Digital Preservation Services: State of the Art Analysis

by

Raivo Ruusalepp and Milena Dobreva

Table of Contents

_Toc324364357

List of tables

List of figures

Executive Summary

This report presents an overview of the state of the art in service provision for digital preservation and curation. Its focus is on the areas where bridging the gaps is needed between e-Infrastructures and efficient and forward-looking digital preservation services.

Based on a desktop study and a rapid analysis of some 190 currently available tools and services for digital preservation, the deliverable provides a high-level view on the range of instruments currently on offer to support various functions within a preservation system. These functions, featuring pre-ingest and transfer, ingest, storage, digital object analysis, preservation planning, access and re-use represent a life-cycle process-oriented approach in preservation.

The analysis in the report shows that some functional entities have a far better offering than others. Amongst the most popular tools and services (over 25% of the total) are those that support digital object analysis (including file format identification and transformation as well as digital object quality analysis and characterisation). Metadata extraction and generation tools which are also essential at ingest and preparation of digital objects for preservation, account for 30% of the identified tools. Currently, the best represented tools play only a supportive role in the digital preservation life-cycle, while the offering of instruments that orchestrate the atomic tools and services into holistic preservation solutions is still low.

The report illustrates typical scenarios of using digital preservation services by memory institutions in some DC-Net partner countries (Italy, Estonia, Hungary).

The general conclusion of the report is that the digital preservation services market is still in its infancy. Although there is no shortage of software tools to aid digital preservation, sustainable services based on well-documented software and offering user support are in short supply. The study identified a lack of widely accepted comparison metrics and service maturity models in the domain of digital preservation that would support selection of tools and services. To address this gap, the report proposes the development of a roadmap that would define digital preservation as an infrastructure service for cultural heritage sector and include benchmarking digital preservation tools as one of its core services.

The report also provides a glimpse into future developments that are likely to influence the digital preservation landscape in the next coming years, including:

1. Transparent enterprise-driven models for digital preservation that will help to identify specific and generic components of the preservation function.
2. Launch of self-preserving objects – initially likely to be simple objects (text, images) that will still make a difference for the cultural heritage sector as the primary caretaker of these data types.
3. Increased flexibility in digital preservation architectures – based on granular or layered structures (e.g. SaaS, PaaS, IaaS) that are easy to adapt to a variety of preservation scenarios.
4. Clearly defined sets of metrics or benchmarks for comparing preservation tools and services and their performance.
5. Terminology and standards are no longer likely to converge along professional community borderlines – besides being interoperability across time, digital preservation will generate pressure for interoperability in real time, bringing along the need to agree on terminology.

1. Introduction: Scope and Purpose

This document is a report ordered by the partners in the DC-NET project.[1] It derives from the projects deliverable D3.1 "Digital Cultural Heritage Services Priorities Report" (Justrell et al., 2011) that highlighted long-term digital preservation as the top priority service where e-Infrastructures and cultural heritage institutions should collaborate.

1.1 Background

The Digital Cultural Heritage (DCH) sector is producing large volumes of digital content that needs to be safely stored, preserved and curated over time to allow for efficient resource discovery and re-use. Recent years have seen an upsurge in common search and retrieval tools for distributed digital collections but in the area of digital preservation cultural heritage institutions or their digitisation programmes are mostly acting as sole players. The DCH sector in general is yet to harness the benefits of shared preservation solutions, like those offered by common e-Infrastructure layers.

The importance of long-term preservation and its complementarity to digitisation efforts was highlighted in report of the Comité des Sages (Reflection group on bringing Europe's cultural heritage online) that clearly stated the DP mandate of memory institutions (The New Renaissance, 2011, p. 6):

- "Preservation is a key aspect in digitisation efforts. Digital preservation is also a core problem for any born digital content. The organisational, legal, technical, and financial dimensions of long term preservation of digitised and born digital material should be given due attention.
- The preservation of digitised and born digital cultural material should be the responsibility of cultural institutions – as it is now for non-digital material.
- To avoid duplication of effort by companies operating across borders and by the cultural institutions, a system could be envisaged by which any material that now needs to be deposited in several countries would only be deposited once. This system would include a workflow for passing on the copy to any institution that has a right to it under national deposit legislation."

In the last decade, the European Commission has supported over a dozen research and development projects that addressed a range of memory institutions' needs: ERPANET, DELOS, DPE, CASPAR, PLANETS, PROTAGE, LiWA, SHAMAN, PrestoPRIME, KEEP, APARSEN, ARCOMEM, BLOGFOREVER, SCAPE. However, statistical data on memory institutions and their involvement in digitisation and preservation, show that almost half of the cultural institutions surveyed do not have preservation plans in place (NUMERIC, 2009, p. 47). While digitisation of collections is widespread, best practices and standards have been identified and co-operation models are well established, digital preservation is not yet in the same position. There is general awareness of its importance, and support from a number of solutions and tools on offer, but a significant proportion of memory institutions still have to find their own way for implementing preservation within their specific business model realities. In the area of digitisation the questions who, how, with which tools and equipment, at what quality, etc. have been largely answered, but in the area of digital preservation these questions still need to be addressed.

Hence it is not surprising that the DC-Net survey (Justrell et al., 2011) identified long-term digital preservation as the highest priority area where new models of co-operation and shared services are required. The current deliverable was produced as a follow up to this survey and answers a clearly identified need across the heritage professionals' community.

1.2 Structure of the Report and Target Audience

This report has seven sections: the introduction is followed by a presentation of the digital preservation landscape looking into the business case for digital preservation and the digital preservation/curation life-cycle; it makes the connection with e-Infrastructures in the humanities and arts, and finally explores

[1] http://www.dc-net.org/

the drivers for new models in automating digital preservation arriving to the innovative area of preservation services. Section 3 presents examples of solutions addressing various digital preservation functional entities, based on a scoping study of 191 software tools. Section 4 explores some case studies from four European countries: Italy, Estonia, Poland and Hungary. Section 5 provides an analysis of maturity of digital preservation services. Section 6 summarises some of the gaps identified throughout the report into a roadmap for the future.

This report is written with several types of readers in mind, representing a range of stakeholders:

- Management (including executive management, information and project management): these readers will find section 6 (Roadmap) of particular interest; if they are not familiar with digital preservation, section 2 introduces the Landscape of Digital preservation. Project managers would find the discussion of service maturity in section 5 of interest.
- National Research and Education Network (NREN) practitioners are the target audience of the complete report. They can explore the preservation-related issues relevant to the services they already provide to the research and education communities or anticipate to provide to DCH.
- Producers/depositors of digital content would benefit most from sections 2 and 3 that discuss digital preservation and its tools.
- Regulators might find particularly useful sections 5 and 6: section 5 addresses maturity of services which is an area of multiple viewpoints and models; section 6 outlines steps for assisting a change in the domain of services for digital preservation for cultural heritage institutions.
- System architects of preservation solutions and e-Infrastructures would benefit from the country examples in Section 4; there are also numerous examples from digital cultural heritage domain across the text which would help these professionals to understand better preservation-related needs of this specific community.
- Solution providers will find in sections 3 and 6 considerations for how to transform their tools into services.

The report does not consider end-user viewpoint of digital cultural heritage collections as it is focussed on preservation activities that are internal to memory institutions operation.

2. The Digital Preservation Landscape

2.1 Introduction to Digital Preservation

2.1.1 Why do digital objects need extra care and preservation?

The importance of preserving digital objects is well understood. Hardware and media obsolescence, lack of support for older computer formats, human error as well as malicious software all can lead to loss of digital objects. Preservation, however, is not concerned only with sustaining single digital objects. To be used meaningfully in the future, digital objects should be preserved in context which makes them understandable to the future users. It is often said that digital preservation is interoperability over time; however this formulation has an element of speculation: we cannot be aware of the future hardware, software and business modalities and the task of long-term preservation thus has, by definition, uncertainty built into it.

Preservation is a complex activity not only because of the increasing complexity of digital objects, but also because the context of use, too, needs to be re-created, which means sustaining not only the data, but also any specific software which was used to work with them, and the technological infrastructure. The gradual expansion of preservation towards various types of objects is presented on Figure 1.

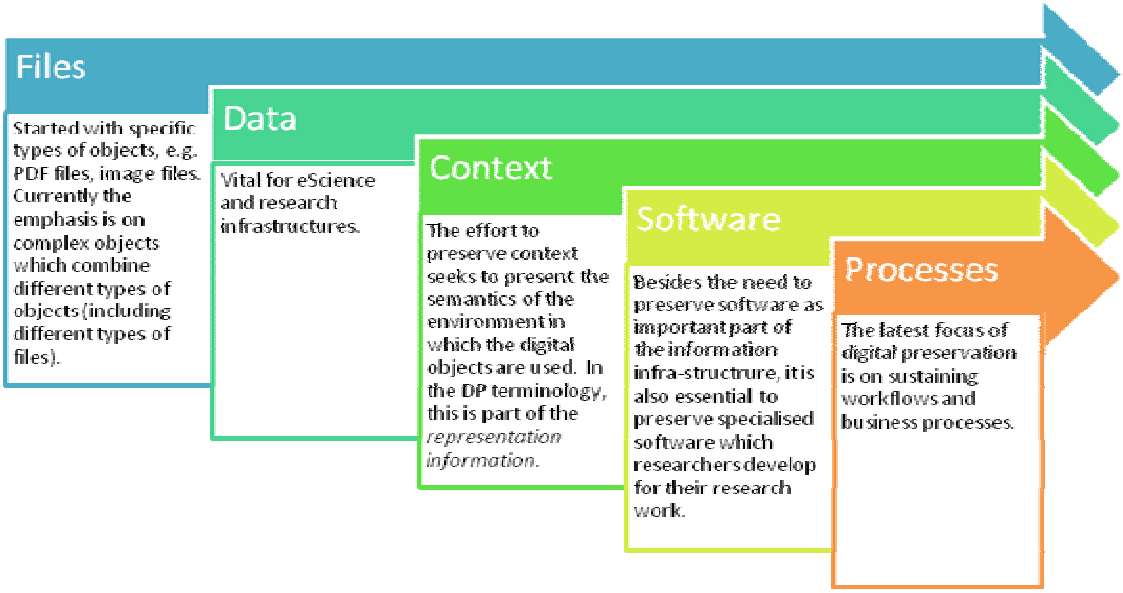

Figure 1: Evolution of digital objects addressed by digital preservation.

All these types of digital objects are relevant for digital preservation within cultural heritage institutions as well as in the humanities research. Although in many cases the emphasis is on the preservation of computer files, it is important to analyse the need to preserve software; the context of digital objects which is necessary for their future use, and any processes which also need to be preserved.

2.1.2 What is Digital Preservation and Curation?

Digital preservation is defined by the DigitalPreservationEurope project as "a set of activities required to make sure digital objects can be located, rendered, used and understood in the future".[2] A more comprehensive term 'digital curation' is often used in parallel with digital preservation. It has wider coverage and involves "maintaining, preserving and adding value to digital data throughout its life-cycle".[3] The key challenge in preserving usability of digital objects over time is overcoming technology obsolescence but a set of other issues around managing collections of digital objects is also involved.

[2] http://www.digitalpreservationeurope.eu/what-is-digital-preservation/
[3] http://www.dcc.ac.uk/digital-curation/what-digital-curation

2.1.3 What are the main strategies for digital preservation?

There are several strategies for sustaining the use of digital objects in the future (see Figure 2).

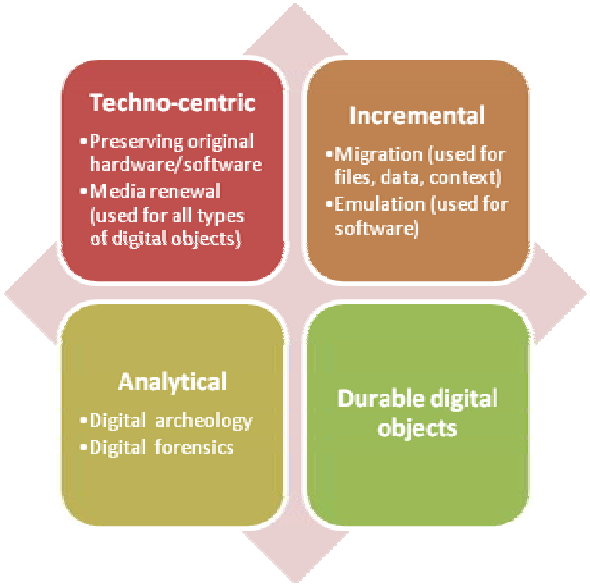

Figure 2: Major strategies for digital preservation.

The techno-centric strategy (Chue Hong, 2012) aims to preserve original hardware and software in a usable state in the future. It involves regular storage *media renewal* to make sure that the physical digital objects are not corrupted. Incremental change relies on constant migration of digital objects into new formats, to avoid format obsolescence. For software products this is done through emulation – which involves recreation of older software environments for newer equipment. Analytical strategies are currently based on techniques used in computer forensics. The underlying logic for this strategy is to apply specialised methods for recovery of objects which are in demand in the future instead of 'mass preservation' which does not seem realistic, having in mind the volume of digital information. The pioneering work in this domain was called *digital archaeology* (see Ross and Gow, 1998). Yet another strategy seeks for ways of changing the formats of the digital objects in a way which allows the objects themselves to invoke preservation actions. Such objects are called *durable digital objects* (see for example Gladney, 2008). The first three strategies require rigorous organisation of processes in organisations; the fourth one is still under development.

It is essential to emphasize that preservation can be seen as passive (bit-level storage that only takes care of sustaining the physical object) and active digital preservation that also takes care of preserving both content and the context of use.

All these strategies outline the principles of preservation; in practice they are implemented within archival lifecycles that integrate various tools and/or services. These lifecycles can be specific to organisations, depending on the type of objects they hold and their target users. The following section looks at one generic life-cycle model to understand the preservation process and discusses the stages that preservation involves. This demonstrates the complexity of the process that digital preservation systems need to support and helps to explain the range of tools and services currently on offer.

2.2 The Digital Preservation Process

2.2.1 Stages of the Digital Archive Life-Cycle

The diversity of digital objects and types of institutions that are responsible for their preservation create a variety on the level of tools used in practice, but the underlying process could be described as universal. The pivotal standard in the domain, ISO 14721:2003 *Space data and information transfer systems – Open archival information system – Reference model*, widely known as the OAIS model, is a

functional framework that presents the main components and the basic data flows within a digital preservation system. It defines six functional entities that synthesise the most essential activities within a digital archive: ingest, preservation planning, archival storage, data management, administration and access. Recently, these six stages have been combined into smaller number of use-cases that preservation systems address, e.g. a report of four major national libraries in Europe looks at three core functions – ingest, retention and access (BL, KB, DNB, NB, 2010).

The OAIS model looks at data stored in the digital archive as a fluid object that can (co-)exist as three types of information packages – submission (SIP) is used to transfer data from the producer to the archive, archival (AIP) is used for the archival storage and preservation, and dissemination (DIP) is used within the access function when consumers request archived materials. Table 1 presents a brief description of each OAIS functional entity; Chapter 3 below maps preservation tools and services to these entities and demonstrates that most of them implement the functional entities only partially.

Table 1: Description of the OAIS functional entities.

Functional entity	Description	Functions implemented within this entity
Ingest	This entity provides the services and functions to accept Submission Information Packages (SIPs) from Producers (or from internal elements under Administration control) and prepare the contents for storage and management within the archive.	Ingest functions include receiving SIPs, performing quality assurance on SIPs, generating an Archival Information Package (AIP) which complies with the archive's data formatting and documentation standards, extracting Descriptive Information from the SIPs for inclusion in the archive database, and coordinating updates to Archival Storage and Data Management.
Archival Storage	This entity provides the services and functions for the storage, maintenance and retrieval of AIPs.	Archival Storage functions include receiving AIPs from Ingest and adding them to permanent storage, managing the storage hierarchy, refreshing the media on which archive holdings are stored, performing routine and special error checking, providing disaster recovery capabilities, and providing AIPs to Access to fulfil orders.
Data Management	This entity provides the services and functions for populating, maintaining, and accessing both Descriptive Information which identifies and documents archive holdings and administrative data used to manage the archive.	Data Management functions include administering the archive database functions (maintaining schema and view definitions, and referential integrity), performing database updates (loading new descriptive information or archive administrative data), performing queries on the data management data to generate result sets, and producing reports from these result sets.
Administration	This entity provides the services and functions for the overall operation of the archive system.	Administration functions include soliciting and negotiating submission agreements with Producers, auditing submissions to ensure that they meet archive standards, and maintaining configuration management of system hardware and software. It also provides system engineering functions to monitor and improve archive operations, and to inventory, report on, and migrate/update the contents of the archive. It is also responsible for establishing and maintaining archive standards and policies, providing customer support, and activating stored requests.
Preservation Planning	This entity provides the services and functions for monitoring the environment of the OAIS and providing recommendations to ensure that the information stored in the OAIS remains even if the original computing environment becomes obsolete.	Preservation Planning functions include evaluating the contents of the archive and periodically recommending archival information updates to migrate current archive holdings, developing recommendations for archive standards and policies, and monitoring changes in the technology environment and in the Designated Community's service requirements and Knowledge Base. Preservation Planning also designs IP templates and provides design assistance and review to specialize these templates into SIPs and AIPs for specific submissions. Preservation Planning also develops detailed Migration plans, software prototypes and test plans to enable implementation of Administration migration goals.

Functional entity	Description	Functions implemented within this entity
Access	This entity provides the services and functions that support Consumers in determining the existence, description, location and availability of information stored in the OAIS, and allowing Consumers to request and receive information products.	Access functions include communicating with Consumers to receive requests, applying controls to limit access to specially protected information, coordinating the execution of requests to successful completion, generating responses (Dissemination Information Packages, result sets, reports) and delivering the responses to Consumers.

As a reference model, the OAIS standard does not imply a specific design or formal method of implementation (cf. Lavoie, 2004). Instead, it is left as an exercise to the reader to develop their own implementation by analysing existing business processes and matching them to OAIS functions. One of the confusing aspects for practical implementation has been the lack of active digital preservation (e.g., migration, emulation) as a separate functional entity.

Other models of the digital preservation process have been developed that do include the active digital preservation processes as well. For example, the recent DCC Digital Curation Life-Cycle Model[4] presents the core digital preservation activities in a wider context that also include appraisal and disposal:

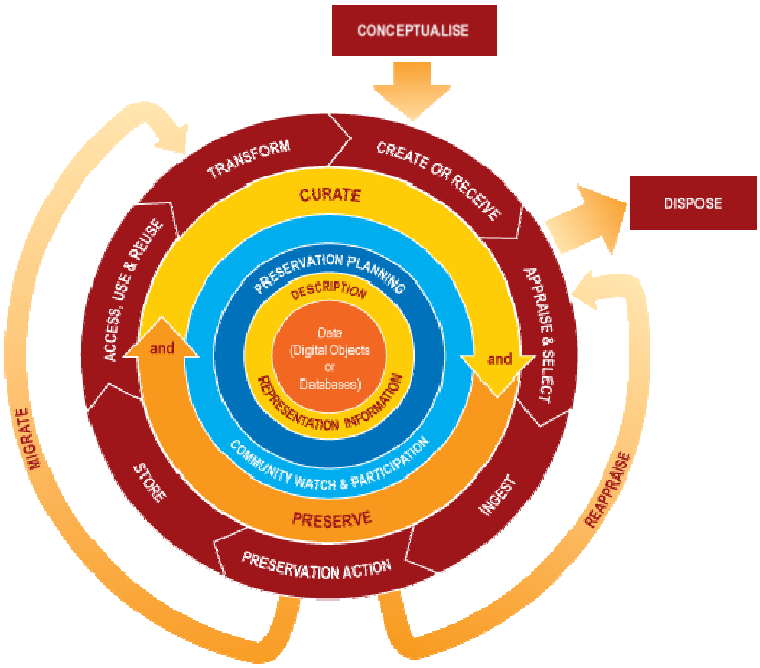

Figure 3: The DCC Digital Curation Life-Cycle Model.

Whichever model is followed, the preservation function is always interconnected with a number of other functions that together form the digital archive. Digital repository software is, thus far, the most common system for supporting digital archive management.

2.2.2 Digital Archive Software

Over the past decade, automation of preservation functions has mainly been seen within the context of holistic software solutions that provide digital collection management as well as digital preservation tools. The digital repository software or digital archive software solutions have dominated the preservation software market while not always providing support for active digital preservation.

Since digital repository software has been available as open source, it has become very popular, especially for research libraries as 'institutional repositories'. Commercial software developers, like

[4] http://www.dcc.ac.uk/resources/curation-lifecycle-model

IBM,[5] Tessella,[6] ExLibris[7] and others, have developed dedicated software systems for digital archive management. While very practical as digital collection management tools, not all repository software solutions offer support for long-term digital preservation: "notably DSpace and Fedora, have promoted support for preservation as a key feature. In contrast, the first software designed for IRs, EPrints, has until now offered less explicit support for preservation" (Hitchcock et al., 2007).

The Preserv2 project summarised that "The market for preservation services for repositories is undeveloped, and is unlikely to be developed until repositories establish clearer requirements, policy frameworks and budgets" (Hitchcock, 2009, 18). Thus the first recommendation of the project was: "Promote a joint approach to repository preservation by repositories and preservation organisations and specialists" (Hitchcock, 2009, 19).

The existing systems, tools and their services where mapped to six entities of the OAIS functional model in 2007 to analyse how well the OAIS concepts were implemented in practice. Table 2 shows the number of systems, projects and tools studied by the CASPAR project[8] that have components matching OAIS functional entities (note that the Administration functional entity was excluded from analysis). The low number of preservation-specific solutions which implement OAIS functional entities is evident.

Table 2: Coverage of OAIS functional entities in digital archive systems (source: CASPAR project).

OAIS functional entity	Specific focus on preservation (11)	Preservation by implication (14)	Other (9)
Ingest	3	3	0
Archival storage	2	7	0
Data management	0	4	0
Preservation planning	5	6	0
Access	0	3	4

Preservation planning is the most popular functional entity in the designated preservation applications; archival storage and preservation planning are the most popular OAIS functional entities within the systems that include preservation by implication.

2.3 Digital Preservation Tools and Services

While the integrated digital repository management software offers little support for actual preservation activities, a range of independent software tools has been developed for individual preservation tasks. Digital preservation is implemented through complex procedures; breaking these life-cycle processes down into smaller manageable tasks is one of the rationales for providing services that address a clear issue and solve it in an efficient way. This approach has become particularly relevant in distributed environments (like e-Infrastructures) and also for smaller institutions or projects that do not have the capacity to develop bespoke solutions covering all preservation functions. Specialised services can also be integrated into in-house preservation systems where they can help to resolve a granular issue without compromising any of the primary functions of the preservation system.

[5] http://www-935.ibm.com/services/nl/dias/
[6] http://www.digital-preservation.com/
[7] http://www.exlibrisgroup.com/category/RosettaOverview
[8] The CASPAR *Review of the State of the Art* (2007) (http://casparpreserves.eu/Members/cclrc/Deliverables/review-of-state-of-the-art-1/at_download/file.pdf) analysed 34 preservation systems, repositories, project efforts and studies. Eleven of these systems had a specific focus on preservation, 14 provided preservation functionality as part of their solution, and 9 included tools/standards which could be implemented in the preservation domain but were not created especially to be used in it.

"Services" is a term with different contexts of use and has lately become very popular through the service oriented architecture (SOA) concept. It is also used to refer to larger sets of consultancy-led activities.[9] Gartner Research has offered a definition for an IT service that is widely used:[10]

> "IT services refers to the application of business and technical expertise to enable organizations in the creation, management, optimization or access to information and business processes."

Terminological caveat

- This report uses the term 'tools' for granular software products.
- The term 'services' are used as unitary software components for which there is a body that offers customer support, access conditions are clearly defined, and user documentation is available.
- When service-oriented architectures (SOA) are used in the design of preservation systems, the term 'services' is used to identify a particular architectural layer.
- In the context of business models, the term 'service' has a more generic meaning, for example in context of Software as a Service (SaaS) – a software distribution model in which vendors host applications and make them available to customers.

An approach that emerged a few years ago and represents a step away from integrated digital archive systems is the one of **microservices**. These allow to flexibly combining specialised solutions for preservation depending on the requirements of the institution. They are defined as follows:[11]

> "Micro-services are an approach to digital curation based on devolving curation function into a set of independent, but interoperable, services that embody curation values and strategies. Since each of the services is small and self-contained, they are collectively easier to develop, deploy, maintain, and enhance. Equally as important, they are more easily replaced when they have outlived their usefulness. Although the individual services are narrowly scoped, the complex function needed for effective curation emerges from the strategic combination of individual services."

Microservices for digital preservation are currently under development at the California Digital Library (Merritt repository) (see Abrams et al., 2010), for the Electronic Records Archives at the US National Archives using iRODS (see Rajasekar et al., 2010, 60), and are also used in the open archival information system Archivematica.[12]

The debate on the applicability of this approach is on-going, for example (Challis, 2010):[13]

> "I'm not convinced the specs for these are well defined enough for general purpose use yet, but I can see the technique in general being very useful. If the same microservices can be called through multiple interfaces (command line, REST, etc.), then it should in theory make them language agnostic."

Work continues on the thorough decomposition and analysis what constitutes a microservice and how various microservices can be orchestrated so that the major requirements for authenticity and integrity of preserved digital objects are not compromised. Further work is needed on elucidating the granularity and the requirements towards various microservices.

The next section of the deliverable will look at the typical services in the digital preservation domain.

---

[9] For example in 2005 the Digital Preservation Coalition in the UK published the second edition of its Directory of digital preservation repositories and services in the UK (Simpson, 2005) which features 6 archives, 11 data services, 6 deposit libraries, 4 libraries, 1 research centre, 5 research councils as well as 3 private sector data services and 4 companies active in consultancy and development.

[10] http://www.webify-services.com:8180/articles/IT%20Services%20Market%20Definition.pdf, p. 13

[11] http://www.cdlib.org/services/uc3/curation/index.html

[12] http://archivematica.org/wiki/index.php?title=Development_roadmap

[13] http://blogs.ecs.soton.ac.uk/webteam/2010/11/15/notes-on-sits-the-scholarly-infrastructure-technical-summit/

3. Overview of Digital Preservation Tools and Services

3.1 Sources Used for the Analysis

On the overall landscape of digital preservation solutions, we could argue that services are currently an experimental area. Through a rapid desk research and analysis of current offering, 191 software tools and services were identified (see the full list in Appendix A). These are typically used in different preservation environments, but it is clear that they are used for an atomic task or a set of tasks. However, it was not always possible to determine whether they meet the criteria of a service (see definitions in previous chapter). Atomic tools can easily be developed into services and for this reason the study also looks at tools that solve a clearly identified problem. The conceptual overlap between tools and services is also clearly demonstrated in recent overviews of digital preservation services.

The main sources for identifying tools and services for this report were:

- A survey of the CAIRO project[14] on tools that can be used to create metadata packages in METS for ingest, identified 54 tools under 15 categories, featuring identifier creator, metadata extractor, digital signature creator, format identifier, as well as very generic tools such as antivirus software (Thomas et al., 2007).
- The National Digital Infrastructure Preservation Program[15] in the U.S. lists a collection of 38 tools and services that were used and/or developed by their partner institutions.
- The U.S. Library of Congress lists 10 tools for preservation metadata implementation supporting PREMIS.[16]
- A mash-up experiment organised by the AQuA project[17] compared 44 tools and services[18] to automatically detect quality issues in digitised collections for both ingest and checking the status of stored objects.
- The blogs of the OpenPlanetsFoundation (OPF)[19] are a valuable source of information on tools and services that are currently being developed and tested. Eight tools discovered on the OPF webpages were included in this study.
- DigiBIC[20] project offers a platform for presenting tools developed by EC-funded research and development projects in the field of digital libraries and content which could be exploited by research centres and small and medium-sized creative businesses. Thus far, the portal lists some 30 tools, five of which are preservation-related and appear under the „Archives/Metadata/Search" heading.
- SourceForge[21] offers a platform for publishing, searching and downloading open source software. The search for terms „preservation", „ingest", „web archiving" returned 81 tools in total, which were downloaded 2450 times during one week.[22] This demonstrates quite a significant interest in open software tools; downloads were made from virtually everywhere in the world excluding African countries. SourceForge was also the source which provided the highest number of tools for this analysis.

Figure 4 shows how many preservation tools/services were discovered using a particular source; the total number is 249 but with some tools repeated across the sources, the number of different individual tools is 191. The most popular tools are DROID and JHOVE – both mentioned in four sources.

[14] The project aimed "to bring together existing tools in a documented, automated, integrated workflow, to produce repository-independent metadata packages, in the form of METS files, that could provide the basis for long term life cycle management" (Thomas, 2008).
[15] NDIIPP Partner Tools and Services Inventory, http://www.digitalpreservation.gov/partners/resources/tools/index.html
[16] http://www.loc.gov/standards/premis/tools.html
[17] http://www.jisc.ac.uk/whatwedo/programmes/inf11/digpres/aqua.aspx
[18] http://wiki.opf-labs.org/display/AQuA/AQuA+Mashup+Tool+List
[19] http://www.openplanetsfoundation.org/
[20] http://www.digibic.eu
[21] http://sourceforge.net/
[22] The tools and the number of downloads were checked during the same week in November, 11-18.11.2011.

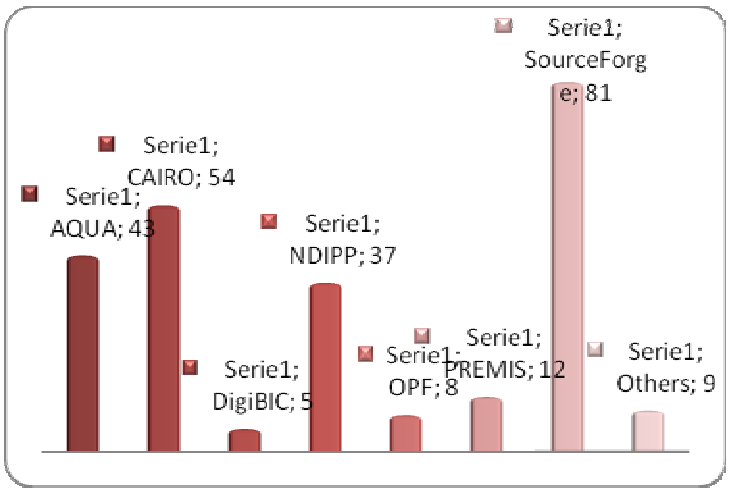

Figure 4: Distribution of digital preservation tools and services across sources.

Consideration Points

- The number of tools and services currently offered for various preservation-related tasks is significant, majority of them offered as open source from the SourceForge.
- In most cases these tools originate from the research community and are made openly available for other members of the community.
- However, this requires institutions willing to adopt the tools to have a considerable technical expertise in order to configure and include them within their existing institutional infrastructure.
- To what extent this usage model suits (smaller) cultural heritage institutions needs to be further investigated.

3.2 Clustering of Tools and Services

The visual representation of all the 191 tools is not easily readable and is included here as an illustration only. Figure 5 shows the tools clustered according to the basic task they support. In some cases these are complete OAIS functional entities, in other cases smaller tasks.

Figure 5: Clustering of digital preservation tools covered in this study.

Note: Areas that the tools are addressing are on the edges of the diagram. The cluster of functional entities is presented in red; smaller atomic tasks in green; various registries, test beds and specifications which are an additional component of the preservation infrastructure are in pink.

The overall picture shows an abundance of tools and services, but there is a significant difference in the decomposition of digital preservation processes into services. For example, the service can be for digital preservation but also specific to a particular institutional repository (e.g., the TIAMAT ingest service for the Tufts Digital Repository (Kumar, Kaplan & Rubinger, 2008)), or the service can be a global

monitoring tool, like the PERPS (Piloting an E-Journals preservation Registry Service) that maintains a registry of electronic articles and tracks if there is a preservation system where the electronic publications are deposited (Burnhill & Guy, 2010).

This study presents an overview of digital preservation tools and services structured along the digital preservation lifecycle (see Section 2.2 above) in five separate clusters:

- Preparation for ingest
- Transfer and ingest
- Archival storage
- Preservation planning
- Access to digital objects.

In defining the categories, the functional entities of the OAIS model were used, with some elaboration based on a recent description of high-level preservation functions in report of four major national libraries in Europe (BL, KB, DNB, NB, 2010). The report uses three core functions – *ingest, retention* (in this context used in the sense of digital storage and associated management functions) and *access*. The OAIS data management and administration functions are not included since these are covered by more generic information technology solutions.

The distribution of identified tools and services shows that most of them deal with very practical tasks. Figure 6 below shows four areas of digital preservation with the highest number of tools on offer. The most popular type of tools is related to metadata, since tools for metadata extraction are used during pre-ingest and ingest, but are not *per se* preservation tools. Digital object analysis is part of both ingest and preservation planning. Web archiving is a popular digital preservation domain at the moment.

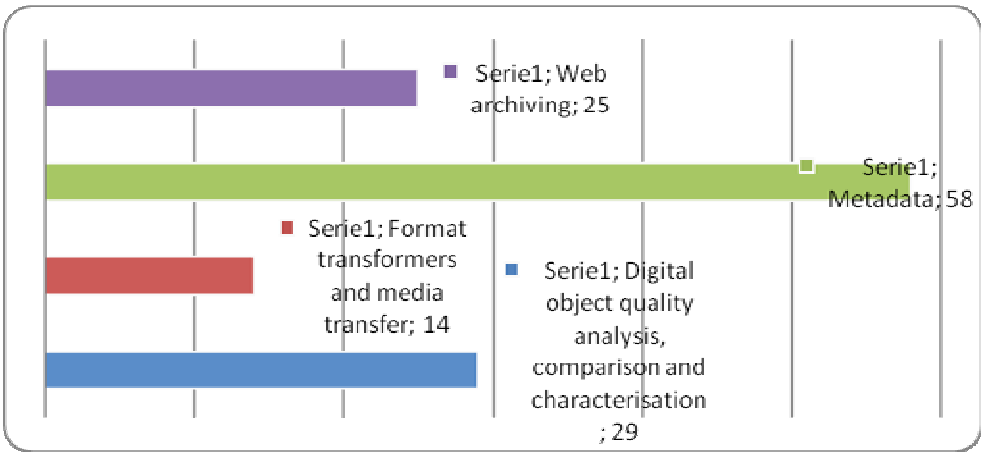

Figure 6: Distribution of digital preservation services according to their purpose.

On the other hand, tools and services addressing major functional entities are not the ones most popular (see Figures 7 and 8):

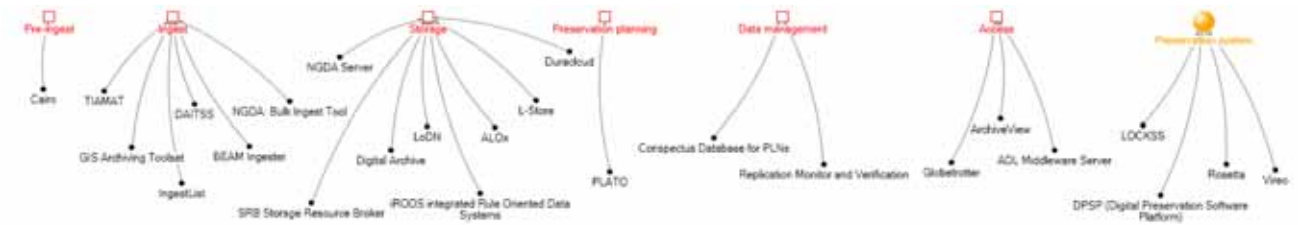

Figure 7: Tools corresponding to OAIS functional entities.

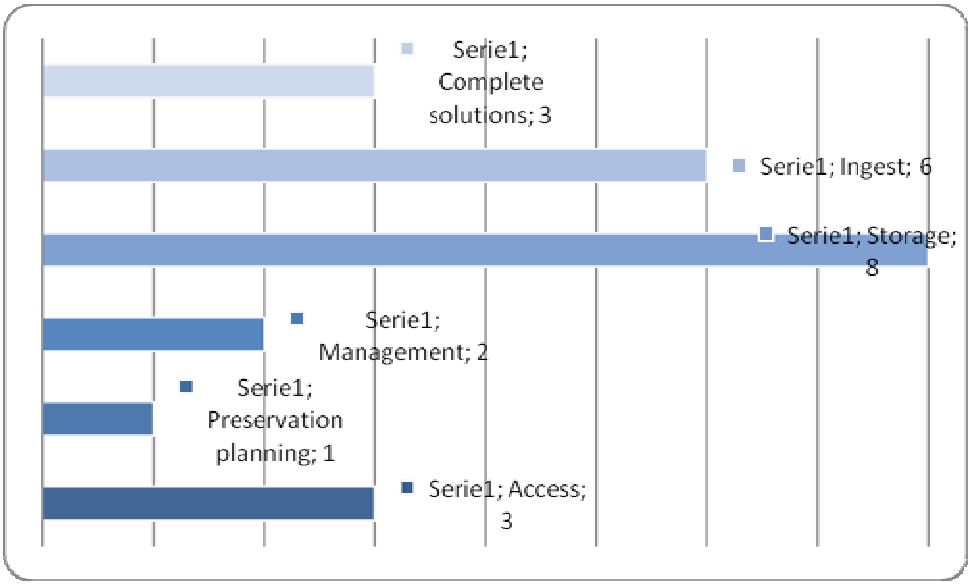

Figure 8: Number of tools corresponding to individual OAIS functional entities.

The following section provides examples of tools that support the core preservation functional entities.

3.3 Digital Preservation Tools and Services

3.3.1 Preparation for Ingest

3.3.1.1 Introduction

In the area of creation and appraisal of digital objects there are three identifiable major work areas:

- Standardisation of the communication between the information producer and an archive;
- Development of tools supporting generation and transformation of metadata; and
- Development of tools for automated or semi-automated appraisal.

These are three very different tasks and typically solved with diverse tools that can be integrated into larger systems for preparation of digital objects at the ingest stage.

3.3.1.2 Examples

3.3.1.2.1Standardisation of the pre-ingest phase

Several standardisation efforts exist to link the OAIS reference model with tasks carried out prior to information is handed to the archive. The Producer Archive Interface specification (PAIMAS, 2004) was developed as a recommendation identifying, defining and providing structure to the relationships and interactions between a Producer and an Archive. It identifies four phases in the process of transferring information, suggests actions which should be carried out during each phase, and provides a general framework which facilitates the identification and/or development of standards and software tools to be used within the ingest process.

A more detailed breakdown of the pre-ingest activities was proposed by the Public Record Office of Northern Ireland (PRONI) (Smyth, 2006) that synthesized their experience. The PRONI approach foresees preliminary research on information needed for archiving of records (including file formats, metadata, migration, appraisal and access), which precedes the four stages suggested in PAIMAS.

3.3.1.2.2Tools supporting generation and transformation of metadata

While the PAIMAS standard looks at the formalised steps in communication between the producer and the archive, it does not detail how the preparation of the digital objects for ingest should take place.

The range of tasks that are needed prior to transfer to an archive depends on the specific situation – what are the digital objects, what metadata they have, is there need to convert objects, their metadata, or both. The need for automating such processes is obvious, because human processing of single objects on mass scale is not viable. The software tools that are used for these tasks are not *per se* preservation tools: they include a range of software that, for example, deal with:

- Quality of a particular type of digital objects, e.g., audio or image files.
- Characterisation tools that can identify the file format of a digital object, check if it is conformant with the format, and extract other technical properties.
- Format migration tools that transform digital objects from a particular format to another format (the tools are usually format-specific).
- Storage media copying and transformation tools.
- Tools for analysing and generating metadata.
- Tools for checking checksums or integrity of objects.
- Tools for capturing web content.

Appendix 1 includes 29 tools for digital object quality analysis, comparison and characterisation; 14 format transformers and media transfer tools; and 58 tools dealing with metadata generation.

For example, The Metadata Extraction tool[23] developed by the National Library of New Zealand extracts technical preservation metadata. The formats currently supported are:

- Images: BMP, GIF, JPEG and TIFF.
- Office documents: MS Word (version 2 to 6), Word Perfect, Open Office (version 1), MS Works, MS Excel, MS PowerPoint, and PDF.
- Audio and video: WAV, MP3 (normal and with ID3Tags), BFW, FLAC.
- Mark-up languages: HTML and XML.
- Internet files: ARC.

Metadata extraction results in elements of digital preservation metadata schemas, but currently covers a very narrow set of popular file formats. A more structured example (see Table 3) presents a typical file format identification service to extract file information.

Table 3: An example of a file characterisation service – FITS.

File Information Tool Set (FITS)[24]
What digital preservation task does it address?
FITS identifies, validates, and extracts technical metadata for various file formats. Such a tool is needed at all stages of a preservation process where automated metadata extraction is needed –most typically during ingest to form an AIP from the received SIP, or when analysing files on demand.
How does it work?
FITS starts consecutively eight open source tools and converts their output into a single XML output document. The feature of combining several tools and providing single output allows for extracting more technical metadata compared with single tools. Detailed information on how FITS resolves cases of conflicting or nonmatching outputs from the tools was unavailable.
How can use it?
FITS can be downloaded from http://code.google.com/p/fits/downloads/list and used under GNU Lesser GPL. Java 1.6 is required with the release 0.3.0 available at time of writing this report. FITS can be used as a command line tool or through java API – both require technical skills for integrating the tool into a particular preservation system.
What support is available for it?
User guide is available on http://code.google.com/p/fits/wiki/user_guide. It is more a technical manual than an end user manual.

[23] http://meta-extractor.sourceforge.net/

[24] Not to be confused with the file format Flexible Image Transport System (FITS).

Does it use other existing services/tools?
Yes, FITS uses JHove, Exiftool, National Libray of New Zealand Metadata Extractor, DROID, FFident.
Developed by
Harvard University Library
Future prospects
Not clearly defined in terms of policy for adding newly emerging tools and future development/support.
Further information
Stern R., McEwen S. (2009) FITS – The File Information Tool Set. In: Open Repositories 09. http://smartech.gatech.edu/jspui/bitstream/1853/28508/2/179-540-1-PB.pdf

This group of tools is numerous, partly because the same tools are used in other information systems. Yet the large offering makes it difficult to analyse the tools and make informed decisions because information on similarities and differences of tools is scarce. Comparable metrics on quality of these tools and their outputs would be an advantage for preservation specialists but is currently lacking.

3.3.1.2.3Tools for automated and semi-automated appraisal

Work on automated appraisal has in recent years advanced on the methodological level. Oliver et al. (2008) provided a summary of automated re-appraisal issues. Harvey and Thompson summarised three requirements for automated appraisal and re-appraisal (Harvey & Thompson, 2010, 319):

> "1) It is necessary to get a sufficient and useful set of metadata from digital objects to make automated technical appraisal worthwhile and meaningful. 2) It is necessary to get sufficient metadata from every format submitted to a technical re-appraisal process. … 3) For automated re-appraisal to be successful, other systems and processes will be required in order to implement the activity of re-appraisal and to act on its outputs. Systems would be required to record the appraisal decision process – for instance, so that comparisons can be made between repositories."

Automated appraisal is also considered in the context of web archiving and an approach based on events had been experimented with in the ARCOMEM project.[25]

3.3.1.3 Summary

Although the creation of digital objects is normally not considered to be part of digital preservation – preservation is mainly concerned with the question how to sustain objects that exist – it is essential to consider the following issues (in particular within the context of digital cultural heritage):

- Decisions taken during digitisation such as formats used and parameters of digitisation are directly related to digital preservation. The practice of creating and storing high quality master files in uncompressed formats is widely used across memory institutions and provides a solid basis for long-term digital preservation.
- The provision of metadata which, if incomplete, can potentially create a serious bottleneck in the collections management. Most frequent problem with the metadata supplied with digital objects is that it does not cater for preservation purposes and significant effort is required to enhance metadata. Nearly a third (58 out of 191) of tools identified for this report serve metadata-related tasks, most offering generation or transformation of metadata. Preservation metadata (as discussed in Caplan (2006)) are an area of continuing development. There are a number of metadata initiatives in this domain, most influential of them is the work on the PREMIS.[26]
- Automated appraisal is a complex activity and depends on the success of other types of tools; this is one of the domains where further developments in the near future can be expected.
- Currently there is no centralised resource that would collect information on existing tools to support decision-making and choice of tools in institutions. The JISC-funded Digital Preservation Console[27] study aimed to "assess the extent to which it might be possible to develop an intuitive

[25] http://ceur-ws.org/Vol-779/derive2011_submission_5.pdf
[26] http://www.loc.gov/standards/premis/
[27] http://www.jisc.ac.uk/whatwedo/programmes/preservation/dpconsolestudy.aspx

graphical user interface (GUI) to enable non-specialist information professionals to undertake a variety of preservation and information management tasks with a minimum of preservation-specific theoretical knowledge." The final report[28] presents some screenshots for such a system and considers 12 existing tools (characterisation and migration only). This represents an initial attempt to offer a system to support the selection of preservation tools but our study shows that the range of tools and their functions are considerably more complex. Since majority of cultural heritage institutions are quite small and possess little technical expertise, further work is needed to present the available options in a clear and intuitive manner for the general heritage specialist.

Consideration Points

- There is a plethora of metadata generation and file format transformation tools, but it is essential to establish solid comparison criteria for their quality and performance. Many tools are developed as research pilots and need to be further adapted for use in institutional practice.
- Automated appraisal depends on detailed metadata. Multiple tools for automated metadata generation are offered but the quality of the metadata they extract need to be studied further.

3.3.2 Transfer and Ingest of Digital Objects

3.3.2.1 Introduction

The OAIS standard describes Ingest as the functional entity that encompasses all functions required for transfer and archiving of digital objects in an archive. This is combined with change of responsibility for the preservation of digital objects – the materials are accepted into archive with contractual or administrative agreements. From this point of view, the ingest covers in parallel two separate activities:

- Technical processing to prepare digital objects for archiving and to transfer them to the archive.
- The contractual coverage of preservation-related activities.

The most important aspects for digital preservation at this stage are:

- File formats that determine the necessary technical steps during ingest;
- Authenticity, integrity and provenance data on digital objects as well as metadata;
- Completeness of metadata accompanying digital objects and the need to enhance metadata;
- Various transformations of objects that may be necessary for the export-import between software systems (the producer and the digital archive).

Further sub-divisions of the ingest process exist, for example, the Cost Model for Digital Preservation (CMDP) project in Denmark (Kejser et al., 2011, 7-8) that uses eight separate ingest tasks. The Ingest framework developed within the SOAPI framework[29] has a different structure and concentrates on the technical side of the Ingest (see Figure 9). There are also examples of looking at ingest as a set of services. A very straightforward definition of ingest services is provided in (BL, KB, DNB, NB, 2010, 13):

> "The term "Ingest Services" is a common denominator for the set of heterogeneous atomic services responsible for marshalling acquired content streams from producers (SIPs), and for performing all required actions against that content to obtain long-term preservation packages of appropriate quality (AIPs) to store in an LTP system."

These examples show that the composition and sequence of ingest activities differs considerably from institution to institution. This makes orchestration of smaller tools into solid workflow solutions a complex task.

[28] http://ie-repository.jisc.ac.uk/548/1/JISCDPConsoleFINAL.pdf
[29] http://www.e-framework.org/Default.aspx?tabid=1007

3.3.2.2 Examples

Among the 191 tools identified for this study:

- A relatively small number of tools (six) implement the ingest function fully.
- There are many tools that implement functionality which could be included in ingest as well as pre-ingest, depending on the particular institutional workflow. For example, the PRONOM register of file formats and their behaviour; and DROID that uses PRONOM to analyse files.
- Digital object characterisation tools are very popular – in total 29 tools were available.

Differences in content type, producer and the way material is submitted require many distinct processes that each employ a number of ingest tools and services in a tailor-made workflow. The (BL, KB, DNB, NB, 2010) suggests the following list of ingest services:

- *Ingest Services receive or take a SIP (metadata and/or content-data from a provider) and transform them into an AIP that meets the requirements of the institution. This includes tasks such as the validation and normalisation of the provided metadata and/or content-data, and metadata creation and enrichment.*
- *Ingest Services work with Catalogue Services to enrich the metadata (or create a link between an item in the LTP system and a catalogue record).*
- *Ingest Services provide persistent identifiers as required.*
- *Ingest Services help operators to monitor and control the ingest procedures.*
- *Ingest Services provide Metadata Services with metadata to hold in perpetuity.*
- *Ingest Services provide Archival Storage Services with files to hold in perpetuity.*
- *Ingest Services may be employed by Preservation Services to execute preservation actions and ingest the resulting package into the LTP system.*
- *Preservation Services work with Metadata Services and Archival Storage Services to enact Preservation Watch and Preservation Planning for the digital formats and metadata formats within the LTP system. Preservation actions will typically involve Delivery Services to retrieve content data and metadata from the store, and Ingest Services to ingest the migrated content-data and metadata back into the store.*

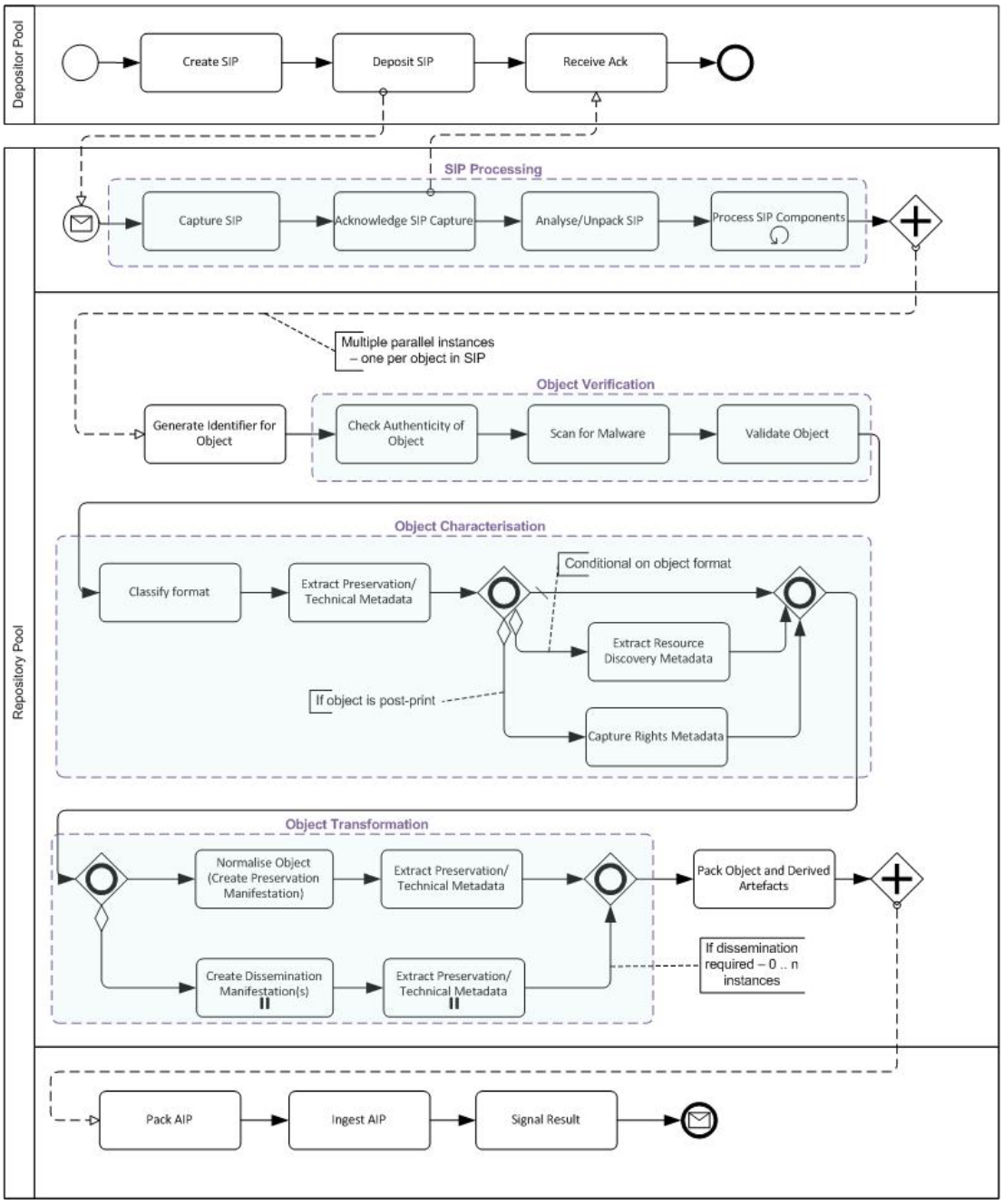

Figure 9: Example of Ingest framework from SOAPI project.

There is also a range of examples of ingest-related microservices (see Table 4):

Table 4: Examples of ingest-related microservices.

Name	Description	Creator	Version
Characterize and extract metadata[30]	Microservice in Archivematica. Identifies and validates formats and extracts object metadata using the File Information Tool Set (FITS). Adds output to the PREMIS files.	Consortium: Artefactual systems, Inc., UNESCO, et al.	
Merritt Characterization service[31]	Provides a mechanism for the automated examination of digital objects to determine their significant properties. Addresses four aspects: Identification, Validation, Feature extraction, and Assessment. Based on JHOVE.	University of California, CDL	

[30] http://archivematica.org/wiki/index.php?title=Micro-services
[31] http://www.cdlib.org/services/uc3/curation/character.html

Merritt Fixity Service[32]	Service checking for file integrity and corruption and related to authenticity. The Fixity service verifies the bit-level integrity by testing two values: filesize and message digest (such as an MD5 checksum).	University of California, CDL	Alpha; not available for download
Merritt Ingest Service[33]	The Ingest micro-service provides a means to add new digital content into the curation environment for active management by the Program. Using terms defined by the Open Archival Information System (OAIS) reference model, the Ingest service accepts Submission Information Packages (SIPs) and converts them into Archival Information Packages (AIPs). This process may involve the use of other micro-services.	University of California, CDL	Alpha; not available for download
msiGetDataObjAIP	iRODS microservice that gets the AIP of a data object in XML format	iRODS	
msiGuessDataType	iRODS microservice which guesses the data type of an object based on its file extension.	iRODS	

3.3.2.3 Summary: transfer and ingest services

In a recent presentation,[34] Adrian Brown from the Parliamentary Archives in the UK stated that "Ingest accounts for up to 90% of digital repository activity". Many issues during ingest are related to problems with digital objects – incomplete metadata which means that the gaps need to be filled in, or need to perform transformations of the objects. While the number of tools implementing ingest as a complete process are very few – 6 in our study, a considerable number, if not most of the tools covered by this study are relevant for ingest. All tools that generate metadata or characterise digital objects could potentially be used at the ingest stage.

Given the differing ingest workflows that institutions have, comparing ingest tools with a single metric is futile. The choice of appropriate tools will have to be made within a particular institutional setting.

Consideration Points

- There are several types of tools offered to support both pre-ingest and ingest: file format identification, metadata extraction and characterisation of digital object are the most popular ones.
- 'Holistic' ingest tools are most typically implemented for a particular repository software, e.g. ingest into DSpace or Fedora systems.

3.3.3 Archival Storage

3.3.3.1 Introduction

Archival storage can be seen as the heart of the preservation system – the functional entity that takes care of sustaining the bit-stream of digital objects and making them available in the future. In addition to storage media failure issues, the storage solution will increasingly have to support large scale digital archives and high speed access to archived material. Linden et al. (2005) identified the following functions of the storage service layer (p. 11):

> "The storage service layer would
> • allocate unique persistent (vendor-independent) identifiers to objects
> • bind each of these identifiers permanently to its object
> • guarantee the authenticity and integrity of objects
> • handle the reliable transport of each object as needed

[32] https://confluence.ucop.edu/display/Curation/Fixity
[33] https://confluence.ucop.edu/display/Curation/Ingest
[34] http://www.dpconline.org/events/details/38-studentconference?xref=38

- recover from any internal failure
- provide full service to users of another site if that other site was unavailable
- provide the means to integrate the physical storage with external systems."

Nowadays the archival storage solutions include grid and cloud architectures. The SHAMAN project looked at a grid-based solution for memory institutions and concentrated on three typical scenarios:[35]

- Indexing and archiving of book-like publications in depot libraries.
- Indexing and archiving of large-scale digitisation work, and
- Scientific publishing and archiving of heterogeneous interlinked material.

3.3.3.2 Examples

Different archiving scenarios are likely to require different storage architectures (e.g., the three scenarios in the SHAMAN project). Figure 10 presents an example of the first SHAMAN scenario that uses a storage architecture with specialised grid storage (iRODS) and involves a web server and several other components that had to be integrated together.

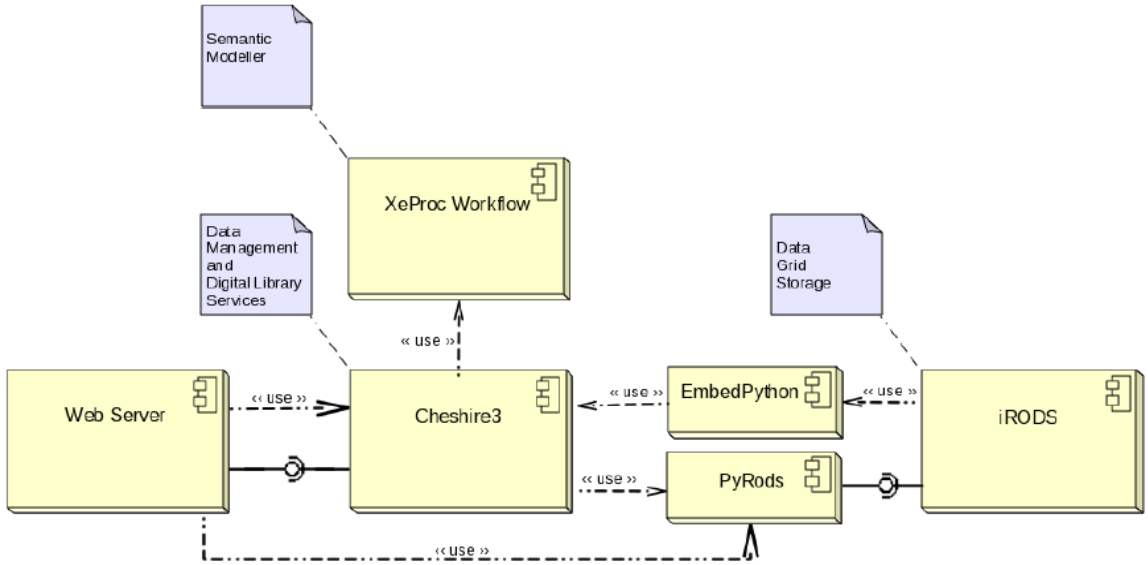

Figure 10: Components of the storage architecture of SHAMAN scenario for archiving book-like publications.

Another detailed view on archival storage components is presented by the PRESERV project where two institutions play the role of storage service providers. The digital preservation function here has a recursive structure and integrates storage and preservation within itself (see Figure 11). This differs from the SHAMAN project example above.

[35] http://shaman-ip.eu/shaman/sites/default/files/SHAMAN%20D11.2_Demonstration%20of%20distributed%20ingestion%20for%20Memory%20Institutions_0.pdf

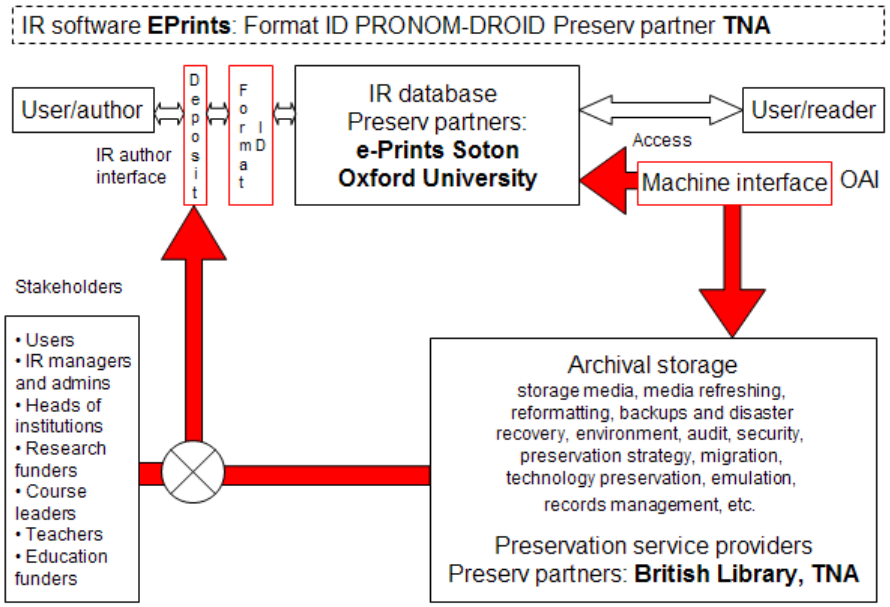

Figure 11: Preserv project's service provider model (Hitchcock et al., 2007)

The German TextGrid project[36] has implemented a grid-based long-term preservation repository (see also Ch. 4.2). Other examples of using the grid technology exist.[37]

3.3.3.3 Summary

There is a modest number (8) of tools that deal specifically with archival storage in particular. The examples provided in this section show that the archival storage is also being decomposed into smaller tasks that can be implemented by a combination of tools, but similarly to ingest, there is no common shared view on the components this functional entity is comprised of.

There are examples of institutions that store their digital collections in the cloud (not as the primary archive, but as a safety copy) and using the grid infrastructure. This number will grow over the coming years and the current "early adopters" will establish a more mature best practice.

Archival storage within digital preservation will continue to follow the trends in general storage hardware and software. Radical technological changes and the introduction of principally new storage solutions would cause more serious needs to migrate digital content and are, thus, unlikely.

Consideration Points

- Archival storage in memory institutions has to already address issues of very large volume and large numbers of objects.
- Grid and cloud storage are at the early adopter stage. More involvement of institutions would help to increase their use for preservation and help establish best practice.
- Archival storage will become more distributed in the future; virtualisation solutions will inevitably have to be considered for modern archives but best practice principles still need to be refined.

[36] http://www.textgrid.de/en/startseite.html

[37] See, for example, the USC Shoah Foundation archive of 52,000 video testimonies stored on the cloud through Nirvanix (http://dornsife.usc.edu/vhi/news/3294), or The COMETA Grid Infrastructure in Sicily (www.indicate-project.eu/getFile.php?id=173).

3.3.4 Preservation planning

3.3.4.1 Introduction

Preservation planning is the functional entity that helps to analyse the objects in a digital archive in order to support preservation risk management and decision making for digital preservation actions. One component of it is the so called technology watch that monitors file formats stored in the digital archive for their obsolescence and initiates active digital preservation tasks. The other component is monitoring the storage media for errors to initiate any passive digital preservation tasks needed to recover data from redundant storage. Preservation planning is closely connected with the institutional context and workflows in the digital archive, and depends on sound methodologies for assessing the risks within the preservation system.

3.3.4.2 Examples

The most substantial work to date on tools for preservation planning is from the Planets project. PLATO[38] is a web-based tool that supports preservation planning workflow – it automates the planning process using a range of registries and services for file format identification and characterisation of digital objects. This is another example of a function that makes use of the same components as pre-ingest and ingest. However, the purpose of using them here is different – to alert the preservation system when a format of a digital object becomes obsolete and to trigger for example migration action.

A core concept of this stage is the preservation plan that 'defines a series of preservation actions to be taken by a responsible institution to address an identified risk for a given set of digital objects' (Becker et al., 2009). Digital preservation tools do not resolve merely technical issues but also need to accommodate knowledge about the goal of preservation, the institutional policies, legal obligations, organisational and technical constraints, user requirements etc. The complexity of preservation planning decisions can be seen on the workflow diagram of PLATO[39] (see Figure 12).

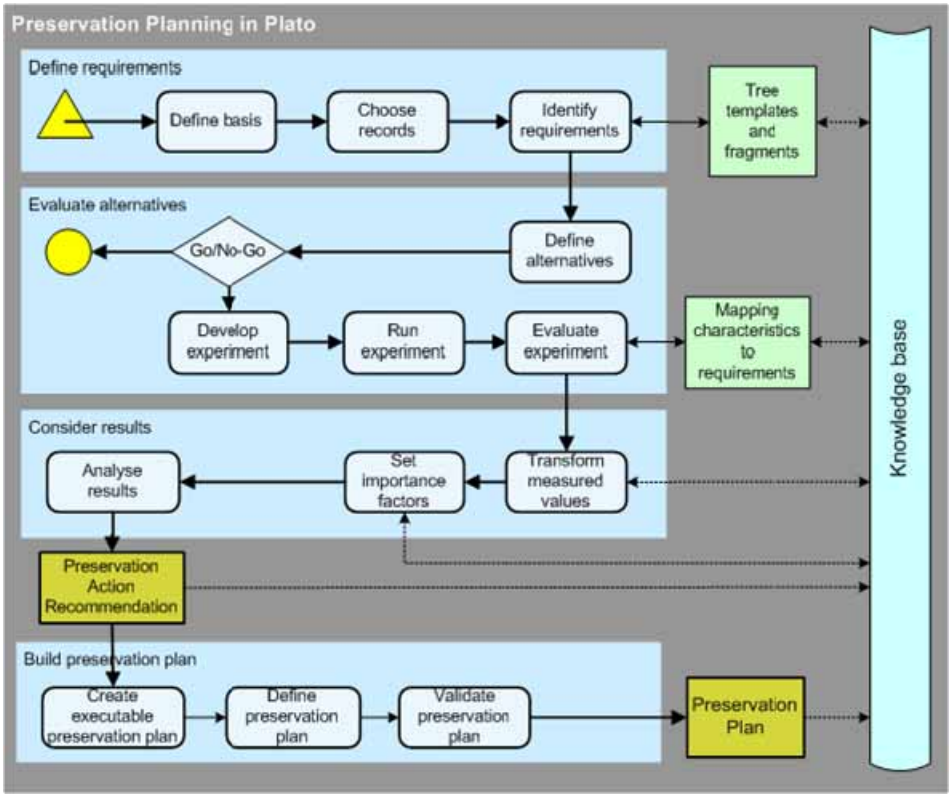

Figure 12: Preservation planning workflow in PLATO.

[38] http://www.ifs.tuwien.ac.at/dp/plato/intro.html
[39] http://plato.ifs.tuwien.ac.at:8080/plato/help/workflow.html

3.3.4.3 Summary

As with other functional entities, preservation planning can be seen as a combination of specialised tools such as characterisation, quality analysis and workflow management components. Currently, the PLATO tool developed by the Planets project is the dominant tool on the market.

Consideration Points

- The growing complexity of digital objects and especially the cases of multiple linked objects will increase the complexity of preservation planning.
- Preservation planning suits particularly well incremental approaches in digital preservation when the idea is not to break the sequence of format migrations. It is not relevant to analytical approaches such as digital archaeology.

3.3.5 Access to Digital Objects

3.3.5.1 Introduction

The access function builds a bridge between the digital archive and its users. The most important aspects of access function are user-friendliness, flexibility and ability to combine objects in repositories with other content (e.g. data with publications, semantic description layers).

The user expectations to digital archives are constantly and currently include for example multilingual access or easy use of aggregated digital resources. This means that access tools need to support not only access to digital objects but also integrate additional features based on the target user group.

3.3.5.2 Examples

Search and retrieval from digital archives usually relies on a catalogue or index that is linked with the stored digital objects. The term 'discovery' is often used for the first part of access (i.e. searching) and 'delivery' for the retrieval part of it. For example, the digital collections infrastructure at the National Library of Australia has a Delivery block which plays the role of stored digital objects (see Figure 13).

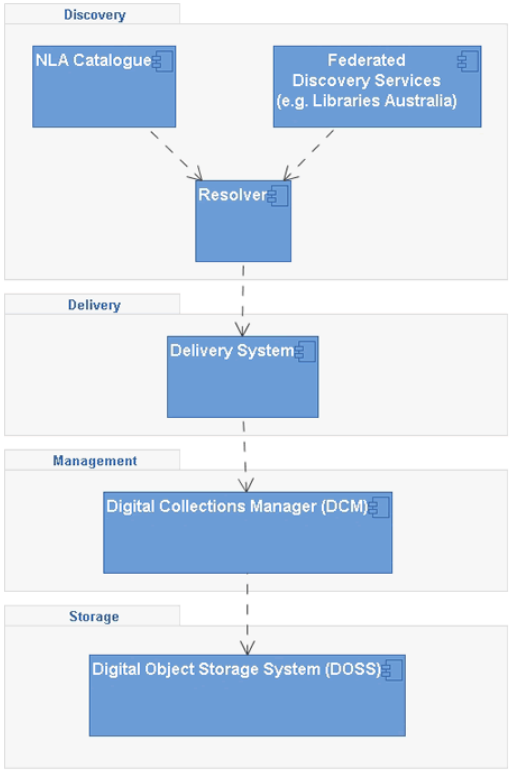

Figure 13: Digital collections infrastructure of NLA.

There are also emerging examples of systems that integrate access tools – for example Preservation, Management, Discovery and Delivery are seen as the three basic building blocks of the Stanford Digital Library System (see Figure 14).

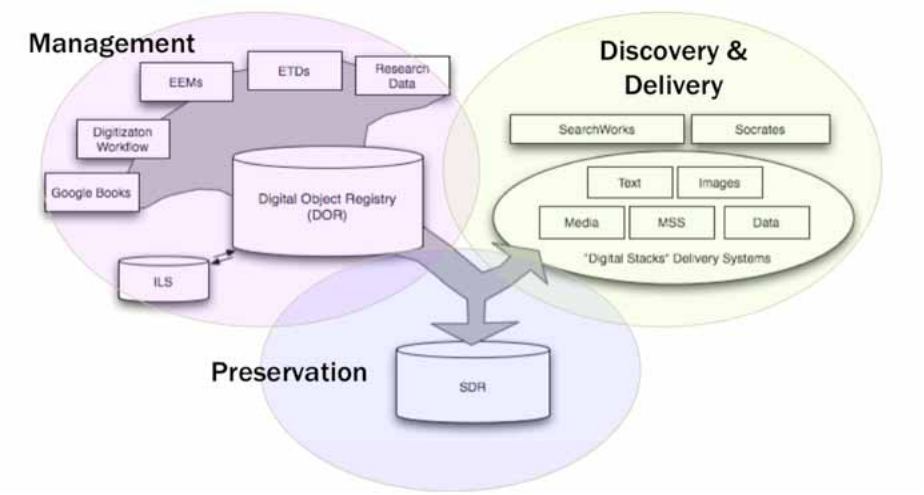

Figure 14: Stanford Digital Library ecosystem.

3.3.5.3 Summary

Access is digital archive's functional component that produces an information package meeting the consumer needs. It involves extraction of the stored digital object and its transformation (e.g. to a new file format) according to the user request. Hence it benefits from a multitude of tools for digital object transformation as well as metadata generation tools. At present access is often merged with resource discovery and delivery tools.

Consideration Points

- The terms 'access' and 'delivery' are often confused; sometimes access is hidden as a layer between the discovery services and the preservation system.

3.3.6 Registries as Preservation Tools

In addition to the tools described above, digital preservation is often making use of public registries. The registries collate and unify available information for easy access and to answer specific questions, for example: What format uses a particular filename extension? What platform can render a particular file format? What formats are supported by a particular software tool?

A **format registry** is a collection of records that characterize existing file formats. For example, a file format entry could include name and version number, characterization elements and links presenting dependencies with other formats. There is no consensus on how this information should be structured. Three examples of format registries with different approaches are PRONOM,[40] Unified Digital Format Registry[41] and IBM Preservation Manager.[42]

Most of the current advanced technologies in automated file format identification rely on some information from an internal or external format registry. The file identification and validation tools automatically generate file format data. The simplest mechanism for identification is to analyse the file extension and consult a registry of file extensions (e.g., File Extensions[43]). One problem with the use of file extensions is that they are not unique - e.g. the popular extension DOC has six possible associated

[40] http://www.nationalarchives.gov.uk/PRONOM/Default.aspx
[41] http://www.udfr.org/
[42] http://www-935.ibm.com/services/nl/dias/is/preservation_manager.html
[43] http://www.file-extensions.org/

formats. In addition, computer users on most platforms can make up their own file extensions or change the existing ones making judgments based on the file extension unreliable.

Another example of a registry is the Trustworthy Online Technical Environment Metadata Database (TOTEM)[44] developed by the University of Portsmouth in the KEEP project. TOTEM allows the checking of dependencies between hardware configurations, operating systems and software tools for the purposes of emulation.

3.4 Summary: Preservation Tools and Services

This section illustrated the variety of tools currently on offer for digital archives. Most tools have a specific and narrow function, e.g. characterisation of a particular file type, generation of particular metadata type, or analysis of the integrity of an object. Some tools exist that implement complete functional entities, as described in the OAIS standard but there is no clear or formal definition what makes a preservation tool OAIS-compatible (see Nicholson & Dobreva, 2009). The discussion of the digital preservation tools and their market is continuing in Chapter 5 below.

Consideration Points

- There is an **overlap** of what services/tools developed in different institutions do – new preservation systems and projects tend to develop their own services anew. For example both ACE and Checksum checker assist in integrity control and Merritt Fixity microservice is used for a similar purpose. The practical testing of software tools was outside the scope of this report but there clearly is an emerging need to better coordinate the effort that goes into tool development.
- **Granularity of services** and microservices needs further clarification – similar tasks are implemented through microservices or services/tools. There is a need for a better formalisation of the functional requirements for the specific services, as well as the risk they would impose on loss of significant properties.
- There are already examples of tools that make **use** of other **existing tools** and services, e.g. FITS. This is likely to lead commodisation of tools and further standardisation.
- The current state of tools and services requires improvement in terms of **ease of use**; user documentation understandable for end-users; and clear development/support plans for the future.
- Some tools are **platform-specific** which makes their use more difficult.
- The digital preservation workflow typically incorporates generic tools, e.g. virus checking, metadata generators or format identifiers, specific preservation services, as well as services that relate to higher-level management, especially distributed preservation environments. Further research is needed to analyse the necessary conditions under which the various services can coexist in a healthy digital preservation ecosystem.
- Most examples of tools and services have been conceived in the **archival or library community** while the museum community seems hardly involved in the development of preservation tools.

[44] http://keep-totem.co.uk/

4. Digital Preservation Stakeholder Landscape

4.1 Introduction

Analysis of the digital preservation stakeholder landscape will have to combine two points of view:

- A user-centric view that concentrates on requirements, development, testing and validation – represented by the digital cultural heritage community in the context of this report;
- A data-centric view using a proven design for generic infrastructure services for persistent storage, access and management – represented by the e-Infrastructures.

The user community consists of a number of members with different range of needs. Table 5 below provides a summarised view on the stakeholders and their interests.

Firstly, *content providers* who need to preserve their content. There are many types of content providers: i) national institutions (e.g. national libraries) with a high level production (digitisation) of digital content, ii) small institutions with little resources but a substantial need to solve their preservation problems ranging from policy recommendations to technical support, iii) publishers and other content creators who have a legal obligation to preserve the digital material they produce.

Secondly, policy makers and programme owners at different levels (national, regional and local) who invest in the development of research infrastructures and need guidance for future actions.

Thirdly, the user community could be final end-users who need the infrastructure to gain safe access to digital objects for their research or cultural consumption. This group will firstly include researchers but potentially all members of the general public.

There is also the research and development community that is active in the digital preservation domain and could either adopt or develop solutions.

Table 5: Stakeholders and interests and involvement in digital preservation.

Type of institution	Interest
Cultural heritage institutions	Management: informed decisions on preservation Curators: understanding and efficient use of tools/services
Institutions that design policies on national and international level	Monitoring, guidance, evaluation of progress
Academic and research institutions	Use digital objects for research and teaching that requires long-term preservation
e-Infrastructures	Preserve distributed resources and provision of services to a variety of stakeholders
R&D institutions active in digital preservation	Develop new solutions in digital preservation
Auditors	Perform evaluation of a number of aspects of digital preservation infrastructures, workflows, policies and procedures
Fund-makers	European Commission and other funding bodies that fund research and development in digital preservation and through this further the state of the art
End-users	End users or consumers are the beneficiaries of reliable digital preservation but this does not mean that they necessarily understand the domain and its issues in detail

Over the last years the number of specialists in digital curation and archiving has been growing but clearly there is a further need for specialists with university degrees and professional training.

4.2 e-Infrastructures and Digital Preservation

Digital preservation has recently been connected with research infrastructures. A row of successful projects to establish a preservation infrastructure for scientific research data include Alliance for Permanent Access (APA),[45] PARSEinsight,[46] Opportunities for Data Exchange (ODE),[47] Scipid-ES,[48] and

[45] http://www.alliancepermanentaccess.org/index.php/about/

Aparsen.[49] A quick survey of e-Infrastructures in the arts and humanities shows that there is a clearly expressed interest towards preservation and curation of research data.

For example, long-term preservation is recognised as an essential aspect of DARIAH[50] consortium: "DARIAH's mission is to foster and support high quality digital research in arts and humanities. Part of this remit involves promotion of the importance of sustaining and preserving data for future use, so that scholars can benefit from and build upon earlier work in their field" (DARIAH, 2010).

TextGrid,[51] a German research infrastructure aiming "to support access to and exchange of data in the arts and humanities by means of modern information technology (the grid)" is part of DARIAH.[52] The project released its TextGrid software v. 1.0 in 2011, after five years of development. The architecture of TextGird is a three-layer service implemented on grid storage.

Another major e-Infrastructure for the humanities, CLARIN[53] has been discussing preservation[54] and issued a short guide on long term preservation.[55]

Not all networks and consortia in the digital arts and humanities domain address preservation issues. For example, the recently started ESF-funded project Network for Digital Methods in the Arts and Humanities (NEDIMAH)[56] is focusing on new research methods but does not address preservation. Another recent example of an e-Infrastructure project is Collaborative EuropeaN Digital Archive Infrastructure (CENDARI)[57] that "will provide and facilitate access to existing archives and resources in Europe for the study of medieval and modern European history through the development of an 'enquiry environment'". This project has ambitions on improving access to historical resources but how it will approach preservation is not clear.

Compared to the infrastructures dealing with science data, digital cultural heritage is a domain where special attention needs to be paid not only to born-digital data but also to digitised objects. As Eric Meyer et al. (2009) highlighted (p. 9):

> "While digitization projects and programmes are somewhat simpler than many advanced Grid infrastructure projects, taken together they are arguably contributing to a growing research infrastructure that supports e-Research in the humanities, whether or not it is formally called e-Research or e-Infrastructure."

A necessary step to consolidate efforts of DCH stakeholders is to design a shared roadmap for the preservation of digital cultural heritage content, focusing on the use of existing e-Infrastructures for the research as a channel for the delivery of services to the digital cultural heritage sector.

---

[46] http://www.parse-insight.eu/
[47] www.alliancepermanentaccess.org/index.php/current-projects/ode/
[48] http://www.scidip-es.eu/
[49] http://www.alliancepermanentaccess.org/index.php/current-projects/aparsen/
[50] Digital Research Infrastructure for the Arts and Humanities. http://www.dariah.eu/; Willemse E., (2007) DARIAH Challenges for arts & humanities data curation. Presentation. http://ipres.las.ac.cn/pdf/Ellen.ppt.pdf
[51] http://www.textgrid.de/
[52] Neuroth, H. (2011) TextGrid – A Virtual Research Environment for the Humanities. Presentation. Available: http://www.clir.org/activities/registration/sympslides2011/neuroth.pdf
[53] Common Language Resources and Technology Infrastructure, http://www.clarin.eu/external/
[54] http://www.clarin.eu/node/293
[55] http://www.clarin.eu/files/preservation-CLARIN-ShortGuide.pdf
[56] http://www.esf.org/activities/research-networking-programmes/humanities-sch/current-esf-research-networking-programmes-in-humanities/nedimah.html
[57] http://www.geschkult.fu-berlin.de/e/fmi/arbeitsbereiche/ab_janz/projekte/Cendari/index.html

Consideration Points

- Major European e-Infrastructures in the arts and humanities are primarily concerned with the preservation of research data and less with digitised content. The efforts in data-intensive infrastructures need to be coordinated with the efforts of infrastructures that are entrusted with the care for digitised resources.
- e-Infrastructures addressing new research methods should also accommodate the preservation of specialised software tools that could be unique to research groups.
- The digital preservation scenarios for single institutions become considerably more complex when it becomes part of an e-Infrastructure. The preservation life-cycles involving e-Infrastructures need further in-depth analysis.

4.3 Country Examples

This chapter presents some example cases from DC-Net partner countries where cultural heritage institutions are using digital preservation services.

4.3.1 Italy

The interest in digital preservation in Italy is prominent and long-standing. In 2003, Italian professionals made an international survey[58] on legislation, rules and policies for the preservation of digital resources. In 2008, a round table on "Digital preservation in Italy: experiences face to face" was hosted by the National Library in Florence and was attended by 90 participants.[59] Rome has hosted workshops, among others, by the PLANETS project in 2008, and the KEEP project in 2011, as well as workshops of LIBER on digital preservation and DELOS summer schools in the same domain. The choice of major EC-funded projects to organise events in this country seems to respond to a well identified interest in this area and a professional community that is eager to follow the most recent developments.

The Digital Stacks project[60] is developing a long term digital preservation system for electronic publications that fall under the legal deposit law. The architecture of the solution is distributed between two sites of the National Library (in Florence and Rome) that allow deposits into the archive[61] and a dark archive for preservation only in Venice. The service is operated by Fondazione Rinascimento Digitale from Florence. The multiple sites offer redundancy of content (altogether six copies of each object is being kept) and the open source platforms used for managing the collections ensure few vendor dependencies.

The Consorzio COMETA[62] – the Sicilian Grid service provider – has developed the gLibrary platform[63] that provides a simple yet powerful system to store, organize, search and retrieve digital assets in repositories built on e-Infrastructures. This effectively hides the underlying technical details of the service from the end users and provides a user-friendly way to archive digital objects.

4.3.2 Estonia

The Estonian Ministry of Culture has been co-ordinating the preservation of digital cultural heritage under the auspices of a national strategy since 2003.[64] A network of competence centres has been established to support both digitisation and digital preservation in memory institutions.[65]

[58] http://eprints.erpanet.org/65/01/Dossier1_English_version_Full.pdf
[59] http://www.digitalpreservationeurope.eu/publications/reports/Report_roundtable_event.pdf
[60] See: http://www.indicate-project.eu/getFile.php?id=233
[61] See: http://www.depositolegale.it/
[62] http://www.consorzio-cometa.it/en/home
[63] http://www.consorzio-cometa.it/en/descrizione
[64] http://www.kul.ee/index.php?path=0x838
[65] http://digiveeb.kul.ee/index.php?id=10429

One of the national competence centres – the Estonian Public Broadcasting Company[66] is offering secure back-up storage to other memory institutions as a secondary storage site. The terms of this service have been negotiated by the Ministry of Culture who also funds the bulk of the cost.

The Restoration Centre "Kanut" is a competence centre for museums, offering specialised digitised services and also digital storage on behalf of museums.[67]

There is a common gateway to digital collections in memory institutions that harvests metadata from a variety of in-house catalogues.[68]

4.3.3 Hungary

NIIF,[69] the developer and operator of the e-Infrastructure in Hungary, together with Hungarnet, the association and representative of the e-Infrastructure users, have jointly been putting emphasis on the co-operation with the academic, research and digital cultural heritage communities. The requirements of these communities have been taken into consideration, as much as possible, when developing the service portfolio. However, approaching a truly service-oriented e-Infrastructure has turned to be possible just more recently, with the advent of virtualization, enabling IaaS, an attractive solution for the majority of e-Infrastructure applications – practically all of them, except the most demanding, most complex ones where the need for special joint treatment of the high complexity research problem and the extraordinary application aspects does not allow utilizing the standard and widely exploitable IaaS solutions.

As far as digital cultural heritage preservation and related activities are concerned, the IaaS is proving to be a promising way of using e-Infrastructures in most cases. NIIFI is well prepared to introduce these new approaches and offer them to cultural heritage sector, including networking, grid, cloud, HPC, and storage.

4.3.4 Poland

Many national e-Infrastructures are already offering basic digital archiving and storage services, including the Poznan Supercomputing and Networking Centre. The PSNC runs the PLATON project[70] that provides archiving and data storage service for research data from scientific and academic community. The services offered by PLATON include:

- Automated and transparent data replication, ensuring the durability of the stored data; file system-level metadata are replicated and protected against disasters, to ensure logical namespace consistency and accessibility.
- Data objects are persistently identified within the logical filesystem namespace; their physical location is masked by the virtual filesystem layer and the underlying logic.
- Abstract, universal data access interface: the virtual filesystem is accessible through WebDAV (over HTTPs), SFTP and GridFTP protocols.

Standard data access methods enable to:

- use the long-term storage service as a networked filesystem or (with additional tools) as the networked drive;
- integrate the service with the Content Management Systems (e.g. dLibra[71]), by using the standard libraries for C, C++, Java and others in order build customized service clients.

The PLATON approach lets users focus on their core business, by outsourcing the following long-term data management issues:

[66] http://arhiiv.err.ee/
[67] http://kanut.ee/index.php/digiteerimine
[68] http://e-kultuur.ee/?locale=en
[69] http://www.niif.hu/en
[70] http://www.man.poznan.pl/online/en/projects/50/PLATON.html
[71] http://www.man.poznan.pl/online/en/projects/20/dLibra.html

- data storage technology migration (e.g. due to the technology development) is performed on the service provider side, transparently to users;
- infrastructure development – equipment and software procurement, deployment and testing;
- infrastructure maintenance and operation – monitoring, failure tracking, handling and resolving.

These are some examples of early adopters of digital preservation as e-Infrastructure in different EU countries. They demonstrate that the traditional models of preservation management are beginning to evolve towards more distributed preservation architectures.

5. Analysis of Preservation Tools and Services

5.1 Comparing Preservation Tools and Services

The description of 191 preservation tools and services in Chapter 3 above clustered them according to stages in the archiving life-cycle. It did not attempt detailed comparison of tools and their functionality, because a task of this scale was not feasible within the scope of this project. One of the reasons why this task is gargantuan, is the lack of universal benchmarks and metrics for evaluating preservation tools.[72] A basic categorisation of these software tools is offered in Appendix 1.

Detailed analysis of most popular preservation software tools has been carried out elsewhere. For example, the SCAPE project carried out an evaluation of five popular format identification tools: DROID 6.0, Fido 0.9, Unix File tool, FITS 0.5 and JHOVE2 (van der Knijff & Wilson, 2011). All tools were tested using two identical datasets – a scientific journals set of 7796 archival packages, containing a total of 11892 files (including both content and associated metadata files) and a set of 8 large files. The evaluation framework included 22 criteria:

- Tool interface
- Language
- Coverage of file formats
- Output format
- Granularity of output
- Comprehensiveness and completeness of reported results
- Ability to deal with nested objects
- User documentation
- Stability
- Provision of event information
- Development activity
- License type
- Platform dependencies
- Extendibility
- Unique output identifiers
- Accuracy of reported results
- Fit to needs of preservation community
- Ability to deal with composite objects
- Computational performance
- Error handling and reporting
- Maturity and development stage
- Existing experience

Practical results from using the SCAPE characterisation tools are reported as they emerge.[73] A further study on criteria for evaluating preservation tools has been published by the SCAPE project that looks into criteria for functionality, performance, compatibility, usability, reliability, security, portability and maintainability (SCAPE D10.1, 2012).

The analysis in the SCAPE project highlights several specific features of existing tools:

- The existing tools often require technical competence for downloading and installation. Many currently available services or tools are far from being “download and play”, especially for not too technically-savvy users.
- Similar tools can have considerably different inputs and outputs.
- When tools make use of other existing tools or services, it is not always clear how quickly they will be able to accommodate any new versions of the dependant services.
- Future development and support plans of tools are not always clear.

The different scope and nature of tools that seemingly use the same approach and address the same issues is also illustrated on the level of microservices. Figure 15 compares three digital archiving systems that rely on microservices and demonstrates the difference in coverage of these services.

[72] See discussion of this in the Technical panel at the Aligning National Approaches to Digital Preservation conference in 2011 (http://www.educopia.org/events/ANADP/presentations and http://digitalplusresearch.blogspot.com/search?q=ANADP).
[73] See for example: http://www.openplanetsfoundation.org/blogs/2012-02-23-identification-tools-evaluation

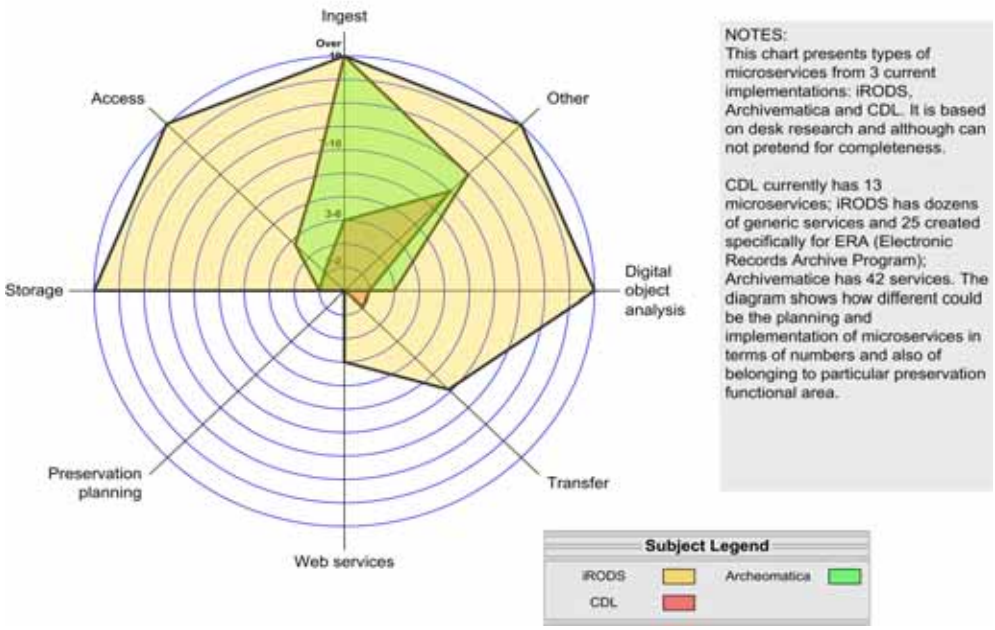

Figure 15: Focus of microservices in three current digital archive systems.

5.2 Enterprise Perspective on Preservation Services

Software services should ideally be flexible, low-risk, granular and address a clear user need. Service models can range from partnerships, memberships, subscription, menu pricing, or be offered on a pay-as-you-go basis.

While previous sections of this report addressed the nature of preservation and major functional entities that preservation systems should support, it is useful to analyse preservation also from the point of view of the organisation's business processes and stakeholders. Since preservation is part of digital objects' lifecycle, it has implications on processes and professionals within the institution.

The organisational structure of cultural heritage institutions varies and understanding the *specific* requirements and their implication on the preservation infrastructure is important. It could be argued that the current state of stand-alone preservation solutions is a result of uniqueness of digital holdings of every institution that require tailor-made approaches. A recent comparison of digital preservation provision across major European national libraries and the German Computer Game museum showed significant differences in the type of holdings which need to be preserved, collection policies, preservation systems and standards used.[74]

The continuing investment into in-house preservation systems is contributing to the lack of interoperability and fragmentation of resources into "digital silos". Stand-alone solutions that are not transferrable and interchangeable lead to fragmentation and do not offer economies of scale. Instead, shared solutions for creation, storage and use of digital resources, including the e-Infrastructures, will become the major component of the future knowledge economy.

In order to move ahead from the current situation where solutions are still quite specific, more work needs to be done in finding ways to define key institutional requirements in a standardised way. The use of enterprise architecture models is one possible approach because enterprise architectures seek to address system complexity while aligning technological developments with the institutional needs. There are a number of approaches for defining enterprise architectures; one of the popular ones is the

[74] The National Library of France develops its in-house preservation system SPAR, OAIS-compliant and based on the use of METS and PREMIS-compliant metadata; The Royal Library of the Netherlands uses the e-Depot system which is based on the IBM DIAS and uses extended Dublin Core bibliographic metadata; The German National Library deployed a combination of tools including kopal-DIAS, koLibRI and has developed its own preservation metadata format, LMER (KEEP, 2009, 54-59).

Open Group Architectural Framework (TOGAF)[75] and its eight-stage Architecture Development Method that help to manage requirements within complex systems, which is illustrated on Figure 16.

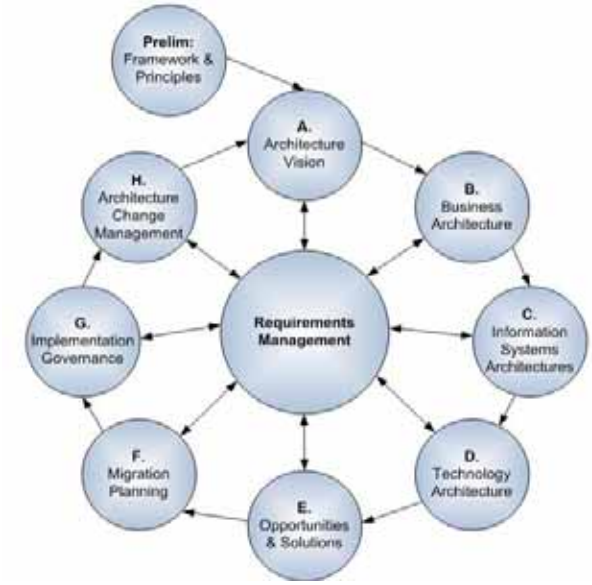

Figure 16. Architecture Development Method, TOGAF.

An earlier framework that looks at the various roles within an organisation and helps to summarise perspectives of various stakeholders on basic modalities of the organisation is the Zachman framework (see Figure 17 for a basic idea about the areas covered by it).[76]

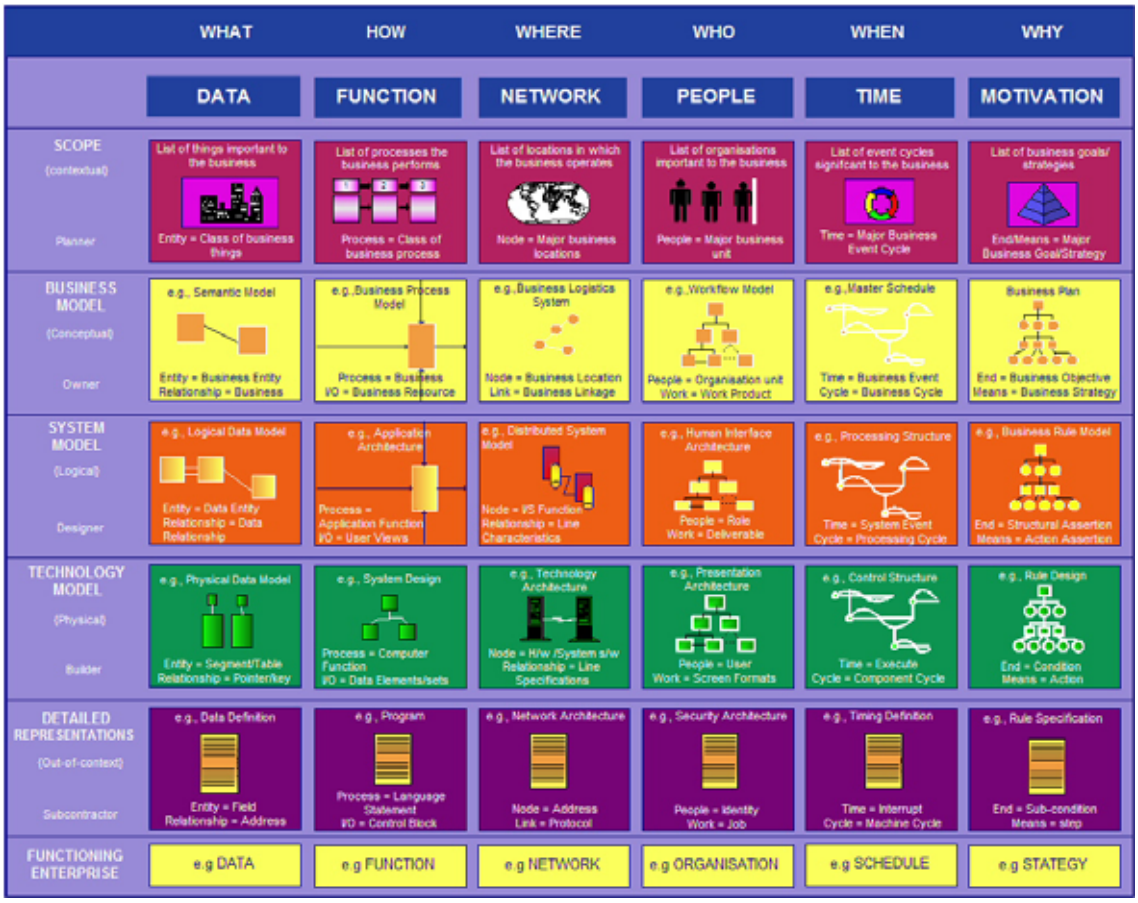

Figure 17. The Zachman framework.

[75] http://www.opengroup.org/togaf/

[76] Zachman, J. *Concise Definition of The Zachman Framework*. http://zachman.com/about-the-zachman-framework

An adaptation of the Zachman Framework into digital preservation domain is presented as a matrix in Table 6 below.

Table 6: Zachman framework adaptation for digital preservation – a generic view.

	What or Data	How or Process	Where or Network	When or Schedule	Who or People	Why or Motivation
Planner's view (contextual level - scope)	A list of objects that the organisation is interested in. *Audit of digital objects helps to identify the digital assets at risk.*	A list of processes or functions that the organisation performs. *This is institution-specific, with some similarities across institutions.*	The business locations. *This includes physical locations and IT-infrastructure details. Use of grid or cloud services would have different implications compared to a stand-alone infrastructure.*	The cycles and events related to each function. *Institution-specific with some similarities across institutions.*	A list of organisations important to the business. *Addressing the level of the institution as well as of aggregators or eInfrastructures.*	A list of business objectives. *This list should address who has the mandate - memory institutions or a designated infrastructure.*
Business owner (conceptual level – enterprise model)	e.g. semantic model	e.g. business process model	e.g. business logistics system	e.g. workflow model	e.g. master schedule	e.g. business plan
Designer (logical level – system model)	e.g. logical data model	e.g. application architecture	e.g. distributed system architecture	e.g. human interface architecture	e.g. process structure	e.g. business rule model
Implementer (physical level – technology model)	e.g. physical data model	e.g. system design	e.g. technology architecture	e.g. presentation architecture	e.g. control structure	e.g. rule design
Subcontractor (detailed representation level)	e.g. data definition	e.g. program	e.g. network architecture	e.g. security architecture	e.g. timing definition	e.g. rule definition
User/customer - Functioning system	e.g. data	e.g. function	e.g. network	e.g. organisation	e.g. schedule	e.g. strategy

The suitability of the TOGAF method and the Zachman Framework to the planning of digital archive workflows has been demonstrated before (Ruusalepp & Aas, 2007; Barateiro et al., 2010). Taking the enterprise view when orchestrating a suite of digital preservation tools, appears to offer an additional, organisational level angle that is lacking when considering the preservation function in isolation.

5.3 Estimating Market Demand

Although the number of preservation tools is substantial, their uptake and use in practice is very hard to measure. This could also be said about the whole market for digital preservation services. The models for evaluating market maturity are very general and do not fit easily the kind of niche area like digital preservation for the moment is.

The Planets project conducted interviews with leading IT companies to explore the emerging marketplace for digital preservation tools and services. Results of this study confirm that engagement is being led by memory institutions and driven primarily by legislation. There is perceived high demand for technology to support automation of digital preservation processes and for consultancy, training, awareness-raising and exchange of best practice, but the overall description of the services market was "market in its infancy" (PLANETS, 2010).

On the demand side the market is falling into broad categories of potential clients with somewhat different needs for preservation tools depending on the purpose and context of their preservation task. The cultural heritage institutions form a segment of this market that is interested in preservation services. The size of this "customer base" is, however, not easy to estimate. The ENUMERATE project[77] is currently collecting statistical data on digitisation, online accessibility and digital preservation in cultural institutions in Europe and has set itself a target of 1,500 respondents. This figure could also be taken as a minimum number for the digital preservation services market size because it is based on the evaluation of institutions active in provision of digital cultural resources. It also roughly matches the number of institutions that currently provide digital objects through the Europeana portal.

Reliable evidence is offered by the download statistics from SourceForge that provides access 81 preservation tools. These were downloaded in total 2450 times over a one week period that was monitored for this report. This shows that there is already a momentum on the market – the tools on offer require time for adapting and installation, yet are quite popular.

Further research is necessary to collect more detailed data on use of tools – their geographic spread as well as the features that make some tools more popular than others. For the time being, digital preservation services can be seen as an emerging yet still a niche market.[78] Market maturity cannot be considered to be significant at this stage neither on national, European or international level since competition is practically non-existent and most tools are offered as open source.

5.4 Evaluating Service Maturity

Although the presentation of preservation tools in section 3 made it clear that service provision is still on pilot projects basis and not yet a standard practice, it is essential to consider how service maturity will be evaluated and monitored in the future. Among the many models that are available for estimating services maturity,[79] the Capability Maturity Model Integration for Services (CMMI-SVC)[80] lends itself for adaptation to the immature market of digital preservation services. An attempt to apply the CMMI-SVC model to preservation services is presented in Table 7 that uses the preservation planning tool Plato as an initial example. The nature and scope of this study did not permit applying such a matrix in depth to all identified tools but it can serve as a basis for further research.

Table 7: Digital preservation service maturity matrix (small example).

Criteria	Metrics	Example: Plato
Maturity	1. Initial 2. Managed 3. Defined 4. Quantitatively managed 5. Optimizing	Managed
Adoption drivers	1. Business 2. Technology	Business: consistent preservation planning Technology: none
Adopters / Main clients / Market share		Large national-level memory institutions with teams of staff involved in digital preservation
Differentiators		None - all similar tools are web-based Reliance on formal knowledge bases for preservation plans (e.g. PRONOM)

The report continues to propose development of a roadmap with concrete steps for systematically gathering existing data on preservation tools and service, and to create the basis for benchmarking of services and their use in digital preservation workflows.

[77] http://www.enumerate.eu/en/about_enumerate/
[78] Cf. http://sbinfocanada.about.com/cs/marketing/g/nichemarket.htm
[79] Cf. http://imagesrv.gartner.com/research/methodologies/methodologies.pdf
[80] See: http://www.sei.cmu.edu/reports/10tr034.pdf

Consideration points

- Finding mechanisms to **evaluating** potential of existing **tools** to become services should be given a high priority.
- **Metrics of service maturity** need to be developed further. For this report available information on the use of services in different implementations was collected but it is difficult to trace all links. The information on numbers of downloads is limited and will require time-series data (e.g., if a tool is gaining in popularity the number of downloads would grow over time). Other aspects could be considered, for example inclusion in professional training programmes in digital preservation and repository management (DROID and JHOVE are likely candidates for the most popular tool in this category).
- A number of **different classifications** of tools and services have emerged. These evaluations are following different methodologies and provide somewhat uneven level of detail.
- This could be overcome by developing a detailed **registry** of preservation services with clearly applied metrics which makes similar tools easy to compare.
- It is important to find **intervention mechanisms** that would help to develop digital preservation services and their market by moving them further along the maturity curve of e-services.
- There is a need to develop a **roadmap** that integrates the achievements to date and outlines steps to make e-Infrastructures preservation-ready.

6. A Roadmap for Digital Preservation e-Infrastructure

6.1 Drivers for New Models in Automating Digital Preservation

There are three major drivers for looking for alternative models to the current repository-centric preservation models and software that supports digital archiving:

1) **Preservation community drivers:** The first set of drivers comes from the preservation community which in the last years has been seeking to integrate preservation better into the overall digital object lifecycle. Examples of this are recent discussion on preservation-ready systems (Borbinha, 2010) and ways for integrating digital preservation components in other existing information systems; and work on new formats of digital objects which help these objects to be self-preserving. The latter is a further development of the idea of durable digital objects; in the last consultation of the EC on areas for future research on digital preservation self-preserving objects were suggested as one of those areas for future research (Billenness, 2011).
2) **New technologies** that are already **being explored** within the preservation community. The second major driver for rethinking digital preservation models is the rapid development of technological solutions for access such as e-Infrastructures, cloud computing and micro-services. For example amongst the Quest Software[81] *Ten predictions for technology trends and practices in 2012*, six are related to developments of cloud services and virtualisation, including: „SaaS growth will help drive wider adoption of cloud services". These technologies will continue to advance and are increasingly deployed by the digital preservation community. For example Askhoj et al. discuss a service model for shared archiving services via the cloud (Askhoj, Nagamori & Sugimoto, 2011).
3) **New technologies not yet adopted** for digital preservation. Another driver are the future technology developments that are forecasted but not yet analysed from the point of view of using them in digital preservation. For example, the Gartner technology hype cycle for 2011[82] includes the „Big Data" domain (marked in red on Figure 18 below) which relates directly to scalability of digital preservation solutions and their ability to handle rapidly growing volumes of digital assets. Other examples include new hardware types, e.g. tablet computers and other hand-held devices that will be used to access content from digital archives, and new software platforms, e.g. Android that will need to be supported and preserved in the future. Others such as augmented reality would influence formats and complexity of digital objects. Systematic analysis of these predicted changes and their consequences for digital preservation are currently too low on the research agenda.

[81] http://www.dataprix.com/files/TEC_Survey_Whitepaper-2011.pdf
[82] http://www.gartner.com/hc/images/215650_0001.gif

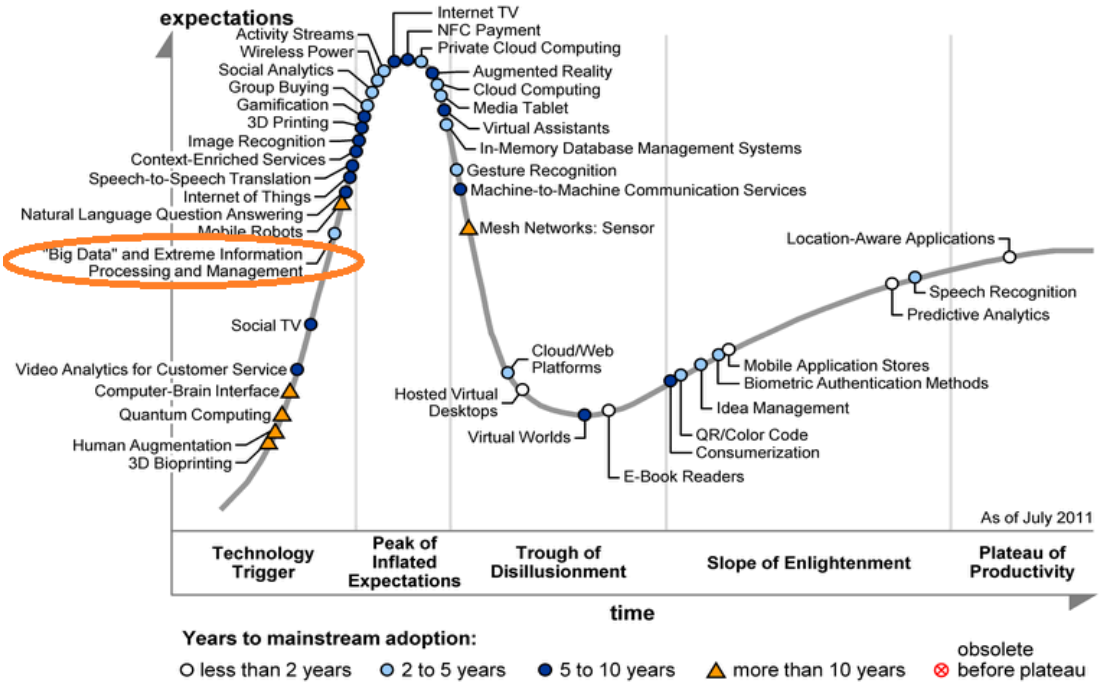

Figure 18: The Gartner Hype Cycle of Emerging Technologies (July 2011).

This report provides evidence that the preservation community is actively adopting for new methods and technologies and is refining existing ones. The preservation models and systems is one area where increased granularity and definition of smaller composite units has been achieved. In 2004 Brain Lavoie and Lorcan Dempsey described a preservation system as (Lavoie & Dempsey, 2004):

> "A digital preservation system can be deconstructed into several functional layers. The bottom layer includes hardware, software, and network infrastructure supporting the storage and distribution of digital content. The next layer includes more specialized services to manage the archived content residing in the system, including metadata creation and management, and validation of materials' authenticity or integrity. Preservation measures are implemented in the next layer of services, including monitoring the repository's environment for changes that could impact the ability to access and use archived content, as well as initiating processes such as migration or emulation to counteract these changes. The top-most layer includes services that support browsing or searching, access requests, validating access permissions, and arranging for delivery. … Determining the extent to which digital preservation can benefit from a division of labor, in the sense of finding 1) a sensible deconstruction of the digital preservation process into a set of more granular services, and 2) the optimal degree of specialization across preserving institutions, is a key issue in the design of digital repository architectures."

Today, examples exist that map the OAIS functional entities to a layered service model, where APIs (Application Programming Interfaces) are the services between PaaS (Platform as a Service) and SaaS (Software as a Service) layers, and, indeed, other layers (e.g., Package layer, Archival and Record Management layer as defined in Askhoj, Nagamori & Sugimoto, 2011). The SaaS models can now be hosted either locally or in the cloud.

The microservices approach that emerged a few years ago also steps away from the encapsulated holistic digital archive systems and, instead allows for combining flexible specialised solutions.

In 2007 a study summarised in total of five models for digital preservation services and admitted that most of them are still in their infancy (Hitchcock et al., 2007):

- *Software model* where preservation features are built into the repository management software.
- *Institutional model* where one institution could have several repositories and one of them is designated for digital preservation purposes.

- *Federated (or distributed) model*, represented by for example LOCKSS[83] and DISTARNET.[84]
- *Networked model* based initially on the use of Storage Resource Broker (SRB), recently also based on the use of iRODS,[85] grid or cloud technologies.
- *Service provider model* where third-party preservation services are used. The model was first described in 2002 and has slowly developed into more detailed models (e.g., Knight, 2005; the Preserv and Preserv2 projects – see Hitchcock, 2009).

The growing use of microservices and increasing reliance on SaaS platforms and cloud technologies have changed these models, paving the way for e-Infrastructures to become part of digital preservation landscape.

6.2 Roadmaps, Digital Preservation and Research Infrastructures

Roadmaps are useful instruments for presenting the scope and coverage of an e-Infrastructure. They are also frequently used within projects and institutions in the digital preservation domain. Some roadmaps can be very detailed as for example the roadmap developed for the UK Parliamentary archives (2008)[86] that presents environmental, policy, preservation, presentation, standards, skills, and communication developments over time. The Open Planets Foundation developed a *Tools and Services Roadmap*[87] to outline their software development plans. The APARSEN project roadmap[88] presents research topics and larger themes, including preservation services as a research topic under the theme of sustainability. Some projects use roadmaps to present various formats, e.g. the PrestoSpace[89] project presents formats for audio-visual material. There are also a number of national roadmaps, especially in the area of research infrastructures that address arts and humanities.[90]

These examples show that the focus of roadmaps can vary and that national roadmaps can coexist with project and institutional level ones. Would another roadmap for a new e-Infrastructure fit into this landscape and could it be co-ordinated with the already existing roadmaps?

The analysis of recent developments in digital preservation services, the emerging synergies between digital cultural heritage sector and e-Infrastructures shows that there are drivers for change that have not quite yet crystallised in terms of understanding the on-going business change, formulating policy frameworks and developing practical tools that could facilitate and drive the process of providing flexible and efficient digital preservation solutions in a distributed environment. A new roadmap that highlights areas that need further elaboration and coordination of actions across a number of stakeholders would help to overcome these shortcomings.

A proposal for developing a new roadmap combines three stages: preparatory stage, development stage and deployment and monitoring stage (see Figure 19 below).

---

[83] LOCKSS (Lots of Copies Keep Stuff Safe) developed at Stanford University Libraries, open-source software that stores multiples copies of the same object and provides migration to new formats on access. See also Skinner & Schultz (2010).

[84] A very recent development, a fully distributed open digital preservation system that executes pre-defined workflows for preservation (Subotic, Schuldt & Rosenthaler, 2011).

[85] One of the recent developments based on iRODS (see Rajasekar et al., 2010) is NASA's planetary data that are stored using iRODS via the cloud (Mattmann et al., 2010).

[86] http://www.parliament.uk/documents/upload/strategy-road-map-final-public.pdf presents the roadmap diagram and http://www.parliament.uk/documents/upload/digital-preservation-strategy-final-public-version.pdf - the justification.

[87] http://www.openplanetsfoundation.org/community/tools-and-services-roadmap

[88] http://www.alliancepermanentaccess.org/index.php/current-projects/aparsen/aparsen-roadmap/

[89] http://wiki.prestospace.org/pmwiki.php?n=Main.Roadmap

[90] See for example the Danish roadmap for RI http://en.fi.dk/publications/2011/danish-roadmap-for-research-infrastructure-2011/uk-roadmap.pdf; Large research (Czech roadmap, 2010) http://www.infrafrontier.eu/docs/national_roadmaps/Roadmap_CR.pdf; Australian humanities infrastructure http://www.paradisec.org.au/blog/2011/03/australian-humanities-research-infrastructure-funding/

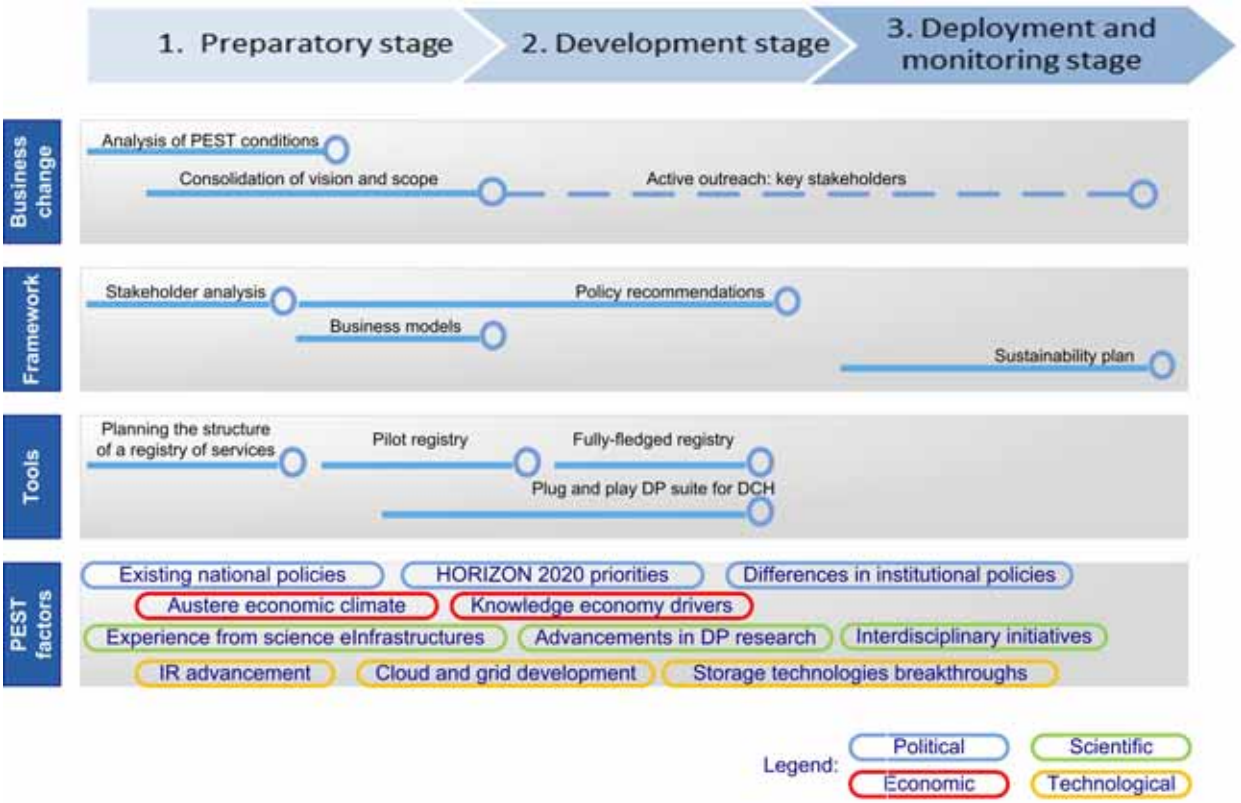

Figure 19. Roadmap - digital preservation services for cultural heritage collections.

The suggested roadmap is integrating three domains of necessary intervention (business change, policy framework and better tools) with the major PEST factors. The necessary future steps are grouped in three phases: Preparatory, Development and Deployment & Monitoring.

1. Preparatory stage

The preparatory phase includes activities such as:

- Defining **essential PEST** (political, economic, societal, technological) **conditions**, with potential extension to ecological impact and conditions. It is necessary to define a common language between various stakeholders and to clearly position the current state of preservation in relation to other impacting factors.
- Defining **vision, scope and boundaries** (as well as other areas that feed into the digital preservation services). This supports outlining clearly what digital preservation services address and what is left beyond the scope of the roadmap. Preservation is part of a wider digital object lifecycle and it is easy to expand its boundaries towards other enterprise components.
- Planning the structure of a **digital preservation services registry.** This study shows that there is abundance of tools, but very little comparison of them can be made based on clearly defined quality criteria. Defining a model that makes existing tools and services comparable will help stakeholders with little technical and preservation know-how. Registries should not only provide structured information, but also an understandable interface to various types of stakeholders, aligned to their primary interests.
- Defining short-, mid-, and long-term measures for **awareness/leadership.** The roadmap should not only capture the current state of digital preservation, but identify milestones for the future.
- Mapping to EC and national **research agendas.** This should help institutions who want to place themselves on the state of the art map of digital preservation.
- Establishing key **partnerships with** relevant **e-Infrastructures.** Collaboration with existing infrastructure projects and initiatives is vital for the success of the roadmap.
- Analysing **innovation drivers.** Identifying expected technology and other changes in the (near) future will help keeping the roadmap up to date.

2. Development

This stage is structured around development of tools, but also of policy instruments that are necessary to achieve efficient intervention in the digital cultural heritage sector. The major actions will be:

- Defining the product that the roadmap focusses on – digital preservation services for digital cultural heritage collections.
- Stakeholders, needs, scenarios, business models.
- Defining critical system requirements, as well as key performance indicators.
- Analysing current technological offers and gaps.
- Developing the repository of tools and services.
- Defining drivers for making a shift in institutional practices.

3. Deployment and monitoring stage

- Effort to fill the identified gaps.
- Promoting the registry toolkit.
- Implementation in a range of institutions while providing guidance and support.
- Evaluating the performance – planning for improvement.

The implementation of this roadmap will allow for achieving business change across the digital cultural heritage sector, and will provide tools (registries) for the wider preservation community.

6.3 Upcoming Significant Trends

There is a large number of drivers and technology pushes that have an impact on how digital preservation develops. Picking out only a few of these as significant may seem unfair. However, in the context of a roadmap to digital preservation e-Infrastructure, the five future trends to follow are:

6. Transparent enterprise-driven models for digital preservation that will help to identify specific and generic components of the preservation function.
7. Launch of self-preserving objects – initially likely to be simple objects (text, images) that will still make a difference for the cultural heritage sector as the primary caretaker of these data types.
8. Increased flexibility in digital preservation architectures – based on granular or layered structures (e.g. SaaS, PaaS, IaaS) that are easy to adapt to a variety of preservation scenarios.
9. Clearly defined sets of metrics or benchmarks for comparing preservation tools and services and their performance.
10. Terminology and standards are no longer likely to converge along professional community borderlines – besides being interoperability across time, digital preservation will generate pressure for interoperability in real time, bringing along the need to agree on terminology.

7. Conclusions

This report reviewed the state of the art of digital preservation tools and services. It concluded that a substantial number of software tools are on offer – a 191 of these were analysed for the report. Yet, only a very small number of these tools are being offered as services; the vast majority have resulted from short-term R&D projects and have been simply been published as open source software after the project has finished, without subsequent user support and often incomplete documentation.

The identified tools were grouped into a taxonomy based on stages of the digital archiving workflow. This allowed demonstrating the areas of preservation work that have most software tools to choose from – metadata extraction, file characterisation and file format identification. Detailed comparison of features of all these tools and testing them in practice was outside the scope of this report. However, the report points out a significant lack of benchmarks and metrics for comparing the preservation tools for both professionals and beginners in the digital preservation business.

The hitherto relatively unchallenged repository-centric model of digital archive is now undergoing changes, primarily thanks to new architectural models emerging in IT – service oriented architecture, software as a service, infrastructure as a service, grid and cloud technologies, and the emergence of microservices. With an ever-broadening range of preservation software tools available, institutions can now combine and tailor digital preservation components according to their specific needs and context. There is also evidence that monolithic, integrated digital archive or repository software systems, many of which have thus far offered little or no support for active digital preservation, are beginning to supplant their workflow-based architectures with external (web)services or microservices for preservation. This beckons to re-define the preservation function from an organisational perspective and place it firmly in the enterprise architecture frameworks that underpin information management systems in organisations.

These developments in technology architectures and provision, combined with drives internal to digital preservation community, lead to a natural next step which is defining digital preservation as an infrastructure service. The success of projects exploring preservation options with e-Infrastructures in neighbouring domains (e.g., scientific or arts and humanities research data) is encouraging. The report, therefore, concludes with a proposal for developing a roadmap for a digital preservation e-Infrastructure in support of cultural heritage institutions. The roadmap would facilitate the development of more efficient and cost-effective service models that would make digital preservation more accessible across the cultural heritage sector, especially for smaller institutions that often lack the necessary resources to acquire, implement and manage in-house software systems.

The analysis in this report is summarised as conclusions or consideration points at the end of each chapter. These sections point out the areas that require further research and could be systematically included in the proposed roadmap. The main issue that the future research topics converge around is benchmarking of existing software tools and creating business scenarios that would ensure sustainability for the tools or their development into maintained e-services.

References

Abrams, S., Kunze, J., Loy, D. (2010) An Emergent Micro-Services Approach to digital Curation Infrastructure. *International Journal of Digital Curation*, 5(1), 172-186

Aschenbrenner, A., Enke, H., Fischer, T., & Ludwig, J. (2011) Diversity and Interoperability of Repositories in aGrid Curation Environment. *Journal of Digital Information*, 12(2). Available: http://journals.tdl.org/jodi/article/view/1896

Askhoj J., Nagamori, M., Sugimoto, S. (2011) Archiving as a service: a model for the provision of shared archiving services using cloud computing, in *Proceedings of the iConference 2011*, 151-158. Available: http://dl.acm.org/citation.cfm?doid=1940761.1940782

Barateiro, J., Antunes, G., Borbinha, J. (2010) *Aligning OAIS with the Enterprise Architecture*. Presentation at the: 8th European Conference on Digital Archiving, 2010 Geneva, Switzerland. Available: http://www.bar.admin.ch/aktuell/00568/00702/00861/01572/index.html?lang=de&download=NHzLpZeg7t,lnp6I0NTU042l2Z6ln1acy4Zn4Z2qZpnO2Yuq2Z6gpJCDdn54fGym162epYbg2c_JjKbNoKSn6A--

Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A., Hofman, H. (2009) Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. *International Journal on Digital Libraries,* 10(4). Available: http://www.ifs.tuwien.ac.at/~becker/pubs/becker-ijdl2009.pdf

Billenness, C. (2011) *The Future of the Past*, Report on the Proceedings of the Workshop, European Commission, Luxembourg, 4-5 May 2011. Available: http://cordis.europa.eu/fp7/ict/telearn-digicult/future-of-the-past_en.pdf

BL, KB, DNB, BN (2010) *Long-Term Preservation Services*. A description of LTP services in a Digital Library environment

Borbinha, J. (2010) Preservation Ready Systems – Digital Preservation and Enterprise Architecture. In: Chanod J.-P., Dobreva, M., Rauber, A., Ross, S. *Proceedings of Dagstuhl Seminar 10291: Automation in Digital Preservation*. Available: http://www.dagstuhl.de/Materials/Files/10/10291/10291.SWM11.ExtAbstract.pdf

Brown, A. (2006) *Automatic Format Identification Using PRONOM and DROID*. Available: http://www.nationalarchives.gov.uk/aboutapps/fileformat/pdf/automatic_format_identification.pdf

Burnhill, P., Guy, F. (2010) Piloting an E-journals Preservation Service (PEPRS). *The Serials Librarian*, 58(1-4), 117-126.

Caplan, P. (2006) *Preservation Metadata*. DCC | Digital Curation Manual. Available: http://www.dcc.ac.uk/sites/default/files/documents/resource/curation-manual/chapters/preservation-metadata/preservation-metadata.pdf

CASPAR (2007) *Review of the State of the Art*. Available: http://casparpreserves.eu/Members/cclrc/Deliverables/review-of-state-of-the-art-1/at_download/file.pdf

Challis, D. (2010) *Notes on SITS – the Scholarly Infrastructure Technical Summit.* Available: http://blogs.ecs.soton.ac.uk/webteam/2010/11/15/notes-on-sits-the-scholarly-infrastructure-technical-summit/

Chue Hong, N. (2012) Digital Preservation and Curation: The Danger of Overlooking Software. In: Anderson, D., Delve, J., Dobreva, M., Baker, D., Billenness, C., Konstantelos, L. (Eds.) *The Preservation of Complex Objects*. *Vol. 1: Visualisations and Simulations*. 24-35.

DARIAH (2010). *Collection Ingest, Management and Preservation*. Available: http://www.dariah.eu/index.php?option=com_docman&task=cat_view&gid=92&Itemid=200

DPimpact (2009) *Socio-Economic Drivers and Impact of Longer Term Digital Preservation*. Final Report. Available: http://cordis.europa.eu/fp7/ict/telearn-digicult/dpimpact-final-report.pdf

Gietz, P., Aschenbrenner, A., Budenbender, S., Jannidis, F., Kuster, M.W., Ludwig, C., Pempe, W., Vitt, T., Wegstein, W., and Zielinski, A. (2006) TextGrid and eHumanities. In: *Proceedings of the Second IEEE International Conference on e-Science and Grid Computing* (E-SCIENCE '06). IEEE Computer Society. Available: http://dl.acm.org/citation.cfm?id=1192614

Gladney, H.M. (2008) *Durable Digital Objects Rather Than Digital Preservation*. ERPANET. Available: http://eprints.erpanet.org/149/01/Durable.pdf

Harvey, R., Thompson, D. (2010) Automating the appraisal of digital materials. In: *Library Hi Tech*, 28(2), 313-322

Hitchcock, S. (2009) *Final report of the (PReservation Eprint SERVices): towards distributed preservation services for repositories Preserve2 project*. Available: http://preserv.eprints.org/JISC-formal/preserv2-progressreport.doc

Hitchcock, S., Brody, T., Hey, J. M. N., & Carr, L. (2007). Digital Preservation Service Provider Models for Institutional Repositories. *D-Lib Magazine*, 13(5-6). Available: http://www.dlib.org/dlib/may07/hitchcock/05hitchcock.html

ISO 14721 (2003): Space data and information transfer systems - Open archival information system - Reference model. Available: http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683

ISO 20652 (2006) Space Data and Information Transfer Systems - Producer-Archive Interface - Methodology Abstract Standard (PAIMAS). Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=39577

Justrell, B., Halling, S., Dessaux, C., Audeval, A., Fresa, A. (2011) *Digital Cultural Heritage Services Priorities Report*. D3.1 deliverable of DC-NET project

KEEP (2009) *Preliminary document analysing and summarizing metadata standards and issues across Europe* (KEEP project deliverable D3.1). Available: http://www.keep-project.eu/ezpub2/index.php?/eng/Products-Results/Public-deliverables

Kejser U.B., Nielsen A.B., Thirifays, A. (2011) *Costs of Digital Preservation Project*. Report for Phase 2. Available: http://www.costmodelfordigitalpreservation.dk/contact/cmdp-2---ingest-and-archival-storage

Knight, G. (2005) SHERPA DP: establishing an OAIS-compliant preservation environment for institutional repositories. In: *Digital repositories: interoperability and common services*, 9th DELOS thematic workshop, Heraklion, Crete, May 11-13, 2005. 43-48

Kumar, A., Kaplan, D. and Rubinger, B. (2008) TIAMAT: An Ingest Service for the Tufts Digital Repository. In: *Third International Conference on Open Repositories 2008*, 1-4 April 2008, Southampton, United Kingdom. Available: http://pubs.or08.ecs.soton.ac.uk/75/

Linden, J., Martin, S., Masters, R., Parker, R. (2005) The large-scale archival storage of digital objects. *DPC Technology Watch Series Report 04-03*. Available: http://www.dpconline.org/advice/technology-watch-reports

Lavoie, B. (2004) The Open Archival Information System Reference Model: Introductory Guide. DPC Technology Watch Report. Available: www.dpconline.org/docs/lavoie_OAIS.pdf

Lavoie, B., Dempsey, L. (2004) Thirteen Ways of Looking at...Digital Preservation. *D-Lib Magazine*, 10(7/8), Available: http://dlib.org/dlib/july04/lavoie/07lavoie.html

Mattmann, C.A., Crichton, D.J., Hart, A.F., Kelly, S.C., Hughes, J.S. (2010) Experiments with Storage and Preservation of NASA's Planetary Data via the Cloud, *IT Professional* , 12(5), 28-35. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5483284&isnumber=5593026

Meyer, E.T., Eccles, K. and Madsen, C. (2009) Digitisation as e-Research infrastructure: Access to materials and research capabilities in the Humanities. In: *Proceedings of the 5th International Conference on e-Social Science*, Cologne, Germany, 24-26 June 2009. Available: http://oxford.academia.edu/KathrynEccles/Papers/608385/Digitisation_as_e-Research_Infrastructure_Access_to_Materials_and_Research_Capabilities_in_the_Humanities

Nicholson, D., Dobreva, M. (2009) Beyond OAIS: towards a Reliable and Consistent Digital Preservation Implementation Framework. In: *16th Int. Conf. on Digital Signal Processing* (IEEE DSP 2009). Santorini, Greece, in 5-7 July 2009. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=05201126

NUMERIC (2009) *Developing a statistical framework for measuring the progress made in the digitisation of cultural materials and content. Study findings and proposals for sustaining the framework* (final report). Available: http://cordis.europa.eu/fp7/ict/telearn-digicult/numeric-study_en.pdf

Oliver, G., Ross, S., Guercio, M., Pala, C. (2008) *Report on automated re-appraisal: managing archives in digital libraries* (Deliverable 6.10.1), DELOS NoE

PAIMAS (2004) *Recommendation for a Producer-Archive Interface Methodology Abstract Standard* (PAIMAS), CCSDS 651.0-B-1, Blue Book. Available: http://public.ccsds.org/publications/archive/651x0m1.pdf

Pickton, M., Morris, D., Meece, S., Coles, S., Hitchcock, S. (2011) Preserving repository content: practical tools for repository managers. *Journal of Digital Information,* 12

PLANETS (2010) *An Emerging Market: Establishing Demand for Digital Preservation Tools and Services*. Available: http://www.planets-project.eu/docs/reports/Planets-VENDOR-White-Paperv4.pdf

Rajasekar, A. et al. (2010) *iRODS Primer: Integrated Rule-Oriented Data System.* Morgan & Claypool

Ross, S. and Gow, A. (1999) *Digital Archaeology: Rescuing Neglected and Damaged Data Resources*, London: Library Information and Technology Centre. Available: http://www.ukoln.ac.uk/services/elib/papers/supporting/#blric

Ruusalepp, R., Aas, K. (2007) *Methodological issues of developing an architecture for digital archive system*. Presentation at the DLM-Forum Members Meeting, Berlin. Available: http://dlmforum.eu/index.php?option=com_jotloader&section=files&task=download&cid=95_6aa356e10757345d51f41f97a2266df6&Itemid=103&lang=en

SCAPE D10.1 (2011) *Identification and selection of large-scale migration tools and services* (draft). Available: http://www.scape-project.eu/wp-content/uploads/2011/09/SCAPE_D10.1_KEEPS_V1.0.pdf

Simpson, D. (2005) *Directory of digital preservation repositories and services in the UK*. DPC

Skinner, K., Schultz, M. (2010) *A Guide to Distributed Digital Preservation*. Available: http://www.metaarchive.org/sites/default/files/GDDP_Educopia.pdf

Smyth, Z. (2006) *Developing a pre-ingest strategy for digital records*, Presentation at the Joint DCC/LUCAS Workshop, 30.11-1.12.2006, Foresight Centre, Liverpool

Strodl, S., Petrov, P., Rauber, A. (2011). *Research on Digital Preservation within projects co-funded by the European Union in the ICT programme*. Available: http://cordis.europa.eu/fp7/ict/telearn-digicult/report-research-digital-preservation_en.pdf

Subotic, I., Schuldt, H., Rosenthaler, L. (2011) The DISTARNET Approach to Reliable Autonomic Long-Term Digital Preservation. In: Yu, J., Kim, M., Unland, R. (eds.) *Database Systems for Advanced Applications*. LNCS 6588, Springer Berlin / Heidelberg, 93-103

The New Renaissance (2011) *The New Renaissance: report of the Comité des Sages. Reflection group on bringing Europe's cultural heritage online.* Available: http://ec.europa.eu/information_society/activities/digital_libraries/doc/reflection_group/final-report-cdS3.pdf

Thomas S. (2008) *CAIRO project final report*. Available: http://ie-repository.jisc.ac.uk/392/1/CairoFinal_0-2.pdf

Thomas, S., Baker, F., Gittens, R., Thompson, D. (2007) *Cairo tools survey. A survey of tools applicable to the preparation of digital archives for ingest into a preservation repository*. Available: http://cairo.paradigm.ac.uk/projectdocs/cairo_tools_listing_pv1.pdf

van der Knijff, J, Wilson, C. (2011) *Evaluation of characterisation tools*. Part 1: Identification. SCAPE deliverable. Available: http://www.openplanetsfoundation.org/system/files/SCAPE_PC_WP1_identification21092011.pdf

List of Abbreviations

AIP	Archival Information Package
API	Application Programming Interface
AQuA	Automated Quality Assurance Project
BRTF	Blue Ribbon Task Force
CDL	California Digital Library
CLARIN	Common Language Resources and Technology Infrastructure
CMDP	Cost Model for Digital Preservation
DAAS	Data as a service
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DCH	Digital Cultural Heritage
DC-NET	Digital Cultural heritage NETwork
DIP	Dissemination Information Package
ERA	Electronic Records Archives
FITS	File Information Tool Set
FOXML	Fedora Object XML
FP	Framework Programe
IR	institutional repository
LOCKSS	Lots of Copies Keep Stuff Safe
LTP	long term preservation
METS	Metadata Encoding and Transmission Standard
NDIPP	National Digital Infrastructure Preservation Program
NLA	National Library of Australia
NREN	National Research and Education Network
OAIS	Open Archival Information System
OPF	Open Planets Foundation
PAAS	Platform as a service
PAIS	Producer-Archive Interface Specification
PAIMAS	Producer-Archive Interface - Methodology Abstract Standard
PEST	Political, Economic, Scientific, Technological
RI	research infrastructure
SAAS	Software as a Service
SIP	Submission Information Package
SOA	Service-oriented architecture
XENA	XML Electronic Normalising for Archives
XML	eXtensible Markup Language

Appendix 1. List of Preservation Tools Analysed in the Report

The tools listed here were gathered from a number of sources (see also Section 3.1):

- The CAIRO project[91]
- The National Digital Infrastructure Preservation Program in the U.S.[92]
- The U.S. Library of Congress list of tools for preservation metadata implementation supporting PREMIS[93]
- AQuA project[94]
- OpenPlanetsFoundation (OPF)[95]
- DigiBIC[96] project
- SourceForge.[97]

The table below presents the number of downloads from Sourceforge for the week of 11-18 November 2011, and the number of sources that mention the same tool (as a very basic indicator of its visibility within the digital preservation community).

Name	Type	N of downloads	N sources
Access (3)			
NGDA/ Alexandria Digital Library: ADL Middleware Server	Access		1
NGDA/ Alexandria Digital Library: Globetrotter	Access		1
NGDA: ArchiveView	Access		1
Digital object quality analysis, comparison and characterisation (29)			
ACE (Audit Control Environment)	Validator of integrity		1
Apache POI	Document analyser		1
BagIt Transfer Utilities	Validation tool	17	2
BWF MetaEdit	Audio quality		1
Checksum Checker	Checksum	34	1
Cue	Text analysis		1
DoxyChecker	Content checker		1
Fiji	Image analysis		1
Fixi	Utility checking cheksums		1
FlashBack - Digital Image Recovery	Digital object recovery (images)	10	1
getID3()	Characterisation of audio		1
ImageMagick	Image analysis		1
itext	Characterisation tool		1
Java Beans MIME Type extractor	Mime-type identifier		1
Java Image Comparison	Image compression		1
Java Mime Magic Library	MIME-type identifier		1
JDeskew	Image quality		1
jHears	Audio quality		1
jpegpixi	Quality of images	2	1

[91] See (Thomas, 2008).
[92] http://www.digitalpreservation.gov/partners/resources/tools/index.html
[93] http://www.loc.gov/standards/premis/tools.html
[94] http://www.jisc.ac.uk/whatwedo/programmes/inf11/digpres/aqua.aspx
[95] http://www.openplanetsfoundation.org/
[96] http://www.digibic.eu
[97] http://sourceforge.net/

jp2StructCheck	Characterisation tool		1
Jpylyzer	JP2 validation + properties extraction		1
KAKADU	JPEG 2000 software developer toolkit (SDK), includes encoder/decoder		1
PDFbox	Characterisation tool		1
pdiff: Perceptual Image Difference utility	Image comparison		1
Sanselan	Visualises image collection consistency		1
Signify	Digital signature creator		1
ssdeep	Duplicate image detection		1
tesseract-ocr	Compare OCR results across various source formats		1
Unpaper	Image quality		1
xcorrSound QA	Tool for audio files comparison		1
STORAGE (8)			
ALOx	Storage		1
Digital Archive	Storage		1
Duracloud	Cloud		1
iRODS integrated Rule Oriented Data Systems	Storage		1
LoDN	Storage		1
L-Store (Logistical Storage)	Storage		1
NGDA: NGDA Server	Storage		1
SRB Storage Resource Broker	Storage		1
COMPLETE SOLUTIONS (3)			
LOCKSS	Complete system		1
Rosetta (ExLibris)	Preservation system		1
Vireo	Complete preservation (ETD)	1	1
CONVERSION (7)			
b2x Translator	Conversion tool	79	2
BagIt Library (BIL)	Digital object transformation		1
pdf2svg	File converter		1
pdf2xml	File converter	75	2
PDFtk	File converter		1
pdftohtml	File converter		1
pstoedit	File converter		1
DATA MANAGEMENT (2)			
Conspectus Database for Private LOCKSS Networks (PLNs)	Data management		1
Replication Monitor and Verification	Data management		1
EMAIL PRESERVATION (2)			
Email Preservation Parser	Email preservation		1
EMARCH (Electronic Mail Archival System)	Email preservation		1
EMULATOR (3)			
Dioscuri - modular emulator	Emulator	33	1
dmklib	Emulator		1
pHash	Encryption		1

FORMAT IDENTIFICTION&MIGRATION (3)			
File Utility	File format identification utility		2
FITS File Information Tool Set	Format identification, metadata extraction		1
TrID File Identifier	Format identifier		1
FORMAT REGISTRY (4)			
PRONOM	Format registry		1
NGDA: Format Registry	Format registry		1
TOTEM	Format registry		1
UDFR	Format registry		1
FORMAT TRANSFORMERS AND MEDIA TRANSFER (14)			
Antiword	Format transformer		1
CatDoc	Format transformer		1
GIMP	Format transformer		1
IngestList (Landesarchiv Baden-Württemberg, Germany)	Transfer tool		1
JODConverter	Format transformer		1
Libwmf	Format transformer		1
Musemaster (Batch Audio Conversion)	Format transformer		1
NIBTOOLS	Media transfer		1
ODF Converter	File transfer		1
OpenJPEG	Format migration		1
PeDALS Email Extractor	File extractor	12	2
Prometheus Digi Pres Workbench	File transfer	1	1
TOM (Typed Object Model)	Format transformer, Format identifier		1
Xena / Xenalite	Format transformer, Metadata extractor	49	2
INGEST (5)			
BEAM Ingester	Ingest tool	5	1
GIS Archiving Toolset	Ingest (GIS objects)		1
IngestList	Ingest	3	1
NGDA: Bulk Ingest Tool	Ingest		1
TIAMAT	Ingest		1
MANAGEMENT (2)			
Dataverse Network	Management: Datasets of quantitative data		1
PERPS	Registry - management		1
METADATA (58)			
Chiba	Metadata creation	7	2
Echodep Hub and Spoke	Metadata creator	1	3
Elated	Metadata creator		1
METS Java Toolkit (Harvard University Library)	Metadata creator		1
OpenExif	Metadata creator		1
OpenTIFF	Metadata creator		1
PDFSSA4MET	Metadata analysis		1
Adobe XMP	Metadata creator (manual), Metadata transformer		1

Archivists Toolkit	Metadata creator (manual) Metadata transformer, Metadata wrapper		2
SHAME	Metadata creator (manual) Tool to implement metadata creator		1
SIP creator	Metadata creator (manual and (automatic) Interface tool to implement metadata creator		1
simple digital object	Metadata embedding		1
Metadata Extraction Tool	Metadata extraction	265	3
Apache Tika	Metadata extractor		1
Extractor	Metadata extractor		1
Aperture	Metadata extractor (automatic) MIME-type identifier		1
BRSoftware - EXIFextractor	Metadata extractor (automatic)		1
Caliph and Emir (or Lira and Caliph and Emir)	Metadata extractor (automatic) Metadata transformer; Metadata creator	146	2
Drew Noakes Metadata Extraction Library/Exif-O-Matic	Metadata extractor (automatic)		1
DROID	Metadata extractor (automatic), Format identifier	152	4
ExifTool	Metadata extractor (automatic)		3
id3lib	Metadata extractor (automatic), Metadata transformer		1
ImageInfo	Metadata extractor (automatic)		1
IPI-Manager	Automated indexing of audio-video content		1
Java Metadata Collection	Metadata extractor (automatic), Metadata transformer		1
Jhead	Metadata extractor (automatic)		1
JHOVE	Metadata extractor (automatic), Format identifier and validator		4
JHOVE2	Metadata extractor (automatic); Format identifier		2
Kaa Media Repository – Kaa Metadata module	Metadata extractor (automatic)		1
Kea	Metadata extractor (automatic), Metadata transformer		1
Libexif	Metadata extractor (automatic), Metadata transformer		1
libextractor	Metadata extractor (automatic); MIME-type identifier		1
Metadata Assistant	Metadata extractor (automatic)		1
Metaphile	Metadata extractor (automatic)		1
Picture Metadata Toolkit	Metadata extractor (automatic) Metadata transformer; Metadata creator		1
Soft Experience - Metadata Miner Catalogue PRO	Metadata extractor (automatic), Metadata transformer		1
wvWare	Metadata extractor (automatic); Format transformer	239	2
Ffident	Metadata extractor (automatic)		1
EMET (Embedded Metadata Extraction Tool)	Metadata extractor (images)	1	2
Statistics New Zealand Prototype PREMIS Creation Tool	Metadata generation		1
BWF MetaEdit	Metadata manipulation	99	1
OMAR Representation Information Repository	Metadata registry		1
DRC Bulk Ingest Tool	Metadata transformer		1
QUEST (Query Electronic Storage)	Metadata transformer		1
UIUC OAI Metadata Harvesting Project	Metadata transformer	8	2
Apache Xerces	Metadata validator		1
Expat	Metadata validator		1

XML Batch Validator	Metadata validator		1
PREMIS in METS Toolbox	Metadata validator and convertor		1
7train METS generation tool	METS generation	5	1
Pedro	Interface tool to implement metadata creator		1
upCast	Interface tool to implement metadata creator		1
PrestoPrime Metadata Semantic Converter	Metadata semantic converter		1
Video Tagging Game PrestoPrime	User tagging		1
PrestoPrime Metadata Semantic Converter	Metadata mapping visualisation		1
ContextMiner	Context collection		1
PRESERVATION PLANNING (1)			
PLATO	Preservation planning		1
WEB ARCHIVING (25)			
archive tweets to pdf + offline browsing	Web archiving	64	1
Archive-It	Web archiving		1
AvantFAX	Web archiving	162	1
ArchiveExplorer	Tool for displaying contents of WARC files		1
ArchiveFS	Tool for mounting ARC and WARC files		1
Heritrix	Web archiving	417	2
MailArchiva	Web archiving	140	1
NutchWAX	Web archiving		1
PLEADE : EAD for the Web	Web archiving	80	1
PROTAGE	Web archiving	0	1
SWAT (Snappy Web Archiving Tool)	Web archiving	15	1
Digital Preservation Recorder	Workflow	1	1
NGDA: Workflow Tool	Workflow		1
Taverna	workflow		1
TubeKit	Web archiving		1
Wayback Machine	Web archiving		1
Web Archive Access Utilities	Web archiving	23	1
Web Archive Extractor	Web archiving	97	1
Web Archives Workbench	Web archiving	3	2
Web Archiving Service	Web archiving		1
Web Harvester	Web archiving		1
wview Weather System	Web archiving	249	1
7Train	Wrapper creator; Metadata transformer		1
LiWARich Media Capture module	Enhancing capturing capabilities of the crawler for streamed multimedia content		1
GTRI XML Validation tool	XML validator		1
Collections, testdbeds, specifications (3)			
Digital Image Collection	Collection		1
FACIT (Federated Archive Cyberinfrastucture Testbed)	Testbed		1
BagIt	Specification		1

Others (19)			
Digital Preservation Services	Combined unspecified	6	1
Digital Preservation Software	Combined unspecified		1
DPSP (Digital Preservation Software Platform)	Mulitple	15	1
DAITSS (Florida Center for Library Automation)	Multiple including ingest		1
The Planets Digital Preservation Suite	Multiple unspecified	24	1
Cairo tool	Pre-ingest (format identification, metadata extraction)		1
Ad-Aware	Spyware/anti-virus		1
Apache Lucene	Search engine		1
DC Proxy	Proxy		1
Digital Tree	Interface		1
Jacksum	Message digest calculator		1
libPST	PST migration library		1
Library of Congress - Recollection	Interface		1
Multivalent	Rendering		1
National Geospatial Digital Archive Tools	Search tool		1
Planets Fedora Integration	Repository objects export		1
PSTViewTool	PST viewer		1
SHA	Message digest calculator		1
Spybot-Search & Destroy	Spyware/anti-virus		1
	TOTAL	2540	