DELIVERABLE

Project Acronym: DCH-RP

Grant Agreement number: 312274

Project Title: Digital Cultural Heritage Roadmap for Preservation – Open Science Infrastructure for DCH in 2020

D5.3 – Report on first Proof of Concept

Revision: Final, v1.0

Authors:

Michel Drescher (EGI.eu)
Eva Toller (RA)
Rosette Vandenbroucke (Belnet)

Authors:

Claudio Prandoni (PROMOTER)

Reviewers:

Raivo Ruusalepp (EVKM)
Borje Justrell (RA)

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

Revision History

Revision	Date	Author	Organisation	Description
Draft	9-9-13	M Drescher	EGI.eu	Initial skeleton
V2	21-9-13	M Drescher	EGI.eu	First sections completed
V3	28-9-13	M Drescher	EGI.eu	Complete draft. Mix-in PoC result reports not included (planned for the final version).
V4	30-9-13	C Prandoni	Promoter	Updated Annex 1
V5	1-10-13	C Prandoni	Promoter	Integrated Rosette's comments
V6	1-10-13	C Prandoni	Promoter	Checked Scenarios numbering and added description of Scenario 1.1
V7	7-10-2013	M Drescher	EGI.eu	Addressed comments from Rosette (author, internal review) Börje Justrell (reviewer) Raivo Ruusalepp (reviewer)
V1.0	8-10-2013	C Prandoni	Promoter	Formal 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This deliverable contains original unpublished work except where clearly indicated otherwise.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

Partners in Work Package 5 have conducted a number of Proofs of Concepts in the first year of the DCH-RP project. The overall objective of these PoCs was to validate in concrete experiments assumptions and concepts expressed in the DCH roadmap to preservation.

Beginning with an initial set of scenarios established in D3.1, WP5 has applied the Scrum project management methodology to the DCH-RP project. These scenarios were swiftly extended into a total of 14 scenarios covering three fundamental concerns of digital preservation, i.e. (1) Operational challenges, (2) End user concerns and (3) New services and infrastructure integration.

Supported by the continuous agile project management partners in WP5 have conducted six Proofs of Concept covering seven of the fourteen scenarios. The outcomes of these PoCs cover a wide spectrum of negative and positive results of tested tools. Irrespective of the individual result, all evidence and experience collected in the PoCs will be taken in to consideration by the members of WP3 to populate the intermediate DCH roadmap in D3.4.

Conducting these PoCs over the last months, and the results that are collected in the individual reports have shown that the DCH community is still very fragmented into national and local solutions and processes. Despite the overall positive assessment of the first year of WP5, the most profound gap that has not been closed by the DCH community is the lack of a vision on a common and international e-Infrastructure suitable to serve the ICT needs of this community in an efficient and accurate manner.

Such a vision is key to the success of the future Proofs of Concept that WP5 will set out to execute during the second year of the project. However, it is relatively easy to declare success to this from the standpoint of having successfully conducted a number of Proofs of Concepts. Such a technical success however is not the real benefit and reason of being of WP5. It is the *applicability* and *meaningfulness* of the results to the DCH roadmap. Without a clear vision expressed in the roadmap, WP5 is in danger of “scope creep” and deviating from the common goal of the project.

2 INTRODUCTION

As stated in the Description of Work for DCH-RP, preservation is one of the most challenging problems of the current digital era, applying to all sectors of society, including the DCH sector. Preservation is a broad concept, but DCH-RP defines preservation as the combination of preserving:

- Data (digitized and born-digital content like databases, catalogues, files, etc.) and
- Information associated with that content (so-called 'infostructure', referred to also as metadata).

DCH-RP deals both with 'long-term preservation' (preserving for an unpredictable long period of access and use) and 'short-term preservation' (preserving for a relatively short period of access and use). The main objective to be achieved by the project is to design a sound roadmap for the implementation of an e-Infrastructure for preservation of DCH content, as part of a more general vision towards an Open Science Infrastructure for DCH in 2020.

Work Package 5, which is responsible for producing and delivering this document, has been charged with coordinating and conducting Proofs of Concept (PoCs) that will inform Work Package 3 with the outcomes of these PoCs so that informed decisions can be made towards further evolution of the DCH roadmap to preservation within Work Package 3.

Deliverable D5.1 [REF], made available in January 2013 to the general public, informs about the planning around the Proofs of Concept; as such it is a prospective document in nature and provides information on the work package management methodology (using SCRUM), the intentions and objectives of the participating national partners, and the general planning and coordination with other Work Packages in the DCH-RP project.

This document, Deliverable D5.3, is retrospective in nature by reporting on the conducted Proofs of Concepts in the past seven months (from 18 February 2013 to 30 September 2013) and prospective in indicating next steps and lessons learnt that would influence the second PoCs phase that will begin immediately after publication of this document.

2.1 OBJECTIVES OF THE DELIVERABLE

This deliverable formally informs the reader about the outcomes of the conducted Proofs of Concept. The primary recipients of this deliverable are the members of Work Package 3, which is charged with maintaining and updating the DCH roadmap based on information contained in this document.

Throughout the operative timespan in the first PoC phase, the work package members have agreed to produce a reporting template, which will be used to inform the roadmap maintainers in Work Package 3. For each "scenario" (see below) that was examined by participating national institutes, a scenario report based on this template was produced and made permanently available in the EGI Document DB. Since the leaders of Work Package 3 and Work Package 4 are closely involved in the activities in Work Package 5 as the "Product Owners" of the PoC activities, this approach was implicitly approved as the means of information transfer to Work Package 3.

Therefore, these scenario reports comprise the core of WP5's feedback to WP3. To avoid duplication of work, this deliverable incorporates these reports by references provided later in this document, thus fulfilling the contractual obligation formulated in the DCH-RP DoW.

2.2 STRUCTURE OF THE DOCUMENT

Reflecting the decisions made during the first phase of Proofs of Concept, this document is structured as follows.

By adopting the Scrum methodology of agile project management for WP5, the team regularly reviewed the last sprint in a “retrospection” session in the sprint planning meetings. Consequently, section 3 provides a retrospective analysis of the experience of applying Scrum to a physically distributed team of project members.

Section 4 summarises how the WP5 participants have started to design the first Proofs of Concept phase; beginning with a cross Work Package discussion with WP3 and WP4, a first list of scenarios that were provided by the DCH roadmap maintainers and the coordination of the individual scenario tests across member institutes in WP5.

Section 4 summarises the activities and results of the individual Proofs of Concept with the common theme “Organisational Challenges”. The concrete and final scenario reports are incorporated by reference in this document and are provided in Annex 4.

Similarly, section 6 covers the Proofs of Concept around the common theme “End User Concerns”.

Section 7 concludes this deliverable summarising the lessons learnt across all member institutes, Proofs of Concepts. It will highlight recommendations to Work Package 3, and indicate the next steps during the second, final Proofs of Concept phase until the conclusion of the DCH-RP project.

Annex 1 provides the discussion document that includes the technical interpretation of the scenarios indicated by Work Package 3, which led to the scenarios that were examined in this first phase of Proofs of Concept.

Annex 2 includes the scenario report template that was used by team members to report back to Work Package 3.

Annex 3 preserves a snapshot of the contents of the online maintained Scrum backlog for long-term reference.

3 REVIEWING THE MANAGEMENT METHODOLGY

As extensively described in D5.1[1], WP5 has adopted the Scrum methodology for managing and running the Proofs of Concept. The methodology, used tools and facilities are also described in D5.1 and will not be repeated here.

A key role in managing a Scrum Product Team working on a specified product (here: the Proofs of Concept and the resulting reports) lies in planning and conducting the so called “Sprints” – focused phases of productivity that subsequently finish with a working end result.

The planning of Scrum sprints revolves around one key tool, the Sprint history. Usually recorded in a spreadsheet, the Sprint history captures key indicators about the Product Team’s progress on finishing tasks that are recorded in the backlog. Annex 3 provides a complete listing of the sprint history of the first PoC phase.

Of these indicators, four are used not only for historic recording of data, but also to *project* the future progress of the team based on the historic data. Hence the Sprint history becomes an intuitive and effective tool to **manage expectations** towards the Product Owner, i.e. the one who oversees and steers toe progress of the team. These key indicators are:

1. **Performance**
   The performance is the aggregation of the points of all tasks that were finished by the team, and accepted by the Product Owner. If a team has not finished any task in a sprint (e.g. Sprint 9), or the Product Owner did not accept any of the finished tasks, then the team performance for that sprint is recorded as 0 (zero).
2. **Velocity**
   The team’s velocity is expressed as an averaging function over the team’s productivity. Typically, the velocity is measured as the average productivity over the team’s last three sprints – and so does this team. Ideally, Productivity and Velocity gradually close in on each other with the team’s gain in experience in the Scrum methodology. The *current* velocity is used as the expected performance of the team for all future sprints.
3. **Remaining**
   The remaining points are calculated as the difference between the total points at the end of a sprint, and the sum of all points of all tasks accomplished in the current and previous sprints. The remaining points are then the basis of projecting into the future (using the team’s velocity) to calculate how many more sprints are needed to accomplish all work encoded in the task points.
4. **Total points**
   The total points are simply the aggregation of all points of all tasks in the backlog.

The following figure provides an overview of the historic development of these four key indicators during the first phase of Proofs of Concepts.

[1] https://documents.egi.eu/document/1544

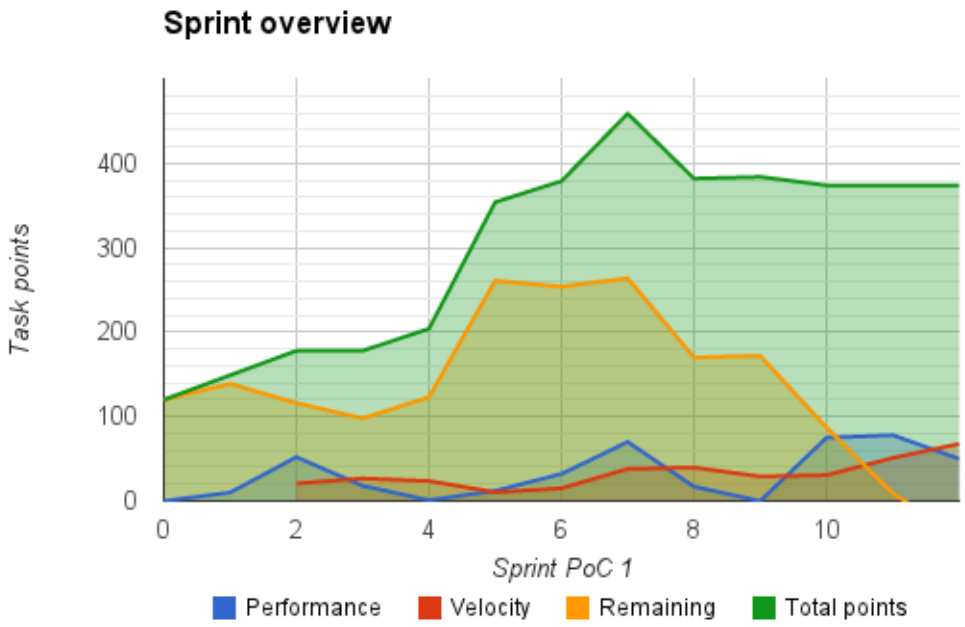


Figure 1: A graphical overview of WP5's sprint history

The graph shows some typical mechanics of how Scrum teams work.

Firstly, the **total points** fluctuate, in this case quite drastically. This is normal for a Scrum Team as WP5 since this methodology was new to all team members, so a learning phase of about three to five Sprints indicates the team getting used to a number of processes within the Scrum methodology. A key process is the allocation of points to tasks, where the task represents a unit of work, and the points reflect the team's expected amount of effort necessary to accomplish this task. With the learning curves increases the tea's accuracy in assigning points – in this case the team had to learn to be more realistic (i.e. reduce expectations and factor in unexpected problems) and allocate more points to the task than initially anticipated.

The steep increase in total points is, however, only partially explained by the initial learning phase. As often seen in Scrum projects, the initial phase is followed by a phase that is best described with "explosion of creativity" in team members, adding many tasks that members feel the need to accomplish to reach the overall goal.

As the team progressed, the projection feature of the Scrum methodology allowed the team to *manage itself* in that the realization began to appear that not everything that the team thought it needed to accomplish in fact should be accomplished during the first phase of the Proofs of Concept. This started to happen in Sprint 7 and onwards, illustrated by the fact that the numbers of total points, remaining points and performance do not add up: The difference is caused by a number of tasks (with story points) that were moved to the second phase of Proofs of concept, as detailed in Annex 3.

The **remaining points** follow the dynamics of total points and performance, except for when the team decided to move tasks to the second PoC phase (see above). This is attributed to the main purpose of the remaining points indicator in Scrum: In the projection for future sprints, the remaining points graph provides the team with feedback as to where they are standing in the middle of the project. It also allows the Product Owner as well as the Team to make an informed decision for project planning. In a situation like DCH-RP, the number of sprints is fixed as determined by the DoW, therefore the project cannot extend the duration to accomplish all planned tasks. Instead, a decision must be taken, a prioritisation

over all remaining tasks in the project. The result is that a number of tasks are cut or in this case, moved to the second PoC phase. This project management situation is perfectly reflected in this graph and explained earlier.

The **velocity** indicator works as expected. Its averaging effect makes it a good progression forecast tool for project management decision. As indicated, the velocity is not available for the first two sprints – this team's velocity is calculated as the average over the last three sprints.

The **performance** graph shows significant peaks over time, with stark troughs in between. Usually, the performance fluctuates in the beginning, and closes in on a steady state over the time of a project. This behavior, however, is typical only for Scrum teams that wok exactly as the Scrum methodology suggests, i.e. a group of participants that are physically located very close to each other – ideally in the same office or even room. *This* product team, however, is quite different from the ideal Scrum setup:

- The team members are distributed over all Europe
- The team members are not working 100% of their time on this project
- The team members act as proxies to local colleagues who often conduct the work.
- Those local colleagues are not included in the Scrum methodology, nor do they work exclusively in this project.

While the troughs can be broadly explained by the vacation periods around Easter and Summer 2013, the main reason of the fluctuations are due to the setup of the team detailed above.

Nonetheless, the team achieved what it had set out to achieve, and the Scrum methodology has proved a useful tool to manage a distributed team of devoted participants. It does not come, therefore, as a surprise that the team decided to continue following the Scrum methodology for the second Proof of Concept phase. However, the means of communications need to be improved. The team expressed its expectations that, with better means of communications during the Scrum sprints, the overall effectiveness of the team could be greatly improved.

4 FROM PROOFS OF CONCEPT TO SCENARIOS

Following the planning of the DoW, WP5 continued to work on how to cover the objectives partners have expressed in D5.1 section 4. In parallel, WP3 worked on a study on the roadmap to preservation for DCH; i.e. D3.1. In that deliverable, section 5 describes a number of gaps between the current state of the art of digital preservation in the CH community, and the current e-Infrastructure that are available in Europe and world-wide. Section 6 in the same deliverable defines a matrix for the roadmap, spanned by four different aspects the roadmap should cover, and a timeline.

The four roadmap aspects are:

1. Harmonisation of data storage and preservation
2. Progress for inter-organisational communication
3. Establishment of conditions for cross-sector integration
4. Governance models for infrastructure integration

The timeline is defined as:

1. Short-term (2014)
2. Medium term (2016)
3. Long-term (2018 and beyond)

Close discussions between WP3, WP4 and WP5 led to including five initial scenarios in the roadmap that, though hypothetical, partners in WP5 could use to kick-start activities. Throughout these discussions, WP5 partners realised that there are indeed more scenarios prevalent in their national CH community; eventually the initial scenarios stemming from D3.1 were extended to cover three different themes with in total 13 scenarios. These thirteen scenarios were further discussed with WP3, through which a technical assessment and requirements document was maintained (see Annex 1). These scenarios were then included in the team backlog (see Annex 2, Table 4) using the Scrum terminus “Epic”, where each individual scenario is represented and identified as an individual Epic. Towards the end, scenario 1.6 was added to the list of scenarios in the Scrum epics, but not further discussed (which is the reason that Annex 1 does not include scenario 1.6 with a technical description).

In short, the scenarios were grouped as follows:

1. **Theme 1 – “Organisational challenges”**
    1. Use specialised DP tools on in-house data
    2. Integrating a new tool into existing infrastructure
    3. Select an existing DP solution at an institute with best effort IT support
    4. Preservation from a consortium of collections on the cloud
    5. Preserving a 3D visualisation
    6. Retrieve archived data
2. **Theme 2 – “End user concerns”**
    1. Researcher discovers a historical database
    2. Research and select a tool serving a specific purpose
    3. Accessing digitised content from schools
    4. Gain access to archived websites
3. **Theme 3 – “New services & infrastructure integration”**
    1. Proof of authenticity in distributed archiving
        a) Extend 3.1 with repository safeguarding policies
    2. Defining new services

3. Integrating new services into existing infrastructure

Quickly WP5 partners realised that, though all scenarios were and are valid by themselves, a decision had to be taken which of these scenarios were the most important to address in the remaining time. Part of these discussions was a technical analysis mainly driven by B. Justrell, RA; the objective was to flesh out the at that moment fairly hypothetical scenarios with more specific requirements and technical details. In that capacity Börje Justrell being the driving force behind the roadmap, acted as the Product Owner towards the activities in WP5 and gave the proper guidance that was asked for. The outcome is a brief document that further served as technical guidance in WP5 is provided in full in Annex 1.

Partners in WP5 then decided that the present landscape in the national DH communities is still too diverse to attempt a joint e-Infrastructure-based first Proof of Concept as they felt that such a PoC would not be achievable in the given timeframe. Instead, an alternative approach was chosen to ease DCH's way into e-Infrastructure usage by giving each partner the freedom to choose and pick which scenarios they found achievable using available local and national resources, as well as with whom they would like to partner. The result was the following mapping of which partner would pick up which scenario for individual Proofs of Concept.

Table 1: Partners in WP5 conducting Proofs of Concept along the Scenario descriptions

Scenario		**Partner**					
Theme	Scenario	Belgium	Estonia	Hungary	Italy	Poland	Sweden
Organisational challenges	1.1	X			X		
	1.2	X			X		
	1.3						X
	1.4	X			X		
	1.5						
	1.6		X				
End user concerns	2.1						
	2.2						X
	2.3						
	2.4						X

The only cross-partner collaboration happened between Belgium and Italy covering Scenario 1.1, scenario 1.2 and scenario 1.4. While Belgium was exclusively concerned with organisational challenges indicating a concern on the sustainability and adequateness of local or national solutions for potentially being used on a European or even global level, the Swedish partners put slightly more focus on the end user's experience on practical preservation activities. The Polish partners already maintain collaboration with the national e-Infrastructure and decided to postpone their participation in the DCH-RP Proofs of Concept to the point where potential synergies with the EUDAT project would yield tangible results to be tested in a DCH-RP context (see section 7). The Estonian partner started working on the new Scenario 1.6. Up until the document production, only the Hungarian partners were not able to actively participate in

the Proof of Concepts even if they started to get in contact with the National Széchényi Library to be ready to participate to the second round of the PoC.

As shown in Table 1 scenarios 1.5, 2.1 and 2.3 as well as all scenarios of theme 3 were not covered by the conducted Proofs of Concept.

5 SCENARIO THEME: "ORGANISATIONAL CHALLENGES"

This section provides a brief summary of the Proofs of Concept conducted for the Scenarios that are part of the first theme, "Organisational challenges". The structure of this section does not follow the individual scenarios but instead illustrates the outcomes of the Proofs of Concept as they were conducted. For example, the Belgian partner covered scenarios 1.2 and 1.4 in one Proof of Concept.

Each subsection briefly describes the main points and references to further documentation and recapitulates the recommendations and grading of each tool, service etc. that have been tested.

5.1 SCENARIO 1.1 – USING SPECIALISED RESEARCH TOOLS

This scenario is documented in the Wiki[2] outlining the synopsis of the PoC. References are given to a more detailed description of the testing procedure and test data.

KIK-IRPA, one of the Belgian DCH organisations, has already a local preservation system for their data. They have described their preservation system in a "Best practices" document that is accepted throughout the organisation. However one of their main concerns is to maintain the integrity of their data.

There exist auditing and certification schemes for trustworthy repositories[3]. Such an audit also equals a risk analysis of the chosen archiving method. How easy this all may sound, real life shows that almost no one ever terminates the whole procedure, hence there is no common "best practices" available.

Doing an audit in a consequent way requests to use the necessary tools.

In this scenario we want to use existing tools and document the auditing process. We will do this on the local data that is in the KIK-IRPA preservation scheme and on data of the Italian partners.

Such an audit is in fact independent of where the data is stored but it is certainly a "tool" that will be very valuable for preservation done on data stored with e-infrastructures or other storage service providers.

Once done it would be useful to execute the procedure on preservation done on grid and cloud.

The original work defined for this PoC included the study of the Drambora auditing and risk assessment scheme, the partially existing implementations and arrive at a user-friendly implementation that would be used for test the integrity of local data at KIK-IRPA in Belgium and ICCU in Italy.

However this work was too huge to be included in the timeframe of the first round of PoCs and the choice went to use the tool Scoremodel developed by DEN[4]. DEN Foundation is the national centre for ICT in cultural heritage.

Recommendations:

The "Scoremodel" is a useful tool to test the integrity of a collection. It does not implement the full Drambora scheme but has the advantage that it is easily understood and usable by DCH people. It can be used in the roadmap as an example for testing the integrity of data.

Grading:

[2] https://wiki.egi.eu/wiki/DCH-RP:PoC_1_Belgium#PoC_1_Audit_and_certification_on_local_data

[3] see: "Trustworthy Repositories – Audit and Certification" http://www.digitalrepositoryauditandcertification.org/pub/Main/ReferenceInputDocuments/trac.pdf and "Risk-analysis for E-depots:DRAMBORA" http://www.repositoryaudit.eu/

[4] http://www.den.nl/standaard/383/

Aspect	Score
Usefulness of the tool	**4**
User friendliness of the tool	**4**

5.2 SCENARIO 1.2 – INTEGRATING A NEW TOOL INTO AN EXISTING INSTITUTIONAL INFRASTRUCTURE; SCENARIO 1.4 – PRESERVATION FROM A CONSORTIUM OF COLLECTIONS ON THE CLOUD

This scenario is documented in the Wiki[5] outlining the synopsis of the PoC. References are given to a more detailed description of the testing procedure and test data.

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

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

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

Recommendations:

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

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

Grading:

Aspect	Score
Using grid storage via e-CSG	**1**
Copy data to grid storage via e-CSG	**0**

[5] https://wiki.egi.eu/wiki/DCH-RP:PoC_1_Belgium#PoC_2_Test_out_download_and_access_of_DCH_data_on_grid_storage

Automatically copy metadata to the e-CSG metadata format	**0**

5.3 SCENARIO 1.3 – SELECTING A DIGITAL PRESERVATION SOLUTION IN THE CASE OF AN INSTITUTION WITH ONLY VOLUNTARY IT SUPPORT

The scenario is documented in the DCH-RP Wiki[6] including a reference and description of the test data, the testing procedures and which tools were used. The third reference, named "Outcome of the tool tests" provides a downloadable version of the report, which is also provided in Annex 4.

The report lists four tools test for typical data preservation activities in a local institute, assessing these for potential pan-European usage.

5.3.1 ROND (Riksarkivet OpeN Data)

This tool is typically used to anonymise data sets before publication or further public usage; for example to anonymise the names of the censors of a film that was indexed and thus banned from public broadcasting. The recommendation and grading is as follows:

Recommendation:

> "If ROND should be recommended as a tool for de-identification of archive information in the DCH sector, the improvements mentioned above should be made first. It should also be pointed out that ROND has a major limitation in that it requires a certain metadata model (ADDML), which is currently only used by Sweden and Norway. However, tools of this *type* could be very useful for publishing huge amounts of archival information as open data; information that otherwise would be locked up in the archives, and much harder to find and obtain for interested parties."

Grades:

On a scale from 1 (very bad) to 5 (very good).

Simplicity of installation:	5
Simplicity of management:	X
Ease of use:	2 - 3
Generality of solution:	1
Quality of result:	5

5.3.2 Archivist's Toolkit (AT)

The Archivist's Toolkit[7] is an Open Source archive management system. It is advertised "[…] to support archival processing and production of access instruments, promote data standardization, increase processing efficiency, and lower training costs."

Recommendation:

> A specific recommendation was not given for AT mostly because the installation was very complicated and beyond the expertise of a typical museum staff.

[6] https://wiki.egi.eu/wiki/DCH-RP:PoC_1_Sweden#Proof_of_Concept_scenarios
[7] http://www.archiviststoolkit.org

Grades:

On a scale from 1 (very bad) to 5 (very good).

Simplicity of installation:	1 – 2
Simplicity of management:	not tested
Ease of use:	not tested
Generality of solution:	not tested
Quality of result:	not tested

5.3.3 XENA

XENA[8] is a file conversion tool that automatically detects the file format of a given file (e.g. an image in GIF, TIFF, PNG, etc. format, or a document in PDF, PDF/A, Word .DOC, Word .DOCX, Open Document format etc.) and converting it into digital objects suitable for digital preservation.

Recommendation:

Xena is very easy to use for batch conversion; when you have supplied default input and output directories, it takes very little effort to normalise/convert a lot of files (or at least, to try to do so). However, it needs a bit of trial-and-error to understand what actually normalisation, binary normalisation, and conversion is about, and what the results will look like.

If you like XML and have formats that are suitable for conversion (for example, TIFF, jpeg, csv, maybe also Open Document Format(s) and MS Office documents), this tool may be useful. For binary conversion, everything(?) seems to work (but you have to figure out how the result can be used later, and how to verify the results). As earlier stated, the results cannot be viewed in a common web browser. Probably, this is because the schema link that can be seen in the “raw xml output”, http://preservation.naa.gov.au/xena/1.0, is out of date.

Xena recognises many more formats than those that have been tested here (see page 15 in http://xena.sourceforge.net/media/How_Xena_ids_file_formats.pdf), but you cannot always trust this. For example, Excel is mentioned as one of the supported formats, and this format was recognised by Xena but impossible to convert. Open Document Format is also supported, but doesn't seem to recognise Swedish characters.

Grades:

On a scale from 1 (very bad) to 5 (very good).

Simplicity of installation:	2 – 3
Simplicity of management:	n/a
Ease of use:	4
Generality of solution:	4
Quality of result:	2

5.3.4 DSPACE

DSPACE[9] describes itself as:

[8] http://xena.sourceforge.net

"DSpace is the software of choice for academic, non-profit, and commercial organizations building open digital repositories. It is free and easy to install "out of the box" and completely customizable to fit the needs of any organization. DSpace preserves and enables easy and open access to all types of digital content including text, images, moving images, mpegs and data sets. And with an ever-growing community of developers, committed to continuously expanding and improving the software, each DSpace installation benefits from the next. […] DSpace is the software of choice for academic, non-profit, and commercial organizations building open digital repositories. It is free and easy to install "out of the box" and completely customizable to fit the needs of any organization. DSpace preserves and enables easy and open access to all types of digital content including text, images, moving images, mpegs and data sets. And with an ever-growing community of developers, committed to continuously expanding and improving the software, each DSpace installation benefits from the next."

Recommendation:

It seems clear from the number of third-party tools, and from the installation instructions themselves, that this is probably a too complicated tool for the actors in Scenario 1.3. However, there is now also a hosted service, DspaceDirect[10], that may be investigated as an alternative.

Grades:

On a scale from 1 (very bad) to 5 (very good).

Simplicity of installation:	1 – 2
Simplicity of management:	not tested
Ease of use:	not tested
Generality of solution:	not tested
Quality of result:	not tested

5.4 SCENARIO 1.6 – ARCHIVED DATA RETRIEVING

The Estonian partners experienced a major restructuring during the first phase of Proofs of Concepts. This caused no activities from the Estonian partners until close to the end of the first phase reported in this document. However, as an initial activity and easing their way into more regular activities in the second phase of PoCs the Estonian partners covered scenario 1.6 in a focused Proof of Concept around existing tools in Estonia.

5.4.1 IBM Tivoli Server Manager/Client Server Version 5, Release 5, Level 2.0

Estonian memory institution (Conservation Centre Kanut) which digitises primarily museum content wants to make backup copy of files to another memory institution's tape library but needs proof that the transferred content is complete and it is possible to transfer a copy of files back to the originating repository when needed. Data retrieval tests will be carried out periodically (quarterly).

IBM Tivoli Storage Manager is a client-server licensed product that provides storage management services in a multiplatform computer environment. The backup-archive client program permits users to back up and archive files from their workstations or file servers to storage, and restore and retrieve backup versions and archive copies of files to their local workstations.

[9] http://www.dspace.org
[10] http://dspacedirect.org/

Conclusions and recommendations:

The tool test is one sample out of hundreds of other possibilities. During the test it was possible to point out the following conclusions:

1. To set up the tool and to use it needs advanced IT expert knowledge.
2. Tool is usable only on the command line and needs (previous) experience. There is no real time support available for this tool for the client. The tool's manual is available and updated file://localhost/ http/::publib.boulder.ibm.com:tividd:td:IBMTivoliDecisionSupportforOS3901.7.html
3. Client side can be modified or upgraded with a great effort and needs previous experience.
4. To provide this service the archive host should be accessible and very responsible, any minor changes or modifications (especially network and server side) can cause problems for end-user client.
5. To log into the TSM server the connection speed should be at least 60 Mb/s, preferably more. Slower connection causes lag and creates confusion when retrieving/restoring copy from archive.
6. The list of archived data is accessible but not as usable as modern tools with a GUI.
7. The tool does not provide immediate response/feedback when errors occurred.
8. Using the ext4 file system neither server nor client tools support file hash function. The success of the archiving or restoring must be manually controlled using tool's log files.
9. There are no limits of the file size or format, although the tool does not have the format recognition features - there is a need for separate tools for that.

Grading:

Simplicity of installation	2
Ease of use	2
Generality of solution	3
Quality of result	5

6 SCENARIO THEME: "END USER CONCERNS"

This section provides a brief summary of the Proofs of Concept conducted for the scenarios that are part of the second theme, "End user concerns". The structure of the section follows that of section 5.

6.1 SCENARIO 2.2 – RESEARCH AND SELECT A TOOL SERVING A SPECIFIC PURPOSE

The scenario is documented in the DCH-RP Wiki[11] including a reference and description of the test data, the testing procedures and which tools were used. The third reference, named "Outcome of the tool tests" provides a downloadable version of the report, which is also provided in Annex 4.

The report lists four tools tested for a specific purpose that is not necessarily related to digital preservation of data.

6.1.1 AVS document converter 2.2

AVS Document Converter 2.2[12] converts files of source formats (PDF, HTML, HTM, MHT, RTF, DOC, DOCX, ODT, PPT, PPTX, TXT, TIFF, TIF, EPUB, MOBI, FB2, DjVu, XPS) into files of target formats (PDF, HTML, MHT, RTF, DOC, DOCX, ODT, TXT, GIF, JPEG, PNG, TIFF, EPUB, MOBI, FB2). It can be downloaded from http://www.avs4you.com/AVS-Document-Converter.aspx.

Note that there is also an AVS Image Converter (http://www.avs4you.com/AVS-Image-Converter.aspx). The reason that the document converter is tested instead is that it supports DjVu as a source format (which the Image Converter does not). However, it may later be necessary to test the AVS Image Converter too, to cover all source and target formats.

Recommendation:

This tool should only be used when you want to convert small amounts of files. It is not reliable for batch conversion. Furthermore, check the result if you convert to files to PDF (the conversion sometimes results only in a white or grey page instead of an image). You may also want to check the quality of the images if you convert to JPEG or PNG (the quality varied between barely acceptable and good).

Note that the PDF files that are generated with this tool are not necessarily in the specific PDF/A-1 format. Even so, they may conform to the rules of PDF/A-1, but this must be tested (see http://www.pdfa.org/2011/08/validating-pdfa). It is possible to convert PDF to PDF/A-1, but not necessarily easy.

Grades:

On a scale from 1 (very bad) to 5 (very good).

Simplicity of installation:	4
Simplicity of management:	n/a
Ease of use:	3
Generality of solution:	4 – 5
Quality of result:	1 – 2

[11] https://wiki.egi.eu/wiki/DCH-RP:PoC_1_Sweden#Proof_of_Concept_scenarios
[12] http://www.avsmedia.com/AVS-Document-Converter.aspx

6.1.2 AVS image converter 2.2

AVS Image Converter 3.0 converts files of source formats (BMP, GIF, JPEG, JPG, JPE, JFIF, PNG, APNG, TIFF, TIF, PCX, TGA, RAS, PSD, CR2, CRW, RAF, DNG, MEF, NEF, ORF, ARW, EMF, WMF, JPEG 2000, SWF) into files of target formats (BMP, GIF, JPEG, JPG, JPE, JFIF, PNG, APNG, TIFF, TIF, PDF, TGA, RAS).

As can be seen in the lists of source formats and targets formats: with respect to the chosen test data formats and Riksarkivet´s allowed formats (see and Section 1.5 in DCH-RP_WP5_Scen-2-2_ID-66-restricted.pdf and Section 3 in DCH-RP_WP5_Scen-2-2_ID-70.pdf, respectively), the conversions that are relevant are the following:

- TIFF to JPEG, PNG, and PDF
- JPEG to TIFF, PNG, and PDF

Recommendation:

As long as you stick to conversion between JPEG and PNG, this tool seems adequate, maybe even for batch conversion (although to make sure of this, it would have to be tested on much larger amounts of files). It is not appropriate at all for conversions to PDF if you want individual documents as target while providing several files as source. Furthermore, the largest TIFF files (407 MB and 154 MB) were impossible to load and/or convert at all.

Note that the PDF files that are generated with this tool are not necessarily in the specific PDF/A-1 format. Even so, they may conform to the rules of PDF/A-1, but this must be tested (see http://www.pdfa.org/2011/08/validating-pdfa/). It is possible to convert PDF to PDF/A-1, but not necessarily easy.

Grades:

On a scale from 1 (very bad) to 5 (very good).

Simplicity of installation:	5
Simplicity of management:	n/a
Ease of use:	3
Generality of solution:	4 – 5
Quality of result:	3

6.1.3 Universal Document Converter

The Universal Document Converter[13] (henceforth called UDC) is a printing service. After installation you can choose UDC as the current printer when you want to convert a file, and then choose the output format. There are eight output formats to choose from, among them JPEG, PNG, TIFF, and PDF. The conversions that will be tested are the following:

- TIFF to JPEG, PNG, and PDF
- JPEG to TIFF, PNG, and PDF
- DjVu to TIFF, JPEG, PNG, and PDF

[13] http://www.print-driver.com/download/

There is no possibility to convert files in batch; you have to do it one-by-one. Thus, even if this service would turn out to give very good results, it can only be a complement to other methods that can handle batch conversion.

Recommendation:

This is not a tool/service that can be used for large amounts of images or documents, since you have to convert each image individually – and especially as you have to supply the name of the output file manually. However, if most image conversion software has the same problem with large TIFF files as the AVS programs had, this can be a complement for converting the (hopefully few) large TIFF files that cannot be handled by other conversion software.

Note that the PDF files that are generated with this tool are not necessarily in the specific PDF/A-1 format. Even so, they may conform to the rules of PDF/A-1, but this must be tested (see http://www.pdfa.org/2011/08/validating-pdfa/). It is possible to convert PDF to PDF/A-1, but not necessarily easy.

Grades:

On a scale from 1 (very bad) to 5 (very good).

Simplicity of installation:	4 – 5
Simplicity of management:	n/a
Ease of use:	3 – 4
Generality of solution:	5
Quality of result:	4

6.1.4 A-PDF DJvu to PDF

A-PDF DjVu to PDF[14] is a fast, affordable utility to allow you to batch convert DjVu (.djvu, déjà vu) into professional-quality documents in the PDF file format.

Recommendation:

This *[I.e. the tool not running on Windows 7 64 bit – the deliverable authors]* was a bit surprising since it is written in the home page that the application can run on Windows 7 (besides XP and Vista). It may be possible that the commercial version can handle 64 bit, but nothing about this is said in the homepage (unless it is deeply buried in a manual). Since there will be fewer and fewer computers with 32-bit Windows, and more and more computers with 64-bit Windows, this does not seem to be a good tool for the future, unless they make a 64-bit version, too.

Grades:

On a scale from 1 (very bad) to 5 (very good).

Simplicity of installation:	0
Simplicity of management:	not tested
Ease of use:	not tested
Generality of solution:	not tested
Quality of result:	not tested

[14] http://www.a-pdf.com/djvu-to-pdf

6.2 SCENARIO 2.4 – GAIN ACCESS TO ARCHIVED WEBSITES

The scenario is documented in the DCH-RP Wiki[15] including a reference and description of the test data, the testing procedures and which tools were used. The third reference, named "Outcome of the tool tests" provides a downloadable version of the report, which is also provided in Annex 4.

The report lists five tools tested to gain access to and manage the content of archived websites.

6.2.1 HTTrack

"HTTrack[16] is a free (GPL, libre/free software) and easy-to-use offline browser utility. It allows you to download a World Wide Web site from the Internet to a local directory, building recursively all directories, getting HTML, images, and other files from the server to your computer."

Recommendation:

Since it was both easy to install and use this tool, and the quality of the result also was good, this should be a suitable tool for the downloading of web sites when the most important aim to give easy access to end users.

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

Grades:

On a scale from 1 (very bad) to 5 (very good).

Simplicity of installation:	5
Simplicity of management:	n/a
Ease of use:	4 – 5
Generality of solution:	5
Quality of result:	5

6.2.2 Snappy Web Archiving Tool (SWAT)

"SWAT[17] (Snappy Web Archiving Tool) is a tool designed for archiving web sites and displaying the archive in a simple way. Besides harvesting all files from the web site, SWAT generates snapshots of each page to TIFF files and describes the entire archive in a METS-file."

Recommendation:

The fact of having to to download, unpack, and install, makes it doubtful if this tool can and/or should be managed by small institutions, that have only one or a few web sites to preserve and present for end-user access. This is also the case with Scenario 2.4; "Linnéjubileet" was a small temporary government agency with only one web site.

Grades:

15 https://wiki.egi.eu/wiki/DCH-RP:PoC_1_Sweden#Proof_of_Concept_scenarios

16 http://www.httrack.com

17 http://swat-archiving.sourceforge.net

On a scale from 1 (very bad) to 5 (very good).

Simplicity of installation: 2

Simplicity of management: not tested

Ease of use: not tested

Generality of solution: not tested

Quality of result: not tested

6.2.3 WARC tools

"The main goal of WARC Tools[18] is to facilitate and promote the adoption of the WARC file format for storing web archives by the mainstream web development community by providing an open source software library, a set of command line tools, web server plug-ins and technical documentation for manipulation and management of web archive files, or WARC files. WARC files are produced by web archiving crawlers, such as Heritrix, the open-source, extensible, Web-scale, archiving quality Web crawler developed by the Internet Archive with the Nordic National Libraries, and Hanzo's own commercial crawlers."

Recommendation:

No real recommendation was given as the installation failed.

Grades:

On a scale from 1 (very bad) to 5 (very good).

Simplicity of installation: 1

Simplicity of management: not tested

Ease of use: not tested

Generality of solution: not tested

Quality of result: not tested

6.2.4 Web Curator Tool

"The Web Curator Tool[19] (WCT) is an open-source workflow management application for selective web archiving. It is designed for use in libraries and other collecting organisations, and supports collection by non-technical users while still allowing complete control of the web harvesting process. It is integrated with the Heritrix web crawler and supports key processes such as permissions, job scheduling, harvesting, quality review, and the collection of descriptive metadata."

Recommendation:

No real recommendation was given as the installation failed.

Grades:

On a scale from 1 (very bad) to 5 (very good).

Simplicity of installation: 1

[18] https://code.google.com/p/warc-tools/

[19] http://webcurator.sourceforge.net/

Simplicity of management:	not tested
Ease of use:	not tested
Generality of solution:	not tested
Quality of result:	not tested

6.2.5 Heritrix

"Heritrix[20] is the Internet Archive's open-source, extensible, web-scale, archival-quality web crawler project."

Recommendation:

It is unclear if the tool can run on other platforms than Linux. According to the FAQ, this has been tried even if it is not supported. A benefit with this tool is that a lot of third-party products don't seem to be required (besides Linux, only Java Runtime Environment is mentioned). However, the installation instructions are not sufficient for an inexperienced user.

Grades:

On a scale from 1 (very bad) to 5 (very good).

Simplicity of installation:	1 - 2
Simplicity of management:	not tested
Ease of use:	not tested
Generality of solution:	not tested
Quality of result:	not tested

[20] https://webarchive.jira.com/wiki/display/Heritrix/Heritrix

7 CONCLUSION

The conclusions and next steps stated in the predecessor deliverable, D5.1 “Technical plan”, were very clear. The following key conclusions were drawn:

a) Interactions between WP3, WP4 and WP5 are iterative by nature
b) Scrum was selected as the methodology to achieve the objectives for WP5.
c) WP5 validates in concrete experiments the concepts established by WP3
d) Results of those experiments may also be negative

Also, the following actions were agreed to be implemented to start activities in WP5:

1. Complete WP5 administration
2. Agile tool chain deployment and usage
3. Project input and material

7.1 APPRAISAL

A common framework of assessing an employee’s performance with respect to her contribution to the overall success of the employer is to (a) divide a calendar year into one or more appraisal periods, (b) define *together* goals (i.e. line manager and employee) the employee sets out to reach during the coming appraisal period, and at the end of said period to compare the achievements to the goals. The same approach will be used to assess the success of WP5 in the covered period.

In terms of the four conclusions reiterated earlier, WP5 activities have taken *all* f them into account as follows:

a) Interactions between WP3, WP4 and WP5 are iterative by nature
Through assigning the leaders of WP3 and WP4 as Product Owners in the Scrum process, close interaction and knowledge exchange between WP3, WP4 and WP5 were ensured. In that respect, the interactions were not only iterative, but also *continuous*.
Also, regular cross-WP Skype calls (where required) immediately after the Sprint planning meetings supported this achievement.

b) Scrum was selected as the methodology to achieve the objectives for WP5.
This is obvious through the usage and uptake of Scum documented in this deliverable and elsewhere (e.g. section 3 in this document).

c) WP5 validates in concrete experiments the concepts established by WP3
WP5 not only ingested the initial five scenarios documented in D3.1, but extended these with additional scenarios that partners found worth exploring. In collaboration with WP3 a technical analysis of the scenarios helped kick starting the experiments conducted in WP5. Sections 4, 5 and 6 give evidence to these.

d) Results of those experiments may also be negative
All conducted Proofs of Concept yielded concrete results, ranging in a wide spectrum from overall positive (e.g. section 6.1.3 6.2.1) to negative (e.g. section 6.2.5), and many variations in between.

Also, the outlined next steps and actions need to be considered:

1. Complete WP5 administration
Key to success here was to assign the roles as described in D5.1 section 3.4.1. All assignments have been completed with the following amendments:
   a. Claudio Prandoni acted as secretary, taking notes for all Sprint planning meetings and generally acted as a proxy for the sponsor, Antonella Fresa.

b. Rosette Vandenbroucke was assigned as second Product Owner; this practical change allowed a Product Owner to be present in almost all Sprint meetings. It also ensured close cross-WP collaboration as described above.

2. **Agile tool chain deployment and usage**
   The tool-chain as described in D5.1 section 3.4.2 was not completely implemented. The reasons were of practical reasons; for example the backlog was not kept as an offline document but as a Google Drive document to allow online collaboration yet benefit from Google Drive's capability of preserving previous versions. However, all tool needs are met, the Wiki, offline documents, A/V conferencing, persistent document store, mailing lists and meeting planning tools were all set up and operative very short after delivering D5.1.
3. **Project input and material**
   This was achieved and ensured though assignment of the Product owner role to the WP3 *and* WP4 leaders, as well as through meetings of the WP3, WP4 and WP5 leaders after the Sprint planning meetings, where required.

7.2 ACHIEVEMENTS & LESSONS LEARNED

Concrete achievements in WP5 are very obvious: Seven scenarios (out of a total of 14) were explored in six concrete experiments conducted by partners in WP5 (see section 4). These results are summarised in sections 5 and 6, with the individual PoC reports attached in Annex 4. These results stand for themselves, however it is important to note that the provided statements and gradings must be considered as subjective and expressions of opinions – fair, perhaps peer-reviewed assessments of tools are out of scope of this project given its funding level.

Beyond these, WP5 contributes to the registry of tools (Task 3.3) with some feedback on tools already included, as well as with a core set of metrics against which these tools should be assessed.

Also, the work of WP5 successfully covered three key gaps identified in D3.1 by modelling these as concrete metrics to be used for tool assessments, informing WP3 for further consideration, and reusable for the registry of tools:

- Easy installation
- Management
- Use of tools

WP5 also gave evidence to a fourth key gap expressed in D3.1, the observed "disjoint of Digital Preservation and existing e-Infrastructures". The Proof of Concept covering scenarios 1.2 and 1.4 involved the use of the EGI Grid as a backend storage system and the e-Cultural Science Gateway hosted by INFN Catania as the frontend service. In a nutshell, the reasons of the failure (or negative result) of this Proof of Concept probably lay in this very gap between the DP ecosystem and existing e-Infrastructures. The reasons for this gap are to some extent also cultural/social discrepancies but even more technical incompatibilities in the understanding of authentication & authorisation infrastructures, and the mechanics of infrastructure-wide support (or non-support) of a certain required feature at a particular point in time: When taking on the decision to integrate with any e-Infrastructure of choice, any Research Community must take into account that infrastructure providers do have feature roadmaps of their technology in place that cannot be changed ad-hoc. Rather, reliable e-Infrastructure providers will have mature change management processes in place to implement the roadmap in the communicated timeframe.

The fundamental lesson the first phase of PoCs has taught DCH-RP is to develop a DCH vision which is expressive yet concise enough to drive which tools and services needed to be tested to support a roadmap towards achieving this vision.

While D3.1 is studying a number of alternative architectures of projects and infrastructures that are related to DCH-RP, the DCH roadmap expressed in D3.4 **must** improve on the community's vision, and give a clear guidance towards the aspired architecture of a DCH e-Infrastructure.

7.3 PLANNING THE NEXT POC PHASE

Work Package 5 expects that D3.4, due 4 months after the publication of this deliverable, will greatly influence the second half of the upcoming Proofs of Concept phases. Planning for further Proofs of Concepts will have to take this into account.

At the time of writing five Proofs of Concept were identified; the general direction being to further improve the coverage of the scenarios described in section 4:

1. **Conduct a Proof of Concept for scenario 1.5**
   In collaboration with Collection Trust, who has access to large collections of 3D visualisations, assess the preservation of 3d data on remote storage locations.
2. **Re-execute the PoC on scenarios 1.1 and 1.4 with an improved e-CSG**
   Several improvements have been implemented in the e-CSG during and after the initial joint Belgian-Italian PoC. Since the underlying framework is due to support Cloud Storage, the PoC may consider switching from Grid to Cloud storage (provided by EGI and Belnet) or support both.
3. **Integrate EUDAT storage services**
   The Polish partner PSNC is an active partner in both EUDAT and DCH-RP. Some EUDAT services are of interest to DCH-RP for future uptake. Depending of the exact definition of this PoC it may cover scenario 3.1 and 3.1.a.
4. **Providing access to preserved data for scientific publishers**
   Together with Elsevier, Editeur and INFN Catania, this Proof of Concept explores the use case of giving a publisher access to scientific data stored in the SCH e-Infrastructure. A possible architecture foresees using the EGI Cloud storage, the e-CSG as the main user facing access, and Elsevier and Editeur as publishers. A partner for the preserved data is yet required.
5. **Value-added metadata analysis services**
   This Proof of Concept will explore the use case of advanced metadata analysis services. By harvesting large amounts of available and accessible metadata collections, new metadata may be generated for consumption of other services. This PoC is interested in exploiting the generic metadata capabilities of CDMI-based Cloud storage services, usability and resolvability of PIDs pointing into datasets (e.g. study supplements) in the Cloud, and the feasibility of cross-referencing studies stored in different repositories (e.g. open access/OpenAIRE and closed access/Elsevier/Springer).
   Potential partners in this PoC are University of Hagen, EGI, Elsevier or any other commercial publisher interested in this, and a supplier of the base data.

ANNEX 1: TECHNICAL INTERPRETATION OF THE SCENARIOS

The following text is a technical analysis of the scenarios (see section 2) on the basis of which the WP5 teams have conducted their Proofs of Concept. The author is Rosette Vandenbroucke, Belspo, with the contribution of Börje Justrell, RA, and Claudio Prandoni, Promoter.

The numbering of the scenarios complies with the work WP5 produced on top of the initial 5 scenarios.

Scenario 1.1

Scenario 1.1 is a complex situation and can be translated into several technical issues that all have their own importance. For more generality we can replace the cloud notion by any use of an e-infrastructure that is external to the DCH organisation.

Given that we start from data that is locally preserved and that this data is accessible via the institute website and via social media channels, the technical issues are the following:

- The issues related to using external tools (provided on the cloud or in general external to the DCH organisation) to local data
- The issue of using tools that require data on e-infrastructure storage.

We can add to these issues:

- Access of data on remote e-infrastructures via the institute website
- Access to the data on remote e-infrastructures via social media (with redirection to the institute website).

A PoC related to this scenario can take into account any combination of issues.

Scenario 1.2

Given that an institute has developed an own preservation infrastructure then the challenge is the use of external tools without comprising the existing solution, issues are:

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

Remark: this will be a very difficult scenario, however scenario 1.2. tackles already part of these issues.

Scenario 1.3

Scenario 1.3 presents the issues of

- Choosing a cost-efficient solution for external preservation
- Choosing a sustainable preservation solution.

A PoC for this scenario will look at different possibilities for external storage of data, its cost and sustainability.

Scenario 1.4

Scenario 1.4 includes the following technical issues:

- Preservation of different data types
- Upload to the preservation storage from different locations
- Preservation of software tools developed or the preserved data
- Copyright issues and/or IPR issues at national and international level

PoCs that take any combination of these issues into account area valuable or the project.

Scenario 1.5

Scenario 1.5 comes down to the preservation of 3D visualisation data together with its running environment (OS, tools, …)

Note that part of scenario 1.5 also pertains to scenario 1.4 (tools and their running environment).

Scenario 2.1

Scenario 2.1 deals with trust issues:

- Authenticity of the data
- Trustworthiness of the data storage
- Transfer errors

It will probably very difficult to realise a PoC for this scenario but we could start an enquiry about the best practices in this matter.

Scenario 2.2

Scenario 2.2 is all about translation of formats. Here the PoC could consist of testing one or more format conversion tools.

Scenario 2.3

This scenario is mostly about retrieval of information, in this case maps, and how some hits “on the fly” on Internet can be transformed into structured searching for knowledge. Metadata is a keyword here and how it is connected to the data to be preserved. In terms of the OAIS model a PoC can be about testing how a SIP can be implemented.

Scenario 2.4

This scenario relates to the problem of “Persistent Identifiers”. There exists standards to cope with the challenge of disappearing links but in order to have such a solution work a lot of coordination is needed at institutional, regional, national and international level.

A PoC for this scenario could consist of taking one of the PID standards and do a coordination exercise between partners of the project.

Scenario 3.1

This scenario is somewhat related to Scenario 1.4 and Scenario 2.1. It is all about proving the trustworthiness of the storage. In the described scenario the grid storage is given as an example but it certainly also pertains to local preservation systems or to cloud based preservation system or any other type of storage used or the preservation.

A PoC in this scenario should look at tools that can describe the trustworthiness of the preserved data (audit, risk analysis).

Scenario 3.1a

This scenario underlines the needed coordination between the e-infrastructure providers (grid, cloud and especially in the terms of storage). It is important to map the requirements at the DCH level to that at the e-infrastructures level.

A PoC for this scenario could be worked out together with a scenario that brings data to the grid or the cloud.

Scenario 3.2

Scenario 3.2 looks at services needed to explore reservation systems on the grid. Probably this could be widened to the cloud preservation solution or any other new preservation solution.

A PoC for this scenario should look at access tools or several preservation solutions (e.; e-CSG, …).

Scenario 3.3

This scenario is mostly about retrieval of information, in this case maps, and how some hits “on the fly” on Internet can be transformed into structured searching for knowledge. Metadata is a keyword here and how it is connected to the data to be preserved.

In terms of the OAIS model a PoC can be about testing how a SIP can be implemented.

ANNEX 2: PRESERVATION OF THE WP5 SCRUM BACKLOG

The Scrum backlog is kept and maintained in an online collaborative document[21] stored in Google's Cloud storage "Google Drive". It is accessible and editable by all WP5 members. This annex is preserving the raw data from the first PoC phase that formed the basis of the day-to-day work in the last months.

Table 2: All "user stories" (tasks) that were accomplished or rejected during the first PoC phase.

Sprint	Epic	ID	Priority	Task	Points	Status
1	1	3	5	Find and confirm up to 4 Italian CH partners for PoCs 1 & 2	0	Accepted
1	1	4	5	Find and confirm up to 4 Polish CH partners for PoCs 1 & 2	0	Accepted
1	2	7	4	Find out EGI NGI International Liaison & deputy for Estonia	1	Accepted
1	2	8	4	Find out EGI NGI International Liaison deputy for Poland	1	Accepted
1	2	9	4	Find out EGI NGI International Liaison & deputy for Sweden	1	Accepted
1	4	10		Confirm BEgrid support through its NIL	0	Accepted
1		14		Confirm BEgrid Grid Resource Providers	0	Accepted
1		20	4	Confirm BEgrid Cloud Resource Providers	3	Accepted
1		40	5	Set up WP5 group and mailing list for DCH-RP	2	Accepted
1		41	5	Find time slot for Sprint 1 review meeting	1	Accepted
1		42	5	Find timeslot for Sprint 2 planning meeting	1	Accepted
2	1	2	5	Find and confirm up to 4 Hungarian CH partners for PoCs 1 & 2	1	Accepted
2	1	5	5	Find and confirm up to 4 Swedish CH partners for PoCs 1 & 2	0	Accepted
2	2	6		Find out EGI NGI International Liaison & deputy for Belgium	1	Accepted
2	2	11	4	Confirm EEnet support through its NIL	2	Accepted
2	2	12	4	Confirm MGKK support through its NIL	2	Accepted
2	2	13	4	Confirm IGI support through its NIL	2	Accepted
2	2	16	4	Confirm MGKK Grid Resource Providers	3	Accepted
2	2	17	4	Confirm IGI Grid Resource Providers	3	Accepted
2	2	18	4	Confirm PL-Grid Grid Resource Providers	3	Accepted
2	2	19	4	Confirm SweGrid Grid Resource Providers	3	Accepted
2	2	21	4	Confirm EEnet Cloud Resource Providers	3	Accepted
2	2	22	4	Confirm MGKK Cloud Resource Providers	3	Accepted
2	2	23	4	Confirm IGI Cloud Resource Providers	3	Accepted
2	2	24	4	Confirm PL-Grid Cloud Resource Providers	3	Accepted
2		30	2	Decide on Cooperation Agreement for Italy	0	Accepted
2		34	5	Provide a link to the installed e-cultural Science gateway on the wiki	1	Accepted
2		43	5	Document what needs to be documented where	3	Accepted
2		44	5	Understand and document Scenario 1.1	5	Accepted
2		45	5	Understand and document Scenario 1.2	5	Accepted
2		46	5	Understand and document Scenario 1.3	5	Accepted

[21] https://docs.google.com/a/egi.eu/spreadsheet/ccc?key=0AuKxnKM_liK-dGpqcGV6ZkVScDhGRjFzZ3gybWp0eVE - gid=0

2		64	5	Provide your Easter holiday details to Michel (covering sprint 3 & 4)	1	Accepted
3	2	25	4	Confirm SweGrid Cloud Resource Providers	3	Accepted
3		35	5	Provide a link to the e-cultural Science gateway documentation on the wiki	1	Accepted
3	3	49	5	Decide on test data for scenario 1.1	2	Accepted
3	5	51	5	Decide on test data for scenario 1.3	2	Accepted
3	5	56	4	Describe testing procedures for scenario 1.3	3	Accepted
3	9	66	5	Decide on test data for scenario 2.2	1	Accepted
3	11	68	5	Decide on test data for scenario 2.4	1	Accepted
3	9	70	4	Describe testing procedures for scenario 2.2	3	Accepted
3	11	72	4	Describe testing procedures for scenario 2.4	2	Accepted
4		33	5	Update https://documents.egi.eu/document/1602 with Epic 1 resilts	1	Accepted
5	3	59	2	Identify required e-Infrastructure resources for scenario 1.1	1	Accepted
5	4	60	3	Identify required e-Infrastructure resources for scenario 1.2	1	Accepted
5	6	62	3	Identify required e-Infrastructure resources for scenario 1.4	1	Accepted
5	3	77	3	Identify required local resources for scenario 1.1	1	Accepted
5	3	78	3	Identify required local resources for scenario 1.2	1	Accepted
5	3	80	3	Identify required local resources for scenario 1.4	1	Accepted
5	5	81	3	Testing de-identification with ROND (Riksarkivet Open Data) Scenario 1.3	1	Accepted
5	5	83	3	Obtaining permission to use test data outside the Preservation Net, scenarios 1.3, 2.2, 2.4	1	Accepted
5		87	3	B PoC2 Ask membership of the VO DCH-RP and Indicate	3	Accepted
5		88	4	B PoC 2 Next Step in mastering the eCSG - attend Webinar 15/5	1	Accepted
6	1	1	5	Find and confirm up to 4 Estonian CH partners for PoCs 1 & 2	1	Accepted
6		26	2	Confirm Cooperation Agreement document & contents	0	Accepted
6		37	5	Organise a training session on the e-CSG	2	Accepted
6	4	50	5	Decide on test data for scenario 1.2	1	Accepted
6	6	52	5	Decide on test data for scenario 1.4	5	Accepted
6	3	54	4	Describe testing procedures for scenario 1.1	5	Accepted
6	4	55	4	Describe testing procedures for scenario 1.2	5	Accepted
6	6	57	4	Describe testing procedures for scenario 1.4	1	Accepted
6	3	79	3	Identify required local resources for scenario 1.3	3	Accepted
6	9	84	3	Identify required local resources for scenario 2.2	3	Accepted
6	9	85	3	Selecting specific images for test data, scenario 2.2	3	Accepted
6	11	86	3	Identify required local resources for scenario 2.4	3	Accepted
7		47	4	Understand and document Scenario 1.4	5	Accepted
7		48	3	Understand and document Scenario 1.5	5	Accepted
7	5	61	3	Identify required e-Infrastructure resources for scenario 1.3	5	Accepted
7	9	74	3	Identify required e-Infrastructure resources for scenario 2.2	3	Accepted
7	11	76	3	Identify required e-Infrastructure resources for scenario 2.4	3	Accepted
7		89	4	B PoC2 Exploit the possibility to have eCSG working with dCache	13	Accepted
7		90	4	Look at the integration of metada	20	Accepted
7		91	3	B PoC2 if dcache works with eCSG add the Belgian gridstorage to the VO DCH-RP	5	Accepted

7		92	3	B PoC 2 if 86 is positive, copy data from KIK to the Belgian gridstorage	5	Accepted
7	9	109	3	S PoC 1 Decide target formats for conversion of images	1	Accepted
7	9	110	3	S PoC 1 Test the AVS Document Converter	3	Accepted
7	5	119		S PoC 1 Metadata: check the status of the Swedish eARD project for the chosen file formats	2	Accepted
8		111		B PoC 2 Italian partners to become member of the VO DCH-RP	3	Accepted
8		112		B PoC 2 Italian partners to check readiness to use the Italian e-CSG	3	Accepted
8		113	2	B PoC 2 Italian partners to make the choice of data to go to the gridstore	5	Accepted
8		114		B PoC 2 Italian partners to check data for transfer to the Italian Grid storage	3	Accepted
8		138		B PoC 1 testing the scoremodel	3	Accepted
10	11	120		S PoC 1 Test WARC Tools	40	Accepted
10	5	121		S PoC 1 Test Archivist's Toolkit	5	Accepted
10	11	122		S PoC 1 Test Web Curator Tool	3	Accepted
10	11	123		S PoC 1 Test SWAT	5	Accepted
10	11	125		S PoC 1 Obtain a test person (like the one described in scenario 2.4)	1	Accepted
10	9	126		S PoC 1 Test Universal Document Converter	3	Accepted
10	5	127		S PoC 1 Test DSpace	3	Accepted
10	9	128		S PoC 1 Test A-PDF DjVu to PDF	3	Accepted
10	11	129		S PoC 1 Test HTTRACK	3	Accepted
10	5	133		S PoC 1 Test XENA	3	Accepted
10	9	136		S PoC 1 Test AVS Image Converter	3	Accepted
10	11	137		S PoC 1 Test Heritrix	3	Accepted
11	2	15	4	Confirm EEnet Grid Resource Providers	3	Accepted
11		93	2	B PoC2 Define the access methods to the grid storage (depends on #139)	5	Accepted
11		94	3	B PoC 2 Define the access measurement tools (depends on #139)	20	Accepted
11		115	1	B PoC 2 Italian partners copy their data to the Italian grid storage	20	Accepted
11		139	1	B PoC 1 Belgian partners to copy their data into the Italian Grid storage	20	Accepted
11		95	3	B POC2 Do the data access tests	5	Accepted
12		104	1	B PoC 1 report the PoC	3	In Progress
12		140	1	B PoC2 Report the PoC	3	In Progress
11	11	130		S PoC 1 Let the test person find the archive Linnéjubileet, open it and evalutate how easy it was to find and use this archive	2	Accepted
11	11	131		S PoC 1 Compare how the access to Linnéjubileet works when you use the original one (www.linne2007.se) and when you use the format/tools chosen for scenario 2.4. (This test must be omitted if the original site ceases to be accessible)	3	Accepted
12		134		S PoC 1 Report the PoC to WP3	3	Accepted
		~~124~~		~~S PoC 1 Register the web site archive for Linnéjubileet in the archive information system NAD ("Nationella arkivdatabasen", the National Archive Database)~~		~~Accepted~~

		~~36~~	~~5~~	~~Provide link to documentation on dARCEO on the Wiki~~		~~Rejected~~
		~~38~~	~~5~~	~~Start the registry of services documentation in the Wiki~~		~~Rejected~~
		~~39~~	~~5~~	~~Add a first set of tools and services to the registry~~		~~Rejected~~
	~~7~~	~~53~~	~~5~~	~~Decide on test data for scenario 1.5~~		
	~~7~~	~~58~~	~~4~~	~~Describe testing procedures for scenario 1.5~~		
	~~7~~	~~63~~	~~3~~	~~Identify required e-Infrastructure resources for scenario 1.5~~		
	~~8~~	~~65~~	~~5~~	~~Decide on test data for scenario 2.1~~		
	~~10~~	~~67~~	~~5~~	~~Decide on test data for scenario 2.3~~		
	~~8~~	~~69~~	~~4~~	~~Describe testing procedures for scenario 2.1~~		
	~~10~~	~~71~~	~~4~~	~~Describe testing procedures for scenario 2.3~~		
	~~8~~	~~73~~	~~3~~	~~Identify required e-Infrastructure resources for scenario 2.1~~		
	~~10~~	~~75~~	~~3~~	~~Identify required e-Infrastructure resources for scenario 2.3~~		
	~~3~~	~~82~~	~~3~~	~~Identify required local resources for scenario 1.5~~		
		~~135~~		~~S PoC 1 Report the PoC to WP3~~		

Table 3: Some tasks were decided to be tackled in the second PoC phase.

Sprint	Epic	ID	Priority	Task	Points	Status
13		27	2	Decide on Cooperation Agreement for Belgium	60	
13		28	2	Decide on Cooperation Agreement for Estonia		
13		29	2	Decide on Cooperation Agreement for Hungary		
13		31	2	Decide on Cooperation Agreement for Poland		
13		32	2	Decide on Cooperation Agreement for Sweden		
13		98	2	B PoC1 definition of documentation of the work	13	In Progress
13		99	4	B-PoC1 choice of audit tool (Drambora or ...) with partners	20	In Progress
13		101	4	B PoC 1 choice of part of the tool if necessary	20	In Progress
13	5	107	3	S PoC 2 Identify Swedish "cloud archive providers"	3	In Progress
13	5	108	3	S PoC 2 Compile a list of requirements for "cloud archive providers"	3	In Progress
14	5	116		S PoC 2 Construct a questionnaire and send it to DCH-RP WP5 and other interested parties for review	3	
14	5	117	3	S PoC 2 Update questionnaire and send it to the "cloud archive providers"	3	
19	5	118	3	S PoC 2 Evaluate the answers to the questionnaire and determine if a commercial "cloud archive provider" is suitable for a small institution like that in scenario 1.3	5	
	5	106	3	S PoC 1 Test and evaluate the ARC Graphical Client (for SweGrid/SweStore)	5	
13		96	1	B PoC2 Exploit the possibility to install eCSG on BEgrid	20	In Progress
13		97	1	B PoC2 Exploit the data access tests on the Belgian eCSG	5	In Progress
13		100	2	B PoC1 definition of the local preserved data on which the audit is done	3	In Progress
13		102	2	B PoC1 definition of the audit procedures	13	In Progress
13		103	1	B PoC 1 execution of the tasks	13	In Progress

13	5	105	3	S PoC1 Obtaining permission to use test data outside Riksarkivet, scenarios 1.3, 2.2, 2.4	3	In Progress
13	5	132		S PoC 1 Test Fedora	3	

Table 4: Each sprint is maintained with start and end dates, and key indicators for planning purposes.

	Start	End	duration [d]	Sprint	Theme	Performance	Velocity	Remaining	Total points
PoC 1	18 Feb 2013	3 Mar 2013	14	0	Setting up Scrum	0	n/a	120	120
	4 Mar 2013	17 Mar 2013	14	1	Connecting with e-Infrastructures	10	n/a	139	149
	18 Mar 2013	4 Apr 2013	18	2	Understanding scenarios	52	20.67	116	178
	4 Apr 2013	21 Apr 2013	18	3	Working on test data and procedures	18	26.67	98	178
	22 Apr 2013	6 May 2013	15	4	Expanding work on PoC scenarios	1	23.67	123	204
	6 May 2013	23 May 2013	18	5	Gather our first concrete results	12	10.33	261	354
	23 May 2013	10 Jun 2013	19	6		32	15.00	254	379
	10 Jun 2013	24 Jun 2013	15	7		70	38.00	264	459
	24 Jun 2013	15 Jul 2013	22	8	Vacation, vacation, vacation ;)	17	39.67	170	382
	15 Jul 2013	29 Jul 2013	15	9	Vacation, vacation, vacation ;)	0	29.00	172	384
	29 Jul 2013	13 Aug 2013	16	10	Finish tests that are in progress	75	30.67	87	374
	13 Aug 2013	9 Sep 2013	28	11	Finish results and draft D5.3	78	51.00	9	374
	9 Sep 2013	30 Sep 2013	22	12	Write and deliver D5.3	50	67.67	-41	374

Table 5: Epics are a means of grouping tasks together in a meaningful way.

Epic ID	Title	Description
1	Establish CH institute cooperation	Find CH institutes that are willing to contribute their data to the PoCs, as well as assume the role of users of the infrastructures set up in the PoCs.
2	Establish EGI resources	EGI is offering Grid and Cloud resources for the DCH-RP project. Establish and confirm these resource on a national & local coordinated way.
3	Scenario 1.1 Using specialised research tools	A major memory institution in France which has its own development team is gradually implementing a solution for digital preservation. It is using local in-house storage. The institution participates in projects which aggregate

		content to Europeana and regularly uses social media channels to engage with the wider public. Thus, the access to its digital collections is either possible through the institutional website, or resource discovery is made via specialised portals and social media which in fact redirect the users to the institutional webserver. Recently, it has happened several times that researchers ask to use specialised document analysis tools that are available through an e-Infrastructure. This raises issues of sharing content outside the institutional storage and preservation facilities on the cloud used by the eInfrastucture, or the use of 'external' tools for processing locally stored documents. Both options raise concerns, and for the time being there is no good solution for the end users.
4	Scenario 1.2 - Integrating a new tool into an existing institutional infrastructure	A major memory institution in Germany had already developed its own preservation infrastructure. A new research project is asking for a newly developed software tool that would save time on checking file formats. However, the integration of this tool with the existing preservation solution cannot compromise any essential preservation features implemented in the local preservation system. The requirement is to analyse the difference that using the new tool will make and how to embed it with other components already in place; or how to run the new tool from a cloud-based provider and integrate this service with the existing preservation solution.
5	Scenario 1.3 - Selecting a digital preservation solution in the case of an institution with only voluntary IT support	A little museum in Malta has a historical library and a digitised personal archive collection. The museum has staff of only 9 and only voluntary IT support. The director of the museum is aware of the need to organise digital preservation for the digitised documents, but is not sure how to do it. He receives periodically offers for long-term storage of digital content, but finds it difficult to select or to make a decision. He has practically no IT competence to rely on for decision-making, but is convinced that the decision should be forward-looking and accommodate the needs of the museum for the next 5 years.
6	Scenario 1.4 - Preservation from a consortium of collections on the cloud	A specialised consortium of several institutions working on a complete digital repository of the works of a modern digital artist who worked and exhibited in 15 different countries has to resolve the issue of preservation of objects that are stored in different location. The works of the digital artist include a variety of digital formats as well as especially developed software tools. The curator of the collection has to identify a cost efficient solution which would also be suitable to store the complex objects in the collection. An additional difficulty is that the copyrights on the objects differ in the countries of origin of the objects.
7	Scenario 1.5 - Preserving a 3D visualisation	A research lab in the UK is collaborating with an archaeological site in Italy to create a 3D visualisation of an ancient building. The visualisation is used as scientific documentation. Both institutions have to agree who will take care for the preservation in usable state of the model. There is also an issue of interoperability of the model with a free visualisation tool which can be used to show the model on a web site which is resolved producing a lower quality visualisation in an additional format. There is an ongoing discussion whether it also needs to be preserved and by whom.
8	Scenario 1.6 - Archived data retrieving	Estonian memory institution (Conservation Centre Kanut) which digitises different content wants to make backup copy of files to another memory institutions tape library but needs proof that content is well preserved and it is possible to receive copy of files if needed. Therefore periodically (quarterly) will be carried out test data retrieving.
9	Scenario 2.1 - Researcher discovers a historical database	Researcher in history discovers a historical database resource presenting parish records. She would like to use the data, but she is also concerned to what extent these data could be trusted (authenticity, error rates introduced, errors caused by any transformations needed).
10	Scenario 2.2 - Research and select a tool serving a	A university lecturer in art history wants to use a collection of digitised art images made 15 years ago. They are stored in a format he is not familiar with. Since there are about 200 images, the researcher is looking for tools

	specific purpose	which would convert them into a format he could easily use in batch mode. He is not sure how to identify a tool or a service which could do this.
11	Scenario 2.3 - Accessing digitised content from schools(?)	Secondary school students are making an assignment looking at historical maps of their village. They already paid a visit to the local museum but discovered some old digitised maps on the internet.
12	Scenario 2.4 - Gain access to archived websites	A history student interested in natural history discovers that Riksarkivet has archived the "Linnéjubilet" web site http://www.riksarkivet.se/default.aspx?id=23153 .He wonders how he can get access to it (the link www.linne2007.se obviously doesn't work anymore).
13	Scenario 3.1 - Proof of authenticity in distributed archiving	The Swedish National Archives takes 10 digitised images of records and ingests them into their national GRID where they undergo a migration cycle or some other processing and the SNA requires a proof of authenticity at the end of this.
14	Scenario 3.2 - Defining new services	A small art gallery looks for the grid infrastructure for storage services that could solve the preservation problems. For that is needed new services not yet defined.
15	Scenario 3.3 - Integrating new services into existing infrastructure	The IT manager of a local art gallery is preserving the digital content using grid X. He attends a workshop on digital preservation where he hears about a new tool for checking the integrity of digital objects. He needs to implement it on the grid-based archiving solution.
16	Scenario 3.1a - extending 3.1 with repository safeguarding policies	During this processing , the GRID provider ask for information about SNA:s requirements for safeguarding a trustworthy repository. The SNA has to describe the methodology and tools they are useing for validating their objectives and methods as well as their management of intrinsic theats and threats originating from the outside of the organisation. The purpose from the GRID providers point of view is to push the SNA to come up with trust critieria for the services it will get from the GRID.

Table 6: WP4 also kept track of which partner examined which scenario in their Proofs of Concept.

Scenario		**Partner**					
Theme	Scenario	BELSPO (Belgium)	EVKM (Estonia)	NIIFI (Hungary)	ICCU (Italy)	PSNC (Poland)	RA (Sweden)
Organisational challenges	Scenario 1.1	X			X		
	Scenario 1.2	X					
	Scenario 1.3						X
	Scenario 1.4	X					
	Scenario 1.5						
	Scenario 1.6		X				
End user concerns	Scenario 2.1						
	Scenario 2.2						X
	Scenario 2.3						
	Scenario 2.4						X
New services,	Scenario						

integrating into infrastructure	3.1						
	Scenario 3.2						
	Scenario 3.3						

Table 7: A key element of Scrum sprints are retrospections of the past sprint, and which (social) issues the team should work on to improve.

Sprint retrospections		
Sprint 1		
Appriases	**Criticisms**	**What to improve**
Keep review and planning in the same meeting	Clarify what needs to be documented where	Keep review and planning in the same meeting
	The menaing of task states is unclear	Better documentation of the Sprint tools
		Remove the "Delivered" state
		Improve communication in sprint execution
Sprint 2		
Appriases	**Criticisms**	**What to improve**
Scenarios are starting to give us more scope, CH institutes are committing.	Compile a list of actions in the minutes, the backlog seem not enough. Perhaps assign tasks to specific people in such way.	Compile a list of actions in the minutes, the backlog seem not enough. Perhaps assign tasks to specific people in such way.
Conference calls get more efficient using less time		Add conf call access details in the invitation Email
		People should join timely to the conference call
		Shorten conference call timeslots to 1.5 hours from now on
Sprint 3		
Appriases	**Criticisms**	**What to improve**
Contributions from Eva and Rosette are good	Roberto has never shown up	Participation of WP members
Those participating were timely	Generally low participation in the WP5 conf calls	Not enough collaboration between WP3 and WP5 (Borje)
	Borje as Product owner has never shown up	
Sprint 4		
Appraises	**Criticisms**	**What to improve**
Eva and Rosette are committed	Much work was accomplished that was	When finishing tasks, team

and proceeed on PoC1	not captured in tasks	members are allowed to adjust the points!
	More fine-grained steps in defining the tasks	Everybody is encouraged to work with and edit the backlog
		Once ytou change the backlog, drop a mail to WP5 notifying of this change

Sprint 5

Appraises	Criticisms	What to improve
Eva and Rosette are committed and proceeed on PoC1		Better circulate the meeting minutes
Lajos joined today, and plans to do so in the future.		

Sprint 6

Appraises	Criticisms	What to improve
Eva and her e-Infrastructure reports are well received and appraised	Still low participation in the sprint planning meetings	Changing scenarios in practice in WP5 should be better synchronised and documented towards WP3
Eva likes that she has fnished something! :-)		
Meeting minutes are now circulated appropriately now.		

Sprint 7

Appraises	Criticisms	What to improve
Eva and her e-Infrastructure reports are well received and appraised	We need to know when people go on vacation	Maintain people's vacation plans in the Backlog
Eva likes that she has fnished something! :-)		Michel to advertise untaken scenarios on DCH-RP mailing list

Sprint 8

Appraises	Criticisms	What to improve
	Very slow responses from INFN and Roberto	

Sprint 9

Appraises	Criticisms	What to improve
	lack of priorities for tasks	Be more "pushy" towards

		Rosette and Borje to provide priorisation
Sprint 11		
Appraises	**Criticisms**	**What to improve**
The SCRUM methodology was appreciated.	Borje was concerned in the beginning whether SCRUM would work. But during the actual PoC1 phase he was convinced that SCRUM is very successful.	Add a PoC report about using SCRUM to D5.3
	Scrum is about the management of a project, not the content. Scrum was not designed for this, so there is a need for a solution that facilitates the open communication between team members during the sprints.	

ANNEX 3: SCENARIO REPORT TEMPLATE

During implementing the Proofs of Concept the members of WP4 decided to report the results of the conducted activities in individual reports, rather collecting them in one deliverable. This approach has several benefits, in that it allows a more individual reporting, focusing on the actual work undertaken, and the respective results. It allows the roadmap evolution activity in WP3 to ingest the results as they see fit. Further, it reflects the more particular nature in the Proofs of Concepts that were not connected to each other, and conducted in a more national fashion (except for Belgian and Italian partners).

This section incorporates the reporting template in a condensed form, highlighting the core benefits and key elements. The full report template is maintained in the persistent document storage (https://documents.egi.eu/document/1892).

A3.1 DOCUMENT METADATA

The first part of the report template contains all necessary metadata about the information contained in the document. This includes enumerating the scenarios and any tools and services that were used in the Proof of Concept:

- Covered scenario, and tools
- Authors
- Revision of the document (including final)
- Revision history
- Dissemination level
- Persistent storage location (in the footer of the document)

A3.2 EXECUTIVE SUMMARY AND GRADING

The first content section provides an executive summary, along with a tabular overview of the grading of a number of aspects that apply to the described Proof of Concept. The template does not impose any aspects to be graded; however each aspect that is included in the summary must be sufficiently explained in the annex of the report. Aspects capture anything that is important for the Proof of Concept and subsequent assessment in roadmap discussions. Aspects may range from functionality (e.g. “Support of Dublin Core metadata”) to non-functional requirements, such as ease of installation.

The executive part of the report also includes specific recommendations the authors wish to highlight to the roadmap evolution team.

A3.3 PROOF OF CONCEPT REPORT

The bulk of the report contains the details of the conducted Proof of Concept. While the template suggests separating the description of the Scenario from the tools tested within the scenario(s), it is intentionally left to the authors how to structure this part of the document: Form follows function, for as long as the key outcomes are reflected accordingly in the executive summary section.

A3.4 ANNEXES

The template provides only one annex, but does not limit the authors to add more annexes where required. The template’s sole required Annex lists all aspects of the involved tools and services that were examined throughout the Proof of Concept.

The description of each aspect is designed to be complete and self-referential within the document, in that it defines:

- Aspect name
- Description
- Grades “n/a”, 1, 2, 3, 4 and 5
- Semantics of each grade

This way, WP4 is able to collect a set of specific aspects that were used and tested during the Proofs of Concept. These may overlap, but the process of reconciliation is expected to be negligible. The resulting set of aspects then can be used for future reference in further Proofs of Concept.

ANNEX 4: PROOF OF CONCEPT REPORTS

This annex preserves the complete reports on the Proofs of Concept conducted in the first phase leading to this deliverable.

Due to document production processes, only the final PDF version of this deliverable will be concatenated with the PDF versions of the PoC reports.

Proofs of Concept
Executive report for Roadmap consideration

Scenario 1.1

Revision:draft

Authors:

Rosette Vandenbroucke (Belspo)
Patrizia Martini (ICCU)
Giovanni Ciccaglioni (ICCU)

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

Revision History

Rev.	Date	Author	Organisation	Description
1	R.Vandenbroucke	M. Drescher	Belspo	Draft description joint PoC Belgian-Italian partners for Scenario 1.1

TABLE OF CONTENTS

1 EXECUTIVE SUMMARY

KIK-IRPA, one of the Belgian DCH organisations, has already a local preservation system for their data. They have described their preservation system in a “Best practices” document that is accepted throughout the organisation. However one of their main concerns is to maintain the integrity of their data.

There exist auditing and certification schemes for trustworthy repositories, see: “Trustworthy Repositories – Audit and Certification” http://www.digitalrepositoryauditandcertification.org/pub/Main/ReferenceInputDocuments/trac.pdf and "Risk-analysis for E-depots:DRAMBORA" http://www.repositoryaudit.eu/. Such an audit also equals a risk analysis of the chosen archiving method. How easy this all may sound, real life shows that almost no one ever terminates the whole procedure, hence there is no common “best practices” available.

Doing an audit in a consequent way requests to use the necessary tools.

In this scenario we want to use existing tools and document the auditing process. We will do this on the local data that is in the KIK-IRPA preservation scheme and on data of the Italian partners..

Such an audit is in fact independent of where the data is stored but it is certainly a “tool” that will be very valuable for preservation done on data stored with e-infrastructures or other storage service providers.

Once done it would be useful to execute the procedure on preservation done on grid and cloud.

The original work defined for this PoC included the study of the Drambora auditing and certification scheme, the partially existing implementations and arrive at a user friendly implementation that woul d be used for test the integrity of local data at KIK-IRPA in Belgium and ICCU in Italy.

However this work was too huge to be included in the timeframe of the first round of PoCs and the choice went to use the tool “Scoremodel” developed by DEN. DEN Foundation is the national center for ICT in cultural heritage.

1.1 GRADING

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

Aspect	Score
- Usefulness of the tool	**4**
- User friendliness of the tool	**4**

1.2 RECOMMENDATIONS

The “Scoremodel” is a useful tool to test the integrity of a collection. It does not implement the full Drambora scheme but has the advantage that it is easily understood and usable by DCH people.It can be used in the roadmap as an example for testing the integrity of data.

2 SCENARIO OVERVIEW

The tool scoremodel will be used to test the integrity of data stored locally at KIK-IRPA in Belgium and at ICCU in Italy.

2.1 DOCUMENT STRUCTURE

Below the tool ”Scoremodel” is introduced together with the results of its use.

2.2 SCENARIO / TOOL TESTING ENVIRONMENT

The “Scoremodel”tool is available via the website www.scoremodel.org. The user needs to register with the website to be enabled to use the tool. The results of the tool are only delivered to the registered user and are not stored on the website or made available to a third party.

3 SCOREMODEL

Scoremodel is a tool to test the integrity of stored data and is mainly built on the Drambora scheme.It aimed at being user friendly while providing a useful result.

3.1 DATA SETS

Local data at KIK-IRPA and ICCU will be used.

3.2 TEST DESCRIPTION

The "scoremode" tool guides the user through the risks and threats to digital materials. It has a series of questions that create a report that points out the strong and weaker points of your digital organization. The report provides recommendations in order to minimise the risks where possible..

These risks are grouped in seven clusters:

1. Organisation and policy: does the preservation of digital files fit the structure and policy of your organisation?
2. Preservation strategy: is it correctly recorded what is being preserved, for whom and in what manner?
3. Expertise and organisation: is the right expertise present in your institution and it it put to good use?
4. Storage management: is the physical storage of the digital files also reliable?
5. Ingest: are the right measures taken whenever a digital object is ingested into your storage system?
6. Planning and control: is the management well prepared? Are all actions retraceable?
7. Access: is access to the digital files properly regulated?

3.3 RESULTS

The tool proved to be user friendly and gave explications and examples during the execution. Both KIK-IRPA and ICCU had positive comments on the tool and appreciated the outcome quoted as "useful and helpful". However ICCU felt that it could not completely judge as the long term preservation is not the first goal, the mission, the core business of Internet Culturale and of the ICCU in general.

4 CONCLUSION

The use of this tool was a good exercise before starting out on a bigger project like exploring the complete Drambora scheme. Of course there are more tools that check the integrity of quality of (preserved) data and it would be fine to include all those tools in the “services registry.

Proofs of Concept
Executive report for Roadmap consideration

Scenario 1.2
Scenario 1.4

Revision: draft

Authors:

Rosette Vandenbroucke (Belspo)
Patrizia Martini (ICCU)
Giovanni Ciccaglioni (ICCU)

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

Revision History

Rev.	Date	Author	Organisation	Description
1	22 Sep 2013	Rosette Vandenbroucke	Belspo	First description of the PoC introduced
2	24 Sep 2013	Rosette vandenbroucke	Belspo	Added information from the Italian partners, refined existing text
3	27 Sep 2013	Giovanni Ciccaglioni Patrizia Martini	ICCU	Information and data from the ICCU PoC

TABLE OF CONTENTS

1 EXECUTIVE SUMMARY

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

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

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

The different steps are:

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

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

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

But within the lessons learned, it should be noted that the integration between E-infrastructures and CH sector has to go through the understanding and awareness of problems existing in adopting the e-infrastructures and tools.

1.1 GRADING

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

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

1.2 RECOMMENDATIONS

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

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

2 SCENARIO OVERVIEW

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

The test scenario was built up as follows:

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

2. Use the eCSG installed in Catania with storage on BEgrid
(for this scenario there is a need to adapt dcache, the storage management system in BEgrid, to eCSG) (Belspo)

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

Execution

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

2.1 DOCUMENT STRUCTURE

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

2.2 SCENARIO / TOOL TESTING ENVIRONMENT

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

3 TESTS TO BE EXECUTED

3.1 CHECK IF BEGRID STORAGE CAN BE USED WITH ECSG

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

Test description

A grid expert checked the characteristics.

Results

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

3.2 COPY DATA TO THE ITALIAN GRID STORAGE

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

Test description

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

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

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

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

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

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

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

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

Results

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

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

3.3 MAKING METADATA AVAILABLE ON THE ECSG

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

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

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

Test description

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

Results

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

4 CONCLUSION

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

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

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

ANNEX 1

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

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

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

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

Proofs of Concept
Executive report for Roadmap consideration

Scenario 1.3

Revision: 0.1 (draft)

Authors:

Eva Toller (Riksarkivet)

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

Revision History

Rev.	Date	Author	Organisation	Description
0.1	2013-08-26	Eva Toller	Riksarkivet, Sweden	First draft

TABLE OF CONTENTS

1 EXECUTIVE SUMMARY

This document describes the results of the tools that were tested in the context of Scenario 1.3. The ultimate goal of the tests was to grade the tools according to different aspects of usefulness. This document contains an overview of the results. For details, see the DCH-RP WP5 document named **DCH-RP_WP5_Scen-1-3_ID-ToolTests.pdf** .

The tested tools were mainly chosen for their estimated suitability for a small cultural heritage institution with little or no IT competence. Sometimes a tool was suitable according to this criterion; in other cases it was highly *un*suitable. However, if the tools were used and/or integrated in a cloud or grid environment, with much more resources and IT competence at hand, the results and conclusions might be different.

The following tools were tested:

- ROND (Riksarkivet OpeN Data, version 1.0)
- Archivist's Toolkit 2.0
- Xena
- DSpace

1.1 GRADING

The aspects that have been graded are shown in the table below. The scores for the individual tools are given in Chapters 3 – 6. For an explanation and the definitions of the various aspects see Annex 1.

Aspect
Simplicity of installation
Ease of use
Generality of solution
Quality of result

The score for each aspect is in the range of 1 (very bad) to 5 (very good).

In **DCH-RP_WP5_Scen-1-3_ID-ToolTests.pdf**, the aspect “Simplicity of management” is also mentioned, but since that aspect has been deemed to be “Not applicable” for all tools, it is omitted here.

Since all the tools could not even be installed, sometimes the only grading that is done is “Simplicity of installation”.

The grades given in this document and the grades given in **DCH-RP_WP5_Scen-1-3_ID-ToolTests.pdf** may differ slightly, since no detailed descriptions or definitions of the grades were made previously. The grades in this document are the more accurate ones.

1.2 CONCLUSIONS AND RECOMMENDATIONS

The tool tests are really only a few samples out of hundreds of possibilities. However, the lessons learned from testing these specific tools were the following:

For local usage:

1. The different software parts necessary for a tool should be *packaged together* as much as possible, so that the end-user will not be required to separately download and install a lot of different tools him/herself.
2. The installation files should be few and very simple to manage; preferably *.msi files* (for Windows), the next best is *.exe files*.
3. The installation instructions should be easy to understand, neither too short nor too long, and be up-to-date.
4. When there is a trial version (limited to 30 days of usage) and a commercial version that you maybe will buy later, the trial version must be *very easy* to install (so the end-user doesn't have to waste time on a tool that may not be appropriate).

For services hosted at an **e-infrastructure provider** (grid or cloud):

1. The provider must be responsible for the packaging of the service so that it is as simple as possible to use (especially if it requires installation on a client by the end-user).
2. If the end-user must install anything his/herself: the installation instructions *must* be up-to-date, and easy to follow.
3. There should be "live" support at the provider, ready to answer all types of questions about the services (also for tools provided by third party).

2 SCENARIO OVERVIEW

“A little museum in Malta has a historical library and a digitised personal archive collection. The museum has staff of only 9 and only voluntary IT support. The director of the museum is aware of the need to organise digital preservation for the digitised documents, but is not sure how to do it. He receives periodically offers for long-term storage of digital content, but finds it difficult to select or to make a decision. He has practically no IT competence to rely on for decision-making, but is convinced that the decision should be forward-looking and accommodate the needs of the museum for the next 5 years.”

For a detailed description of the data that was used for this scenario, see **DCH-RP_WP5_Scen-1-3_ID-51.pdf** .

2.1 DOCUMENT STRUCTURE

Chapter 3 and the following chapters (“X”) are structured in the following way:

In the beginning of Chapter X, a short description is given of the tool and how it works.

In sections X.1, the data set(s) that the tool will be tested on is described.

In sections X.2, the execution of the tests are described.

In sections X.3, the results of the tests (if any) are described. General comments are given about the tool and its usability for digital cultural heritage preservation, dissemination et c. (This section may be skipped if it was not possible to install and/or run the tool).

For a more detailed description of the tests, see **DCH-RP_WP5_Scen-1-3_ID-ToolTests.pdf** .

2.2 SCENARIO / TOOL TESTING ENVIRONMENT

The test environment was a PC (Personal Computer) with Windows 7 Professional, processor Intel(R) 2,7 GHz, and 8 GB working memory (RAM).

3 ROND (RIKSARKIVET OPEN DATA)

ROND (Riksarkivet OpeN Data) is a tool for de-identifying data sets. It is dependent on Riksarkivet's chosen meta data format for structured text files, **ADDML** : http://xml.ra.se/addml/ (in Swedish).

3.1 DATA SETS

The data set that was used to test ROND is called “Filmregistret” (the Film Records Collection). These are not records of films *per se*, but of the censorship that has been performed for some films containing scenes that are violent or offensive in other ways. The actual cuts are not included in this data set.

There are four separate files in “Filmregistret”. The file affected by the de-identification is named *Filmregistret.csv*. It contains general and administrative information about the films and the censoring process (including the name of the censors and the technicians).

3.2 TEST DESCRIPTION

The purpose of the test was to de-identify the names of the censors and the technicians (in the file *Filmregistret.csv)*. All letters in these names would then be replaced with x's. (Note that the names of the directors and actors were not candidates for de-identification; firstly, they are already widely known, and secondly, to remove them would greatly decrease the usability of the data set).

3.3 RESULTS

The program behaved as expected and gave the correct results. However, there may be a problem with this program concerning usability. Even though it was possible to do the de-identifications without a user manual, some of the text in the graphical user interface is quite misleading, and it is also unnecessarily cumbersome to choose the files you want to work with. It must also be pointed out that ROND has a major limitation in that it requires a certain metadata model (ADDML), which is currently only used by Sweden and Norway. However, tools of this *type* could be very useful for publishing huge amounts of archival information as open data (information that otherwise would be locked up in the archives, and much harder to find and obtain for interested parties).

Grades

Simplicity of installation: 5

Ease of use: 3 (previously 2 – 3)

Generality of solution: 1

Quality of result: 5

4 ARCHIVIST'S TOOLKIT 2.0

The Archivists' Toolkit™ is intended for a wide range of archival repositories. The main goals of the AT are to support archival processing and production of access instruments, promote data standardization, increase processing efficiency, and lower training costs.

4.1 DATA SETS

Not applicable (no tests were run).

4.2 TEST DESCRIPTION

The purpose was to test the functionality of Archivists' Toolkit. This could not be achieved since a small error in the first installation attempt made it impossible to re-install the tool (at least on the same computer).

4.3 RESULTS

If this tool is to be recommended for small Cultural Heritage institutions, the installation problem should be solved first.

Grades

Simplicity of installation: 1 – 2

5 XENA

Xena performs two important tasks: detecting the file formats of digital objects and converting digital objects into open formats for preservation.

5.1 DATA SETS

Samples from test data for Scenario 1.3 (see **DCH-RP_WP5_Scen-1-3_ID-51.pdf**) and from Scenario 2.2 (see **DCH-RP_WP5_Scen-2-2_ID-66-restricted.pdf**).

5.2 TEST DESCRIPTION

The purpose of the test was mainly to test “normalisation”; that is, converting documents of different types of formats to XML. The format of the input files were the following: TIFF, DjVu, Excel, ODT, CSV, JPEG.

5.3 RESULTS

The results varied much for different types of input files. Xena worked best for TIFF and JPEG as input formats, worse for ODT and CSV, and not at all for DjVu and Excel.

Xena is easy to use for batch conversion; when you have supplied default input and output directories, it takes very little effort to normalise/convert a lot of files (that is, if the input formats are “good” ones). However, since no XML schema is generated, and the results cannot be viewed in a common web browser, it is hard to verify the result in depth.

Grades

Simplicity of installation: 3 (previously 2 – 3)

Ease of use: 4

Generality of solution: 4

Quality of result: 2

6 DSPACE

DSpace preserves and enables access to digital content like text, images, moving images, mpegs and data sets.

6.1 DATA SETS

Not applicable (no tests were run).

6.2 TEST DESCRIPTION

The purpose was to test the functionality of DSpace. This could not be achieved since it seems clear from the number of third-party tools, and from the installation instructions themselves, that this is probably a too complicated tool for the actors in Scenario 1.3.

6.3 RESULTS

If this tool is to be recommended for small Cultural Heritage institutions, the many software parts should be *packaged* in a way that makes it easy for a novice user to make the installation.

However, there is now also a hosted service, ***DspaceDirect***, that may be investigated as an alternative.

Grades

Simplicity of installation: 1 – 2

ANNEX 1

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

Aspect: Simplicity of installation	
Definition: How complicated was it to download the tool? Did you have to register to get the download? Was it obvious which download version you should choose? If the download was packaged in a compressed file, how easy was it to unpack it? Were there any installation instructions, either on the download site or in the download itself? Was it necessary to install databases or other large third-party tools? In all, how many separate programs were necessary to install? How many mandatory parameter values had to be given during installation? If the first installation try failed, was it easy to install the tool anew?	
Grade	**Description (only most important criteria listed)**
1	The tool is virtually impossible to install.
2	The tool is very hard to install **and/or** depends of many third-party products.
3	The tool is of medium difficulty to install **and/or** depends of some third-party products.
4	The tool is relatively easy to install **and/or** depends on very few third-party products.
5	The tool is extremely easy to install.

Aspect: Ease of use	
Definition: Was there a user manual or in-built help? Was it obvious what to do without a user manual? Was the graphical user interface self-explanatory? Was it necessary to give initial values to any parameters? When browsing for input files/saving output files, did the tool "remember" the latest used input/output directory? Did the tool itself suggest suitable file names for output? Did the tool work reasonably fast, with respect to the complexity of the type of task it performed?	
Grade	**Description**
1	The tool is virtually impossible to use.
2	The tool is very hard to use.
3	The tool is of medium difficulty to use.
4	The tool is relatively easy to use.
5	The tool is extremely easy to use.

Aspect: Generality of solution	
Definition: Was it possible to run the tool on several platforms, including the most common platforms? Were the file formats that the tool could use as input/output well-known and general formats? What languages could you choose for the graphical user interface? Were the "big" languages represented? Did you need a lot of less-well-known and/or obscure third-party software? Was it possible to do batch processing on large collections of files?	
Grade	**Description (only most important criteria listed)**
1	The tool is only relevant for the institution that developed it.
2	The tool may be relevant for a few institutions in a few countries **and/or** some obscure third-party tools are needed.
3	The tool can run on at least one common platform **and/or** some obscure third-party tools are needed.
4	The tool can be run on the most common platforms, is relevant in many countries, **and** none or few obscure third-party tools are needed.
5	The tool can be run on virtually any platform, is relevant in most countries, **and** no obscure third-party tools are needed.

Aspect: Quality of result (applicable when the tool does any kind of format conversion)	
Definition: Were the converted items of the same quality as the corresponding input items? Were converted images of reasonably good quality to "the naked eye"? Was it possible to convert huge files? For huge input files, could the converted items be reduced in size with preserved quality?	
Grade	**Description (only most important criteria listed)**
1	Almost no items could be converted **and/or** converted items were of very bad quality.
2	Most items could not be converted **and/or** converted items were of bad quality.
3	A reasonable amount of the items could be converted **and** converted items were of acceptable quality.
4	Most of the items could be converted **and** converted items were of good quality.
5	Almost all items could be converted **and** converted items were of very good quality.

Proofs of Concept
Executive report for Roadmap consideration

Scenario 1.6

Revision: 0.1 (draft)

Authors:

Andres Uueni (Estonian Ministry of Culture, Conservation Centre KANUT)

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

Revision History

Rev.	Date	Author	Organisation	Description
0.1	27 Aug 2013	A. Uueni	EVKM, CC Kanut, Estonia	First draft

TABLE OF CONTENTS

1 EXECUTIVE SUMMARY

This document describes the results of the tool that were tested in the context of Scenario 1.6. The ultimate goal of the tests was to grade the tool according to different aspects of usefulness. This document contains an overview of the results.

The tested tool was chosen as it is in daily use for a memory institution with high IT competence. The tool was suitable according to this criterion; in other cases it is highly unsuitable. However, if the tools were used and/or integrated in a cloud or grid environment, with much more tools, resources and IT competence at hand, the results and conclusions might be different.

The following tool was tested:

- IBM Tivoli Server Manager/Client

1.1 GRADING

The grades given in this document and the grades given in may differ slightly, since no detailed descriptions or definitions of the grades were made previously. The grades in this document are the more accurate ones.

For an explanation and the definitions of the various aspects see Annex 1. the aspect “Simplicity of management” is also mentioned, but since that aspect has been deemed to be “Not applicable” for all tools, it is omitted here.

The score for each aspect is in the range of 1 (very bad) to 5 (very good).

Simplicity of installation
Ease of use
Generality of solution
Quality of result

1.2 CONCLUSIONS AND RECOMMENDATIONS

The tool test is one sample out of hundreds of other possibilities. During the test was possible to point out following conclusions :

1. To set up the tool and to use it needs advanced IT expert knowledge's.
2. Tool is usable only on the command line and needs (previous) experience. There aren't any live support for this tool for the client. Tool's manual is available in man page and updated version URL: http://pu*blib.boulder.ibm.com/tividd/td/IBMTivoliDecisionSupportforOS3901.7.html*
3. Client side can be modified or upgrade with a great aware and needs previous experience.

4. To provide this service the archive host should be accessible and very responsible, any minor changes or modifications (especially network and server side) can cause problems for end-user client.
5. To log into the TSM server should be the connection speed is at least 60 Mb/s, preferably more. Slower connection causes lag and creates confusion when retrieving/restoring copy from archive
6. The list of archived data is accessible but not as usable as modern tools with a GUI.
7. The tool does not provide immediate response/feedback when errors occurred.
8. Using ext4 file system neither server or client tool does not support file hash function.
   The success of the archiving or restoring should be manually controlled using tool's log files.
9. There is no limits of the file size or format, although the tool does not have also the format recognitions features - there is a need for separate tools to use that.

2 SCENARIO OVERVIEW

Estonian memory institution (Conservation Centre Kanut) which digitises different content wants to make backup copy of files to another memory institutions tape library but needs proof that content is well preserved and it is possible to receive copy of files if needed. Therefore periodically (quarterly) will be carried out test data retrieving.

2.1 DOCUMENT STRUCTURE

Chapter 3 and the following chapters (“X”) are structured in the following way:

In the beginning of Chapter X, a short description is given of the tool and how it works.

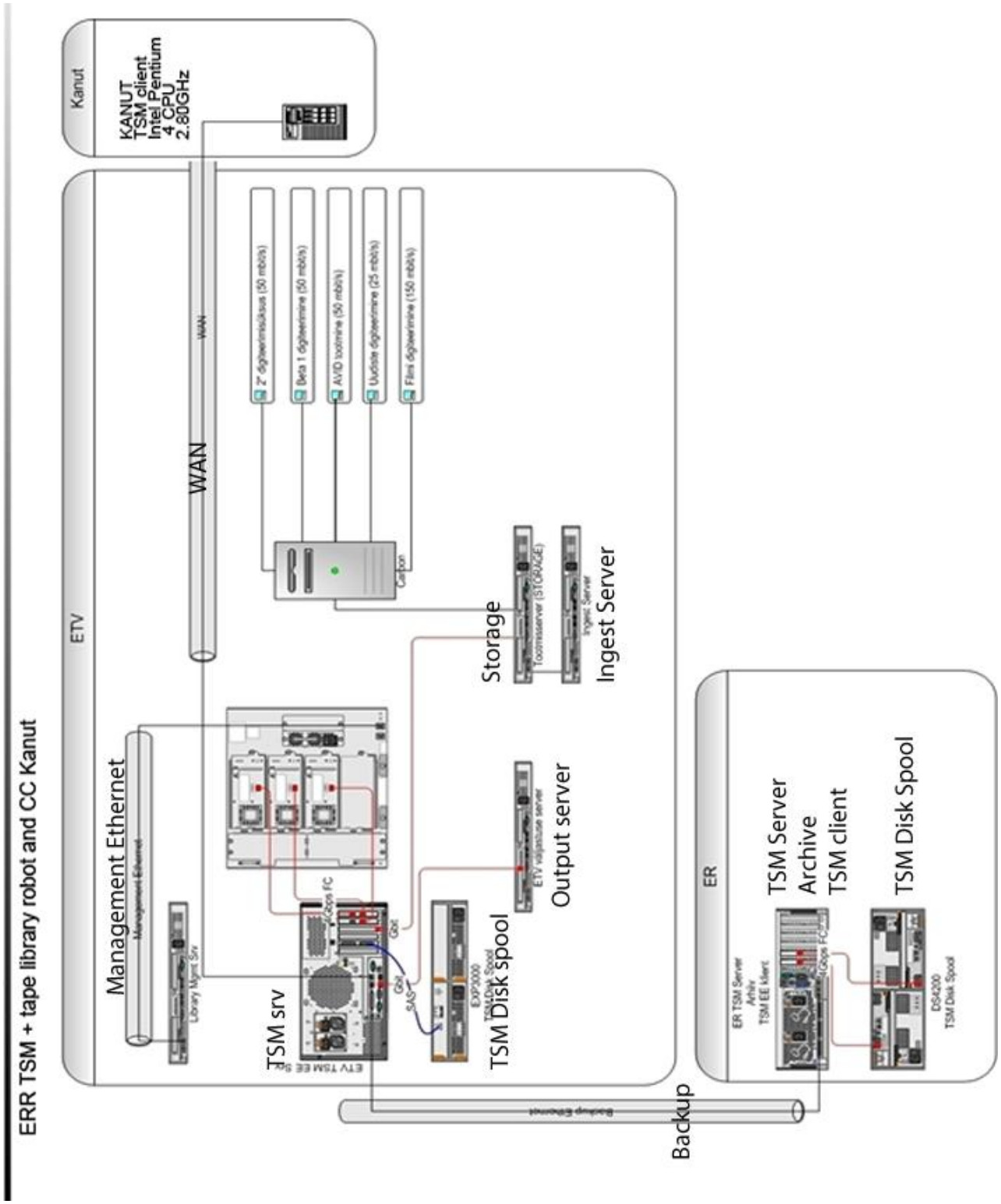
In sections X.1, the data set(s) that the tool will be tested on is described.

In sections X.2, the execution of the tests are described.

In sections X.3, the results of the tests (if any) are described. General comments are given about the tool and its usability for digital cultural heritage preservation, dissemination et c. (This section may be skipped if it was not possible to install and/or run the tool).

2.2 SCENARIO / TOOL TESTING ENVIRONMENT

The test environment was a server with Debian 5.0 processor Intel(R) 2,8 GHz, and 6 GB working memory (RAM).


ERR TSM + tape library robot and CC Kanut
Kanut
KANUT
TSM client
Intel Pentium
4 CPU
2.80GHz
ETV
WAN
Management Ethernet
TSM srv
TSM Disk spool
Storage
Ingest Server
Output server
ER
TSM Server
Archive
TSM client
TSM Disk Spool
Backup

3 IBM TIVOLI SERVER MANAGER/CLIENT

SERVER VERSION 5, RELEASE 5, LEVEL 2.0

IBM Tivoli Storage Manager is a client-server licensed product that provides storage management services in a multiplatform computer environment. The backup-archive client program permits users to back up and archive files from their workstations or file servers to storage, and restore and retrieve backup versions and archive copies of files to their local workstations.

3.1 DATA SETS

The data set that was used to test IBM TSM is conservation work description and digitised object images. Conservation work report is represented in PDF and DOC format, images are in JPEG and TIFF format. There are 107 separate files in this set.

3.2 TEST DESCRIPTION

The purpose of the test was to get a proof that data stored several years ago (4 years) to another memory institution's tape library are preserved, accessible and copy from these files can be created to owners institution. After creating a copy all the files were checked.

3.3 RESULTS

The tool behaved as expected and gave the correct results. However, there may be a problem with this program concerning usability and functionalities. Without previous experience with this tool it is rather difficult to set it up or to use it, especially there are some commands different than usual UNIX command line.

2009/06/26 was uploaded to the archive tape library (un-erasable LTO-4 Ultrium WORM) a data containing conservation work description and digitised object images (nr 07T019). The amount of the data: 736.24Mb, file formats: PDF, JPEG, TIF, DOC.

2013/08/27 data was searched and using a retrieving method were made a copy to the client server. All the files were complete and didn't had any errors.

When retrieving of the data (736.24Mb), it is relatively fast, in this test data transfer time: 67.75 sec and elapsed processing time: 00:02:47.

Connection details:

Using public broadband network of data communication between government institutions or simply EEBone. The connection is established through direct WAN VPN (ipsec) channel between IBM Tivoli client and server.

Network data transfer rate: 11,127.49 KB/sec

Aggregate data transfer rate: 4,511.00 KB/sec

Average client upload speed: 61.51 Mbit/sec

Average client download speed: 42.76 Mbit/sec

Equipment:

IBM Tivoli Server: Juniper SSG-550,

IBM Tivoli Client: Cisco router 1811

Server and Client OS: Debian 5.0

As it is a server - client tool, it is quite important to have trustful host-client relationship.
Although the tool is not very easy to use but based on this test this tool can is reliable, but needs some additional tools to control and see the whole archiving process.

Through this system is very difficult to make directly this data as open data and provide any public access.

Grades

Simplicity of installation:	2
Ease of use:	2
Generality of solution:	3
Quality of result:	5

ANNEX 1

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

Aspect: Simplicity of installation	
Definition: How complicated was it to download the tool? Did you have to register to get the download? Was it obvious which download version you should choose? If the download was packaged in a compressed file, how easy was it to unpack it? Were there any installation instructions, either on the download site or in the download itself? Was it necessary to install databases or other large third-party tools? In all, how many separate programs were necessary to install? How many mandatory parameter values had to be given during installation? If the first installation try failed, was it easy to install the tool anew?	
Grade	**Description (only most important criteria listed)**
1	The tool is virtually impossible to install.
2	The tool is very hard to install **and/or** depends of many third-party products.
3	The tool is of medium difficulty to install **and/or** depends of some third-party products.
4	The tool is relatively easy to install **and/or** depends on very few third-party products.
5	The tool is extremely easy to install.

Aspect: Ease of use	
Definition: Was there a user manual or in-built help? Was it obvious what to do without a user manual? Was the graphical user interface self-explanatory? Was it necessary to give initial values to any parameters? When browsing for input files/saving output files, did the tool “remember” the latest used input/output directory? Did the tool itself suggest suitable file names for output? Did the tool work reasonably fast, with respect to the complexity of the type of task it performed?	
Grade	**Description**
1	The tool is virtually impossible to use.
2	The tool is very hard to use.
3	The tool is of medium difficulty to use.
4	The tool is relatively easy to use.
5	The tool is extremely easy to use.

Aspect: Generality of solution	
Definition: Was it possible to run the tool on several platforms, including the most common platforms? Were the file formats that the tool could use as input/output well-known and general formats? What languages could you choose for the graphical user interface? Were the "big" languages represented? Did you need a lot of less-well-known and/or obscure third-party software? Was it possible to do batch processing on large collections of files?	
Grade	**Description (only most important criteria listed)**
1	The tool is only relevant for the institution that developed it.
2	The tool may be relevant for a few institutions in a few countries **and/or** some obscure third-party tools are needed.
3	The tool can run on at least one common platform **and/or** some obscure third-party tools are needed.
4	The tool can be run on the most common platforms, is relevant in many countries, **and** none or few obscure third-party tools are needed.
5	The tool can be run on virtually any platform, is relevant in most countries, **and** no obscure third-party tools are needed.

Aspect: Quality of result (applicable when the tool does any kind of format conversion)	
Definition: Were the converted items of the same quality as the corresponding input items? Were converted images of reasonably good quality to "the naked eye"? Was it possible to convert huge files? For huge input files, could the converted items be reduced in size with preserved quality?	
Grade	**Description (only most important criteria listed)**
1	Almost no items could be converted **and/or** converted items were of very bad quality.
2	Most items could not be converted **and/or** converted items were of bad quality.
3	A reasonable amount of the items could be converted **and** converted items were of acceptable quality.
4	Most of the items could be converted **and** converted items were of good quality.
5	Almost all items could be converted **and** converted items were of very good quality.

Proofs of Concept
Executive report for Roadmap consideration

Scenario 2.2

Revision: 0.1 (draft)

Authors:

Eva Toller (Riksarkivet)

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

Revision History

Rev.	Date	Author	Organisation	Description
0.1	2013-08-26	Eva Toller	Riksarkivet, Sweden	First draft

TABLE OF CONTENTS

1 EXECUTIVE SUMMARY

This document describes the results of the tools that were tested in the context of Scenario 2.2. The ultimate goal of the tests was to grade the tools according to different aspects of usefulness. This document contains an overview of the results. For details, see the DCH-RP WP5 document named **DCH-RP_WP5_Scen-2-2_ID-ToolTests.pdf** .

The tested tools were mainly chosen for their estimated suitability for a small cultural heritage institution with little or no IT competence. Sometimes a tool was suitable according to this criterion; in other cases it was highly *un*suitable. However, if the tools were used and/or integrated in a cloud or grid environment, with much more resources and IT competence at hand, the results and conclusions might be different.

The following tools were tested:

- AVS Document Converter 2.2
- AVS Image Converter 3.0
- Universal Document Converter
- A-PDF DjVu to PDF

1.1 GRADING

The aspects that have been graded are shown in the table below. The scores for the individual tools are given in Chapters 3 – 6. For an explanation and the definitions of the various aspects see Annex 1.

Aspect
Simplicity of installation
Ease of use
Generality of solution
Quality of result

The score for each aspect is in the range of 1 (very bad) to 5 (very good).

In **DCH-RP_WP5_Scen-2-2_ID-ToolTests.pdf**, the aspect “Simplicity of management” is also mentioned, but since that aspect has been deemed to be “Not applicable” for all tools, it is omitted here.

Since all the tools could not even be installed, sometimes the only grading that is done is “Simplicity of installation”.

The grades given in this document and the grades given in **DCH-RP_WP5_Scen-2-2_ID-ToolTests.pdf** may differ slightly, since no detailed descriptions or definitions of the grades were made previously. The grades in this document are the more accurate ones.

1.2 CONCLUSIONS AND RECOMMENDATIONS

The tool tests are really only a few samples out of hundreds of possibilities. However, the lessons learned from testing these specific tools were the following:

For local usage:

1. When there is a trial version (limited to 30 days of usage) and a commercial version that you maybe will buy later, the trial version must be *very easy* to install (so the end-user doesn't have to waste time on a tool that may not be appropriate).
2. There should preferably be some benchmarking for the performance of the tool, and a description of the limits it can work within (for example, that it can not handle files larger than ***N*** MegaBytes, and that batches larger than ***N*** files will be process very slow or may even fail).

For services hosted at an **e-infrastructure provider** (grid or cloud):

1. The provider must be responsible for the packaging of the service so that it is as simple as possible to use (especially if it requires installation on a client by the end-user).
2. If the end-user must install anything his/herself: the installation instructions *must* be up-to-date, and easy to follow.
3. There should be "live" support at the provider, ready to answer all types of questions about the services (also for tools provided by third party).

2 SCENARIO OVERVIEW

“A university lecturer in art history wants to use a collection of digitised art images made 15 years ago. They are stored in a format he is not familiar with. Since there are about 200 images, the researcher is looking for tools which would convert them into a format he could easily use in batch mode. He is not sure how to identify a tool or a service which could do this.”

For a detailed description of the data that was used for this scenario, see **DCH-RP_WP5_Scen-2-2_ID-66-restricted.pdf**.

2.1 DOCUMENT STRUCTURE

Chapter 3 and the following chapters (“X”) are structured in the following way:

In the beginning of Chapter X, a short description is given of the tool and how it works.

In sections X.1, the data set(s) that the tool will be tested on is described.

In sections X.2, the execution of the tests are described.

In sections X.3, the results of the tests (if any) are described. General comments are given about the tool and its usability for digital cultural heritage preservation, dissemination et c. (This section may be skipped if it was not possible to install and/or run the tool).

For a more detailed description of the tests, see **DCH-RP_WP5_Scen-2-2_ID-ToolTests.pdf** .

2.2 SCENARIO / TOOL TESTING ENVIRONMENT

The test environment was a PC (Personal Computer) with Windows 7 Professional, processor Intel(R) 2,7 GHz, and 8 GB working memory (RAM).

3 AVS DOCUMENT CONVERTER 2.2

AVS Document Converter converts files of ***source formats*** PDF, HTML, HTM, MHT, RTF, DOC, DOCX, ODT, PPT, PPTX, TXT, TIFF, TIF, EPUB, MOBI, FB2, DjVu, XPS into files of ***target formats*** PDF, HTML, MHT, RTF, DOC, DOCX, ODT, TXT, GIF, JPEG, PNG, TIFF, EPUB, MOBI, FB2.

3.1 DATA SETS

The data sets that were used to test the AVS Document Converter were seven TIFF images and 106 DjVu images. The sizes of the images were between 26 KB (the smallest DjVu image) and 397 MB (the largest TIFF image).

3.2 TEST DESCRIPTION

The purpose of the test was to make the following conversions, with acceptable output quality: TIFF to PDF, DjVu to PDF, DjVu to JPEG, and DjVu to PNG.

3.3 RESULTS

As a whole, the behaviour of the program was not acceptable, and the required results were only partly obtained. It was not even possible to load all seven TIFF files simultaneously, and it was altogether impossible to load the largest TIFF file.

Furthermore, this tool should only be used when you want to convert *small* amounts of files. It is not altogether reliable for batch conversion. Also check the result if you convert to files to PDF (the conversion sometimes results only in a white or grey page instead of an image). You may also want to check the quality of the images if you convert to JPEG or PNG (the quality varied between barely acceptable and good).

Grades

Simplicity of installation: 4

Ease of use: 3

Generality of solution: 4 – 5

Quality of result: 2 – 3 (previously 1 – 2)

4 AVS IMAGE CONVERTER 3.0

AVS Image Converter converts files of ***source formats*** BMP, GIF, ***JPEG***, JPG, JPE, JFIF, ***PNG***, APNG, ***TIFF***, TIF, PCX, TGA, RAS, PSD, CR2, CRW, RAF, DNG, MEF, NEF, ORF, ARW, EMF, WMF, JPEG 2000, SWF into files of ***target formats*** BMP, GIF, ***JPEG***, JPG, JPE, JFIF, ***PNG***, APNG, ***TIFF***, TIF, ***PDF***, TGA, RAS.

4.1 DATA SETS

The data sets that were used to test the AVS Document Converter were seven TIFF images and twelve JPEG images. The sizes of the images were between 122 KB (the smallest JPEG image) and 397 MB (the largest TIFF image).

4.2 TEST DESCRIPTION

The purpose of the test was to make the following conversions, with acceptable output quality: TIFF to JPEG, PNG, and PDF; JPEG to TIFF, PNG, and PDF.

4.3 RESULTS

As a whole, the behaviour of the program was not acceptable, and the required results were only partly obtained. It was not even possible to load all seven TIFF files simultaneously, and it was altogether impossible to load the largest TIFF file.

However, for conversion between JPEG and PNG, this tool seems adequate, maybe even for batch conversion (although to make sure of this, it would have to be tested on much larger amounts of files). On the other hand, it is not appropriate at all for conversions to PDF if you want individual documents as target while providing several files as source.

Grades

Simplicity of installation: 5

Ease of use: 3

Generality of solution: 4 – 5

Quality of result: 3

5 UNIVERSAL DOCUMENT CONVERTER

The Universal Document Converter (henceforth called **UDC**) is a ***printing service***. After installation, you can choose UDC as the current printer when you want to convert a file, and then choose the output format (TIFF, BMP, DCX, GIF, JPEG, PCX, PNG, or PDF).

5.1 DATA SETS

The data sets that were used to test the Universal Document Converter were seven TIFF images, three JPEG images, six DjVu images. The sizes were between 26 KB (the smallest DjVu image) and 397 MB (the largest TIFF image).

5.2 TEST DESCRIPTION

The purpose of the test was to make the following conversions, with acceptable output quality: TIFF to JPEG, PNG, and PDF; JPEG to TIFF, PNG, and PDF; DjVu to TIFF, JPEG, PNG, and PDF.

5.3 RESULTS

As a whole, the behaviour of the program was acceptable, and all required results were obtained (although with inferior quality in some cases, especially for large DjVu files). However, it is a drawback that the tool cannot do batch conversion (at least not if you use the graphical user interface directly). It may be used as a complement to convert “difficult” files that cannot be converted otherwise (for example, very large TIFF files).

Grades

Simplicity of installation: 4 – 5

Ease of use: 3 – 4

Generality of solution: 5

Quality of result: 4

6 A-PDF DJVU TO PDF

A-PDF DjVu to PDF allows you to batch convert DjVu into the PDF file format.

6.1 DATA SETS

Not applicable (no tests were run).

6.2 TEST DESCRIPTION

The purpose was to test the functionality of A-PDF DjVu to PDF. This could not be achieved since the program could not run in 64-bytes Windows (although it was said to work on Windows 7).

6.3 RESULTS

If this tool is to be recommended for small Cultural Heritage institutions, a version that works in 64-bytes Windows should be made.

Grades

Not applicable.

ANNEX 1

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

Aspect: Simplicity of installation	
Definition: How complicated was it to download the tool? Did you have to register to get the download? Was it obvious which download version you should choose? If the download was packaged in a compressed file, how easy was it to unpack it? Were there any installation instructions, either on the download site or in the download itself? Was it necessary to install databases or other large third-party tools? In all, how many separate programs were necessary to install? How many mandatory parameter values had to be given during installation? If the first installation try failed, was it easy to install the tool anew?	
Grade	**Description (only most important criteria listed)**
1	The tool is virtually impossible to install.
2	The tool is very hard to install **and/or** depends of many third-party products.
3	The tool is of medium difficulty to install **and/or** depends of some third-party products.
4	The tool is relatively easy to install **and/or** depends on very few third-party products.
5	The tool is extremely easy to install.

Aspect: Ease of use	
Definition: Was there a user manual or in-built help? Was it obvious what to do without a user manual? Was the graphical user interface self-explanatory? Was it necessary to give initial values to any parameters? When browsing for input files/saving output files, did the tool “remember” the latest used input/output directory? Did the tool itself suggest suitable file names for output? Did the tool work reasonably fast, with respect to the complexity of the type of task it performed?	
Grade	**Description**
1	The tool is virtually impossible to use.
2	The tool is very hard to use.
3	The tool is of medium difficulty to use.
4	The tool is relatively easy to use.
5	The tool is extremely easy to use.

Aspect: Generality of solution	
Definition: Was it possible to run the tool on several platforms, including the most common platforms? Were the file formats that the tool could use as input/output well-known and general formats? What languages could you choose for the graphical user interface? Were the "big" languages represented? Did you need a lot of less-well-known and/or obscure third-party software? Was it possible to do batch processing on large collections of files?	
Grade	**Description (only most important criteria listed)**
1	The tool is only relevant for the institution that developed it.
2	The tool may be relevant for a few institutions in a few countries **and/or** some obscure third-party tools are needed.
3	The tool can run on at least one common platform **and/or** some obscure third-party tools are needed.
4	The tool can be run on the most common platforms, is relevant in many countries, **and** none or few obscure third-party tools are needed.
5	The tool can be run on virtually any platform, is relevant in most countries, **and** no obscure third-party tools are needed.

Aspect: Quality of result (applicable when the tool does any kind of format conversion)	
Definition: Were the converted items of the same quality as the corresponding input items? Were converted images of reasonably good quality to "the naked eye"? Was it possible to convert huge files? For huge input files, could the converted items be reduced in size with preserved quality?	
Grade	**Description (only most important criteria listed)**
1	Almost no items could be converted **and/or** converted items were of very bad quality.
2	Most items could not be converted **and/or** converted items were of bad quality.
3	A reasonable amount of the items could be converted **and** converted items were of acceptable quality.
4	Most of the items could be converted **and** converted items were of good quality.
5	Almost all items could be converted **and** converted items were of very good quality.

Proofs of Concept
Executive report for Roadmap consideration

Scenario 2.4

Revision: 0.1 (draft)

Authors:

Eva Toller (Riksarkivet)

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

Revision History

Rev.	Date	Author	Organisation	Description
0.1	2013-08-26	Eva Toller	Riksarkivet, Sweden	First draft

TABLE OF CONTENTS

1 EXECUTIVE SUMMARY

This document describes the results of the tools that were tested in the context of Scenario 2.4. The ultimate goal of the tests was to grade the tools according to different aspects of usefulness. This document contains an overview of the results. For details, see the DCH-RP WP5 document named **DCH-RP_WP5_Scen-2-4_ID-ToolTests.pdf**.

The tested tools were mainly chosen for their estimated suitability for a small cultural heritage institution with little or no IT competence. Sometimes a tool was suitable according to this criterion; in other cases it was highly *un*suitable. However, if the tools were used and/or integrated in a cloud or grid environment, with much more resources and IT competence at hand, the results and conclusions might be different.

The following tools were tested:

- HTTrack 3.47-21
- SWAT (Snappy Web Archiving Tool, version 1.0)
- WARC Tools
- Web Curator Tool (WCT)
- Heritrix

1.1 GRADING

The aspects that have been graded are shown in the table below. The scores for the individual tools are given in Chapters 3 – 7. For an explanation and the definitions of the various aspects see Annex 1.

Aspect
Simplicity of installation
Ease of use
Generality of solution
Quality of result

The score for each aspect is in the range of 1 (very bad) to 5 (very good).

In **DCH-RP_WP5_Scen-2-4_ID-ToolTests.pdf**, the aspect “Simplicity of management” is also mentioned, but since that aspect has been deemed to be “Not applicable” for all tools, it is omitted here.

Since all the tools could not even be installed, sometimes the only grading that is done is “Simplicity of installation”.

The grades given in this document and the grades given in **DCH-RP_WP5_Scen-2-4_ID-ToolTests.pdf** may differ slightly, since no detailed descriptions or definitions of the grades were made previously. The grades in this document are the more accurate ones.

1.2 CONCLUSIONS AND RECOMMENDATIONS

The tool tests are really only a few samples out of hundreds of possibilities. However, the lessons learned from testing these specific tools were the following:

For local usage:

1. The different software parts necessary for a tool should be *packaged together* as much as possible, so that the end-user will not be required to separately download and install a lot of different tools him/herself.
2. The installation files should be few and very simple to manage; preferably *.msi files* (for Windows), the next best is *.exe files*.
3. The installation instructions should be easy to understand, neither too short nor too long, and be up-to-date.

For services hosted at an **e-infrastructure provider** (grid or cloud):

1. The provider must be responsible for the packaging of the service so that it is as simple as possible to use (especially if it requires installation on a client by the end-user).
2. If the end-user must install anything his/herself: the installation instructions *must* be up-to-date, and easy to follow.
3. There should be "live" support at the provider, ready to answer all types of questions about the services (also for tools provided by third party).

2 SCENARIO OVERVIEW

“A history student interested in natural history discovers that Riksarkivet has archived the "Linnéjubilet" web site http://www.riksarkivet.se/default.aspx?id=23153 .He wonders how he can get access to it (the link www.linne2007.se obviously doesn't work anymore).“

For a detailed description of the data that was used for this scenario, see **DCH-RP_WP5_Scen-2-4_ID-68.pdf** .

2.1 DOCUMENT STRUCTURE

Chapter 3 and the following chapters (“X”) are structured in the following way:

In the beginning of Chapter X, a short description is given of the tool and how it works.

In sections X.1, the data set(s) that the tool will be tested on is described.

In sections X.2, the execution of the tests are described.

In sections X.3, the results of the tests (if any) are described. General comments are given about the tool and its usability for digital cultural heritage preservation, dissemination et c. (This section may be skipped if it was not possible to install and/or run the tool).

For a more detailed description of the tests, see **DCH-RP_WP5_Scen-2-4_ID-ToolTests.pdf** .

2.2 SCENARIO / TOOL TESTING ENVIRONMENT

The test environment was a PC (Personal Computer) with Windows 7 Professional, processor Intel(R) 2,7 GHz, and 8 GB working memory (RAM).

3 HTTRACK 3.47-21

HTTrack is a browser utility. It allows you to download a web site from the Internet to a local directory (recursively building all directories.

3.1 DATA SETS

The web site “Linnéjubiléet” (Linné Anniversary) was constructed for the 300th anniversary of the birth of Carl von Linné. It contains information about specific anniversary activities, about Linné and his life, about gardens and exhibitions, and much more. The site consists of 4058 files altogether, and the total size is 469 MegaByte. For details, see: **DCH-RP_WP5_Scen-2-4_ID-68.pdf**

3.2 TEST DESCRIPTION

The purpose of the test was to:

1) download the web site “Linnéjubiléet” as completely and correctly as possible;
2) let a test person look at the downloaded site and compare it with the original one (acceptance test).

3.3 RESULTS

The quality of the downloaded web site seemed, in the first examination, to be as good as the original web site. However, during the acceptance test, it was detected that the English version of the site had not been downloaded. The original Linné site was not reachable during the period of the acceptance test, so it was not possible to compare the downloaded site to the original site, but he test person thought that it was very believable that this could be the real (original) site. For details of the acceptance test, see: **DCH-RP_WP5_Scen-2-4_ID-72.pdf** .

Grades (for the downloading part only)

Simplicity of installation: 5

Ease of use: 4 – 5

Generality of solution: 5

Quality of result: 4 (previously 5)

4 SWAT (SNAPPY WEB ARCHIVING TOOL, VERSION 1.0)

SWAT is a tool designed for archiving web sites and displaying the archive in a simple way. Besides harvesting all files from the web site, SWAT generates snapshots of each page to TIFF files and describes the entire archive in a METS-file.

4.1 DATA SETS

Not applicable (no tests were run).

4.2 TEST DESCRIPTION

The purpose was to test the functionality of SWAT. This could not be achieved since it seems clear from the number of third-party tools that this is probably a too complicated tool for a small cultural heritage institution (especially if they only have a single web site to download and preserve).

4.3 RESULTS

If this tool is to be recommended for small Cultural Heritage institutions, the many software parts should be *packaged* in a way that makes it easy for a novice user to make the installation. (The developer also suggested that there may now exist newer tools that do approximately the same thing as SWAT, but in a simpler way).

Grades

Simplicity of installation: 2

5 WARC TOOLS

The main goal of WARC Tools is to facilitate the adoption of the WARC file format for storing web archives by providing an open source software library, a set of command line tools, web server plug-ins and technical documentation for manipulation and management of WARC files.

5.1 DATA SETS

Not applicable (no tests were run).

5.2 TEST DESCRIPTION

The purpose was to test the functionality of WARC Tools. This could not be achieved since it seems extremely hard to install this tool even if you follow the instructions.

5.3 RESULTS

If this tool is to be recommended for small Cultural Heritage institutions, the installation instructions should be re-written and tested.

Grades

Simplicity of installation: 1

6 WEB CURATOR TOOL (WCT)

The Web Curator Tool (WCT) is an open-source workflow management application for selective web archiving. It is designed for use in libraries and other collecting organisations.

6.1 DATA SETS

Not applicable (no tests were run).

6.2 TEST DESCRIPTION

The purpose was to test the functionality of Web Curator Tool. This could not be achieved since there were no up-to-date installation instructions.

6.3 RESULTS

If this tool is to be recommended for small Cultural Heritage institutions, the installation instructions should be re-written and tested. The many software parts should also be *packaged* in a way that makes it easy for a novice user to make the installation.

Grades

Simplicity of installation: 1

7 HERITRIX

Heritrix is the Internet Archive's open-source web crawler project.

7.1 DATA SETS

Not applicable (no tests were run).

7.2 TEST DESCRIPTION

The purpose was to test the functionality of Web Curator Tool. This could not be achieved since there were installation instructions that sufficient for a novice user.

7.3 RESULTS

If this tool is to be recommended for small Cultural Heritage institutions, there should be more extensive installation instructions. It would also be an advantage if the tool cold be run on more platforms than Linux.

Grades

Simplicity of installation: 1 – 2

ANNEX 1

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

Aspect: Simplicity of installation	
Definition: How complicated was it to download the tool? Did you have to register to get the download? Was it obvious which download version you should choose? If the download was packaged in a compressed file, how easy was it to unpack it? Were there any installation instructions, either on the download site or in the download itself? Was it necessary to install databases or other large third-party tools? In all, how many separate programs were necessary to install? How many mandatory parameter values had to be given during installation? If the first installation try failed, was it easy to install the tool anew?	
Grade	**Description (only most important criteria listed)**
1	The tool is virtually impossible to install.
2	The tool is very hard to install **and/or** depends of many third-party products.
3	The tool is of medium difficulty to install **and/or** depends of some third-party products.
4	The tool is relatively easy to install **and/or** depends on very few third-party products.
5	The tool is extremely easy to install.

Aspect: Ease of use	
Definition: Was there a user manual or in-built help? Was it obvious what to do without a user manual? Was the graphical user interface self-explanatory? Was it necessary to give initial values to any parameters? When browsing for input files/saving output files, did the tool “remember” the latest used input/output directory? Did the tool itself suggest suitable file names for output? Did the tool work reasonably fast, with respect to the complexity of the type of task it performed?	
Grade	**Description**
1	The tool is virtually impossible to use.
2	The tool is very hard to use.
3	The tool is of medium difficulty to use.
4	The tool is relatively easy to use.
5	The tool is extremely easy to use.

Aspect: Generality of solution	
Definition: Was it possible to run the tool on several platforms, including the most common platforms? Were the file formats that the tool could use as input/output well-known and general formats? What languages could you choose for the graphical user interface? Were the "big" languages represented? Did you need a lot of less-well-known and/or obscure third-party software? Was it possible to do batch processing on large collections of files?	
Grade	**Description (only most important criteria listed)**
1	The tool is only relevant for the institution that developed it.
2	The tool may be relevant for a few institutions in a few countries **and/or** some obscure third-party tools are needed.
3	The tool can run on at least one common platform **and/or** some obscure third-party tools are needed.
4	The tool can be run on the most common platforms, is relevant in many countries, **and** none or few obscure third-party tools are needed.
5	The tool can be run on virtually any platform, is relevant in most countries, **and** no obscure third-party tools are needed.

Aspect: Quality of result (applicable when the tool does any kind of format conversion)	
Definition: Were the converted items of the same quality as the corresponding input items? Were converted images of reasonably good quality to "the naked eye"? Was it possible to convert huge files? For huge input files, could the converted items be reduced in size with preserved quality?	
Grade	**Description (only most important criteria listed)**
1	Almost no items could be converted **and/or** converted items were of very bad quality.
2	Most items could not be converted **and/or** converted items were of bad quality.
3	A reasonable amount of the items could be converted **and** converted items were of acceptable quality.
4	Most of the items could be converted **and** converted items were of good quality.
5	Almost all items could be converted **and** converted items were of very good quality.