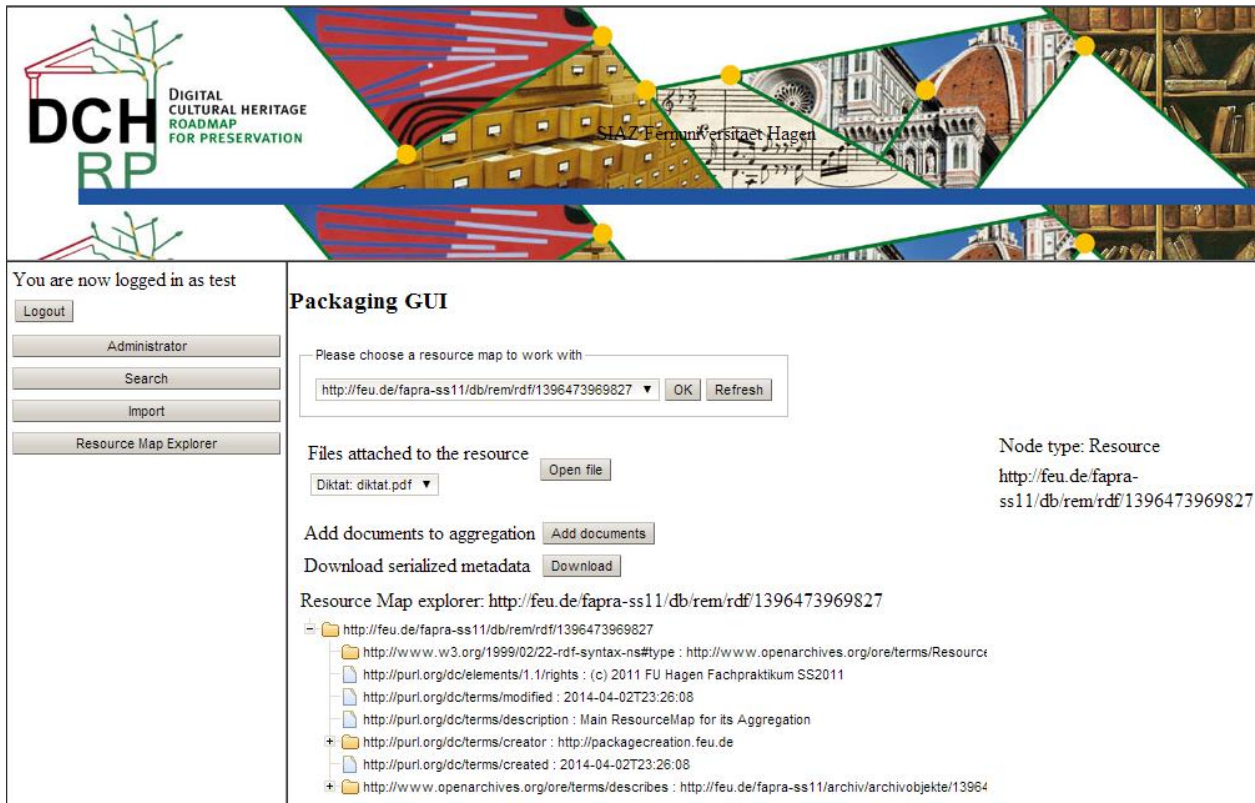


Figure 10: The envisaged GUI for packaging datasets for archival

7.2.1 Prerequisites

To install the current system you need to run a virtual machine based on Linux, preferably Ubuntu LTS.

The following software is required to build and run the services:

Software	Description	Installation	Version
Subversion	Check out source code	apt-get install subversion	
Maven	Build binaries	apt-get install maven	
Zip	Create zip files (BPEL)	apt-get install zip	
Tomcat	Application server that run the service	apt-get install tomcat6	
MySQL	Database that holds AIPs	apt-get install mysql	

Perl	Execute scripts	installation	apt-get install perl	
------	-----------------	--------------	----------------------	--

7.2.2 Installation

The following installation steps are necessary:

- Checkout the code:
 - `svn checkout svn://svn.lgmmia.fernuni-hagen.de/fapra/siaz-ss2012/Install/trunk/ Install`
- Change to working directory:
 - `cd Install;`
- Run the command, which uses maven to deploy the system:
 - `./fapra preinstall build install restart all`
- Further adaptations and configurations to host name, URLs etc. are currently necessary.

7.2.3 Modules

Further modules need to be installed to provide certain functionalities:

- Virus checking: ClamAV (ClamAV.net)
- File & Format checking: JHove (jhove.sourceforge.net/)

7.3 EXPERIMENTATION

As first step we investigated and installed our packaging and access system, as it needed to be adapted due to software and hardware updates. We documented the installation process and conducted a functional testing in a virtual machine. The system is up and running at: <http://kokum.fernuni-hagen.de:8080/JSFGui/start.jsf>

7.4 RESULTS & NEXT STEPS

We applied the first use case with our use case partner OpenAire. OpenAire is a metadata repository service and provides search and access to a variety of resources. The following steps were be investigated:

- Harvest a collection of data objects including meta-data and supplementary data consisting of PDF documents via OAI-PMH.
- Focus on OAIS compliant metadata packaging.

We harvested a collection of 214 documents and their metadata. The records are then packaged into an OAIS compliant package for long-term preservation. The packages were tested within our storage system and can be searched and downloaded.

Following this experiment we will investigate with other use case partners more complex supplementary data objects like 3D visualisations, which need a different ingest processing than ordinary PDF documents.

8 NATIONAL EXPERIMENTS

During the project lifetime, some national side-line projects were undertaken, that ran along or affiliated with experiments coordinated through DCH-RP. To acknowledge the effort and outcomes of these activities this deliverable includes descriptions and outcomes of these in this chapter.

8.1 IDENTITY FEDERATION EXPERIMENT (ICCU, GARR)

To ease uptake and support locally at ICCU for conducting experiment 4, a side-project was conducted to establish and configure an Identity provider (IdP) service at ICCU with the help of INFN Catania and GARR. This IdP service provides authentication services so that institutional users can keep using their institutional authentication credentials for using remote services.

Specifically, this would allow curators working at ICCU to upload digital assets (i.e., data and metadata) through the eCSG using their institutional credentials. Authentication and authorization are then decoupled: the former is done by the user's organisation (ICCU in this specific case), while the latter is done by the Service Provider (the eCSG in this specific case).

Using "IdP in the cloud", GARR provided an IdP as a service to ICCU, populated it with a subset of ICCU staff information, and linked to ICCU's backend credential management system. Also, this IdP been registered in the Italian Identity Federation IDEM (www.idem.garr.it), which is also managed and operated by GARR. Through IDEM this ICCU IdP has also been registered in the eduGAIN (www.edugain.org) inter-identity provider federation.

Figures 4 and 5 show, respectively, the selection of the ICCU IdP in the Where Are You From (WAYF) service of the IDEM federation and its login page where users are redirected to when they sign in on the eCSG.

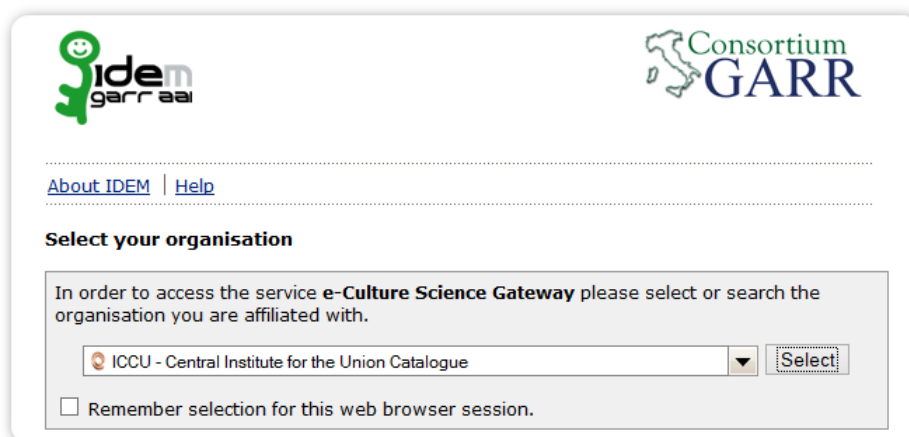
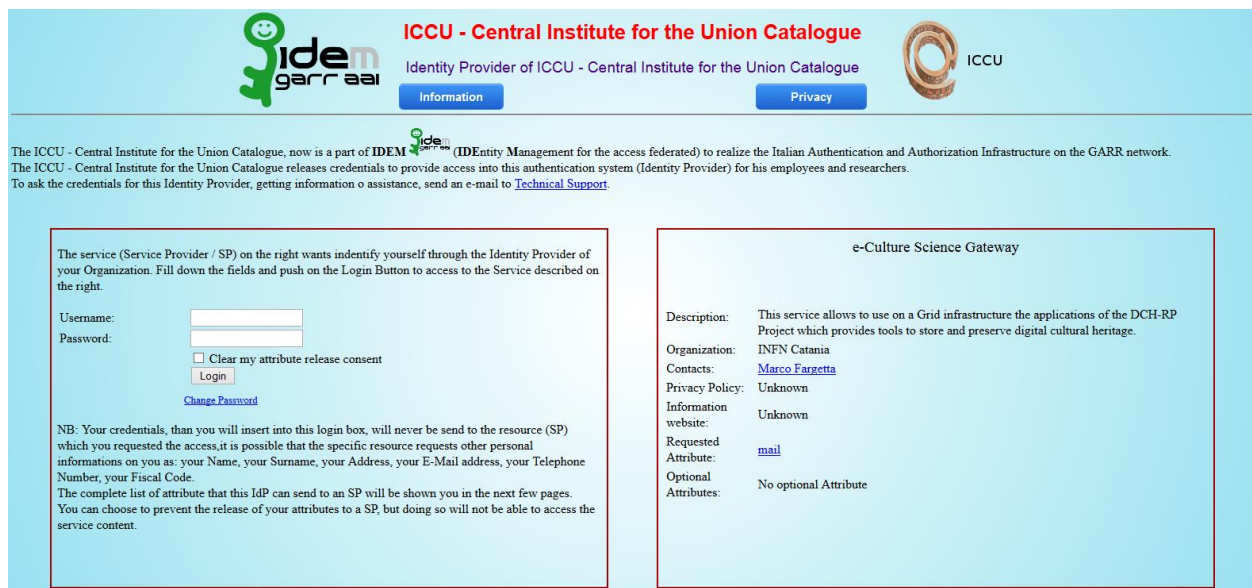


Figure 11: The ICCU IdP in the WAYF service of the IDEM Federation



IDEM
garr aai

ICCU - Central Institute for the Union Catalogue
Identity Provider of ICCU - Central Institute for the Union Catalogue

Information Privacy

The ICCU - Central Institute for the Union Catalogue, now is a part of **IDEM** (IDEntry Management for the access federated) to realize the Italian Authentication and Authorization Infrastructure on the GARR network. The ICCU - Central Institute for the Union Catalogue releases credentials to provide access into this authentication system (Identity Provider) for his employees and researchers. To ask the credentials for this Identity Provider, getting information o assistance, send an e-mail to [Technical Support](#).

The service (Service Provider / SP) on the right wants identify yourself through the Identity Provider of your Organization. Fill down the fields and push on the Login Button to access to the Service described on the right.

Username:

Password:

Clear my attribute release consent

Login

[Change Password](#)

NB: Your credentials, than you will insert into this login box, will never be send to the resource (SP) which you requested the access, it is possible that the specific resource requests other personal informations on you as: your Name, your Surname, your Address, your E-Mail address, your Telephone Number, your Fiscal Code.
The complete list of attribute that this IdP can send to an SP will be shown you in the next few pages. You can choose to prevent the release of your attributes to a SP, but doing so will not be able to access the service content.

e-Culture Science Gateway

Description: This service allows to use on a Grid infrastructure the applications of the DCH-RP Project which provides tools to store and preserve digital cultural heritage.

Organization: INFN Catania

Contacts: [Marco Fargetta](#)

Privacy Policy: Unknown

Information website: Unknown

Requested Attribute: [mail](#)

Optional Attributes: No optional Attribute

Figure 12: Login page of the ICCU IdP re-directed from the eCSG

8.2 EXPERIMENTING WITH NATIONAL E-INFRASTRUCTURES (PSNC, SDL)

This subsection provides a summary of the experiment activities. For the full description please refer to section 12.

8.2.1 Experiment description

Participants and motivation

The experiment involved Polish DCH institution: Silesian Digital Library (<http://www.sbc.org.pl>) and Polish e-Infrastructure services namely Archival Services of the PLATON - Science Services Platform (<http://storage.pionier.net.pl/en>).

Silesian Digital Library (SDL) is the second largest regional digital library in Poland (100 000 items). The content creators include public libraries, academic and educational institutions, cultural institutions, publishers and archives, museums and Protestant commune. The assets includes institutional collections: regional heritage, rare materials, educational materials, scientific and research publications, doctoral theses, periodicals and special collections as well as private collections.

The motivations for this proof of concept are the following:

- While the capability of the SDL infrastructure addresses today's needs, it is predicted than in several years, the volume of the digitized content exceeds current capacity of the SDL infrastructure.
- The level of the data protection on the physical level, however high compared to most other DCH institutions, must be improved in future, in order to preserve data even from local disasters.
- Third, collecting the data from distributed locations is still partially manual. While most institutions upload the data to SDL servers already, quite a few contributors still provides the data by sending the storage media (disks, DVDs, Blue-Rays) using a surface mail or courier service.

- Even large data sets located in Katowice are remotely accessed by the collaborators. Complete replication of data and/or their caching closer to users would improve performance and user experience while interacting with large datasets and limit the load on the central infrastructure.
- In addition to dealing with high-resolution versions, preparation and management of the presentation versions of the digital assets should be supported by appropriate software platform. This is already addressed by dLibra package, developed and maintained by PSNC, used by most of the Polish Digital Libraries.

8.2.2 Application to Archival Services and National Data Storage tools to DP processes

Within the PoC, services and tools configuration was worked out in order to address the SDL use-case needs related to DP. Several kinds of tests were conducted to assess the usability of the solutions provided by the national e-Infrastructure for implementing DP workflows.

Archival Services

For the purposes of the PoC, a profile was created for SDL in the Archival Services. It was shared by institutions participating in PoCs in order to collaboratively store, access and share the data.

Archival Services provided the virtually unlimited storage space which addresses the limitations of the storage space available at SDL. In addition replication of data and meta-data implemented transparently to end-users by Archival Services internal mechanisms ensured the data safety and long-term availability even despite possible local disasters. Finally, data integrity control guaranteed that possible media failures are automatically detected and data replicas are recovered based on the additional replicas maintained by the system.

NDS2 clients

Participants of the PoC were given a NDS2 project client-side tools for Windows and Linux.

Virtual drives enabled intuitive access to shared, secure and safe storage space, despite putting the data transparently into the remote SFTP servers of the Archival Services. Applying NDS2 virtual drives enabled using even legacy applications for data management and processing on top of the remote storage space. For instance, it was possible to save the data to the virtual filesystem directly from the image editing application. NDS2 clients also supported browsing the share directory structure using typical tools such as Windows Explorer, Total Commander or Nautilus. In turn GUI enabled managing massive data uploads and downloads of the digitalized cultural assets.

8.2.3 PoC organisation

PoC was organised in two phases: evaluating the e-Infrastructure usability from the SDL perspective and testing the data storage and access solutions by the end-users in contributing institutions. For the purposes of PoC, testing subset of digital assets was selected. Also the group of users in the SDL and contributing institutions was chosen.

SDL tests

SDL test included two kinds of aspects. First, it was tested if the data structure including directories and files hierarchy can be implemented in Archival Services. For this purpose, the existing archive directory structure was snapshotted (using tar tool) and “unpacked” on the Archival Services side. During the tests no issues with replicating the data structure was encountered.

Second, selected testing data sets were transferred from the SDL premises to the Access Nodes of the Archival Services in PSNC, Poznań. For the transmission, NDS2 GUI client was used, including its massive data transfer monitoring and control functionality.

Third, users in SDL partnering institutions were asked to access the data sets transferred to the shared data store configured in Archival Services.

Feedback and discussion

During the PoC, DCH institution feedback was collected. It is discussed in this section.

SDL feedback

As discussed earlier, SDL is on one hand the provider of the resources and services to the DCH community in two voivodeships, and is the e-Infrastructure services user on the other.

Application of PLATON Archival Services to host the collaborative data storage space performed within this PoC, however limited to several datasets, enabled drawing conclusions on overall usability of this e-Infrastructure from DP processes perspective.

SDL workers appreciated the improved reliability, storage capacity and efficiency of the nation-wide storage infrastructure, compared to their local infrastructure. Compatibility of the outsourced storage service with the data organisation convention was also confirmed.

Possible usage models, considered by the institution include the full outsourcing model as well as hybrid model. In the former, all the source and high quality data will be stored in the external infrastructure. The local system will be used mainly for acquiring, processing and presenting the data through CMS platforms such as dLibra. In the latter, selected subset of the digital assets will be stored locally in SDL. Extra replicas will be kept remotely in Archival Services in order to ensure their long-term safety and resistance to local disasters. In addition, virtually unlimited storage space of Archival Services will be used in order to increase the archival storage capacity available to SDL and its users.

End-users feedback

End-users experience during the tests was positive in general. Ease of use of the virtual filesystem interfaces was appreciated, especially by voluntary participants of Social Digitization Workshop. Also the performance of the data storage and access tools considered as acceptable. This applies to both highly interactive tasks such as browsing the filesystem structure and massive amounts of the data where average upload and download speed counts. As uploading large data sets effectively and reliability in the environment where the network quality is poor (this happens especially in case of small partnering institutions) poses a challenge, NDS2 GUI's support for monitoring and managing upload and download jobs was considered useful.

8.2.4 Conclusions

Experiment involving Silesian Digital Library and Archival Services of the PLATON project in Poland prove that proper application of the e-Infrastructure services to implementing digital preservation processes may be effective and have limited negative impact on the user experience. Usage of data replication functionality of the Archival Services enabled improving the data durability, safety and availability. Virtually unlimited storage capacity of PLATON infrastructure enabled extending the storage space available to SDL and its partnering institutions. NDS2 tools enabled users to keep their methods and habits related to storing and accessing data, while performing data acquisition, processing and preparation for archival. Performance offered by the remote storage system was acceptable for the use case.

Important observations related to the role of the Silesian Digital Library in the e-Infrastructure services take up. SDL is the is the example of the very open, collaborating however demanding community side-partner of the e-Infrastructure providers. Thanks to its technical competence as well as awareness of the opportunities brought by the e-Infrastructures it is an early adopter of the services provided by e-Infrastructures. It also helps defining high-level and real-life requirements of the solutions to be provided in order to address DP processes.

SDL is also effective promoter of e-Infrastructure services usage in the DCH institutions environment. SDL also transferred the related know-how to the partnering institutions.

We argue, that partnership among e-Infrastructure and community leaders and visionaries greatly stimulates and contributes to the take up of cloud services in the DCH domain.

e-Infrastructures must also offer the adopters and users of their services a flexibility. PoC conducted in collaboration with SDL shows, that putting efforts into customizing the service for the needs of large, partnering institutions, pays off, as it help in the services take up and integration with the DP processes existing in the DCH institutions.

9 CONCLUSION

Comparing the first Proofs of Concept phase with the second phase one cannot but identify significant progress in the DCH community on several levels.

Firstly, the first experiments in the project were very much focused around low-level tools that were already known *within* the DCH-RP consortium, mainly the memory institutes and partners in direct contact with these (e.g. ICCU, RA, BELSPO). This can be attributed to the project being in its early phase, and a roadmap to align the activities not yet being in place, though planned this way (c.f. the project's DoW). Yet, the experiments of the first Proofs of Concept phase yielded sufficient results to contribute to the subsequent intermediate Roadmap [R 2] evolving from the study provided early on in the project.

The second Proofs of Concept experiments in contrast focus more on usable solutions and services that have the potential of being integrated into existing solutions due to promising functionality – all the way towards experimenting with assembling a preservation platform fit for purpose. The results already indicate a much easier and valuable mapping of requirements and use cases described in the roadmap into functional and non-functional capabilities and potential solutions for them in a future version of the roadmap. The intermediate roadmap provided for the first time a specific list of functional capabilities that must be satisfied in a conceptual preservation platform, such as: OAIS compliance, automatic metadata capture and extraction, authenticity and integrity of data, distributed storage systems; the second Proofs of Concept phase responded with experimenting with specific solutions and services addressing some of these capabilities. In turn, the expected results allow the next iteration of the roadmap (to be published as D3.5 [R 3]) to further specify and concretise specific short-term and medium-term plans for the DCH community.

Second, the project partners significantly improved the yield of “funding-to-experimentation” ratio through developing and signing strategic MoUs with those projects that developed tools and services, and are the curators and stewards of data used or planned to be used in the second round of experiments. This is most visible in the experiments 1, 2, 4 and 5 where partners or data owned by these partners were involved through the projects SCAPE, SCIDIP-ES, ARIADNE and APARSEN. Compared with the mostly internally conducted experiments in the first PoC phase one cannot but note a significantly improved traction in the experiments themselves.

Third, it is well worth revisiting past experiments that failed when compared to the expected results, and improve on the tools and experimentation design through open and genuine analysis of the issues that were at hand – experiment 4 demonstrates this principle very well.

Fourth, it may prove beneficial to identify and select a few DCH or DP institutes that are *open, technology-savvy, yet demanding* in terms of their needs, and are *open and willing* to engage with e-Infrastructures in terms of collaboration towards arriving at a service model that provides benefit to either party involved. Such institutes are able and have the capacity to engage in a more formalised requirements engineering and service acceptance testing cycle hence provide a secondary benefit to all those DCH and DP community members who do not have this capability but wish to engage anyhow. Such

institutes, once having decided to take up the developed solution, often also serve as multipliers into the target community for further take up of e-Infrastructure services.

Based on the results and next steps provided in earlier sections the following conclusions may already be drawn based on the progress made so far:

The “Matchbox” tool developed in the SCAPE project is clearly not designed for end users. However, preliminary results and existing integration with other platforms and services may well indicate that integrating a parallelised Matchbox tool into a higher-level preservation platform at reasonable cost has the potential of providing a scalable service for duplication filtering across large archives of digital data.

Experimenting with B2SHARE and eCSG both so far demonstrated that current implementations are only not entirely suitable for mass-upload of data into a preservation archive. This perhaps indicates that initial ingestion of data into an archive might best be designed as a separate activity in the overall preservation lifecycle even though specialised uploader portlets for the eCSG used in experiment 4 were key to the progress of that experiment.

This points to a more general issue concerning metadata. The current landscape of metadata standards does not support any indication of convergence towards a common standard across domains. Tool developers need to take this into account when designing software – unless they specifically target one market segment. This project’s experience with the eCSG points into that direction that pluggable metadata parsers and handlers seem the better way over general-purpose elements.

Even though all but one experiment were finished, the following recommendations for the DCH roadmap may be concluded as follows:

Recommendation 1: Tools designed for installation on end user IT equipment, and intended for installation by end users, should be *as easy as possible to install* – ideally by a single action. It should be as easy as copying a number of files into one directory, followed by double-clicking an icon. Exemplar applications are the Eclipse Foundation’s IDE “drops”, or Firefox releases that literally require little more than copying a number of files into a directory of choice, or on a platform level, the Mac OS X application installation process comprising of one simple dragging the application icon to drop it over the system’s Applications folder.

Recommendation 2: Tools integrating with typical Linux package management systems such as apt-get for Debian based Linux distributions or yum for Red-Hat based systems must provide an appropriate package for all supported hardware architectures (32bit and 64bit), including a well-defined and well-managed dependency manifest, so that, after downloading the package, a single command to install that package automatically installs any missing dependency without further unnecessary interaction.

Recommendation 3: Ideally, tools identified as suitable for inclusion in the DP roadmap should have active maintainers for the used/desired target platforms who ensure that recommendations 1 and 2 are adequately met, so that installing an application, tool, or service requires little more than issuing a command similar to “`sudo apt-get install scape-matchbox`”.

Recommendation 4: If some software does not entirely match DCH requirements, investigate whether it has a modular design, preferably including well-documented extension interfaces (c.f. “plug-in” and “connector” design), for which DCH-specific extensions might be developed at greatly reduced cost. Aim to find partners and communities in the same market segment that might join in the maintenance effort for either the entire tool, or specific plugins.

Recommendation 5: Aim to avoid vendor lock-in by developing a service-oriented architecture for the DCH digital preservation landscape (or a desired “Preservation-as-a-Service” platform) including strategically placed and mandated publicly defined standards governing the interfaces between the services within the platform. Aim to avoid or reduce to an absolute minimum second-level dependencies such as one service directly depending on one or more specific instances of other services – operational maintenance and reliable rollout is next to impossible in an entangled network of dependencies. Ideally, an SOA with the right abstraction level and service scoping allows upgrading one service entirely independently from any other service.

Recommendation 6: Before defining the technical architecture of the preservation services and platforms, define and agree on the business process(es) you wish to implement in the technical architecture. Good business process modelling results not only in a business process architecture satisfying the requirements, but allows changes in its orchestration and sequences without redefining or altering the defined activities.

Recommendation 7: In the process of further developing the roadmap, describe each service that is required, and which capabilities it is expected to implement. For example, describing a storage service the roadmap might attach the following capabilities to it:

- Bit-level preservation of each digital object stored in and managed through this service;
- Data access and modification policies: Read-only, copy-on-write, transactional, or version-controlled;
- Self-service configuration of object replicas
- Self-service configuration of geospatial distribution of replicas
- Central or distributed data access points
- Transparent storage medium obsolescence management

These recommendations are arguably very technical in their nature. However, describing the overall results from both Proof of Concepts phases in the project one observation is key: Those experiments that were conducted with help and support of technical domain experts (Experiment 3, 4, 5) thrived well, installation was done smoothly and provided more crisp results. This is not to negate the value of the other experiments. The point is to reinforce the observation made already in the first Proofs of Concept phase in the project: *CH users are neither IT experts (or savvy with IT management and operation) – nor are they supposed to be.*

When engaging with e-Infrastructures, user communities and especially the CH community needs to be aware of different mandates hence different objectives of e-Infrastructures and customers, which will inevitably result in a gap analysis of “services needed vs services provided”. While it is clear that e-Infrastructures are supposed to support research and scientific communities in Europe it is not clearly stated nor mandated *how exactly* this has to happen. While e-Infrastructures such as PRACE,

EUDAT have a clear mission and mandate bestowed upon them by their members, these are targeting specific communities in Europe hence able to provide more focused services towards these communities. EGI, on the other hand, has a clear mission to scale out its support from High-energy Physics towards essentially *any* research community in Europe.

Regardless, there is a clear gap emerging from the experiments conducted in the entire project, which we wish to convey as the last two but not least recommendation:

Recommendation 8: The DCH community relies very heavily on appropriate ICT support geared towards real end users. This again is an observation, not a judgement, which needs to be appropriately taken into account. When engaging with e-Infrastructures, a third stakeholder must be considered for inclusion: The first stakeholder is clearly the DCH community as the consumer of any ICT services provided to them. The second stakeholders are the e-Infrastructures in Europe (and potentially worldwide) that provide a certain set of underpinning ICT services. The third, possibly new, stakeholders are service integrators and platform providers offering services tailored to the DCH community. The business relationships and value chain up to the memory institutes most likely will look like this:

1. Service consumer – Memory institutes, digital libraries, etc.
2. Service provider – ICT experts who are domain experts in the CH field
3. E-Infrastructure suppliers – Providing general-purpose infrastructure services on-demand and at scale to service providers.

Recommendation 9: Regardless of who is taking up the task of doing so, the strategy of sketching, developing, refining and eventually executing a strategy of providing a preservation as a service Cloud platform to the DCH community, the involved stakeholders need to be very clear in who their target audience is, and which institutes among these are suitable for early adoption and serve as multipliers into the “market” of DCH and DP.

10 REFERENCES

R 1	DCH-RP D3.1 “Study on a Roadmap for preservation”, R. Ruusalepp (EVKM), M. Dobrova (EVKM), March 2013. http://www.dch-rp.eu/getFile.php?id=114
R 2	DCH-RP D3.4 “Intermediate version of the Roadmap”, B. Justrell (RA), L. Balint (NIIFI), E. Toller (RA), R. Ruusalepp (EVKM), January 2014. http://www.dch-rp.eu/getFile.php?id=221
R 3	DCH-RP D3.5 “Final version of the Roadmap”, to be published,
R 4	DCH-RP D4.1 “Trust building report”, R. Ruusalepp (EVKM), B. Justrell (RA), L. Florio (TERENA), April 2014. http://www.dch-rp.eu/getFile.php?id=274
R 5	DCH-RP D5.3 “Report on the first proof of concept”, M. Drescher (EGI.eu), E. Toller (RA), R. Vandenbroucke (BELSPO), September 2013, http://www.dch-rp.eu/getFile.php?id=198
R 6	Fourth DCH-RP plenary meeting, Catania, Italy, 20 – 21 January 2014; http://www.dch-rp.eu/getFile.php?id=340 (reserved space, may require login)
R 7	Fifth DCH-RP plenary meeting, Tallinn, Estonia, 24 – 25 April 2014; http://www.dch-rp.eu/getFile.php?id=339 (reserved space, may require login)

11 ANNEX 1: INSTALLING AND TESTING MATCHBOX

During the PoC2 test was target to install and test some tools created during the Scalable Preservation Environments project - SCAPE project (<http://www.scape-project.eu/>). Although there are available amd64 compiled package it is not possible to change the bit signature of a binary. It has to be compiled for a certain architecture where the system will be in use. Because of the technical issues the only tested tool was Matchbox - Duplicate image detection tool.

There are several important tools/modules to install:

1. Install GCC, the GNU Compiler Collection

The GNU Compiler Collection includes front ends for C, C++, Objective-C, Fortran, Java, Ada, and Go, as well as libraries for these languages (libstdc++, libgccj,...). GCC was originally written as the compiler for the GNU operating system. The GNU system was developed to be 100% free software, free in the sense that it respects the user's freedom.

2. Install “build-essentials”

This package contains an informational list of packages, which are considered essential for building Debian packages. This package also depends on the packages on that list, to make it easy to have the build-essential packages installed.

3. Install g++

Released by the Free Software Foundation, g++ is a *nix-based C++ compiler usually operated via the command line. It often comes distributed with a *nix installation, so if you are running Unix or a Linux variant you likely have it on your system.

4. Install CMake

CMake is the cross-platform, open-source build system. CMake is a family of tools designed to build, test and package software. CMake is used to control the software compilation process using simple platform and compiler independent configuration files. CMake generates native makefiles and workspaces that can be used in the compiler environment of your choice.

CMake has the following dependencies that need to be satisfied:

- libarchive-3.1.2 (<http://www.linuxfromscratch.org/blfs/view/svn/general/libarchive.html>)
- curl-7.36.0 (<http://www.linuxfromscratch.org/blfs/view/svn/basicnet/curl.html>)
- Lib Boost (<http://ubuntuforums.org/showthread.php?t=1725216>)

and optionally

- Cmake GUI (<http://www.linuxfromscratch.org/blfs/view/svn/x/qt4.html>)
and https://secure.mash-project.eu/wiki/index.php/CMake:Quick_Start_Guide

More information can be found at:

- <http://www.cmake.org/cmake/resources/software.html>
- <http://www.linuxfromscratch.org/blfs/view/svn/general/cmake.html>

5. Install Python

Python is a widely used general-purpose, high-level programming language. Its design philosophy emphasizes code readability, and its syntax allows programmers to express concepts in fewer lines of code than would be possible in languages such as C. Ubuntu 14.04 LTS has available Python 3.4.0, but for the Matchtool is necessary Python 2.7.

Python's important dependencies are: `libsqlite3-dev`, `sqlite3`, `bzip2` and `libbz2-dev`.

More information is available at <https://www.python.org/>

6. Install OpenCV

OpenCV is the most popular and advanced code library for Computer Vision related applications today, spanning from many very basic tasks (capture and pre-processing of image data) to high-level algorithms (feature extraction, motion tracking, machine learning). It is free software and provides a rich API in C, C++, Java and Python. Other wrappers are available. The library itself is platform-independent and often used for real-time image processing and computer vision. OpenCV has already lot of interesting developments like face detection, similar object finder and etc. , see also screenshots below.

Creating and compiling the OpenCV is one of the most important step.

OpenCV has quite a long list of dependencies, mostly supporting the various image filtering and detection algorithms: *build-essential*, *libgtk2.0-dev*, *libjpeg-dev*, *libtiff4-dev*, *libjasper-dev*, *libopenexr-dev*, *cmake*, *python-dev*, *python-numpy*, *python-tk*, *libtbb-dev*, *libeigen2-dev*, *yasm*, *libfaac-dev*, *libopencore-amrnb-dev*, *libopencore-amrwb-dev*, *libtheora-dev*, *libvorbis-dev*, *libxvidcore-dev*, *libx264-dev*, *libqt4-dev*, *libqt4-opengl-dev*, *sphinx-common*, *texlive-latex-extra*, *libv4l-dev*, *libdc1394-22-dev*, *libavcodec-dev*, *libavformat-dev*, *libswscale-dev*.

Final configuration of the OpenCV needs the following settings:

```
-- Video I/O
--   DC1394 1.x:                NO
--   DC1394 2.x:                YES (ver 2.2.1)
--   FFMPEG:                    YES
--   codec:                      YES (ver 54.35.0)
--   format:                     YES (ver 54.20.4)
--   util:                       YES (ver 52.3.0)
--   swscale:                    YES (ver 2.1.1)
--   gentoo-style:              YES
```

```

-- GStreamer: NO
-- OpenNI: NO
-- OpenNI PrimeSensor Modules: NO
-- PvAPI: NO
-- GigEVisionSDK: NO
-- UniCap: NO
-- UniCap ucil: NO
-- V4L/V4L2: Using libv4l (ver 1.0.1)
-- XIMEA: NO
-- Xine: NO
--
-- Other third-party libraries:
-- Use IPP: NO
-- Use Eigen: YES (ver 2.0.17)
-- Use TBB: YES (ver 4.2 interface 7000)
-- Use OpenMP: NO
-- Use GCD: NO
-- Use Concurrency: NO
-- Use C=: NO
-- Use Cuda: NO
-- Use OpenCL: YES
--
-- OpenCL:
-- Version: dynamic
-- Include path: /home/OpenCV/opencv-2.4.9/3rdparty/include/opencvcl/1.2
-- Use AMD FFT: NO
-- Use AMD BLAS: NO
--
-- Python:
-- Interpreter: /usr/bin/python2 (ver 2.7.6)
-- Libraries: /usr/lib/i386-linux-gnu/libpython2.7.so (ver 2.7.6)
-- numpy: /usr/lib/python2.7/dist-packages/numpy/core/include (ver
1.8.1)
-- packages path: lib/python2.7/dist-packages
--
-- Java:
-- ant: NO
-- JNI: NO
-- Java tests: NO
--
-- Documentation:
-- Build Documentation: YES
-- Sphinx: /usr/bin/sphinx-build (ver 1.2.2)
-- PdfLaTeX compiler: /usr/bin/pdflatex
--
-- Tests and samples:
-- Tests: YES
-- Performance tests: YES

```



```
-- C/C++ Examples:          YES
--
-- Install path:            /usr/local
--
-- cvconfig.h is in:       /home/OpenCV/opencv-2.4.9/build
-----
--
-- Configuring done
-- Generating done
-- Build files have been written to: /home/anz/Downloads/OpenCV/opencv-2.4.9/build
```

More information:

<https://help.ubuntu.com/community/OpenCV>

<https://github.com/jayrambhia/Install-OpenCV/tree/master/Ubuntu>

<http://milog.blogspot.com/2012/12/install-opencv-ubuntu-linux.html>

7. Install Matchbox

The idea of the tool is that there are numerous situations in which you may need to identify duplicate images in collections, for example:

- to ensure that a page or book has not been digitised twice
- to discover whether a master and service set of digitised images represent the same set of originals
- to confirm that all scans have gone through post-scan image processing.

Checking to identify duplicates manually is a very time-consuming and error-prone process.

In the Readme file provided on the Github repository there are information about how to compile and set up the Matchbox toolset. There is an image of a virtual machine, that we use for training within the SCAPE project but during the test there were no necessary tools to use it.

However the preparation and starting the application from scratch is requiring several ICT skills and development knowledge, especially when there is a need to use any Linux desktop version without any development tools and specific libraries, modules and programs.

Below are written down necessary steps including information about to make Matchbox to work.

Building CMAKE:

```
$ ./configure
The C compiler identification is GNU 4.9.0
The CXX compiler identification is GNU 4.9.0
Check for working C compiler: /usr/bin/cc
```

```
Check for working C compiler: /usr/bin/cc -- works
Detecting C compiler ABI info
Detecting C compiler ABI info - done
Check for working CXX compiler: /usr/bin/c++
Check for working CXX compiler: /usr/bin/c++ -- works
Detecting CXX compiler ABI info
Detecting CXX compiler ABI info - done
COMPARE: Opencv found.
Boost version: 1.54.0
Found the following Boost libraries:
serialization
Configuring done
$
```

Compiling Matchbox:

```
$ make
...
[ 95%] Building CXX object
DPQA_Compare/CMakeFiles/mb_compare.dir/DPQA_Compare.cpp.o
Linking CXX executable mb_compare
[ 95%] Built target mb_compare
Scanning dependencies of target mb_extractfeatures
[ 97%] Building CXX object
DPQA_ExtractFeatures/CMakeFiles/mb_extractfeatures.dir/DPQA_Ext
ractFeatures.cpp.o
Linking CXX executable mb_extractfeatures
[ 97%] Built target mb_extractfeatures
Scanning dependencies of target mb_train
[100%] Building CXX object
DPQA_Train/CMakeFiles/mb_train.dir/DPQA_Train.cpp.o
Linking CXX executable mb_train
[100%] Built target mb_train

$ ls
CMakeCache.txt  cmake_install.cmake  CPackSourceConfig.cmake
DPQA_ExtractFeatures  DPQA_Train
CMakeFiles      CPackConfig.cmake   DPQA_Compare  DPQAlib
Makefile

$ sudo make install
[ 93%] Built target DPQAlib
[ 95%] Built target mb_compare
[ 97%] Built target mb_extractfeatures
[100%] Built target mb_train
Install the project...
-- Install configuration: ""
```

```
-- Installing: /usr/bin/FindDuplicates.py
-- Installing: /usr/bin/MatchboxLib.py
-- Installing: /usr/lib/libDPQAlib.so
-- Removed runtime path from "/usr/lib/libDPQAlib.so"
-- Installing: /usr/bin/mb_compare
-- Removed runtime path from "/usr/bin/mb_compare"
-- Installing: /usr/bin/mb_extractfeatures
-- Removed runtime path from "/usr/bin/mb_extractfeatures"
-- Installing: /usr/bin/mb_train
-- Removed runtime path from "/usr/bin/mb_train"
```

More information is available at <https://github.com/openplanets/matchbox>.

12 ANNEX 2: EXPERIMENTING WITH NATIONAL E- INFRASTRUCTURE IN POLAND

12.1 EXPERIMENT DESCRIPTION

12.1.1 Participants and motivation

The experiment involved Polish DCH institution: Silesian Digital Library (<http://www.sbc.org.pl>) and Polish e-Infrastructure services namely Archival Services of the PLATON - Science Services Platform (<http://storage.pionier.net.pl/en>).

Silesian Digital Library (SDL) is the second largest regional digital library in Poland (100 000 items). The content creators include public libraries, academic and educational institutions, cultural institutions, publishers and archives, museums and Protestant commune. The assets includes institutional collections: regional heritage, rare materials, educational materials, scientific and research publications, doctoral theses, periodicals and special collections as well as private collections.

A specific feature of DP process implemented at SDL is that it is distributed across over 60 autonomous contributing institutions. In fact they are located in two administrative districts in Poland - Silesian voivodeship and Lower-Silesian voivodeship.

SDL acts as the coordinator of this distributed digitization activity. It also provides the infrastructure, know-how and support to the partnering institutions. SDL runs its own mass storage service, made available to the digitization contributors. Currently it is capable of storing 300TB of data on disk arrays, with tape-based backup.

From this point of view SDL can be considered as the community-side service and resource provider. However, in order to ensure scalability and long-term sustainability of the solutions and the processes SDL coordinates, collaboration with national e-Infrastructure is necessary.

Silesian Digital Library presented its view on the e-Infrastructure support for DP processes at DCH-RP Concertation Workshop in Tallin¹⁶ (23-24.04.2014). Over the course of this meeting PSNC and SDL discussed and agreed to run a proof of concept of the long-term archival, storage and sharing service for massive amounts of data the library collects and protects in behalf of its collaborators. The experiment was implemented from June to August 2014.

The motivations for this proof of concept are the following. First, while the capability of the SDL infrastructure, addresses today's needs, it is predicted than in several years, the volume of the digitized content exceeds current capacity of the SDL infrastructure. Second, the level of the data protection on the physical level, however high compared to most other DCH institutions, must be improved in future, in order to preserve data even from local disasters. Data sets should be replicated geographically beyond the SDL premises, ideally to other, distant regions. Third, collecting the data from distributed locations is partially manual. Most institutions upload the data to SDL servers, thanks to ongoing improvements of network connectivity, however a lot of contributors still provides

¹⁶ Presentation at: http://www.digitalmeetsculture.net/wp-content/uploads/2014/04/Silesian-Digital-Library-R.Lis_.pdf

the data by sending the storage media (disks, DVDs, Blue-Rays) using a surface mail or courier services. This is inconvenient, error-prone and does not scale, due to the administrative burden in the central location.

Other needs of the institution that can be addressed by e-Infrastructure include the following. First, currently, even large data sets are accessed by the collaborators from the central location in Katowice. Complete replication of data and/or their caching closer to users would improve performance and user experience while interacting with large datasets and limit the load on the central infrastructure. Second, in addition to dealing with high-resolution versions, preparation and management of the presentation versions of the digital assets should be supported by appropriate software platform. This is already addressed by dLibra package, developed and maintained by PSNC, used by most of the Polish Digital Libraries.

12.2 PROOF OF CONCEPT SCOPE

12.2.1 Scope

Having in mind the timeframe and resource availability, SDL and PSNC agreed to conduct a proof of concept related to addressing the first group of issues mentioned, i.e.: scaling the archived content storage capacity, improving the physical protection level of the data as well as providing a platform that enables efficient and reliable storage, collecting and sharing massive data sets including high resolution source and archival versions of the assets.

12.2.2 Data organisation at SDL

Organisation of the distributed digitisation process has impact on the solution provided, therefore it is shortly discussed SDL offers its partnering institutions a storage system with file system interface. Storage space is organised into directories according to structure agreed among institutions. Top level of the structure includes directories dedicated to digital assets types. For instance “cza”, “kar” and “gra” are abbreviations and stand for “czasopisma” (eng. periodicals), kartografia (eng. “cartography”) and grafika (eng. graphics). Listing of the top level directory is presented in Picture 1.

```
root@serw210:/mnt/master# ls -la
...
drwxrwxrwx  20 leser leser   4096 2014-01-07  cza
drwxrwxrwx  374 leser leser  26624 2013-12-30  dru
drwxrwxrwx 1314 leser leser 167936 03-18 13:37  fot
drwxrwxrwx  745 leser leser   67584 03-18 13:45  gra
drwxrwxrwx   58 leser leser    6144 2013-08-22  ink
drwxrwxrwx  434 leser leser   32768 2013-12-12  kar
drwxrwxrwx  351 leser leser   63488 04-15 07:34  poc
drwxrwxrwx   33 leser leser    4096 04-15 07:32  rek
drwxrwxrwx  418 leser leser   34816 04-15 07:28  rtl
drwxrwxrwx  548 leser leser   49152 04-14 11:28  sta
```

Figure 13: Top level of the data structure at SDL

Names convention adopted in SDL enables to include the assets meta-data within the filenames. The file paths include: siglum of the library, space name, type of asset, signature, issue year, issue number and the finally the filename with the extension. The filename itself is composed from: signature, issue year, year number, 4-digit page number (calculated within the unit) separated with by a dash sign. Extension determines the format of the digital asset. Example full path and name to the digital asset file is presented in Picture 2.

```
//bzz/master/cza/ab1234/1922/12/ab1234-1922-12-0001.tif
```

Figure 14: Example file path and name in SDL.

While the data organisation and file and directory naming convention enables including meta-data into directory and filenames, it is also tailored to the technical features of most popular file systems. This facilitates implementation of the convention in various environments, including Windows and Linux servers as well as facilitates possible future migration of the data. For instance, usage of national characters is not allowed as they might be not supported in several environments. Similarly, the length of the path is limited.

12.2.3 Digitization process organisation

The data organisation presented above supports the distributed acquisition of the assets. Assets acquired by particular institutions are put to directories according to the convention agreed. Different assets and collections are put into separate areas of the name space. This enables independent, parallel storage of the data while keeping the possibility to access all data sets organised into unified structure. It also helps preventing duplicated efforts - already digitized assets are visible in the shared storage so that users can quickly realise that a copy of data already exist (in fact users also refer to the digitization plans published by particular institutions).

12.2.4 Social Digitization Workshop

Interesting initiative within SDL is the Social Digitization Workshop. 50+ volunteers including senior citizens, students and employees of the Cultural Institutions participates in the digitization of cultural assets. Specific requirements on the solutions and tools result from this way of supporting the digitization process conducted by professionals. They include need for great user friendliness of data storage and access interfaces, and possibility to integrate standard tools with the offered data storage and handling solution.

12.2.5 Archival Services of the PLATON project

Archival Services are the distributed data storage services for the backup and long-term archival purposes. The service is coordinated by PSNC and implemented by 10 partnering institutions including universities and MANs across the country. The services offers its users tape storage capacity of 12,5 PB and 2PB of the disk cache, distributed in 10 locations.

The Archival Services mechanisms implement automatic data and meta-data replication, as well as periodic, system-side data integrity control. The storage space is available by a set of protocols including SFTP, WebDAV and GridFTP. Popular SFTP clients such as WinSCP or WebDrive can be used under Windows and several sftp tools can be exploited under Linux in order to store and access the data in the service.

Data and users are organised into so-called profiles. A profile is a logically separated storage space available to authorized users. Users are typically authenticated based on X.509 certificates, SSH keys or logins and passwords. The service is in its production phase since 2011, currently holding the total of more than 3,5 PBs of data belonging to 200 profiles.

12.2.6 National Data Storage project data storage and access tools

Archival Services provide the server-side data storage access interfaces including SFTP. In the National Data Storage project¹⁷, several end-user interfaces and tools were developed in order to facilitate the usage of Archival Services as well as improve data security and safety.

NDS2 tools include virtual drive for Windows, virtual filesystem for Linux as well as portable Java application. Mobile clients are also offered for Android devices. Windows mobile client is currently being implemented. iOS service is on the roadmap. These tools enable intuitive, easy, efficient and secure access to the data. For instance virtual filesystem interface enables storing and accessing the data in Archival Services, in the way that mimics the local drive behaviour. Picture below shows the clients available for the National Data Storage.



Figure 15: NDS2 Secure and safe storage and access clients

12.2.7 Virtual drive interface

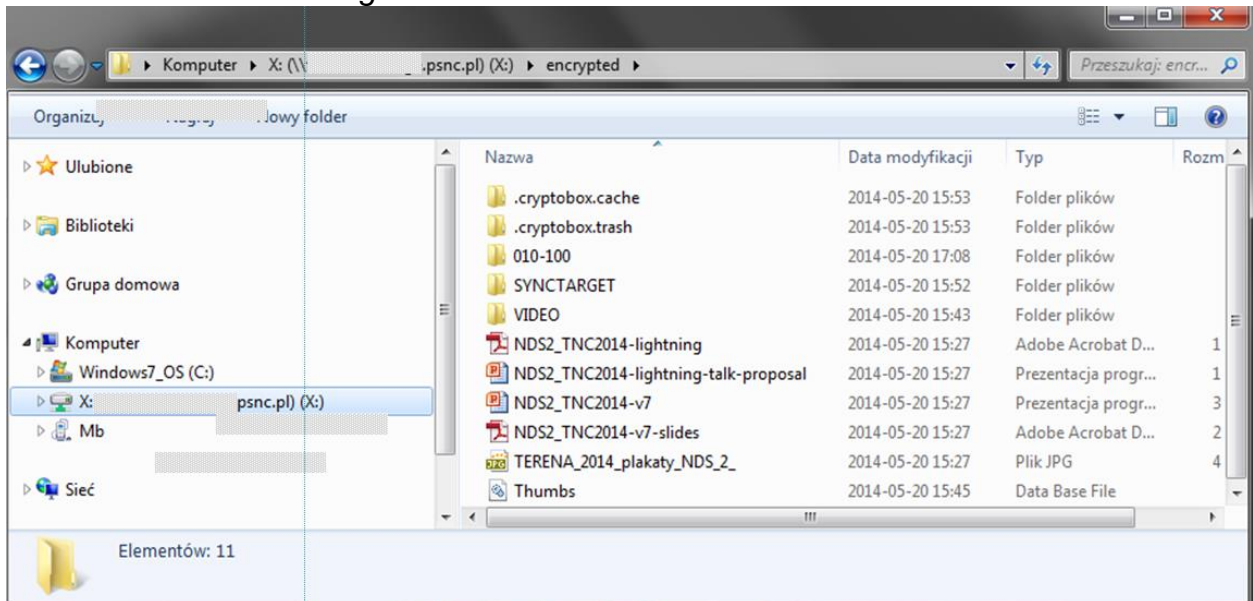
Virtual drive and file system interfaces provide a local filesystem-like experience to users. Picture 4 shows how the Windows client appears to users after configuring the connection and mounting the storage space in the Archival Services to the local workstation.

Users may interact with the remote storage space in the e-Infrastructure in especially easy and intuitive way. This interface hides from them the complexity of the storage infrastructure and the mechanisms implemented in it such as automated data replication or integrity checks.

Therefore this tool is especially useful while offering the users the interactive access to data. As such it was tested within this PoC as the method for interacting with digitized data sets for data management and curation purposes including acquisition data from scanners, harvesting pictures image editors etc.

¹⁷ <http://nds.psnc.pl>

Figure 16: NDS2 Virtual Drive for Windows



12.2.8 Java GUI client

Portable GUI client of SFTP protocol enables interacting with remote storage system in a browser like mode, similarly to Windows Explorer or WinSCP tools. In addition it also supports management of the massive data upload and download jobs. Picture 5 shows the main application window. Picture 6 present the multiple file upload/download progress bar.

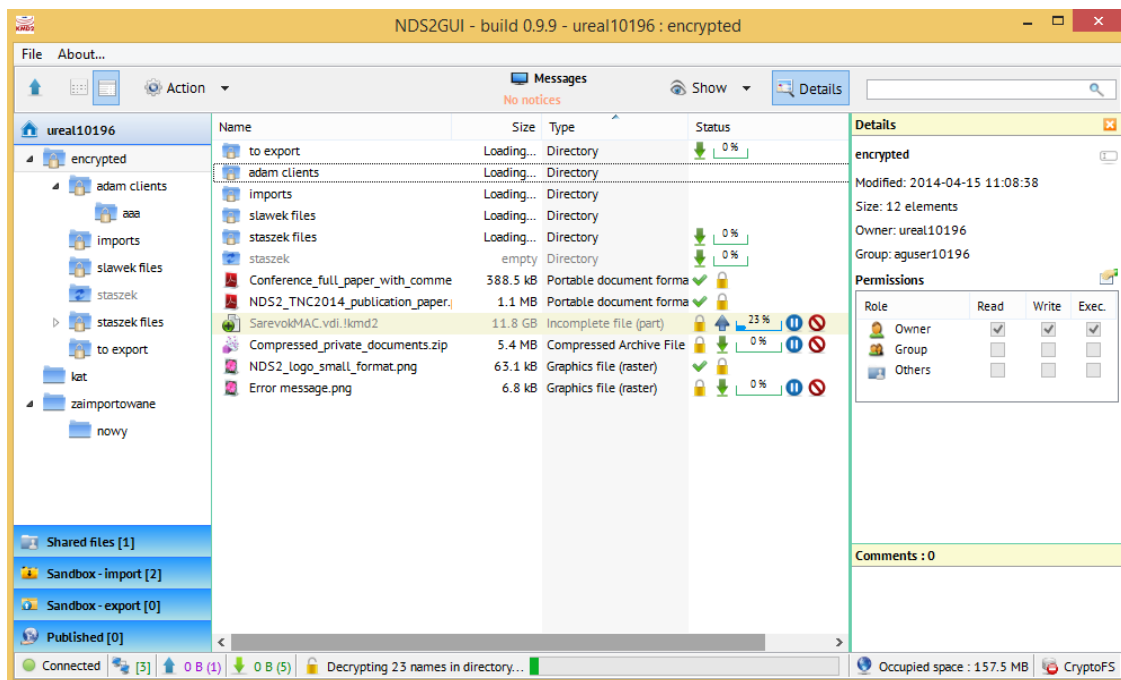


Figure 17: NDS2 GUI Application main window

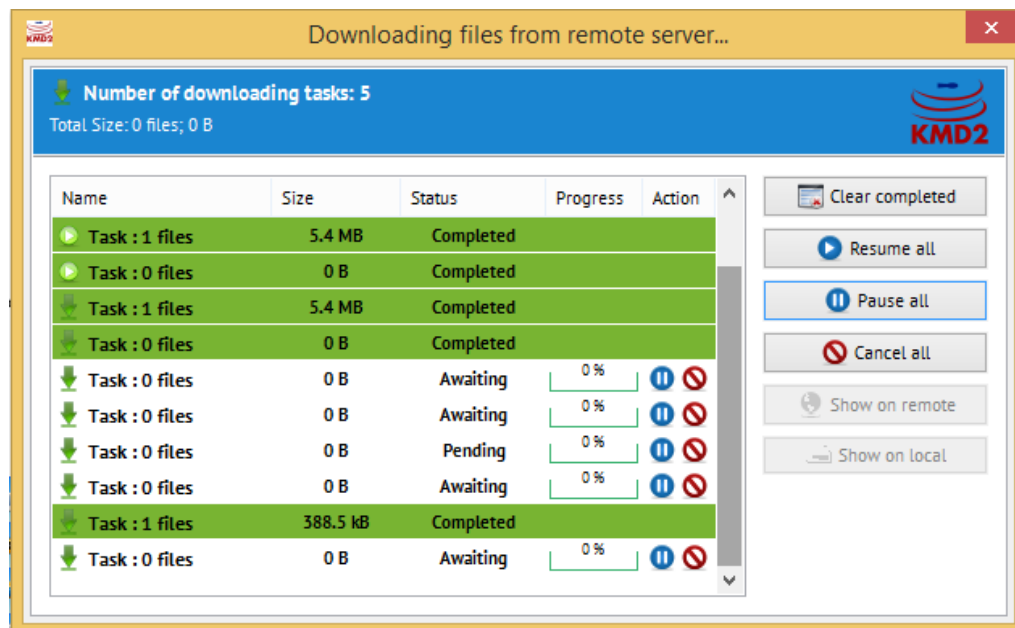


Figure 18: NDS2 GUI Application download task progress monitoring window

12.3 APPLICATION TO ARCHIVAL SERVICES AND NATIONAL DATA STORAGE TOOLS TO DP PROCESSES

Within the PoC, services and tools configuration was worked out in order to address the SDL use-case needs related to DP. Several kinds of tests were conducted to assess the usability of the solutions provided by the national e-Infrastructure for implementing DP workflows.

12.3.1 Achival Services

For the purposes of the PoC, a profile was created for SDL in the Archival Services. It was shared by institutions participating in PoCs in order to collaboratively store, access and share the data.

Archival Services provided the virtually unlimited storage space which addresses the limitations of the storage space available at SDL. In addition replication of data and meta-data implemented transparently to end-users by Archival Services internal mechanisms ensured the data safety and long-term availability even despite possible local disasters. Finally, data integrity control guaranteed that possible media failures are automatically detected and data replicas are recovered based on the additional replicas maintained by the system.

12.3.2 NDS2 clients

Participants of the PoC were given a NDS2 project client-side tools for Windows and Linux.

Virtual drives enabled intuitive access to shared, secure and safe storage space, despite putting the data transparently into the remote SFTP servers of the Archival Services. Applying NDS2 virtual drives enabled using even legacy applications for data management and processing on top of the remote storage space. For instance, it was

possible to save the data to the virtual filesystem directly from the image editing application. NDS2 clients also supported browsing the share directory structure using typical tools such as Windows Explorer, Total Commander or Nautilus. In turn GUI enabled managing massive data uploads and downloads of the digitalized cultural assets.

12.4 POC ORGANISATION

PoC was organised in two phases: evaluating the e-Infrastructure usability from the SDL perspective and testing the data storage and access solutions by the end-users in contributing institutions. For the purposes of PoC, testing subset of digital assets was selected. Also the group of users in the SDL and contributing institutions was chosen.

12.4.1 SDL tests

SDL test included two kinds of aspects. First, it was tested if the data structure including directories and files hierarchy can be implemented in Archival Services. For this purpose, the existing archive directory structure was snapshotted (using tar tool) and “unpacked” on the Archival Services side. During the tests no issues with replicating the data structure was encountered.

Second, selected testing data sets were transferred from the SDL premises to the Access Nodes of the Archival Services in PSNC, Poznań. For the transmission, NDS2 GUI client was used, including its massive data transfer monitoring and control functionality.

Third, users in SDL partnering institutions were asked to access the data sets transferred to the shared data store configured in Archival Services.

12.4.2 Feedback and discussion

During the PoC, DCH institution feedback was collected. It is discussed in this section.

SDL feedback

As discussed earlier, SDL is on one hand the provider of the resources and services to the DCH community in two voivodeships, and is the e-Infrastructure services user on the other.

Application of PLATON Archival Services to host the collaborative data storage space performed within this PoC, however limited to several datasets, enabled drawing conclusions on overall usability of this e-Infrastructure from DP processes perspective.

SDL workers appreciated the improved reliability, storage capacity and efficiency of the nation-wide storage infrastructure, compared to their local infrastructure. Compatibility of the outsourced storage service with the data organisation convention was also confirmed.

Possible usage models, considered by the institution include the full outsourcing model as well as hybrid model. In the former, all the source and high quality data will be stored in the external infrastructure. The local system will be used mainly for acquiring, processing and presenting the data through CMS platforms such as dLibra. In the latter, selected subset of the digital assets will be stored locally in SDL. Extra replicas will be kept remotely in Archival Services in order to ensure their long-term safety and resistance to local disasters. In addition, virtually unlimited storage space of Archival Services will be used in order to increase the archival storage capacity available to SDL and its users.

End-users feedback

End-users experience during the tests was positive in general. Ease of use of the virtual filesystem interfaces was appreciated, especially by voluntary participants of Social Digitization Workshop. Also the performance of the data storage and access tools considered as acceptable. This applies to both highly interactive tasks such as browsing the filesystem structure and massive amounts of the data where average upload and download speed counts. As uploading large data sets effectively and reliability in the environment where the network quality is poor (this happens especially in case of small partnering institutions) poses a challenge, NDS2 GUI's support for monitoring and managing upload and download jobs was considered useful.

12.5 CONCLUSIONS

Experiment involving Silesian Digital Library and Archival Services of the PLATON project in Poland prove that proper application of the e-Infrastructure services to implementing digital preservation processes may be effective and have limited negative impact on the user experience. Usage of data replication functionality of the Archival Services enabled improving the data durability, safety and availability. Virtually unlimited storage capacity of PLATON infrastructure enabled extending the storage space available to SDL and its partnering institutions. NDS2 tools enabled users to keep their methods and habits related to storing and accessing data, while performing data acquisition, processing and preparation for archival. Performance offered by the remote storage system was acceptable for the use case.

Important observations related to the role of the Silesian Digital Library in the e-Infrastructure services take up. SDL is the is the example of the very open, collaborating however demanding community side-partner of the e-Infrastructure providers. Thanks to its technical competence as well as awareness of the opportunities brought by the e-Infrastructures it is an early adopter of the services provided by e-Infrastructures. It also helps defining high-level and real-life requirements of the solutions to be provided in order to address DP processes.

SDL is also effective promoter of e-Infrastructure services usage in the DCH institutions environment. SDL also transferred the related know-how to the partnering institutions.

We argue, that partnership among e-Infrastructure and community leaders and visionaries greatly stimulates and contributes to the take up of cloud services in the DCH domain.

e-Infrastructures must also offer the adopters and users of their services a flexibility. PoC conducted in collaboration with SDL shows, that putting efforts into customizing the service for the needs of large, partnering institutions, pays off, as it help in the services take up and integration with the DP processes existing in the DCH institutions.