DELIVERABLE

Project Acronym:	**DCH-RP**
Grant Agreement number:	**312274**
Project Title:	**Digital Cultural Heritage Roadmap for Preservation – Open Science Infrastructure for DCH in 2020**

D3.3 – Registry of services

Revision: Final

Authors:

Maciej Brzeźniak (PSNC)
Michał Jankowski (PSNC)
Norbert Meyer (PSNC)
Borje Justrel (RA)
Tim Devenport (EDItEUR)

Reviewers:

Rosette Vandenbroucke (BELSPO)
Raivo Russalepp (EVKM)
Claudio Prandoni (PROMOTER)

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	9-9-2013	Maciej Brzeźniak	PSNC	Draft for comments and discussion during DCH-RP meeting
Draft	1-10-2013	Maciej Brzeźniak	PSNC	2nd draft including meeting discussion conclusions and findings
Draft	7-10-2013	Maciej Brzeźniak	PSNC	Final version including reviewers' comments
Final	8-10-2013	Claudio 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This deliverable contains original unpublished work except where clearly indicated otherwise.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

In this document we present the structure and initial content of a Registry of Services for digital preservation purposes. The registry collects and describes information and knowledge related to tools, technologies and systems that can be applied for the purposes of digital cultural heritage preservation. It also reviews existing and emerging services developed and offered by R&D projects, public organisations and commercial solution vendors.

Whilst providing a broad overview of the existing solutions, the registry initiative focuses on analysing those services and tools that can enable cultural heritage institutions to benefit from the capacities of e-Infrastructures including cloud and grid systems.

Tools and services are categorized by purpose, technologies required, resource formats supported and domain-specific application, among many other criteria. Alongside this functional description, an attempt has been made (for a subset of the tools and services covered) to provide assessments of each. In the first iteration, assessment criteria chosen have been: popularity, support level, portability, scalability, licensing model, and modularity/openness of architecture.

The registry concept, and in particular the assessment mechanism, has been “road tested” in conjunction with DCH-RP Proofs of Concept, drawn from WP5 of the project’s work.

Finally, proposals have been developed for future improvements and for a series of actions to encourage the future sustainability and usefulness of the registry beyond the timescale of the project.

The document summarizes the work of Task 3.3 of the DCH-RP project.

2 INTRODUCTION

The aims of the DCH-RP project are to explore how best to harmonise data preservation policies - both across Europe and in the digital cultural heritage (DCH) sector - and to promote dialog and collaboration among DCH institutions, e-Infrastructure providers and research and private organisations. An important part of this process is to review and analyse existing services and tools and to understand how they can most effectively be applied for the purposes of DCH preservation.

Special attention is paid to those services and tools that can be deployed in cloud or grid arrangements, since these have the potential to enable cultural heritage (CH) institutions to exploit both existing and future data processing, storage, preservation, presentation and delivery capabilities of e-Infrastructures.

Within Task 3.3, DCH-RP project participants carried out an analysis of existing and emerging tools and services, and used this to develop a tools and services registry. The registry provides basic information on the functionality, possible usage and applications of these tools and services as well as additional information related to their maintenance status, licensing and technologies used in them. For selected tools, DCH-RP participants also provided some quality metrics based on practical evaluation performed within WP5's proofs of concept. We also developed a mechanism for updating and extending the registry as well as assessing the tools and services it contains.

This deliverable document summarizes this discussion and work performed in the Task 3.3. aimed at developing the DCH-RP registry of data preservation tools and service.

2.1 STRUCTURE OF THE DOCUMENT

In the following paragraphs of this section we present the context and aims of this work and discuss the main assumptions employed during the development of the DCH-RP registry. We also explain the focus of the registry, its target audience and position the T3.3. work versus remaining activities in the project. Section 3 presents our methodology in detail while Section 4 describes the structure of the registry and its mechanisms. In section 5 we present the registry sustainability plan. Section 6 contains final discussion and conclusions.

2.2 CONTEXT

Within the Task 3.1 Preservation Services Architecture summarized by Deliverable D3.1 "Study on a Roadmap for Preservation", DCH-RP analysed key characteristics and requirements of digital preservation (DP) in DCH institutions and investigated how they could be linked with e-Infrastructures. This work provided an overview of the available tools and services suitable for running in e-Infrastructures as well as describing the existing and emerging e-Infrastructures that may 'host' these tools and services. The same work also identified several gaps among the existing services and tools for digital preservation and e-Infrastructures.

An important observation is that although there are hundreds of software tools and services that can be used to support automation and preservation tasks, the support status, interoperability, quality, reliability and level of documentation of these tools and services is not at all clear. And there appears to be no central, reliable knowledge base that could act as a source of information for decision makers engaged in choosing technologies, systems, services and tools.

The aim of the DCH-RP tools & services registry is to address these gaps by providing an authoritative list of solutions that can be used for DCH preservation together with information related to their support

status, portability, and interoperability (support for standards), as well as assessments of their quality, reliability and the possibility of running each of them in e-Infrastructures.

While a number of tools and services registries already exist, they rarely consider tools and services run in an outsourced model. This is probably because most organisation that perform digitization or produce digitally-born assets still run digital preservation processes in their own premises within a user-controlled environment. At the same time, however, there is a need to store and preserve a growing number of digital objects and increasing data volumes while achieving high reliability and advanced functionality and keeping costs low. These factors naturally direct the attention of DCH institutions towards e-Infrastructures.

DCH-RP's Task 3.1 observed that this new situation brings both opportunities and challenges. Outsourcing DP processes to external organisations requires a clear decision to rely on other parties, establishing a trust relationship as well as formal and legal guarantees and agreements. Moreover, usage of externally provided resources may require changing the technical and organisational aspects of the DP process within the DCH institutions themselves.

The aim of the DCH-RP tools & services registry is to help decision makers in reorganising the DCH preservation process in the context of the new situation. It provides not only a list of tools and services but also comments on their usability in remote outsourced models, portability, compliance to standards and scalability. It also contains the solutions assessment and rating mechanisms that can be used by project participants and registry users in order to collect real-life experience related to tools and services deployment and usage.

Problems that may arise while using externally provided resources can be illustrated using the following example. A museum employee performs certain process using licensed software on a dedicated workstation. Usage of remote resources provided by e-Infrastructure may improve the performance of the process (e.g. by using the computing power of the cluster for performing file format transformations for large collections). However the user must accept the increased latency of the tool's graphical interface and the fact that someone else provides and manages his working environment, including control (physical access) over the user's data.

More complicated situations can occur if part of the process must organised using another approach or technology (software or hardware) in order to make it possible to use external resources. Yet another problem may arise when attempting to run software from vendors that do not allow their software to be run in a multi-tenancy cloud environment. In such cases migration of the DP process to an outsourced model might be less straightforward.

Overall, there are always opportunities and risks related to using resources provided by external parties. The CH organisation itself must carry out a risk assessment to evaluate the pros and cons of running services in e-Infrastructures. Also the e-Infrastructure providers must understand the requirements and specifics of the DP process and technical and organisational aspects of domain-specific tools and services that may be required.

As the situation in which CH institutions may use resources and services provided by e-Infrastructures is relatively new, there is not yet suitable methodology for performing the risk assessment. Such a view is shared by the authors of Deliverable D3.1 of the DCH-RP project:

> *"There continues to be inadequate support for decision-making, selecting, testing and benchmarking tools for preservation (i.e., the process from analysing local needs, picking the best combination of available tools to implementing a robust solution)."*

The DCH-RP tools & services registry provides the information to support DCH institutions and e-Infrastructure providers in assessing the risks and making educated decisions related to reorganising DCH preservation processes.

Reliable risk assessment require clear, well-defined assessment criteria (metrics). Without these metrics it is difficult to make educated decisions on the selection or adoption of particular tools and services for DP processes. Existing metrics are rarely useful for evaluating potential gains and risks related to running the tools and services on externally-provided resources, such as e-Infrastructures. This slows down dialog and hinders the development of collaboration between DC institutions and e-Infrastructure providers.

The DCH-RP tools & services registry structure and mechanisms support providing assessments and evaluations of tools and services related to quality, scalability, reliability, portability, and interoperability in accordance with agreed standards. This information will support the dialog between CH institutions and e-Infrastructure providers related to migrating DP processes from in-house infrastructure towards outsourced resources and services.

2.3 OUR AIMS IN DEVELOPING THE REGISTRY

As already articulated in the previous section, our aims in developing the registry include providing a list of existing tools and services as well as assessments of each, according to some defined metrics related to the usability of these solutions for digital preservation of DCH assets and their potential for exploiting the capabilities and features of e-Infrastructures.

The DCH-RP registry of tools and services will help CH institutions in selecting quality, mature, sustainable (maintained) and portable tools and solutions that may be used for addressing particular, isolated DP-related needs or may be combined to support partial or full DP workflows.

The registry will also help in making decisions related to the deployment of DP processes on top of e-Infrastructures, in order to offer reliable and trusted tools and services in outsourced models.

2.4. Focus

The DCH-RP tools and services registry provides a list of many well-known tools and services, enriched with information on their selected feature. They include characteristics interesting from the point of view of the general usability of these solutions, including basic functionality, generality of solution, ease of use and quality of their products and results. Projected registry structure will also consider the DP solutions aspects related to their potential of being deployed in an outsourced model using cloud and grid environments, e.g the technology used for tools and services development, portability, compliance to standards and licensing limitations.

As there are hundreds of potentially relevant tools and services, assessing and examining all of them was impossible within the project. Therefore we decided to provide quality-related assessments for selected solutions that we used with Proofs of Concept (PoCs) conducted in DCH-RP's Work Package 5. We also developed quality assessment functionality for the registry. It can be used by users for sharing their experience on deployment and usage of tools and services, thus supporting the process of building the knowledge base related to DP solutions.

2.5. Target audience of the registry

The target audience of the registry includes digital preservation professionals who are trying to find reliable, high-quality services and tools for implementing DP processes and to understand if and how these solutions can be run outsourced using cloud and grid resources of e-Infrastructures. The registry will also be useful for institutions and initiatives providing e-Infrastructure or planning the provision of DP

services for current and prospective infrastructure users. The registry is also the input for the development of DCH-RP's data preservation roadmap.

2.6. Scope and boundaries of the work

Taken together, this deliverable document and the registry itself provide a snapshot of the current situation and an overview of emerging solutions. However, drawing a comprehensive picture of future trends and planning the actions necessary to bring DCH institutions and e-Infrastructures together for the benefit of DCH preservation is out of the scope of this work.

The registry covers two important areas related to services and tools for digital preservation. It provides information on existing tools and services that are widely known and used for implementing DP processes. In this context, the unique feature of our registry is that it also includes the results of assessments of selected solutions performed in collaboration with CH institutions themselves, which we believe should make it particularly useful for the CH sector.

But in addition to existing tools and services, the registry includes emerging tools and services being developed and established within various international and national projects and initiatives. This part of the registry is important for DCH-RP roadmap development, which naturally aims to be forward looking. On the one hand it makes visible to DC institutions initiatives aiming at providing solutions for data preservation undertaken using e-Infrastructures. On the other hand it may also be used by e-Infrastructure owners and operators in order to understand what services and tools they should offer in order to meet the sector-specific requirements of DCH institutions.

3 METHODOLOGY

Since a number of services and tools registries already exist, we devised a methodology for the DCH-RP registry to ensure that it both contributes to the state of the art and is directly useful for project purposes. Among other aspects this covers the sources of information related to the services and tools and the way in which we process this information. Other important parts of the methodology include selection criteria for the services and tools included in the registry and definition of measures and evaluation methods applied to selected solutions.

In this section we summarize the discussion we conducted within the project and decisions we have made related to the methodology of registry building and maintenance.

3.1 SOURCES OF INFORMATION

Obvious sources of the basic information about available DP solutions are existing (and frequently more general) registries of services and tools. A list of the registries taken into account is presented in Table 1.

Registry name	**Organisation responsible**	**Link**
Library of Congress Tools Showcase	Library of Congress, USA	http://www.digitalpreservation.gov/tools/
AQuA Mashup Tool List	AQuA project	http://wiki.opf-labs.org/display/AQuA/AQuA+Mashup+Tool+List
List of preservation tools registries	British Library, UK	http://wiki.opf-labs.org/display/SPR/Digital+Preservation+Tools
The Technical Registry PRONOM	National Archives, UK	http://www.nationalarchives.gov.uk/PRONOM/Default.aspx
Digital Preservation Tool Registry	Open Planets Foundation	http://wiki.opf-labs.org/display/TR/Home
DC-NET project list of	List of Preservation Tools	http://www.dc-net.org/getFile.php?id=467
Tools & Services registry	Digital Curation Centre	http://www.dcc.ac.uk/resources/external/tools-services
Evidence-based Digital Preservation Tools Repository	APARSEN	http://www.alliancepermanentaccess.org/index.php/knowledge-base/tools/

Table 1. Existing tools and services registries used as initial sources of information for the DCH-RP registry

These information sources were only the starting point for our work and were used in the first stage of building the DCH-RP registry. Our aim was to provide really new value in this area as well as supporting

the development of the DCH-RP roadmap. Therefore we included in the registry additional tools and services not covered in other registries, including solutions developed by DCH-RP project participants as well as new, emerging tools and services worked from a variety of R&D projects. By design, the added values of our registry include quality assessments as well as information useful particularly in the DCH sector, whilst for selected tools and services we included feedback related to particular solutions acquired during the proofs of concept exercises in DCH-RP's Work Package 5. In order to include ongoing work in the picture, we reviewed the documentation and publications available for the projects and initiatives shown in Table 2.

Project / initiative name	**Material type**	**Responsible / contact organisation**	**Link**
CAIRO project, UK	Tools survey	JISC	http://cairo.paradigm.ac.uk/projectdocs/cairo_tools_listing_pv1.pdf
NDIIPP programme, USA,	NDIIPP Partner Tools and Services Inventory	Library of Congress	http://www.digitalpreservation.gov/tools/index.html
DC-NET project	Project report	DC-NET consortium	http://www.dc-net.org/getFile.php?id=467
INDICATE project	Report D5.1	INDICATE consortium	http://www.indicate-project.eu/gestFile.php?id=339
EUDAT project	Documentation	EUDAT consortium	www.eudat.eu
SCAPE project	Deliverable document:	SCAPE consortium	http://www.scape-project.eu/wp-content/uploads/2011/09/SCAPE_D10.1_KEEPS_V1.0.pdf

Table 2. Projects and initiatives reviewed as sources of information for the DCH-RP registry

In addition to a review of documentation, we conducted a dialogue with our participants in the projects and initiatives concerned, in order to learn in more depth about the work conducted and the current and projected tools and services involved.

We have also developed a mechanism that enables the wider public (including digital preservation professionals) to contribute to the on-line version of the registry. This mechanism will make the registry a living list of tools and services. It will enable authors from within and outside the DCH-RP project to contribute to the registry by feeding, updating and commenting on its contents.

3.2 SERVICES & TOOLS SELECTION CRITERIA

As there are hundreds and tools and services that may be considered as useful for DP processes, **we defined some basic selection criteria**. By applying these criteria, we selected the services and tools to be included in the first, pilot version of the registry.

In particular we tried - where possible, based on the available documentation - to evaluate several **features of the tools and services such as popularity, support level, portability between environments, licensing models, scalability, openness and modularity of the architecture, as well as the overall quality and maturity of the solution.**

By applying the above mentioned selection criteria to both existing and emerging services and tools we created a list of solutions that have the potential to leverage the resources and capabilities of e-Infrastructures for the benefit of CH institutions. More detail on the selection and evaluation criteria is provided below.

3.2.1 Initial list of the selection and assessment criteria

One of the basic assumptions adopted is that selection criteria for the tools and services registry should be common with general, widely used software evaluation criteria. At the same time metrics defined and assessments provided using these metrics should support DCH institutions in the process of choosing solutions appropriate for their DP processes and the capacities of e-Infrastructures and organising the process of migrating these processes to e-Infrastructures for increased scalability, reliability, functionality and cost efficiency.

The initial list of criteria to be considered in the DCH-RP registry was created during discussions conducted within the DCH-RP project consortium. It is based on the experience, knowledge and intuition of the project participants. As they represent DCH institutions and coordination bodies as well as e-Infrastructure operators, managers and funders, these criteria and requirements are felt to be generally representative of those in the domain as a whole. Importantly, DCH institutions represented in the project performed a practical verification of the importance and practical usage of these criteria, within the WP5's proofs of concept.

The list is presented in Table 3 below. The criteria are grouped into three main categories: quality, usability and scalability. The division of the criteria into these categories in not strict, as particular aspects of these categories overlap - e.g. quality of results and applicability or usefulness of the solution for the purpose. Against each criterion, the table provides comments and signals potential issues related to the usage of each in evaluating tools and services for DP.

Overall, while these criteria are intuitive and easy to understand, their application to all interesting tools and services is not straightforward for the following reasons. Firstly, research based on documentation may fail to provide enough information, while practical evaluation of all tools and services included in the registry (more than 140 in total) lies behind the practical scope or capabilities of the DCH-RP project itself. Secondly, even if the consortium were able to conduct practical evaluation of all interesting tools and services, the objectivity of its results and the general applicability of its assessments could still be questioned. Thirdly, some aspects of the criteria do not require a very formal assessment process. Thanks to this simplicity it is possible for regular users to provide useful assessments. At the same time, however, the assessments made using these criteria are not very formal, therefore their 'portability' to other scenarios and usability for a wider public is not obvious. Fourthly, assessing the solutions in the context of DCH applications requires domain-specific knowledge and end-user type experience with the

evaluated systems. Within the constraints of the project we were able to make carry out such assessments on a limited scale. Therefore we designed a mechanism for providing evaluation and feedback that can be used by registry end-users.

Criteria / group	Comments	Issues
1. Quality		
a. usefulness for the purpose	• user point of view • requires domain knowledge	• difficult to evaluate doc-based • main purpose / application of the tool must be taken into account while making evaluation / assessments
b. quality of the results	• used when applicable, e.g. for format conversion tools	• difficult to evaluate paper-based • requires practical evaluation
c. generality of solution	• means usability for various use-cases • and/or means that solutions spans/ covers a wide part of the DP process	• generality vs specialisation • not always an advantage
2. Usability		
a. usefulness for the purpose	• user point of view	• difficult to evaluate doc-based
b. affordability for small / big institution	• technical and economical affordability • technical affordability depends on staff skills; does not simply map to institution size	• economical affordability hard to assess; lots of aspects impact total cost: purchase, maintenance, licenses, staff training, etc.
c. ease of use / simplicity; level documentation	• user point of view	• subjective, may be perceived differently by particular users
3. Scalability		
a. scalability, modularity of the architecture	• may be evaluated based on documentation	• documentation of the architecture/ internals not available for some solutions
b. technical sophistication	• possible to evaluate doc-based to some extent (see above)	• hard to evaluate by end-users • requires technical knowledge
c. scale of the service possible to achieve	• number of objects, volumes of data possible to handle with the tool • institution-, regional- national-scale	• reliable assessment may require practical deployment in large scale

Table 3. Initial list of the tools and services selection and assessment criteria

We conclude our discussion with the statement that, while applying the criteria listed above does not require a very formal evaluation process and the generality of the assessments made using them may be put into question, they are nonetheless practical and informative enough to be useful for DCH institutions and e-Infrastructure providers. While there are alternative ways for evaluating the quality of services and tools, we found them to be unaffordable within the timeframe of the project and the resources available. Some known alternatives are discussed in the following section.

3.2.2. Alternative approaches to services and tools assessment

Interesting guidelines for evaluating software quality are provided by the ISO 25010:2011 standard[1]. This approach goes beyond functional suitability and usability, and provides more definitions of characteristics related to software quality. It also extends the intuitive understanding of the terms and features listed in the previous section. Table 4 below presents selected non-functional aspects of software products as defined in ISO 25010:2001.

Features group:	**Example features / characteristics:**
Performance efficiency	• response, processing times and throughput rates of a system; • the amounts and types of resources used by a system, when performing its functions; • the maximum limits of a product or system parameter.
Compatibility	• product can perform its functions efficiently while sharing environment and resources with other products; • system can exchange information with other systems and use the information exchanged
Reliability	• system is operational and accessible when required for use; • system meets needs for reliability under normal operation; • system operates as intended despite the presence of hardware or software faults; • system can recover data affected and re-establish the desired state of the system in case of an interruption or a failure.
Security	• system ensures that data are accessible only to those authorized to have access; • system prevents unauthorized access to, or modification of, computer programs or data; • actions or events can be proven to have taken place, so that the events or actions cannot be repudiated later; • actions of an entity can be traced uniquely to the entity; • the identity of a subject or resource can be proved to be the one claimed.
Maintainability	• system is composed of components such that a change to one component has minimal impact on other components; • an asset can be used in more than one system, or in building other assets; • it is possible to assess the impact of an intended change (analysability); • system can be effectively and efficiently modified without introducing defects or degrading existing product quality; • test criteria can be established for a system.
Portability	• system can effectively and efficiently be adapted for different or evolving hardware, software or usage environments; • system can be successfully installed and/or uninstalled in an efficient way; • product can be replaced by another specified software product for the same purpose in the same environment.

Table 4. Software quality assessment criteria defined in ISO 25010:2001

[1] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=35733

While these features and characteristics provide good measures of software quality, it was not possible to perform a full evaluation of the 100+ DP solutions included in the DCH-RP registry, having in mind available resources and the timeframe. Therefore, after intensive discussion and brainstorming, we decided not to apply the guidelines defined in the standard to our registry.

Instead we decided to develop the simplified list of assessment criteria, discussed in the next section. We also developed further the registry update and tools and services rating mechanisms; these can be used within the project time frame and beyond in order to provide new entries to the registry as well as to enrich entries with feedback and quality assessment information.

We believe that these means are both affordable and useful enough to support the dialog among DCH institutions and e-Infrastructure providers and help end-users and resource providers in planning the migration of DP processes into cloud and grid environments.

3.2.3. DCH-RP project-specific selection and assessment criteria

While working on the selection and assessment criteria for the DCH-RP registry of services and tools we followed the assumption that it should support the development of the DCH-RP roadmap towards operating digital preservation processes in cloud and grid e-Infrastructures.

Therefore, we included in our consideration features of DP solutions important from the point of view of feasibility of running tools and services in outsourced models using clouds and grids.

Table 5 below summarizes criteria discussed by project participants and lists the most important comments related to them and identified issues related to their practical usage.

Criteria / group	Comments	Issues
4. Deployability in cloud/grid		
a. ease of deployment	• user or infrastructure/service provide point of view	• may be perceived differently depending on users' skills or objectivity?
b. licensing limitations	• might be a hard limitation • licensing models/ conditions can possibly be negotiated	• negotiations are out of scope of the project; registry provides current / known licensing conditions
c. architectural limitations	• importance depends on desired scale of the deployment	• architectural limitations may be invisible at document-based analysis stage
d. technical limitations	• might be a hard limitation (e.g. software may require particular hardware - e.g. a license key)	• some technical limitations may be invisible at the stage of document-based analysis
e. portability	• tools supporting various platforms are welcome • not a strict criterion as virtualisation may overcome limits of software;	• declared vs real portability? • real portability may be assessed during practical evaluation
f. modularity	• software modularity may support mixed / elastic deployment models	• evaluating products vs this criterion may require more in-depth analysis than it's possible to make using public documentation

Table 5. DP process outsourcing feasibility-related criteria

3.2.4. Final list of the selection and assessment criteria

The final list of the services and tools selection criteria represents a compromise between the need for in-depth analysis of the DP solutions and the wish to keep the registry structure simple and evaluation processes possible to implement within the time limits and resources of the project. We also wanted the registry to be a living, sustainable source of information, therefore we decided to keep its structure simple to enable wide audience to both use it and contribute to it.

On one hand we wanted to bring new value compared with that of existing registries that do not provide indications on the quality of services and tools and/or their usability in outsourced models. On the other hand we are aware of our limitations and have deliberately limited the assessments provided by our registry to the factual information available either in public documentation or from feedback acquired from WP5's proofs of concept. The tools and services evaluation criteria adopted for the initial version of the registry are listed in Table 6 below.

Criteria / group	Comments	Issues:
a. popularity	• size of the user base • freq of applying the solution to a purpose	• hard to determine • domain-specific knowledge
b. support level	• maintenance level, latest developments • availability of end-user support/ helpdesk	• possible to assess • hard to assess without practical usage of the tool
c. portability	• possibility to run the solution in various computing environments: Windows, Linux, workstations, servers, clusters, clouds, grids	• possible to assess to some extent
d. scalability	• scalability to high volumes, large I/O traffic, lots of queries • can the solution exploit capacities of e-Infrastructures: e.g. grids/clouds	• hard to assess without careful testing
e. licensing model	• open source/free vs licensed • possible to run in the cloud/grid environment	• possible to assess
f. modularity, openness of the architecture	• is the solution pluggable, modular and open • can the solution be integrated with larger workflows; is it monolithic or pluggable	• applies to complicated solutions: frameworks, toolkits service • partially do-able based on document analysis

Table 6. Final list of the assessment criteria

Starting from this list of features and assessment criteria we developed a basic structure for the registry. Details of the registry structure are discussed in the next section of this document.

We applied this structure to more than 140 tools and services. Information for this fundamental set of criteria was collected based on publicly available documentation.

In addition to basic features of DP solutions we have defined and applied quality-related criteria. These criteria include usefulness, generality of solution, user-friendliness and quality of results (see Table 8 in Section 4.2). They are limited to user-centric aspects of digital preservation tools and services and are applied only to a subset of the solutions described in the registry. These limitations result from several reasons.

First, our priority was to address the needs of the DC institutions, which are interested mainly in the user-centric aspects of solutions. Adding assessment criteria and content interesting from the e-Infrastructure point of view represents a possible future extension of the registry.

Second, we assumed that technical aspects of the solutions, such as portability, architecture and scalability, are easy to understand for e-Infrastructure providers while they may not obvious for DCH institutions. Similarly, evaluating the domain-specific usefulness of the tools and services is quite feasible for DCH institutions but difficult for e-Infrastructure providers. We decided to focus on user-centric assessments, as they are useful both for DCH institutions and for e-Infrastructures. The former will make use them in order to find solutions for implementing their DP processes. The latter may use the registry while planning their service portfolio for the DCH sector.

Third, we wanted to provide objective and trustworthy assessments for defined quality criteria. For this reason we rated only those tools and services that were evaluated within DCH-RP's WP5 Proofs of Concept. Obviously, broadening the coverage of services and tools assessments using current and possibly additional criteria is a potential next step in registry development. It is also a part of the registry sustainability plan presented in Section 5.

Fourth, the focus of T3.3 was to develop the structure of the registry and its mechanisms. In this context, the current content of the registry i.e. descriptions of 140+ tools and detailed assessments of 15+ solutions constitutes a practical evaluation of the registry structure and the update and quality assessment mechanisms.

4 REGISTRY STRUCTURE, MECHANISMS AND CONTENTS

This section presents the features, organisation and structure of the registry of services and tools as well as the mechanisms of the registry that might be used in order to improve its contents.

4.1. Basic structure of the registry

The DCH-RP tools and services registry contains more than 140 solutions, ranging from basic tools such as a data format identification application to complex frameworks and services enabling implementation of all phases of the DP process.

The registry is grouped into several service categories based on functional areas, stage of the digital preservation process where the tools are used (based on the OAIS model), content type of the digital objects supported, etc. This information enables the user to find the solution appropriate to the process they want to implement.

Current structure covers these features of systems that are possible to determine within the timeframe and resources of the project and are limited to factual information that it is possible to acquire based on publicly available documentation.

The registry provides information related to the tools and services licensing models and their availability (open source, free, paid). In addition, the solution status covers the level of maintenance of the software tool or service including information on latest updates and developments related to the solution (based on information found in the public domain, e.g. sourceforge, github or google code portals). This information helps users in determining the level of maturity of the solution and the type and scope of the support he may expect for it. The registry provides information that may help users determine possible models for implementing the services and tools. Currently this type of information is limited to factual statements related to the implementation technology and licensing model.

The list of basic registry fields as well as the possible values of each are presented in Table 7 below. In addition to basic information on tools and services, categorized using the structure discussed above, quality assessments are provided for a subset of the solutions, based on the outcomes of Proofs of Concept conducted in WP5. This part of the registry content is discussed further in Section 4.2.

Field name	**Meaning**	**Comments:**
Name	Name of the solution	
Web	Link to the website	Link to website, wiki or repository page describing the solution
Functionality	Description of the tool's functionality	Possible values: ● file processing tool ● web application/service ● search/indexing tool ● CMS/webpage ● framework / system ● other

Table 7a. DCH-RP registry structure

Field name	Meaning	Comments:
OAIS stage	OAIS model stage where the tool is used. Classification based on[2]	Possible values: • ingest • data management • archival storage • access • preservation planning • administration
Usage	Type of action that can be performed on digital objects using the tool or service concerned. In some cases it is related to the stage of the OAIS model.	Possible values: • conversion • integrity check • acquisition • metadata extraction • metadata assignment • search • data similarity check • other
Content type	Type of digital object content that can be handled using a particular tool or service.	Possible values: • audio • documents • disk images • web pages • video • database • text • other
Category	Category/type of the tool: simple tool vs toolkit, application vs web system	Possible values: • file processing tool • web application/service • search/indexing tool • CMS/webpage • framework / system • other
Technology	Technologies used to implement and/or deploy and offer the tool/service	• Programming language used for implementation • Run-time environments: e.g. tomcat, relational DB etc. • Deployment environment: cloud, hosted service etc.
Status	Status of the solution	• maintained / not maintained • in progress / unknown
Status comment	Some details on the status	Optionally may provide: • time of last / recent updates • maturity information (e.g. alpha version, proof of concept, etc.)
Licence	General licensing model	• open source • free (but not open) • paid
Licence details	Detailed type of the license applied to the solution	FreeBSD, (L)GPL etc.
Project / institution	Project, institution or initiative responsible for the tools or service	

Table 7b. DCH-RP registry structure (continued)

[2] Reference model for an open archival information system (OAIS). http://public.ccsds.org/publications/archive/650x0m2.pdf

4.1 QUALITY ASSESSMENTS IN THE REGISTRY

In order to make the registry a more useful source of information, we decided to include in its structure several quality assessment metrics, extending the basic set of features described in the registry. These criteria were derived from the assumptions and results of WP5 Proofs of Concept. Out of five aspects considered by WP5 we included four aspects in the registry (see Table 8 below).

Criteria / group	Comments
Usefulness	Usefulness of the solution for particular purpose. Assessed from the domain user perspective. In the WP5 PoC context the value of this measure reflects the perception of the actual functionality of the solution, in the context of its declared target application and features.
Generality of the solution	Possibility to address broad range of specific scenarios from particular application domain. For instance for document conversion tool it means support for many data formats.
User-friendliness	Subjective measure of ease of deployment, installation and use of particular solutions. For instance the low value of the measure may reflect the complication of installation process (lots of dependencies on 3rd party software, poor documentation etc.)
Quality of results	Domain-expert provided evaluation of the quality of results produced by a given solution. For instance for file conversion utilities it reflects the quality of output images or documents.

Table 8. Final list of the quality assessment criteria

While the registry provides descriptions for 140+ tools and services, quality-related metrics are provided for 15+ solutions. This results from the wish to keep the assessment reliable, objective and based on the actual, practical evaluation of the solutions by DP domain experts. In our case this meant to rely on the results of the WP5's proofs of concepts conducted in close collaboration with CH institutions from various countries.

The quality assessment criteria we selected may look simplistic, but we deliberately wanted to keep them easy to understand and use by registry users.

More advanced assessments of particular types of tools and services (e.g. tools vs frameworks, simple applications vs complex services) would require analyzing many various aspects of the DP solutions. Similarly, providing more detailed and complete evaluation statements would require in-depth analysis and practical experience with use of the tools. Providing even basic quality assessments for all the solutions listed in the registry is not realistic within the project.

Importantly though, we believe that the registry (including quality assessments) can be further improved in future. The range of the solutions covered may be broadened using the in-built update and assessment mechanisms. Also the scope of the features of the tools and services described in the registry may be extended towards the e-Infrastructure-oriented assessment criteria. This may require updating the registry structure, in order to accommodate new type of information.

4.2 ONLINE VERSION OF THE REGISTRY

Based on the structure defined for the registry and the concept of the mechanisms planned for it, Task 3.3 of the DCH-RP project developed a pilot version of the on-line registry.

During the course of the project, we decided to use two “flavours” of the on-line registry. We started with a collaborative platform for document editing in order to work out the initial version of the registry presented in this report. This enabled all DCH-RP participants to access, edit and discuss the registry content. Thanks to this approach, representatives of all stakeholders in the digital preservation process had an opportunity to contribute to the registry. We followed this approach over several months. The contents of the registry document - as of the end of September 2013 - is presented in Annex 1 of this deliverable.

Based on our experience with the internal usage of the online registry, we developed an on-line version of the registry to be used by a wider public. Currently this is based on the Viewshare[3] application. This provides an easy and functional access interface to the registry data that supports multiple entry points to the information (discussed in the next section of this chapter).

In addition to publishing the static version of the registry contents, we have also incorporated a registry update mechanism that enables end-users to contribute to correcting, updating, extending and enriching the registry contents. Assumptions adopted for the update mechanism and the mechanism itself are described in Section 4.3.2.

Besides providing new content for the registry - e.g. entries related to tools and services not yet covered - the registry concept assumes that its users will be equipped with a mechanism for providing assessments of the tools and services according to defined evaluation criteria. This mechanism is crucial to ensure that the DCH-RP registry will advance and bring new value to the state of the art and that it will contain more in-depth information on DP solutions than other registries. Solutions rating and assessment mechanisms are discussed in detail in Section 4.3.3.

4.2.1 Registry entry points and views

One of the fundamental objectives of the DCH-RP registry was to provide a source of information about DP tools and services, as well as DP processes implementation options, for various types of end users including DCH institutional employees and e-Infrastructure providers. The knowledge and technical awareness of these end users may differ depending on their experience and type, and on the particular business focus of their organisations and technical sophistication of the solutions they use for implementing DP processes.

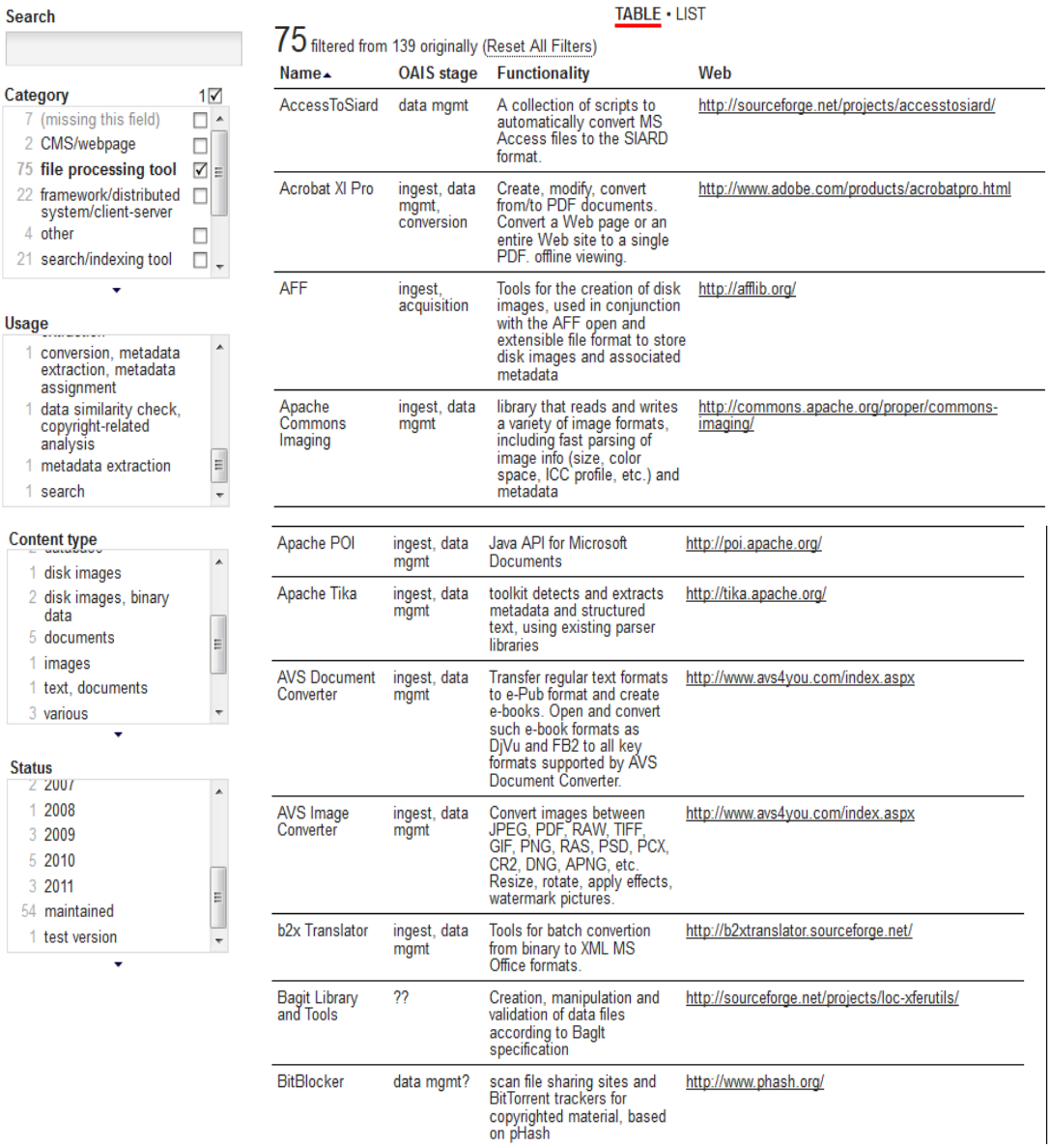
In order to make the information gathered in the registry easily available and the registry useful for a wider public, the registry provides multiple entry points to the underlying data. These enable users with various technical backgrounds and experience in the DP domain effectively to explore the registry content. Three examples will illustrate this approach.

First, users may simply browse the complete contents of the registry. This option is useful for users that want to learn about typical tools and services for implementing DP processes.

Second, the registry offers content filtering based on predefined categories. These include tool or service category and usage as well as the content type the solution can deal with. Filters can be easily applied to the content - the view of the registry is automatically filtered when users mark them in the navigation

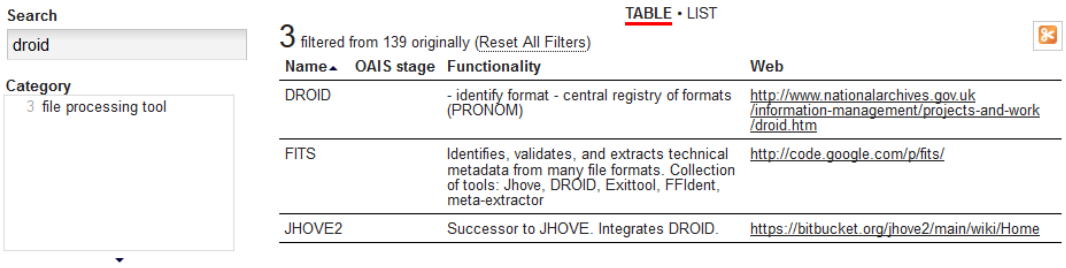
[3] Viewshare. Interfaces to our heritage. http://viewshare.org.

panel. Picture 1 below shows the registry table view with one filter applied (in the example presented, the user is interested in file processing tools). Filters may be combined in order to narrow the range of solutions shown in the registry. These functions are especially useful for more advanced users (the ones who know what are they looking for).



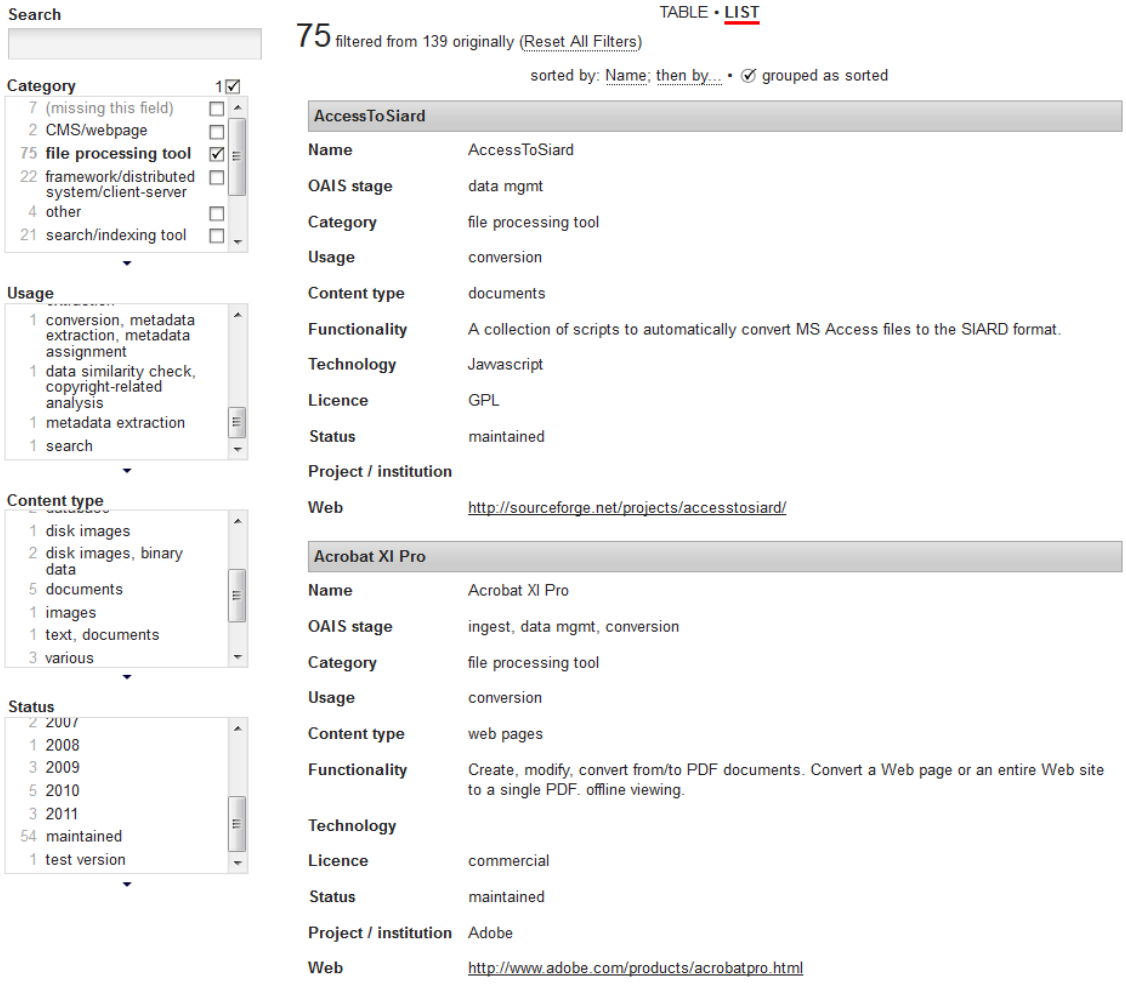
Picture 1. Interface of the on-line version of DCH-RP registry

Third, in addition to content filtering, it is possible to perform a keyword-based tools search. The result of the search is a filtered view of the registry that shows only the entries containing keywords specified by the user. Picture 2 depicts the results of searching on the keyword “droid”, the name of a popular data format identification tool.



Picture 2. Keyword based search results

In order to make the registry content easily accessible its interface provides different views on the data. The pictures above show simplified, table views of the registry, containing only basic information on tool name, OAIS stage, functionality and web link. Users may also switch to a list view that provides more exhaustive information on the tools and services, including all the fields and columns defined in the registry. An example of the list view is presented in Picture 3 below.



Picture 3. List view of the DCH-RP registry

4.2.2 Registry update mechanism

An important aim in developing the DCH-RP registry was to enable users coming from both the DCH-RP consortium and outside to update and maintain the contents of the registry. For this purpose the on-line version of the registry supports submitting information on new products not included in the registry and updating, correcting or completing the existing material.

We considered various approaches to registry editing such as wiki-style editing, where the contributions from users are immediately visible in the official version of the registry and peer-reviewed editing, where contributions and updates are reviewed by the project partners and made available in the public version after acceptance. Eventually, we decided to adopt the peer-reviewed editing model. We have secured human resources to deal with the contributions via the review process for the lifetime of the project and beyond it.

Posting new information and data updates are both possible using web forms. This enables the submission of a new or updated entry containing all or only part of the fields included in the registry structure. Submitted information is reviewed and verified before publication in the registry.

4.2.3 Tools and services evaluation mechanisms

In addition to the registry contents update mechanism, users are given the means for evaluating the tools and services included in the registry. Submitting quality assessments, as well as pros and cons of solutions, will be possible using web forms enabling free-text comments or selecting values for predefined measures (e.g. 'for ease of use' measure: hard, moderate, easy) etc. The feasibility of particular ways of implementing the evaluation mechanism will be examined during the registry exploitation phase in the later stage of the project.

Quality measures currently supported by the evaluation mechanism include aspects of digital preservation solutions examined within WP5's proofs of concept (PoCs). A list of these aspects, along with comments on evaluation performed within WP5, are included in Table 8 below. Explanation of the grading scheme of WP5's proofs of concept is discussed in Annex 1.

Field name	**Comments**
Simplicity of installation	May be subjective. However results of WP5 PoCs show that some tools clearly lack documentation that is usable by non-IT experts.
Simplicity of management	PoCs led to similar observations. There are tools that are too complicated to be used and managed by DCH institutions that do not have IT specialists.
Ease of use	PoCs led to similar observations as the ones discussed above.
Generality of solution	Should be understand in the context of the main application of the solution. E.g. for file processing tools, the scope of supported file formats determines its generality.
Quality of result	Can be evaluated by domain experts. E.g. for image processing tools applied to paintings actual quality can be assessed only by experts in digital imaging solutions. For text conversion, national character sets might constitute an issue for some tools.

Table 8. Aspects of DCH-RP solutions to be graded using evaluation mechanisms of DCH-RP registry

5 NEXT STEPS, REGISTRY SUSTAINABILITY

Task 3.3 of the DCH-RP project focused on defining the structure of the registry and developing its pilot version. Discussions conducted in the DCH-RP consortium related to the assumptions, aims and fundamental features and mechanisms of the registry are summarized in Sections 2-4.

In this section we present steps, mechanisms and actions that will be undertaken during the second year of the project, aiming at ensuring the registry's usage, popularity and sustainability.

The aim of developing the DCH-RP registry was to support the development of the DCH-RP roadmap and to provide DCH institutions and e-Infrastructure providers with a valuable source of information helping them in planning the migration of DP processes to clouds and grids.

Therefore the work of Task 3.3 was not limited to simply defining the registry structure and providing a pilot version. Within Task 3.3. we also planned actions necessary to make the registry useful beyond the DCH-RP consortium and sustainable beyond the lifetime of the project.

The DCH-RP registry already provides added value compared to other registries. It contains 140+ tools and services that may be applied for digital preservation purposes. The list of tools described in our registry is a superset of those in many existing registries.

The on-line version of the registry provides an easy and functional interface, enabling browsing, content filtering and searching the appropriate entries based on categories or keywords.

The registry is also equipped with mechanisms that enables its content to be enriched by contributing new registry entries, updating and correcting existing descriptions, and providing solution assessments and ratings. During the DCH-RP project these mechanisms were used by project participants, in order to capture assessments acquired during WP5 PoCs. This effort will be continued during the second year of the project. The registry update and solution assessment mechanisms may also be used by a wider public for contributing to the registry in future and improving its contents and quality.

Alongside the concrete steps already undertaken, it is of course necessary for the DCH-RP consortium to put in train various actions to assure the long-term sustainability of the registry. These include promoting the DCH-RP registry and encouraging its usage in the DCH and e-Infrastructure environments as well as securing resources for its long-term availability, development and improvement. These aspects are discussed in the following sections of this chapter.

Section 5.1 discusses the details of the DCH-RP promotion plan. Section 5.2 presents a series of other steps aiming at ensuring the sustainability of the registry.

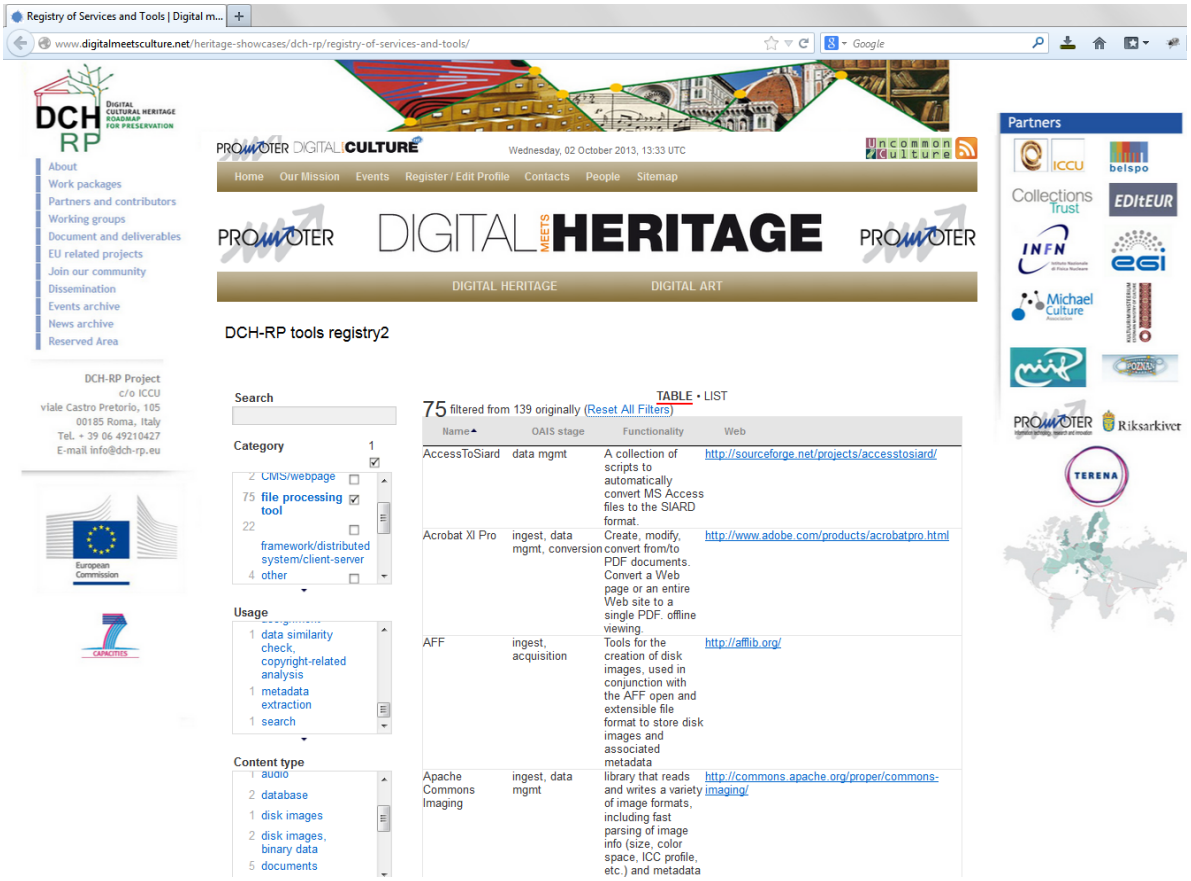
5.1 REGISTRY PROMOTION AND DISSEMINATION

The DCH-RP registry will be promoted using various project dissemination mechanisms, supported by mechanisms and PR capabilities of participating and collaborative organisations.

We will be trying to demonstrate the registry's value to the DCH community as an ongoing, living resource that is continuously improved and extended. We will be presenting the registry to e-Infrastructure providers as a preferred source of information useful for adapting their service portfolios for DCH domain-specific needs.

The dissemination plan includes publishing news items on the DCH related portals, embedding links to the registry in these portals, promoting the registry via events organised by the consortium and collaborating with relevant initiatives for DCH institutions and e-Infrastructure providers.

The online version of the registry will be embedded in web portals managed by DCH-RP consortium participants. At the time of writing we have the registry embedded in the Digital Meets Heritage portal. Picture 4 below contains a web browser page showing the embedded registry.



Picture 4. DCH-RP registry of tools and services embedded in the DigitalMeetsHeritage portal

http://www.digitalmeetsculture.net/heritage-showcases/dch-rp/registry-of-services-and-tools/

We will also try to position the registry in popular web search engines so that people looking for such information will be given the chance to visit our registry. Adding the DCH-RP registry to the lists of existing registries will also be attempted in order to increase its visibility.

In addition to typical promotion and dissemination methods, we will popularize the registry in the environments of CH institutions and e-Infrastructure providers. For this purpose we will explore existing contacts to these organisations on the part of members of the DCH-RP consortium as well as looking to discover and exploit new contacts.

5.2 OTHER ACTIONS FOR REGISTRY SUSTAINABILITY

In addition to registry promotion and dissemination the consortium will undertake a series of actions aiming at maintaining and improving the quality of the registry.

Within these activities we will involve the cultural institutions and e-Infrastructure, grid and cloud resource providers in the process of using, editing and improving the registry. This may be implemented e.g. within the DCH-RP proofs of concept in WP5 of the project.

On the basis of these collaborations we will build the user base for the registry, collect useful feedback and information for improving the content of the registry, and establish relationships with organisations and experts that can act as reliable sources of information on DP solutions and e-Infrastructure. This information can be then used for further extending the list of solutions covered in the registry, providing quality assessments, etc.

Updating and extending the content and the solution assessment mechanisms available in the on-line version of the registry will play an important role in the process of improving the registry. These mechanisms are crucial for ensuring its long-term sustainability and usability, bearing in mind the dynamic situation in the DCH sector and rapid developments in the capabilities of e-Infrastructures. In the next section we provide some specific ideas on possible improvements to the registry.

Improvements in registry content should, in turn, raise the interest in using the registry. We hope that this may create a self-sustained impetus for long-term registry development. Importantly, a critical mass of users and information must be established for this to happen; therefore the process of user involvement must be supported by registry promotion mechanisms.

We recognize the necessity of securing resources and funding for the long-term maintenance of the registry. Therefore we will seek sponsorship from leading DC institutions involved in the DCH-RP project or, indeed, outside the consortium. We will be also investigate whether a leading organizatoin active in the area could be invited to take the registry under their wing and provide administrative and hosting support. For now, Promoter, the technical coordinator of the DCH-RP project, has expressed its commitment to maintaining the registry available online on the DigitalMeetsCulture portal beyond the timeframe of the project. However we will also explore other possibilities.

5.3 POSSIBLE REGISTRY IMPROVEMENTS

The registry in its current form already brings some new value to the arena of digital preservation solutions registries. However several aspects of the registry can be improved in future in order to make it even more useful and sustainable in the longer term.

The registry structure can be extended in order to support assessing additional aspects of DP tools and services, for instance evaluation of the compliance of DP solutions to domain standards as well as assessing e-Infrastructure oriented aspects of tools and services.

Including compliance to standards in the criteria would also be an important improvement as this aspect of DP tools and services impacts their usefulness as well as the portability and long-term usability of the results of DP processes. Compliance to standards on both the DP solutions end and the e-Infrastructure side may also help in promoting collaboration between DCH institutions and e-Infrastructures. Therefore selection of eligible standards and assessing the compliance of tools and services described in the registry to these standards constitute a high-priority improvement to the DCH-RP registry. This will be performed in collaboration with the whole WP3 team.

Adding criteria interesting from the e-Infrastructure point of view would make the registry more useful for e-Infrastructure providers, as currently considered criteria are mainly user centric. As we aim to provide reliable assessments of the tools and services, including e-Infrastructure related aspects, this will require us to adapt somewhat the planning of the second round of WP5 proofs of concept, so that e-Infrastructure-related assessments can be acquired. For instance, selected tools and services can be evaluated in the context of their CPU, memory and I/O processing resource consumption and their scalability understood as the ability to benefit from increased capacities compared to local, workstation-based working environments.

In conclusion, broadening the scope of the solutions described and increasing the coverage of quality assessments represent obvious possible improvements of the registry and both are planned within the second year of the project. As constantly improving of the registry content requires substantial efforts it will require involving DCH-RP project partners during the project duration and encouraging end-users from CH institutions and e-Infrastructure environment to join the initiative and continue it beyond the timeframe of the project. Particular options for implementing this process and making it self-sustainable or supported by some leading organisation are discussed in previous sections of this chapter.

6 DISCUSSION AND CONCLUSIONS

In this document we presented the results of Task 3.3 “Registry of services and tools” of DCH-RP project. The aim of this task was to define the structure of a registry of services and tools used by the stakeholder community for digital preservation for DCH as well as putting in place a pilot registry.

We defined the structure of the registry including basic characteristic and features of the digital preservation tools and services. We have also tested this structure by providing and verifying the descriptions of 140+ tools and services including well-known solutions as well as emerging services and tools. We have also defined additional assessment criteria for DP solutions and tested their usability against 15+ solutions within WP5’s proofs of concepts. This extra content is one of the fundamental added values of the DCH-RP registry as it may act as guideline for DCH institutions in (re)organising their DP processes towards the e-Infrastructure provided services and resources. It may also be used by e-Infrastructure managers as the source of information on the tools and services needed by DCH community, thus helping in planning the e-Infrastructures service portfolios.

Importantly, the registry structure and content is the product of discussion and brainstorming that involved all stakeholders of the future digital preservation process, namely DCH institutions, coordination and funding bodies and e-Infrastructure providers. Therefore it is believed to provide a good snapshot of the currently available and emerging options for implementing DP processes.

We have also developed a pilot, on-line version of the registry. It is embedded into one of the DCH communities portals and thus is already available to public. It provides several entry points enabling users with different background to explore, search and use the information gathered in the registry.

While the registry provides already descriptions for 140+ tools and services, we assume that the work performed within T3.3 of DCH-RP constitutes first step of the process of making the registry used, authoritative, recognised and sustainable for long time. We recognise the need of providing to ourselves, i.e. the DCH-RP consortium as well as to wide public the technical and organisational means to discover, explore, assess and improve and extend the registry content. Therefore we planned the registry promotion and dissemination actions to be performed beyond the scope of Task 3.3, during the second year of the project. We have also put in hand of end-users the mechanisms enabling to update the registry content, submit new entries to the registry as well as evaluate the tools and services using defined assessment criteria. Part of the registry development plan is also further usage of the registry assessment mechanisms while performing the proofs of concept in WP5. We have also secured resources necessary for implementing peer review-based process of registry development and maintain the registry on-line beyond the timeline of the project.

One of the challenges related to developing and sustaining the registries of tools and services is to make them both popular and trusted. We believe that the current registry structure and content makes it already useful for DCH institutions and e-Infrastructure providers while registry improvement and dissemination mechanisms will help making the registry content up to date and reliable.

ANNEXES:

ANNEX 1: REGISTRY CONTENT

This annex provides snapshot of the registry content in end of September 2013. Table below shows the descriptions of 140+ tools spanning the basic characteristics and features of the solutions. For better readability selected columns are removed. Full material can be examined in the on-line version of the registry.

Name	Category	Functionality	Content type	Usage	Technology	Status	Status comment	Licence type	Licence details	Project / institution
AccessToSiard	file processing tool	A collection of scripts to automatically convert MS Access files to the SIARD format.	documents	conversion	Javvascript	maintained		open source	GPL	
ACE (Audit Control Environment)	framework / system	Set of tools to help archives monitor the integrity of collections. It provides a mechanism to allow a 3rd party to independently verify a collections integrity.	*various*	integrity check	Java, Python, MySQL, Tomcat	maintained		open source		ADAPT project
Acrobat XI Pro	file processing tool	Create, modify, convert from/to PDF documents. Convert a Web page or an entire Web site to a single PDF. offline viewing.	web pages	conversion		maintained		paid		Adobe
AFF	file processing tool	Tools for the creation of disk images, used in conjunction with the AFF open and extensible file format to store disk images and associated metadata	disk images, binary data	acquisition	C/C++, Java	maintained		open source	BSD	Naval Postgraduate School
Apache Commons Imaging	file processing tool	library that reads and writes a variety of image formats, including fast parsing of image info (size, color space, ICC profile, etc.) and metadata	disk images, binary data	conversion, metadata extraction	Java	maintained		open source	Apache	Apache
Apache Lucene	file processing tool	text search engine library	*documents*	search	Java	maintained		open source		Apache
Apache POI	file processing tool	Java API for Microsoft Documents	documents	acquisition	Java	maintained		open source		Apache
Apache Tika	file processing tool	toolkit detects and extracts metadata and structured text, using existing parser libraries	web pages	metadata extraction	Java	maintained		open source		Apache
Archive-It	web application/service	Web archiving service to capture, build, and manage collections of web content.	web pages	acquisition, other	only service, soft not available	maintained		open source		Archive-It
Archivematica	framework / system	Digital preservation system that is designed to maintain standards-based (OAIS, DC, METS, PREMIS), long-term access to collections of digital objects. Micro-services design pattern to provide an integrated suite of software tools. Users monitor and control the micro-services via a web dashboard.	*various*	*various*	only service, soft not available	in progress	Beta version available (08.2013)	free		Archivematica

Archivists Toolkit	framework / system	Archival data management system. Supports accessioning and describing archival materials; establishing names and subjects associated with archival materials, including the names of donors; managing locations for the materials; and exporting EAD finding aids, MARCXML, METS, MODS, Dublin Core records.	*various*	acquisition, metadata extraction	Java, MySQL/MS SQL/Oracle	maintained		open source		Five Colleges, Inc., New York University Libraries, and the UC San Diego Libraries
Archon	framework / system	Platform for archival description and access. Record information about collections and digital objects and view, search, browse that infomation in a fully-functional public web site. Export MARC and EAD records.	*various*		PHP, MySQL/MS SQL Server	unknown	last modification: 2011	open source	Illinois Open Source License	University of Illinois
AudioScout	search/indexing tool	Distributed audio content indexing system. It can index a large collection of audio content for the purpose of later recognition of unknown signals. Robust to noise, different encodings and other types of distortion, based on pHash	audio	acquisition, indexing	C++	maintained		paid		Evan Klinger & David Starkweathe
AVS Document Converter	file processing tool	Transfer regular text formats to e-Pub format and create e-books. Open and convert such e-book formats as DjVu and FB2 to all key formats supported by AVS Document Converter.	documents	conversion	only executable available	maintained		paid	commercial, free version for test purposes	AVS4YOU® Featured Software
AVS Image Converter	file processing tool	Convert images between JPEG, PDF, RAW, TIFF, GIF, PNG, RAS, PSD, PCX, CR2, DNG, APNG, etc. Resize, rotate, apply effects, watermark pictures.	images	conversion	only executable available	maintained		paid	commercial, free version for tests	AVS4YOU® Featured Software
b2x Translator	file processing tool	Tools for batch convertion from binary to XML MS Office formats.	documents	acquisition, conversion	Visual Studio	unknown	last modification: 2009	open source	BSD	
Bagit Library and Tools	file processing tool	Creation, manipulation and validation of data files according to BagIt specification	*various*	*other*	Java	maintained	maintained	free		Library of Congress
BitBlocker	file processing tool	scan file sharing sites and BitTorrent trackers for copyrighted material, based on pHash	*various*	*data similarity check, copyright-related analysis*	C++	maintained		paid		Evan Klinger & David Starkweathe
BitCurator	file processing tool	Stack of software. Features: Pre-imaging data triage. Forensic disk imaging. Filesystem analysis and reporting. Identification of private and individually identifying information. Export of technical and other metadata.	disk images	acquisition, conversion, meta-data extraction		in progress	test version available (08.2013)	open source		Bit Curator
BWF MetaEdit	file processing tool	Embedding, editing, and exporting of metadata in Broadcast WAVE Format (BWF) files. Enforce metadata guidelines developed by the Federal Agencies Audio-Visual Working Group, the European Broadcasting Union (EBU), Microsoft, and IBM.	audio	conversion, metadata extraction, metadata assignment		unknown	last modification: 2011	open source		Federal Agencies Digitization Guidelines Initiative

C3PO	file processing tool	Uses meta data extracted from files of a digital collection as input to generate a profile of the content set. Supports FITS and Apache TIKA.	*various*	*ingest*	Java, MongoDB	maintained		open source	Apache2	
CASPAR tools	framework / system	Set of loosely coupled applications that follow OAIS model.	*various*	*various*		not maintained	last modificati: 2009, software rep.unavailable	unknown		CASPAR
CHRONOS	file processing tool	Database Retirement, Partial and Ongoing Database Archiving, Application Retirement.	database	acquisition, conversion, metadata extraction		maintained		paid		CSP
ClipSeekr	search/indexing tool	Real time video monitoring solution that can be used to identify video clips contained in a video stream, basd on pHash	*various*	*data similarity check, copyright-related analysis*	C++	maintained		paid		Evan Klinger & David Starkweathe
CONTENTdm	framework / system	Handles the storage, management, and delivery of library digital collection to the web. Web-based digital collection tool -prepare items. Server -store collection. Web-based discovery interface. Self-service tool to upload the metadata to WorldCat using the Digital Collection Gateway, integration with OCLC, harvesting from the Web sites and adding long-term preservation.	*various*	*various*		maintained		paid		OCLC
ContextMiner	web application/service	Framework to collect, analyze, and present the contextual information. Tools to collect data and metadata off the Web by automated crawls (blogs, YouTube, Flickr, Twitter, and open Web).	web		only service, soft not available	maintained		free	CC Attr.-Noncommercial-Share Alike 3.0 US License	Rutgers University
Cue	file processing tool	Java library for simple text analysis - counting strings, identifying languages, and removing stop words. Support for many languages.	text, documents	acquisition, ...	Java	unknown	last modification: 2011	open source	Apache	IBM Research's Visual Communication Lab.
Cumulix	search/indexing tool	cloud-based image search and retrieval system that runs on top of Neo4j	text, documents	*ingest*	Java	maintained	maintained	paid		Evan Klinger & David Starkweathe
dArceo	framework / system	System for long-term preservation of source data (e.g. master files), primarily focused on textual, graphical and audiovisual content. It makes migration of source data possible with respect to the OAIS model. Additionally, dArceo provides conversion and source data delivery functions	text, graphics, audio, video		Java, PostgreSQL, GlassFish	maintained		paid	commercial, free version for test purposes license includes support	PSNC

DeepArc	file processing tool	Graphical editor for mapping an existing relational data models and XML Schema. Export the database content into an XML.	database	acquisition, conversion	Java	maintained		open source	GPL	
Dependency Discovery Tool	file processing tool	Searches through binary office files (.doc, .xls and .ppt) and tries to find any documents or files that are linked to the document.	*documents*	*acquisition*	Java	maintained		open source	Apache	
dLab	framework / system	Robust, flexible and extendible system for digitisation process management	*documents*	acquisition, conversion etc.	Java, relational DB, Tomcat	maintained		paid	commercial, free version for test, license incl. support	PSNC
dLibra	framework / system	Dedicated software tool to build digital libraries. Ingest, store, retrieve digial objects independant of format. Import from other formats and systems. Metadata management. Authorization. Publication.	various	various	Java, relational DB, Tomcat	maintained		paid	commercial, free version for test, license incl. support	PSNC
dMuseion	framework / system	Complex system for building digital museums. It allows accessing digital versions of cultural heritage monuments via the Internet. Dedicated mainly to visual arts holdings, including 3D objects.	various	various	Java, relational DB, Tomcat	maintained		paid	commercial, free version for test, license incl. support	PSNC
DPSP	file processing tool	Set of tools: Manifest Maker, checksum checker, Xena (convert to standard format) and Digtial Preservation Recorder (workflow engine, that runs the other tools)	various	various	Java	maintained		open source	GPL	National Archives of Australia
DRAMBORA	other	This toolkit is intended to facilitate internal audit by providing repository administrators with a means to assess their capabilities, identify their weaknesses, and recognise their strengths			online and offline versions	unknown	last modificati on: 2010	open source	download requires registering	DRAMBORA
DRB	file processing tool	Data Request Broker -Java API for different file types.	webpages, zip	metadata extraction	Java	unknown	last modificati on: 2009	open source	LGPL	GAEL Consultant
DROID	file processing tool	- identify format - central registry of formats (PRONOM)	various	format recognition	Java	maintained		open source		UK National Archives
DSpace	framework / system	Software platform that allows capture and describe digital material using a submission workflow module, or a variety of programmatic ingest options, distribute an organisation's digital assets over the web through a search and retrieval system, preserve digital assets over the long term			Java, Tomcat, Perl, relational DB	maintained		open source		DuraSpace
DuraCloud	framework / system	Storage cloud. Can be used with DSpace. Provides automated storage to multiple cloud providers with integrity checks.			Java	maintained		open source		DuraSpace
eCulture Science gateway	web application/servi ce	A portal, that enable transparent access to Digital Cultural Heritage contents. It uses user authentication and authorization with federated access.	webpages	view, search	only service, software not available	in progress	proof of concept implement	free	only service, software not available	DCH-RP project

							ation (09.2013)			
Email Preservation Parser	file processing tool	Parse e-mails in MBOX format and convert to XML	e-mail	conversion	Squeak (Smalltalk)	unknown	last modification: 2008	open source	MIT licence, CC Attribution Non-comm. Share Alike Licence	The Rockefeller Archive Center, Smithsonian Institution Archives
EMET	file processing tool	extract metadata embedded in JPEG and TIFF	grephics	metadata extraction		unknown	last modification: 2007	open source	GPL	ARTstor
Exhibit	CMS/webpage	Create web pages with advanced text search and filtering functionalities, with interactive maps, timelines, and other visualizations from Excel, Google spreadsheet, mediaWiki, BibTex, JSONP and embed in Wordpress or Drupal.	documents, webpages	view, search	JavaScript, CSS	maintained		open source	BSD	Massachusetts Institute of Technology
ExifTool	file processing tool	Perl library and command-line application for reading, writing and editing meta information. Formats: EXIF, GPS, IPTC, XMP, JFIF, GeoTIFF, ICC Profile, Photoshop IRB, FlashPix, AFCP, ID3, maker notes of digital cameras of many camera manufacturers.	grephics	metadata extraction	Perl	maintained		open source	GPL	Phil Harvey
Extractor	file processing tool	Extract metadata from: PNG, TIFF, DOCX, PDF, SVG, CGM, MP3, WAV, BMP, PBM, PCD, PCX, PICT, PPM, PSD,TGA, XBM, XPM, JPEG, JP2, GIF, HTML, AIF	grephics	metadata extraction	C++, Qt	unknown	last modification: 2010	open source		Planets -XCL project
EZID	web application/service	Create and manage unique, persistent identifiers. Store citation metadata for identifiers in a variety of formats. Update current URL locations so citation links are never broken. Provides QR codes for identifiers.			only service, software not available	maintained		free	Annual subscription fee, hosted service,	University of California
Fedora Commons	framework / system	Core repository service (storing, managing, and accessing digital content in the form of digital object) and supporting services and applications: search, OAI-PMH, messaging, administrative clients, etc.	various	storage, search	Java, Tomcat, relational DB	maintained		open source	Apache2	FedoraCommons
ffident	file processing tool	Java library to extract information from files and identify their formats		format recognition, metadata extraction	Java	not maintained	last modification: 2005	open source	LGPL	
FFmpeg	search/indexing tool	Record, convert and stream audio and video. It includes libavcodec - the leading audio/video codec library.	video	acquisition, conversion, view		maintained		open source	LGPL	
FIDO	file processing tool	Tool to identify the file formats of digital objects. It is designed for simple integration into automated work-flows.	various	format recognition	Python	maintained		open source	Apache	Open Planets Foundation
Fiji	file processing tool	image processing package	graphics	conversion	Java, ImageJ	maintained		open source	GPL	

file	file processing tool	Recognize file format	various	format recognition		maintained		open source		Ian F. Darwin
FITS	file processing tool	Identifies, validates, and extracts technical metadata from many file formats. Collection of tools: Jhove, DROID, Exittool, FFIdent, meta-extractor	various	format recognition, metadata extraction		maintained		open source		Harvard University Library Office for Information Systems
Fixi	search/indexing tool	Command line, indexes, verifies, and updates checksum information for collections of files.	various	integrity check	Ruby, SQLite	maintained		open source	Apache	Chris Wilper
getID3()	file processing tool	PHP script that extracts useful information from MP3s & other multimedia file formats	various	format recognition, metadata extraction	PHP	maintained		open source	GPL, gCL	
GIMP	file processing tool	Image manipulation program. GUI and command line. Image modification, format conversion, etc.	graphics	conversion	c++	maintained		open source	GPL	GNU
Grab-A-Site		Offline Browser, filtering file types, prepare CD with grabbed data, etc.	webpages	acquisition, search		maintained		paid		Blue Squirrel
Gumshoe	file processing tool	Search interface for metadata extracted from disk images	disk images	search, metadata extraction	Ruby, Rails	maintained		open source	Apache	Mark A. Matienzo
Heritrix	search/indexing tool	Extensible, web-scale, archival-quality web crawler project.	webpages	acquisition, search	Java	maintained		open source	Apache	
HTTrack	search/indexing tool	Download a World Wide Web site from the Internet to a local directory, building recursively all directories, getting HTML, images, and other files. GUI, commandline, C-API.	webpages	acquisition	C	maintained		open source	GPL	HTTrack
ImageMagick	file processing tool	Create, edit, compose, or convert bitmap images. Support > 100 formats. Command line and libraries for many languages	graphics	conversion, acquisition		maintained		open source	Apache	
ImageVerifier	file processing tool	traverses a hierarchy of folders looking for image files to verify. It can verify TIFFs, JPEGs. PSDs, DNGs, and non-DNG raws (e.g., NEF, CR2).	graphics	searcch, format recognition		maintained		paid		Marc Rochkind
iRODS	framework / system	Integrated Rule-Oriented Data-management System to organize, share, protect, and preserve large sets of computer files.			Java, PHP, relational DB	maintained		open source	BSD	Regents of the University of California, the University of North Carolina, and the Data Intensive Cyberinfrastructure Foundation
itext	file processing tool	library to create and manipulate PDF documents	documents	conversion, metadata extraction	Java, C#	maintained		open source	AGPL, paid if AGPL conditions	iText Software Corp. (California, USA) and 1T3XT

								need to be released	BVBA (Ghent, Belgium)	
JHOVE	file processing tool	Format-specific identification, validation, and characterization	various	format recognition, metadata extraction	Java	maintained		open source	GNU GPL	JSTOR and the Harvard University Library
JHOVE2	file processing tool	Successor to JHOVE. Integrates DROID.	various	format recognition, metadata extraction	Java	maintained		open source	GNU GPL	Library of Congress, California Digital Library, Stanford University
jj2000	file processing tool	JPEG 2000 encoder/decode	graphics	conversion	Java	not maintained	last modification: 2002	open source	LGPL	Google
JODConverter	file processing tool	automates conversions office document formats (ODF, PDF, RTF, HTML, Word, Excel, PowerPoint, and Flash)	documents, webpages	conversion	Java	maintained		open source	Lesser GPL	
jpylyzer	file processing tool	JPEG 2000 Part 1 validator and properties extractor.	graphics	format verification, metadata extraction	Python	maintained		open source	Apache	Open Planets Foundation
Kakadu	file processing tool	Complete implementation of the JPEG2000 standard. Java and c++ libraries.	graphics	conversion	Java, C++	maintained		open source	Non-commercial, commercial	University of New South Wales, Sydney, Australia
KOST tools	file processing tool	SIARD-Val: validate SIARD container. SIP-Val: validate SIP. TIFF-Val: validate TIFF files. SIP-browser. csv2siard converter.	graphics, other	data validation, conversion	C/C++	maintained		open source	GPL	Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen
libPST	file processing tool	tools for extracting messages and contacts from MsOutlook, convert to MBOX	e-mail	conversion		not maintained	last modification: 2004	open source	GNU GPL	Software Group
LOCKSS	web application/service	Library-led digital preservation system. Allows exchange of publicattion of electronic materials between libraries.	documents, graphics	acqiusition		maintained		open source	Creative Commons Attribution 3.0 Unported License.	Stanford University Libraries
Matchbox	search/indexing tool	Duplicate detection tool for digital document collections. Identifies duplicated content, even where files are different, eg in format, size, rotated, cropped, colour-enhanced. Assembling collections from multiple sources, and identifying missing files.	graphics	comparsion, search	C++, Python	maintained		open source	open source	Austrian Institute of Technology

Mediainfo	file processing tool	Supplies technical and tag information about a video or audio file. There is a GUI for most OSes and CLI for all.	video, audio	metadata extraction	C++	maintained		open source	BSD	
Meta-extractor	file processing tool	extracts preservation-related metadata and outputs in XML	documents, webpages, graphics, audio, video	metadata extraction	Java	unknown	last modification: 2010	open source	Apache	National Library of New Zealand
MINT	web application/service	Web based platform that facilitates the aggregation of cultural heritage metadata. Support ingestion, semantic alignment and aggregation of metadata records, and proceeds to implement a variety of remediation approaches for the resulting repository.			Java	maintained		open source	Affero GPL	MINT
MPG321	file processing tool	command-line mp3 player and decoder	audio	viewer		maintained		open source	GPL	
Multivalent	file processing tool	Browse, view, compress, encode, validate etc. document files (scanned paper, PDF, HTML, UNIX manual pages, TeX DVI, and more)	documents	viewer, format recognition	Java	unknown	last modification: 2009	open source	GNU, commercial	University of Liverpool
NARA File Analyzer and Metadata Harvester	search/indexing tool	File Analyzer and Metadata Harvester allows a user to analyze the contents of a file system or external drive and generates statistics about the contents of the contained directories.	filesystem/disk image	acquisition	Java	maintained		open source	NARA Open Source	US National Archives
NARA Video Frame Analyzer	file processing tool	Analyze technical properties of individual frames of a video file in order to detect quality issues within digitized video files.	video	quality check	Java	maintained		open source	NARA Open Source	US National Archives
Nesstar	framework / system	Enable publishers and users of social and economic data (and other similarly structured data) to exploit the new generation of web technologies. Publish statistics (search, browse, analyse, download, metadata), authorization, semantic data web. Server, publisher application, web application.	documents	extract metadata	Java, Tomcat, relational DB	maintained		open source	commercial	Norsk samfunnsvitenskapelig datatjenest
NetarchiveSuite	search/indexing tool	Web archiving software package.	webpages	acquisition	Java	maintained		open source	LGPL	NetarchiveSuite
NutchWAX	search/indexing tool	Searches web archive collections. Adaptation of the Nutch fetcher step to go against web archives rather than crawl the open net.	webpages	search		unknown	last modification: 2010	open source	GPL	NutchWAX
ODF Converter	file processing tool	conversion of office document formats (ODF, PDF, RTF, HTML, Word, Excel, PowerPoint, and Flash), command line, OpenOffice integration	documents, webpages	conversion		unknown	last modification: 2010	open source	BSD	Clever Age, DIaLOGIKa, Sonata Software Ltd
ODF Toolkit	file processing tool	Java modules that allow programmatic creation, scanning and manipulation of Open Document Format	documents, webpages	conversion	Java	maintained		open source	Apache	Apache Software Foundation
Omeka	CMS/webpage	Web-publishing platform for library, museum, archives, and scholarly collections and exhibitions. Manage data collections and objects. Support for Dublin Core, OAI-PMH and many other plugins.			PHP, MySQL, CSS	maintained		open source	open source	George Mason University

OpenJPEG	file processing tool	JPEG 2000 codec	graphics	conversion	C++	maintained		open source	BSD	Université catholique de Louvain
Pagelyzer	search/indexing tool	Suite of tools for detecting changes in web pages and their rendering	webpages	acquisition	Ruby	maintained		open source	Apache	Université Pierre et Marie Curie
PAIRTREE	search/indexing tool	library that supports the mapping between identifiers and filepaths according to the Pairtree Curation Microservices Specification			Java	maintained		open source	free	Library of Congress
PDF Tools	file processing tool	Create, parse, search, etc. with PDF files.	documents	search, acquisition	Python	maintained		open source	open source	Didier Stevens Labs
PDF/A Manager	file processing tool	Converts PDF documents to PDF/A (Level A and B) documents, or validate a PDF file against the PDF/A specification. Command line and development kit (C#, VB.net, C/C++, Java, Perl ...)	documents	conversion, integrity checking				open source	commercial	PDFTron
pdf2svg	file processing tool	Convert PDF to SVG	documents, graphics	conversion		maintained		open source	open source	David Barton
pdf2xml	file processing tool	pdf to xml conversion text extraction vectorial instruction extraction	documents	conversion	C++			open source	GPL	Xerox Research Centre Europe
PDFbox	file processing tool	Creation of new PDF documents, manipulation of existing documents and the ability to extract content from documents. Library and command line.	documents, text, graphics	conversion, metadata extraction	Java	maintained		open source	Apache	Apache Software Foundation
PDFSSA4MET	file processing tool	metadata extraction and tagging based on structural and syntactic analysis of PDF content in XML	documents	metadata extraction	Python	unknown	last modification: 2010	open source	free	
PDFtk	file processing tool	Manipulating pdf files: text an dmetadata extract, rotate, background, unpack, fill form, attach file etc.	documents	conversion	C	maintained		open source	GPL	Steward and Lee, LLC
pdftohtml	file processing tool	converts PDF files into HTML and XML, uses xpdf	documents, webpages	conversion	C++	maintained		open source	GPL	Lincoln & Co
pdiff: Perceptual Image Difference utility	file processing tool	image comparison utility that makes use of a computational model of the human visual system to compare two images.	graphics	comparsion	C++	not maintained	last modification: 2006	open source	GPL	
PeDALS	file processing tool	Outlook mail extractor	e-mail	acquisition		maintained		open source	free	
peepdf	file processing tool	Check PDF to find potentially harmful elements. Create and modify PDF files.	documents	conversion, security checks	Python	maintained		open source	GPL	Jose Miguel Esparza
pHash	other	perceptual hashing library (perceptual hash is a fingerprint of a multimedia file derived from various features from its content but not content itself)	graphics	comparsion	C++	maintained		open source	GPL, proprietary	Evan Klinger & David Starkweathe
PiM Toolbox	file processing tool	PREMIS:METS conversion and validation, support the implementation of PREMIS in the METS container format.	documents	metadata conversion		maintained		open source	public domain	Library of Congress and the Florida Center for Library

										Automation
PLATO	framework / system	Framework with atomic preservation services: Identify, Characterize, Compare, Modify, Migrate, and View.	various	conversion, view, format recognition	Java Webservices	maintained		open source	Apache	SCAPE
PLATTER	other	Planning Tool for Trusted Electronic Repositories provides a basis for a digital repository to plan the development of its goals, objectives and performance targets over the course of its lifetime in a manner which will contribute to the repository establishing trusted status amongst its stakeholders.				unknown	last modification: 2010	unknown		DPE
pstoedit	file processing tool	Convert PS and PDF graphics to other vector formats	graphics	conversion	C++	maintained		open source	GPL	
PSTViewTool	file processing tool	Browse the internal structures of a PST file.	e-mail	view	MFC/C++	unknown	last modification: 2010	open source	Apache	
PRONOM	web application/service	Database of file formats, software products and other technical components required to support long-term access to electronic records and other digital objects.	various	conversion, format recognition	only service, software not available	maintained		free	free service, software not available	
RODA	framework / system	Complete digital repository, functionality for all the main units of the OAIS model. Ingesting, managing and providing access to the various types of digital content. Standards: OAIS, METS, EAD and PREMIS.	various	storage, replication	Java, VirtualBox	maintained		open source	open source	KEEP Solutions
SAFE	framework / system	System for policy driven collaborative archival replication. It creates an overlay network across other storage systems. Currently, it supports LOCKSS.	various	storage, replication	Java	maintained		open source	Apache	Library of Congress
SCIDIP-ES	framework / system	Set of tools from CASPAR tailored for earth science.	various	acquisition, search, metadata extraction		in progress	ongoing			SCIDIP-ES
SIARD Suite	file processing tool	Converts databases into a collection of easy-to-handle XML files in SIARD standard (and back), preserving content, relations and metadata. It allows to view the primary data and to edit the metadata according to each organization's individual policy.	databases, documents	conversion, acquisition, metadata extraction		maintained		free	free, but closed and redistribution not allowed, support paid	Swiss Federal Archives
Sleuth Kit and Autopsy	file processing tool	A library, framework, and set of command line tools to analyze disk images. Autopsy is GUI for Steuth Kit.	disk images	acquisition	C/C++	maintained		open source	open source	Brian Carrier
SPAR	framework / system	Distributed Archiving and Preservation System". Data replication and monitoring of possible corruptions. File format transformations. Digital signatures. External access.	various	acqiusition, integrity check, conversion, storage, replication,		maintained		paid	soft written for BNF commercially, probably not available	National Library of France (BNF)

				view						
ssdeep	other	Computing context triggered piecewise (fuzzy) hashes (CTPH). Helps finding similar sequences in different strings.	text, binary	comparsion	C++	maintained		open source	GPL	ManTech International Corporation
Taverna	framework / system	Workflow Management System – a suite of tools used to design and execute scientific workflows and aid in silico experimentation.	various	various	Webservices	maintained		open source	LGPL	myGrid
TeleportPro, TeleportUltra, TeleportExec	search/indexing tool	Download Web pages locally. Up to ten simultaneous retrieval threads, access password-protected sites, filter files by size and type, search for keywords, and much more.	webpages	search, acquisition		maintained		open source	commercial, free evaluation version	Tennyson Maxwell Information Systems, Inc.
tesseract-ocr	file processing tool	Optical character recognition (OCR) tool. it can read a wide variety of image formats and convert them to text in over 60 languages.	text, graphics	conversion	C++	maintained		open source	Apache	Most of the work on Tesseract is sponsored by Google
TextGrid	framework / system	Virtual research environment humanities scholars. Tools and services for the creation, analysis, editing, and publication of texts and images.	text, graphics	various	Java, Eclipse	maintained		open source	open source	TextGrid
TubeKit	search/indexing tool	Extract YouTube video links from any webpage, extract video data, collect text comments for videos, extract a YouTube users' profile data	video	acquisition, metadata extraction	PHP, MySQL, Python	maintained		open source	CC Attribution-Noncommercial-Share Alike 3.0 US License	
Universal Document Converter	file processing tool	Conversion of documents into PDF, JPEG, TIFF or other graphical files. The underlying basis of the program is the technology of virtual printing.	documents, graphics	conversion	only executable available	maintained		open source	commercial, free version for test purposes avialable	fCoder Group, Inc
Unpaper	file processing tool	Post-processing tool for scanned sheets of paper, especially for book pages that have been scanned from previously created photocopies. The main purpose is to make scanned book pages better readable on screen after conversion to PDF	graphics	conversion	C	unknown	last modification: 2007	open source	GPL	Jens Gulden
Viewshare	web application/service	Platform for storing data in different formats, including metadata and creating views of the data to be used on a webpage.	documents, text, graphics	view, conversion	WWW	maintained		open source	only running service available, software unavailable	Library of Congress
Voyant Tools	web application/servi	web-based reading and analysis environment for digital texts (statistics)	text	acquisition,search	only service, soft not	unknown	last info from 2009	unknown		Hermeneuti.ca

	ce				available					
Warc Manager	search/indexing tool	Browse, search, and analyze archives of web crawl data. Lightweight database web application, indexes and provides browsing interface to a collection of warc data.	webpages	search, acquisition, metadata extraction	Java, Tomcat, MySQL	not maintained	last modification: 2002	open source		AdAPT project
Wayback Machine	search/indexing tool	Digital archive of the Web and other information on the Internet.	webpages	search, acquisition, metadata extraction	Perl, Java	unknown	last modification: 2011	open source		Internet Archive
Web Curator (WCT)	search/indexing tool	Workflow management application for selective web archiving. Uses Hettric.	webpages	search, acquisition, metadata extraction	Java, Tomcat, relational DB	maintained		open source	Apache	NL of New Zealand, British Library, IIPC
WebWhacker	search/indexing tool	Offline Browser, filtering file types, prepare CD with grabbed data, etc. More features than Grab-A-Site	webpages	search, acquisition		maintained		paid		Blue Squirrel
wget	search/indexing tool	non-interactive download of files from the Web	webpages	acquisition	C	maintained		open source	GPL	GNU
xcorrSound	file processing tool	Compares sound waves using cross correlation.	audio	comparsion	C++	maintained		open source	GPL	Open Planets Foundation
Xena	file processing tool	Detecting the file formats of digital objects; converting digital objects into open formats for preservation.	documents, graphics, text, webpage, e-mail, audio	format conversion and detection	Java	maintained		open source	GPL	National Archives of Australia
xpdf	file processing tool	PDF viewer, text extractor, PDF-to-PostScript converter, and various other utilities.	documents, text	view, extraction, conversion	C++	unknown	last modification: 2011	open source	GPL, commercal	Glyph & Cog, LLC
EUDAT PID service	web application/service	Persistent Identifier Service for data objects stored in EDUAT infrastructure. Based on handle system. Provided among others by SARA in Netherlands.	various	storage, replication	Java	in progress	pilot/pre-production)	open source	TBC	EUDAT project
EUDAT Safe replication service	framework / system	Data replication service for reliable long term storage and availability of the data. Exploits iRODS and microservices for data replication, implementing policy rules and meta-data handling (e.g. PIDs). Targeted to large user communities.	various	storage, replication	iRods, Java, PHP	in progress	production version available (09.2013)	open source	TBC	EUDAT project
EUDAT Simple store	framework / system	Researcher's data storage service with a user friendly interface and meta-data support. Youtube-like interface. Targeted to individual scientists.	various	storage, replication	iRods, Java, PHP	in progress	pilot version available (09.2013)	open source	TBC	EUDAT project
EUDAT Metadata service	framework / system	Metadata service supporing the EUDAT registered domain of data and communities with well-described and stable metadata offers. Will be extended to other meta-data providers.	various	metadata extraction	iRods, Java, PHP	in progress	pilot version available (09.2013)	open source	TBC	EUDAT project

EUDAT AAI	web application/service	AAI system supporting access to EUDAT services an federated identity systems.	various	access control	Java	in progress	pilot version available (09.2013)	open source	TBC	EUDAT project
EUDAT Data staging	framework / system	Supports researchers in transferring large data collections from EUDAT storage to HPC facilities for data processing purposes. It offers simple to use tools for managing data transfers. Computation results can be re-ingested to EUDAT infrastructure.	various	storage, replication, exploitation	iRods, Java, PHP	in progress	production version available (09.2013)	open source	TBC	EUDAT project

Table 9. DCH-RP tools and services registry snapshot (basic structure, some fields removed for greater readability)

Table below depicts the usage of the extender registry structure for 15+ tools in order to provide evaluations for additional assessment criteria. For better readability selected columns are removed. Full content of the registry can be examined in the online version.

Name	Usefulness (1-5)	Generality of solution (1-5)	User-friendliness (1-5)	Quality of results (1-5)	Assessment source
Archivists Toolkit	not tested (problems with installation)	not assessed: (problems with installation)	1: Complicated installation - beyond the expertise of non-IT staff	not assessed: (problems with installation)	DCH-RP WP5 PoCs: RA
AVS Document Converter	4: Usefull for small files. Batch conversion not reliable.	4-5: supports many documents formats;	3-4: Easy to install (4), easy to use (3)	2-3: Issues with PDF files (conformance to PDF/A needs to be checked). Quality of JPEGs and PNGs must be checked after conversion.	DCH-RP WP5 PoCs: RA
AVS Image Converter	3-4: Deals well with typical convertions (e.g. JPEG to PNG). Issues with large files (>100MB TIFFs)	4-5: supports many image formats	3-4: Easy to install (4). Ease of use (3).	3: issues with convertion to PDF	DCH-RP WP5 PoCs: RA
DSpace	2: usefull for preserving and enabling easy and open access to digital content including text, images, moving images, mpegs and data sets	Not assessed (problems with installation)]	1: difficult to install	Not assessed (problems with installation)]	
eCulture Science gateway	1: usefull for single file uploads; copying data collections not doable; no support for automated meta-data association	Not assessed	4: easy to use	Not assessed	DCH-RP WP5 PoCs: Belspo, ICCU
Heritrix	1: Works only on Linux.	Not assesed	1: Difficult to install.	Not assesed	DCH-RP WP5 PoCs: RA
HTTrack	4-5: Suitable for downloading websites and giving access to them for users. Not tested in large scale.	5: suitable for archiving most types of websites.	4-5: Simple to install and use.	4-5: Archived websites very similar to orignal versions.	DCH-RP WP5 PoCs: RA
Universal Document Converter	3-4: Userfull for single files. Batch conversion impossible. Not able to handle lots of objects. Problems with large TIFF files.	5: supports many formats	3-4: Ease of use for simple cases: 4. Integrates with the system as a virtual printer. Batch processing difficut (3).	Results of PDF conversion not necessarily in the specific PDF/A-1 format. Need to be tested using additional tools.	DCH-RP WP5 PoCs
Warc Manager	Not assessed - problems with installation	Not assessed - problems with installation	1: Difficult to install	Not assessed - problems with installation	DCH-RP WP5 PoCs: RA
Warc Tools	Not assessed - problems with installation	Not assessed - problems with installation	1: Difficult to install	Not assessed - problems with installation	DCH-RP WP5 PoCs: RA

Web Curator Tool (WCT)	Not assessed - problems with installation	Not assessed - problems with installation	1: Difficult to install.	Not assessed - problems with installation	
Xena		4: works for many documents formats	3-4: Easy to use for batch conversion. However requires some trial and error effort to nderstand how it works.	2: Usefull for XML, TIFF, jpeg, csv, ODF, MS office documents. Possibly not useful for binary data. Not all formats are converted properly. E.g. there were problems with converting Excel and ODF files. Issues with national characters.	DCH-RP WP5 PoCs: RA
Scoremodel	4: usefull for performing simple auditing of the collection integrity		4: user friendly; simpler than the DRAMBORA	Not assessed	DCH-RP WP5 PoCs: KIK-IRPA, ICCU
ROND (Riksarkivet OpeN Data)	4: overall general usablity of this type of tool is high; see notes on generality	1: this tool is bound to particular meta-data model (ADDML) with limited usage scope (Sweden, Norway);	3-5: avg. easy to use (3); easy to install (5)	5: The program behaved as expected and gave the correct results	DCH-RP WP5 PoCs: RA
IBM Tivoli Storage Manager	2: Usefull for bit-level preservation of files. Limited functionality from DP perspective. Simplistic interface. No format recognition. No data hashes exposed to user. Manual verification of data consistency necessary.	1. not general purpose tool	1: Difficult to install.	Provides good results (reliable storage).	DCH-RP WP5 PoCs: EVKM
A-PDF DJvu to PDF	Not tested - does not work on Windows	1: Does not work on Windows 7 as declared on website. Only linux version.	Not tested - does not work on Windows	Not tested - does not work on Windows	DCH-RP WP5 PoCs: RA
Snappy Wev Archiving Tool (SWAT)	Not assessed - problems with installation	Not assessed - problems with installation	1: Difficult to install	Not assessed - problems with installation	DCH-RP WP5 PoCs: RA

Table 10. DCH-RP tools and services registry's assessment criteria and measures applied to selected tools (some fields removed for greater readability)

ANNEX 2: WP5'S PROOF OF CONCEPT ASPECTS AND GRADES

Tables below present the tools and services assessment criteria (aspects) considered by Proofs of Concept conducted in collaboration with DCH institutions in the confines of DCH-RP's WP5.

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.