Project no. 269977

APARSEN
Alliance for Permanent Access to the Records of Science Network

Instrument: Network of Excellence

Thematic Priority: ICT 6-4.1 – Digital Libraries and Digital Preservation

D21.1 Overview of Preservation Services

Document identifier:	**APARSEN**-REP-D21_1-01-2_1
Due Date:	2013-02-28
Submission Date:	2014-05-05 **(for this revised version)**
Work package:	WP21
Partners:	APA, CSC, DNB, KNAW-DANS, KB, CINI, FORTH, FRD, UNITN, Tessella, SBA
WP Lead Partner:	STFC
Document status	Released
URN	urn:nbn:de:101-2014051694

Delivery Type	Report
Author(s)	Simon Lambert and others (see Status Sheet below)
Approval	David Giaretta, Maurizio Lunghi (internal reviewer of V1.0)
Summary	Report on study of services related to long-term digital preservation, the needs for them and approaches to classifying them
Keyword List	Software services, Digital preservation
Availability	☒ Public

Document Status Sheet

Issue	Date	Comment	Author
0.1	2013-01-25	First draft of document based on preceding work within WP21	Simon Lambert (STFC)
0.2	2013-02-15	Expanded after discussion with other WP partners	Simon Lambert (STFC), Stefan Hein (DNB)
0.3	2013-02-28	Added conclusions and executive summary	Simon Lambert
1.0	2013-03-04	Final additions and corrections following internal review by M. Lunghi	Simon Lambert
1.1	2013-04-29	Introduced revisions following recommendations from project review in March 2013	Simon Lambert
1.2	2013-06-28	Completion of revisions as above	Simon Lambert, Barbara Bazzanella (UNITN), Stefan Hein, Stefan Proell (SBA), Stephan Strodl (SBA)
2.0	2013-12-31	Expansion and revision to maintain the report as a living document through the lifetime of APARSEN	Simon Lambert
2.1	2014-05-05	Completion with some additional information missing from previous versions	Simon Lambert

Abstract: This deliverable reports on the work done in APARSEN on preservation services, regarded as an aspect of sustainability for long-term digital preservation. An approach to the definition of high-level services is formulated applied, and a representative list of services is assembled and put into perspective, allowing some conclusions about the landscape of services.

Project information

Project acronym:	**APARSEN**
Project full title:	**Alliance for Permanent Access to the Records of Science Network**
Proposal/Contract no.:	**269977**

Project coordinator: Simon Lambert/David Giaretta

Address:	STFC, Rutherford Appleton Laboratory Chilton, Didcot, Oxon OX11 0QX, UK
Phone:	+44 1235 446235
Fax:	+44 1235 446362
Mobile:	+44 (0) 7770326304
E-mail:	simon.lambert@stfc.ac.uk / david.giaretta@stfc.ac.uk

CONTENT

EXECUTIVE SUMMARY **5**

1 INTRODUCTION **6**

2 HIGH-LEVEL SERVICES: UNDERSTANDING THE SPACE **9**

2.1 INTRODUCING HIGH-LEVEL SERVICES 9
2.2 SOURCES OF HIGH-LEVEL SERVICES 10
2.3 TOWARDS A CHARACTERISATION OF HIGH-LEVEL SERVICES 15

3 ELABORATING THE HIGH-LEVEL SERVICES **18**

3.1 CHARACTERIZATION OF SIPS 19
3.2 QUALITY ASSURANCE OF SIPS 21
3.3 POLICY-BASED ASSESSMENT OF SIPS 22
3.4 ACQUISITION AND MAINTENANCE OF REPRESENTATION INFORMATION 23
3.5 AUTOMATED METADATA CREATION/MAINTENANCE 25
3.6 METADATA MIGRATION 26
3.7 ENVIRONMENT MONITORING (PRESERVATION WATCH) 27
3.8 KNOWLEDGE MODEL COMPARISON 28
3.9 PRESERVATION PLAN FORMULATION 29
3.10 OBSOLESCENCE SUBSTITUTION 31
3.11 DEPENDENCY MANAGEMENT 32
3.12 AUTHENTICITY EVIDENCE MANAGEMENT 33
3.13 APPRAISAL OF COLLECTIONS 36
3.14 DRM CLEARINGHOUSE 37
3.15 BROKERAGE BETWEEN REPOSITORIES 38
3.16 LONG-TERM ARCHIVING 39
3.17 INTEGRITY CHECKING 41
3.18 CLOUD STORAGE FOR PRESERVATION 42
3.19 PRESERVATION POLICY CONSTRUCTION 43
3.20 ANALYSIS OF AUTHENTICITY MANAGEMENT POLICIES 44
3.21 FORMAT TRANSFORMATION 46
3.22 FINDING AIDS 47
3.23 FEDERATED SEARCH 48
3.24 PID RESOLVER 49
3.25 EMULATION FACILITIES 52
3.26 FULL REPOSITORY SERVICE 53
3.27 AUDIT AND CERTIFICATION OF REPOSITORIES 54

4 CLASSIFYING SERVICES AND MAPPING THE LANDSCAPE **55**

4.1 SERVICES AGAINST THE STANDARD FOR AUDIT AND CERTIFICATION OF REPOSITORIES 55
4.2 SERVICES AND PRESERVATION SCENARIOS 61

5 CONCLUSIONS FOR PRESERVATION SERVICES **67**

EXECUTIVE SUMMARY

Preservation services are services offered to an organisation (repository, archive) carrying out long-term preservation of digital material, and assisting that organisation in carrying out the tasks relating to preservation. The services might be services in the information technology sense (as in 'web services'), or they might be human services such as consultancy or audit. Indeed it is this wide spectrum of possibilities that makes the landscape of services so varied and in need of study.

The party providing the services might be external to the preserving organisation (a commercial supplier, for example), though it need not be—or even if the services originate outside the organisation, they might be configured and operated entirely within it. Services (in the software sense) may be contrasted with tools, which are regarded as standalone special-purpose software typically running on the desktop, though inevitably there is some overlap. In any case services gain their benefit from being widely applicable across a user base, meaning that development effort is for the benefit of all.

Examples of preservation services include:

- a registry repository of representation information, available for use by different repositories in the same domain
- a fully functional preservation platform covering ingest operations, backups and integrity checking, and probably other operations as well
- carrying out audit and certification of a repository to give confidence in its long-term ability to preserve its holdings.

Services may be viewed from the demand side—what is needed as part of the operations of a digital repository, and where the benefits of corresponding services would be most felt—and from the supply side—what services are under development or already available. Services cover the whole spectrum of operations associated with preservation, and so a systematic approach to their analysis and positioning with respect to each other is desirable.

Such an approach was taken in the work that led to this report. The participants in the work package collected a representative set of services, and described them in a systematic way. These services were then positioned according to two schemes:

- with respect to the ISO and CCSDS standard on audit and certification of trustworthy digital repositories;
- with respect to the scenarios developed in APARSEN's Work Package 14 on testing environments.

On the basis of the above analyses, it was possible to draw some conclusions, first about the landscape of preservation services—which will form part of the common vision that APARSEN is constructing—then specifically relating to the Virtual Centre of Excellence. Three types of conclusion, expressed as recommendations for further investigation, have been formulated:

- Areas of the landscape where there seems to be less coverage of services. The reasons for this uneven coverage are diverse, but it seems clear that there are opportunities for exploring the potential for service development—in 'Definition of Designated Community and associated knowledge bases' to give just one example.
- Limitations, or areas that have not been deeply explored. Particularly the domain specific nature of services seems not to have attracted much attention, and yet must be a factor in conditioning which services are suitable for the needs of a particular repository.
- Deeper description of services. For example, an ontology of services that formalizes the relationships between the main entities linked to services, enabling a more comprehensive model that would allow services to be understood as a potential collaboration, rather than just a disparate set of non-interacting offerings.

1 INTRODUCTION

Services for digital preservation are an aspect of *sustainability*: a contribution to the ability of digital preservation systems to continue their existence into the future. The existence of any kind of infrastructure for preservation of digital materials immediately implies the existence of services—understood as activities with a specific function that can help organisations preserve their digital holdings against threats such as changes in hardware, software, environment and designated communities, and that have a scope and applicability wider than a single digital repository. Unless a repository is attempting to conduct all of its preservation activities alone and unaided, there is potential for some kind of services to be of benefit. Indeed this is the motivation for constructing an infrastructure: the potential for sharing of effort and expertise, for economies of scale, and for greater robustness.

Before going further, we will consider what is meant by a preservation service. In the universe of information technology, a service is a clearly understood concept, defined for example by W3C thus:

> 'A service is an abstract resource that represents a capability of performing tasks that form a coherent functionality from the point of view of provider entities and requester entities. To be used, a service must be realized by a concrete provider agent.'[1]

This makes it clear that a service involves at least two parties, the provider and the requester, and that it embodies a set of tasks that in some way are regarded as a coherent package.

In the context of APARSEN, we understand a preservation service to have the following characteristics:

- It can be shared across repositories and organisations
- It is distinguished from a tool, though the distinction might not be completely clear-cut
- It could in principle be performed by an external party, though it need not be

These are all consistent with the W3C definition. However highly specific operations such as ingest of a particular data type into a particular repository will probably not qualify as services. There are blurred boundaries between tools and services: a tool generally runs as a standalone system on an individual desktop, but many tools may be generalised to services if the need and desire is present. The distinction is that a tool in itself does not have a provider in the sense above: of course the tool will have been provided by someone, but it is not invoked from the provider but rather run by the requester alone. Examples of tools (which are nonetheless of wide applicability) are Apache Tika (detection and extraction of metadata) and DROID (content profiling). In some cases in the overview in this document, tools are listed under the relevant headings where they are of wide usage and clear importance.

It is important to note that we consider both human and computer-implemented services as within scope of the definition; we are not restricting ourselves to automated services. Thus certain types of advisory service or consultancy will qualify as preservation services.

This work package within APARSEN aims to characterise and examine a range of services covering the spectrum of needs of digital preservation. The approach is a pragmatic one of examining what services already exist, are under development, or have potential for development, identifying both general, infrastructure-type components as well as delimiting the applicability of specific services and techniques and taking into account the requirements of different domains. The results will provide input for the evolving model of the Virtual Centre of Excellence as well as the common vision of digital preservation.

[1] http://www.w3.org/TR/ws-gloss/

The overall approach is typical of many APARSEN work packages and shown in Figure 1.

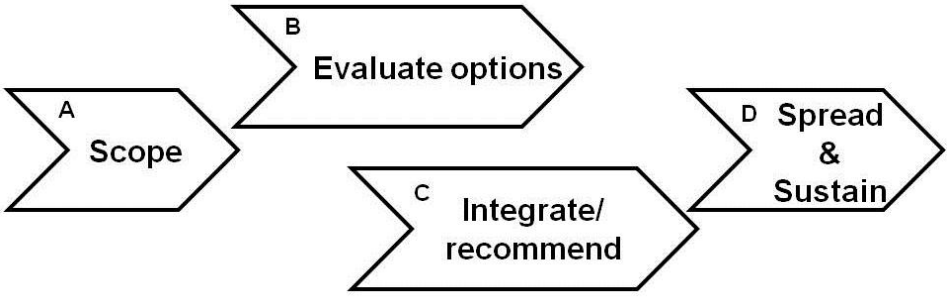

Figure 1 General approach of APARSEN work packages

In order to give some structure to the overview of preservation services, the first step was to identify a set of *high-level services*, which can also be seen as corresponding to requirements for the functionality of services. These high-level services are identified not bottom-up but by examining which activities of digital preservation could benefit from support by services. Specifically, a number of starting points were selected to help identify these high-level services:

- List from Neil Grindley of JISC
- A framework for services produced by several national libraries
- OAIS Functional Model
- DCC Curation Lifecycle Model

Section 2 presents these high-level services. Then in section 3 each of these is expanded with particular reference to *implementations* of the service. This captures the supply side, presenting actual developments that have been produced.

There is some arbitrariness in deciding how far to go with assembling lists of services. The effort available is limited, and there is no particular benefit in exhaustiveness, provided that the selection is representative. Therefore we have tried to assemble a reasonable collection with which the APARSEN partners are familiar, taking advantage of the large network and the involvement in other projects, past and present, such as PLANETS, SCAPE and SCIDIP-ES. The wide membership of the consortium means that a range of 'supply-side' expertise is represented, with various degrees of development from deployment to research projects. The work completed within the 'trust' topic of APARSEN is also a source of (potential) services.

Contacts with other initiatives have also resulted in extensions to the list of high-level services. In particular, discussions with the DCH-RP project (DCH-RP Digital Cultural Heritage Roadmap for Preservation) have been helpful.

It is intended to continue to update the list of services through the remaining duration of APARSEN, to provide a living resource for the community.

This work package, in common with all, is intended to contribute to the configuration and creation of the Virtual Centre of Excellence that is to be the lasting outcome of APARSEN. Specifically, we expect the following contributions to the knowledge base and activities of the VCoE.

- Understanding the landscape of preservation services will enable recommendations to be made about what services are available and appropriate (or not) for organisations (both within the VCoE and outside) seeking advice on preservation approaches technology.
- Identification of gaps or weaknesses in current service provision will enable targeted research and development to augment the range of services available.

- Specific services offered by APARSEN partners may become part of the offering of the VCoE.

This work package has the following relations to others in APARSEN.

- **WP11 Common vision.** WP21 uncovers, analyses and proposes constituents of both the common vision (for example, important research areas) and the Virtual Centre of Excellence (as just outlined above). After the end of the work package, effot continues in WP11 on integration of the findings into the common vision and specifically the topic of sustainability.
- **WP14 Common testing environments.** WP14 will provide a framework for systematic evaluation of preservation services that have been studied in WP21. Furthermore, the scenarios of WP14 have been used as an additional means of labelling and grouping services.
- **WP16 Common tools, software repository and market place.** As noted above, there is an overlap between services and tools in this context. It is expected that the repository of WP16 will expand to cover preservation services as well.
- **WP22 Identifiers and citability.** The persistent identification of resources by its very nature requires an infrastructure to support it. If it is to be more than local identification, it will require some kind of service support available to a range of users.
- **WP23 Storage solutions.** It is evident that storage solutions may be (and are) offered as a service.
- **WP25 Interoperability and intelligibility.** Interoperability is a key requirement for preservation services that are to have general applicability. Particularly if services are to work effectively together, interoperability is essential at the level of (for example) OAIS functional roles.
- **WP32 Cost/benefit data collection and modelling.** Costings for long-term digital preservation will depend on the decisions made on use of external services rather than internal implementations. There is already work on cost models for cloud storage, for example.
- **WP36 Business cases.** As for WP32, the formulation and evaluation of business cases will depend on the place of services to supply some of the operations of long-term preservation.

2 HIGH-LEVEL SERVICES: UNDERSTANDING THE SPACE

2.1 INTRODUCING HIGH-LEVEL SERVICES

High-level services arise from considering the need for services rather than the actual provision of services. They are expressed in very general terms, and relate directly to the challenges and activities of digital preservation. The question 'Where is there a need for preservation services?' is really of two parts:

- 'Where is there a need for assistance in some preservation activity?', implying that a repository/archive is aware of an activity that it alone is not competent or comfortable in carrying out.
- 'Can this assistance be provided in the form of a service?', asking whether the criteria of generality mentioned above are met.

One fundamental starting point is the roadmap of PARSE.Insight[2]. This important document, published in 2010, aims to provide an overview and initial details of a number of specific components, both technical and non-technical, which would be needed to supplement existing and already planned infrastructures for science data and extend into digital preservation. Science data infrastructure was taken to mean those things, technical, organizational and financial, which are usable across communities to help in the preservation, re-use and (open) access of digital holdings.

The basis of the PARSE.Insight roadmap is a number of high-level threats to long-term preservation, and the responses that mitigate those threats. The reality of the threats was supported by an extensive survey conducted by PARSE.Insight[3]. The threats and the corresponding responses are as follows:

Threat	**Requirements for solution**
Users may be unable to understand or use the data e.g. the semantics, format, processes or algorithms involved	Ability to create and maintain adequate Representation Information
Non-maintainability of essential hardware, software or support environment may make the information inaccessible	Ability to share information about the availability of hardware and software and their replacements/substitutes
The chain of evidence may be lost and there may be lack of certainty of provenance or authenticity	Ability to bring together evidence from diverse sources about the authenticity of a digital object
Access and use restrictions may make it difficult to reuse data, or alternatively may not be respected in future	Ability to deal with Digital Rights correctly in a changing and evolving environment
Loss of ability to identify the location of data	An ID resolver which is really persistent
The current custodian of the data, whether an organisation or project, may cease to exist at some point in the future	Brokering of organisations to hold data and the ability to package together the information needed to transfer information between organisations ready for long term preservation
The ones we trust to look after the digital holdings may let us down	Certification process so that one can have confidence about whom to trust to preserve data holdings over the long term

[2] 'Science data infrastructure roadmap', PARSE.Insight project, 2010, available from http://www.parse-insight.eu/.

[3] 'Insight into digital preservation of research output in Europe: survey report', PARSE.Insight project, 2009, available from http://www.parse-insight.eu/.

It is clear that there are a number of high-level services implied in the solutions, and also a way of describing specific services in terms of which threat they help to address.

2.2 SOURCES OF HIGH-LEVEL SERVICES

For the purposes of this APARSEN work package, and within this conceptual framework, a number of starting points have been chosen to identify high-level services.

- List from Neil Grindley of JISC
- A framework for services produced by several national libraries
- OAIS Functional Model
- DCC Curation Lifecycle Model

Neil Grindley of JISC has produced an overview of 'What might organisations need in terms of preservation services?'[4]. This is stimulated by looking at curation tasks and identifying those that could realistically be offered as services.

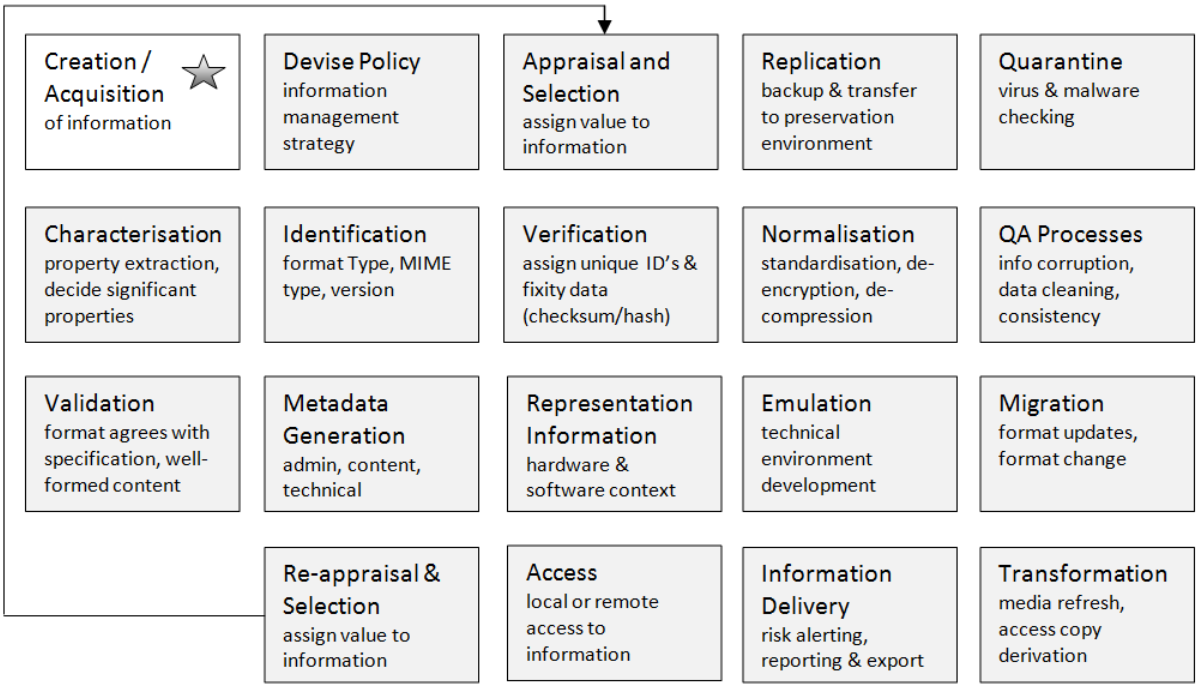

Figure 2 Possible preservation services (Neil Grindley, JISC)

The report 'Long-Term Preservation Services: A description of LTP services in a Digital Library environment' authored by the national libraries of the UK, the Netherlands, Germany and Norway[5] is of course oriented towards the preservation roles of a library, but takes a general approach, looking at the three core functions of a preservation system (to ingest, retain, and provide access to digital content) and showing how the services work together to fulfil each function. The framework for

[4] Neil Grindley, 'Value From Data Now and Into the Future', presentation at APA Conference, November 2012.

[5] 'Long-Term Preservation Services: a description of LTP services in a Digital Library environment', The British Library, Koninklijke Bibliotheek, Deutsche Nationalbibliothek, Nasjonalbiblioteket, 2010.

services is shown in Figure 3; the report goes on to specify in detail the functions of each of the types of service.

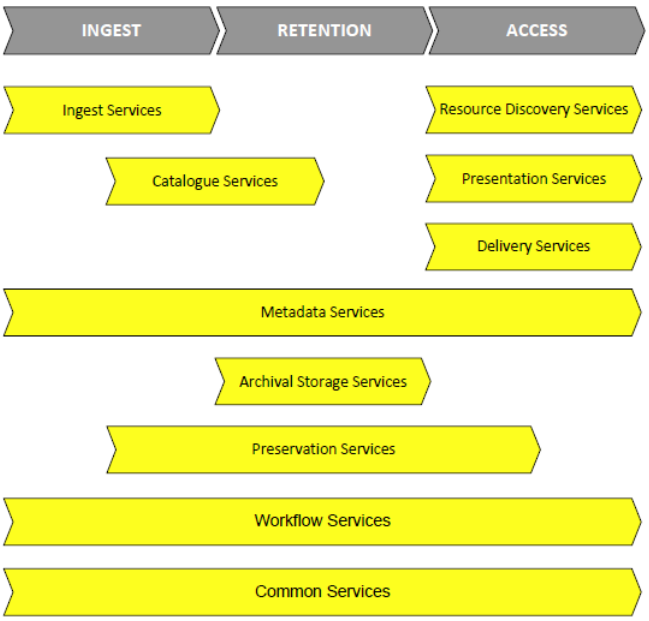

Figure 3 Contextual framework for preservation services (national libraries)

Another reference model to frame preservation services is the OAIS model[6] that is a functional framework, which represents the main concepts and the basic data flow within a digital preservation system. The model uses six functional entities to represent the key activities within a digital archive: ingest, preservation planning, archival storage, data management, administration and access.

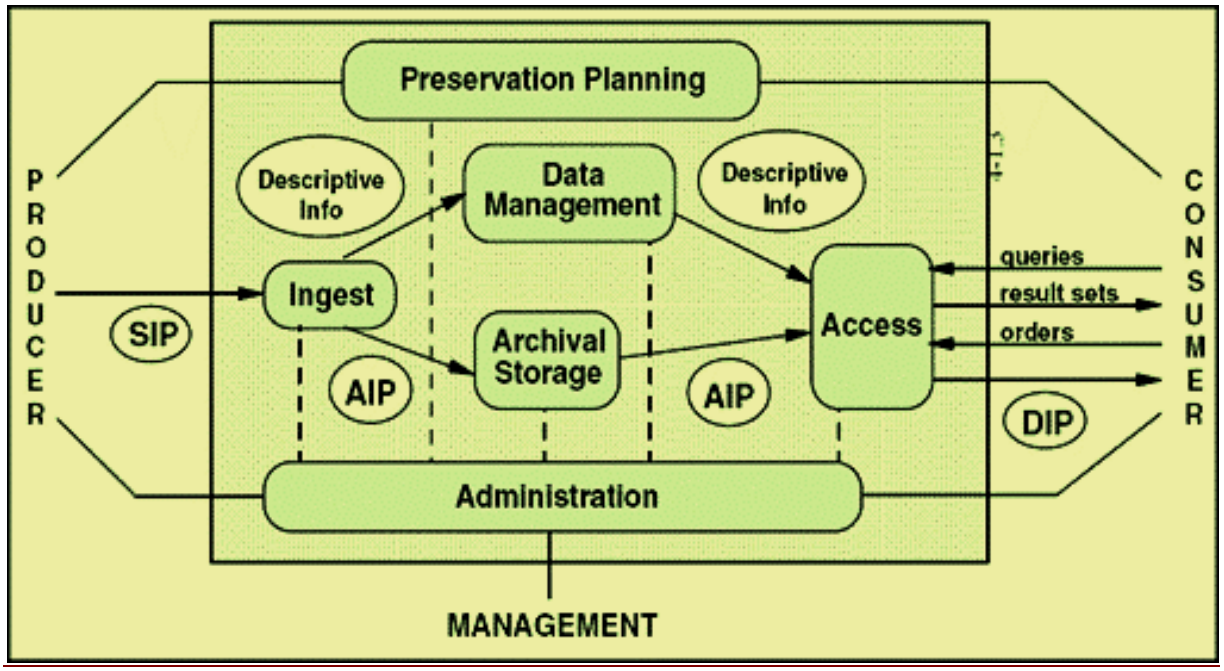

Figure 4 OAIS Functional Model

Another way to classify preservation services is in terms of stages of the digital curation lifecycle. The DCC curation lifecycle model[7] provides a useful reference model in this respect. The model identifies the high-level stages required for successful curation and preservation of data from initial conceptualisation or receipt through the iterative curation cycle. Compared to the OAIS model, the DCC model positions digital preservation activities in a wider context within the digital archive management process (including conceptualization, disposal and appraisal) and indicates that the preservation function is always interconnected with a number of other functions addressed by digital archive management systems. The model suggests that a joint approach to develop digital repository management services and digital preservation services should be beneficial in the curation lifecycle. However, currently the DP market is still in its infancy and many management tool and software do not support specifically preservation activities[8]. A graphical representation of the key elements of the model is presented in the Figure 5.

[6] 'Reference model for an Open Archival Information System', CCSDS, 2012, available at http://public.ccsds.org/publications/archive/650x0m2.pdf.

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

[8] 'An Emerging Market: Establishing Demand for Digital Preservation Tools and Services', PLANETS project, http://www.planets-project.eu/docs/reports/Planets-VENDOR-White-Paperv4.pdf.

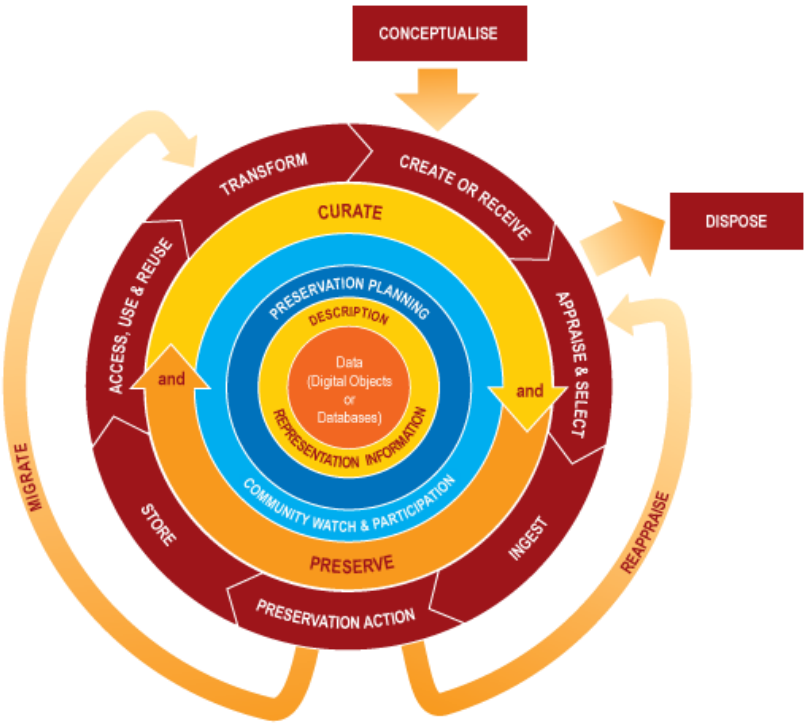

Figure 5 DCC Curation Lifecycle Model

The above entry points into the space of possible services all originate from 'user needs' at some level: the PARSE.Insight survey of researchers across many fields, the experience of JISC in directing and funding programmes of research on digital preservation, the self-reflection of a group of libraries, and so on.

A further way of thinking about preservation services is in terms of the actors involved. At the simplest level, there is one actor that provides the service and another that uses it (takes its output). There might also be a third actor that provides the input for the service. These actors might be human actors, or automated components of a digital preservation system. The report 'Long-Term Preservation Services: A description of LTP services in a Digital Library environment' recognises several (human) actors involved in the interaction with a long-term preservation system:

- *Producer* is the role played by those who provide the information to be preserved. The producer may be a direct creator of information or a publisher.
- *Consumer* is the role played by those who interact with services to find preserved information of interest and to access that information in detail.
- *Manager* is the role played by those who set overall policy.
- *Operator* is the role played by those who operate and manage the system.

These actors are of course all users of services, though some might also double as providers of input. This simple scheme has the advantage of emphasizing the motivations of the different actors for using services. It could however be refined by reference to the OAIS functional model, reproduced as Figure 4, with such a refinement reflecting the functional breakdown of the model:

- Producer
- Consumer
- Ingest

- Preservation planning
- Administration
- Data management
- Archival storage
- Access

The distinction is that these actors refer to functional entities rather than human agents. Even the producer and consumer need not be people, as the labels refer to the roles they play. As the OAIS standard makes clear in its definition of terminology, a producer signifies 'The role played by those persons or client systems that provide the information to be preserved. This can include other OAISes or internal OAIS persons or systems.' The value of thinking in terms of actors at this level is that these entities have defined roles to play, and so those roles could in principle be performed, supplemented or assisted by services.

2.3 TOWARDS A CHARACTERISATION OF HIGH-LEVEL SERVICES

The above sources of preservation evidently have much in common, though with divergence in detail. How then to characterise and assemble a coherent set of services that can be taken as a basis for examining availability and applicability? It might be possible to compare the above sources, eliminating apparent duplicates, to produce an unstructured list that would at least have the merit of comprehensiveness. However, this would be an unsatisfying approach. It would tend towards confusion of terminology, with names coming from different sources, and it would have no underlying principle allowing for extension and revision.

The approach chosen has been to refer back to the OAIS Functional Model as the point of reference for positioning preservation services. This model, as can be seen from Figure 4, comprises six fundamental functional entities:

- Ingest
- Preservation Planning
- Data management
- Archival storage
- Administration
- Access

The responsibilities of each of these are clearly defined within the OAIS standard. Indeed a master diagram in Annex A of that standard shows the sub-functions of each of these entities and their relationships in terms of information passed between them. Here we have a well-developed framework with a consistent terminology (on which the APARSEN glossary is based), framed in terms of necessary functions of parts of a repository, which therefore is a suitable basis for a framework to identify and position preservation services.

The challenge lies in identifying which of the responsibilities of these functional entities could be considered as a high-level service. It might be tempting to simply choose the full set of sub-functions as depicted in the previously mentioned master diagram, but that would be too pedantic and counter-productive. Though these sub-functions do specify quite precisely the tasks of a repository, few of them map easily to what we could consider preservation services in the sense of this document. Examples include ‘Receive submission’ (within Ingest), ‘Generate DIP’ (within Access), and ‘Archival information update’ (within Administration), all of which are too generic to be thought of as services in themselves. Some of the sub-functions might however correspond to services (‘Error checking’ within Archival storage), but not all.

It is here that a pragmatic approach, taking account of what the APARSEN consortium knows of the state of development of services, must be employed. The high-level services will be derived by analysing the underlying functions implied by the OAIS diagram, with a number of criteria applied: the aim is to make a judgement on the balance of these criteria as to what should be included.

- **The function is specifically related to preservation.** There are some functions such as ‘Perform queries’ within Data Management that will be part of the activity of a repository but are not really preservation-specific; perhaps they could be thought of in terms of services, but it would probably not be very illuminating to do so.
- **The function is sufficiently precise to be thought of and implemented as a single service, albeit at an abstract level.** Thus ‘Ingest’ itself is not a service, because it comprises many sub-functions. There is ambiguity in applying this criterion, mitigated by the possibility of allowing sub-services (see below). It should be noted that a single service certainly does not mean a single implementation.
- **The function is not likely to be highly domain-specific or repository-specific.** This refers back to the original characterisation of services presented in the introduction.

It is worth repeating that there is ambiguity in applying these criteria, and for that reason the list of high-level services, and their categorisation in this framework, should not be seen as definitive but as subject to evolution and refinement.

In particular, the possibility of sub-services is admitted. For example, the high-level service 'Environment monitoring' (also known as preservation watch) has the sub-service 'Knowledge model comparison' which is one facet of monitoring, concerned with the ability to compare knowledge models of the designated community to determine the consequences of shifts in the knowledge base. There are many other aspects of environment monitoring, such as tracking changes in file formats and software availability, so knowledge model comparison is one aspect, but one that is worth singling out, and that because it is known to be an area of research interest in which significant developments are being made.

Because the starting point of the exercise was the six functional entities of the OAIS model, it might be thought that a hierarchical organisation of the high-level services would result. This is not so, however, and it is more flexible to tag each service with one or more of the entities to which it contributes. This is simply to allow for the possibility that a single high-level service might have a place in more than part of the OAIS model. For example, format transformation is applicable both on ingest (to convert non-standard file types to acceptable types) and later (technically under administration) when preservation planning has determined that an update is required.

The following table represents the structure of preservation services developed according to the above principles. Black circles indicate where a service is a key contributor to the corresponding OAIS functional entity; while circles indicate possible or marginal relevance.

High-level service	IN	PP	DM	AS	AD	AC
Characterisation of SIPs	●					
Quality assurance of SIPs	●					
Policy-based assessment of SIPs	●					
Acquisition and maintenance of rep info	●	●			❍	
Automated metadata creation/maint	●				●	
Metadata migration					●	
Environment monitoring (preservation watch)		●				
Knowledge model comparison		●				
Preservation plan formulation		●				
Obsolescence substitution		●				
Dependency management		●				
Authenticity evidence management		●				
Appraisal of collections		●	❍	❍	❍	
DRM clearinghouse		●				●
Brokerage between repositories		●				
Long-term archiving	❍		❍	●	❍	❍
Integrity checking				●		
Cloud storage for preservation				●		
Preservation policy construction					●	
Analysis of authenticity management policies					●	
Format transformation	●				●	
Finding aids						●
Federated search						●
PID resolver						●
Emulation facilities		❍				●
Full repository service	●	●	●	●	●	●
Audit and certification of repositories	●	●	●	●	●	●

3 ELABORATING THE HIGH-LEVEL SERVICES

Having set up the list of high-level services in the previous section, the next step was to examine them in more detail, with particular reference to their implementations. This was a bottom-up exercise, with the various participants in the work package assembling contributions on the basis of their own knowledge of the field and their familiarity with particular projects or areas of work.

Each service is described under the following headings:

Name of service	A short descriptive name for reference.
Service description	Summary of the function that the service fulfils
Scenario	Illustrative scenario of usage showing the benefits of the service
Selected implementations	An overview of a range of implementations of the service; there might of course be multiple implementations, particularly where a service has a relatively broad scope
Other relevant initiatives	Any additional references, perhaps not implementations, or of only tangential relevance but nonetheless worth mentioning
Next steps	Expected next steps towards full implementation and deployment of the service
Final destination	Vision of the service when fully developed and deployed
Date of update	The date at which the above information was entered or updated

Note that the sub-services are listed on the same level as (but directly following) their parent services.

3.1 CHARACTERIZATION OF SIPS

Name of service	Characterization of SIPs
Service description	Examination of digital objects (files) regarding criteria like file format identification, detection of protection measures in online publications, format validation, extracting technical metadata for digital preservation purposes.
Scenario	The service addresses some threats of the PARSE.Insight Roadmap: how to understand or use the data, and detection of access and use restrictions. It helps to identify erroneous objects which may not be able to be presented in the future by format validation and it delivers valuable information for the representation information registry. Without file format identification the archiving institution has no information what kinds of objects are archived. This information is essential to establish an efficient preservation strategy. Preservation decisions will be based on the format and its characteristics. Moreover, format validation helps to sort out erroneous objects which may not be presented in the future. The service also allows the detection of protected data, e.g., password protected, and report about the kind of the protection measure. Protected data is likely to cause difficulties throughout the preservation process and access now and in the future. A password protected file is hard to convert into another format. Also the success of preservation actions by emulation is threatened.
Selected implementations	**Didigo (DNB internal service).** DNB is considering whether it can offer it as an external web service. Didigo uses mainly the File Information Tool Set (FITS). The core of didigo was developed in the project dp4lib as a module in the koLibRI-toolkit. **DROID**[9], a software tool developed by The National Archives (UK), performs automated batch identification of file formats. It can be used to identify the exact format of all stored digital objects in a digital repository. **FITS (File Information Tool Set)**[10] is able to identify, validate, and extract technical metadata for various file formats. **C3PO (Clever, Crafty Content Profiling of Objects)**[11] is a software tool prototype that uses FITS generated data of a digital collection and generates a profile of the content set. **FIDO**[12] is a Python command line tool to identify file formats of digital objects. It is hosted by the Open Planets Foundation. **UNIX file**[13] is a standard UNIX program with which the type of data

[9] http://droid.sourceforge.net/

[10] http://code.google.com/p/fits/

[11] https://github.com/peshkira/c3po

[12] http://www.openplanetsfoundation.org/blogs/2010-11-03-fido-%E2%80%93-high-performance-format-identifier-digital-objects

	contained in a file can be recognized. **JHOVE/JHOVE2 (JSTOR/Harvard Object Validation Environment)** is a project to develop an extensible framework for format validation. JHOVE2 has similar aims but a completely separate code base. **Xena**[14] is a file processing tool for detecting the file formats of digital objects; converting digital objects into open formats for preservation. It was developed by the National Archives of Australia.
Other relevant initiatives	Some tools/services for automated metadata creation/maintenance also have a characterization function.
Next steps	Offering a web service for the characterization service, if there is not any available already, would be the next step for the tools that do not implement this feature. Constantly evolving and supporting the detection of new file formats is important for all tools. More concrete steps are: For C3PO there will be case studies of different content types and an integration with Plato and repositories and automated monitoring services. JHOVE2 will implement modules to support important preservation-related processes.
Final destination	With the support of a characterization service that is able to identify the file format of any data, we will be able to support the preservation planning and select the appropriate preservation actions.
Date of update	December 2013

[13] e.g., http://pubs.opengroup.org/onlinepubs/9699919799/utilities/file.html

[14] http://xena.sourceforge.net/index.php

3.2 QUALITY ASSURANCE OF SIPS

Name of service	Quality assurance of SIPS
Service description	Quality assurance of Submission Information Packages is closely related to characterization (service 3.1), though the emphasis is on validation of the submitted material rather than its identification—that is, checking that the file formats (for example) are as they should be.
Scenario	The scenario is very similar to that for characterization, though the motivating force is the need to establish that material entering the repository is valid. This is a part of establishing its authenticity, as well as being an essential for a trustworthy digital repository.
Selected implementations	Many of the implementations of Characterization of SIPs and Integrity checking also fulfil the role of quality assurance on ingest.
Other relevant initiatives	The services for integrity checking of collections are also related to quality assurance.
Next steps	As for characterization of SIPs.
Final destination	A full range of quality assurance/validation services for a wide range of file formats, able to be incorporated into repositories within diverse domains and scopes.
Date of update	April 2014

3.3 POLICY-BASED ASSESSMENT OF SIPS

Name of service	Policy-based assessment of SIPs
Service description	Online service for content and metadata validation against well-defined (and certified) policies and standards.
Scenario	For metadata, a typical usage scenario would be: • Take a METS document containing PREMIS elements and validate it according to the PREMIS in METS Best Practice Guidelines • Accept a PREMIS XML document and create a METS document with embedded PREMIS elements • Accept a METS XML document and create a PREMIS document from any embedded PREMIS • One can also generate PREMIS XML document for an uploaded file For objects, the scenario is rather different: Format validation is only one aspect of a quality assessment workflow. Digitisation guidelines typically impose various constraints on the technical characteristics of preservation and access images. For example, they may state that a preservation master must be losslessly compressed, and that its progression order must be RPCL. A format profile is a set of such technical constraints. The process that compares the technical characteristics of a file against a format profile is sometimes called Policy Driven Validation. This corresponds to what JHOVE2 refers to as Assessment.
Selected implementations	**Schematron** and **Jpylyzer:** Metadata validation/conversion service[15]
Other relevant initiatives	Gap Manager originating in CASPAR project[16].Access through a dedicated SPARQL endpoint will be provided in future.
Next steps	Unified view of the scope of policy-driven validation for both metadata and digital objects.
Final destination	Integration with more general preservation policies at different levels.
Date of update	April 2014

[15] http://pim.fcla.edu/

[16] http://athena.ics.forth.gr:9090/Applications/GapManager/

3.4 ACQUISITION AND MAINTENANCE OF REPRESENTATION INFORMATION

Name of service	Acquisition and maintenance of representation information
Service description	The information needed to understand and use a digital object is termed, in OAIS, 'Representation Information'. This is a catch-all term which includes information about a digital object's format, semantics, software, algorithms, processes and indeed anything else needed. A Registry/Repository of Representation Information contains Representation Information of all kinds, and each piece of Rep Info has its own Rep Info so that a Representation Information Network (RIN) can be encoded. Ideally digital objects provided by archives will have an identifier associated; that identifier points to the start point of the Rep Info Network. The registry can also be searched in order to help create RINs from existing Rep Info.
Scenario	**Scenario 1:** A user needs additional Rep Info in order to understand a piece of digitally encoded information. Alternatively a repository wants to avoid creating Rep Info afresh by reusing existing Rep Info. In both cases the Registry can provide the required Rep Info. Note that the Rep Info must cover Structure, Semantics and any Other types. In addition the Rep Info for the Rep Info should also be available (recursion). Fuller version: A dataset created by one researcher may need to be used by a second, either contemporaneously or at some later time. This second researcher may come from a different discipline and use different analysis tools. In order to avoid producing misleading results he/she must be able to understand what the data means. For example, given an astronomical image in the FITS format, with its several variants, the researcher would need to be able to extract the values of the pixels of the image from what may be quite a complex and highly tailored digital object. In order to use an analysis tool one would need to know how to deal with these pixel values, their units, their coordinates on the sky and the way in which the photons have been selected e.g. the bandpass of the filters used. Representation Information is the OAIS term for everything that is needed in order to understand a digital object. A registry would help to ensure that the required Representation Information is available in the future and across disciplines. **Scenario 2:** An organisation maintains both document and CAD based digitised data. It uses Representation Information Networks (RIN) from its local National Archive (NA) in relation to applying preservation policies to document file formats, but they do not hold any specific policy information relating to the CAD file formats. In order to overcome this limitation, the organisation's instance of Linked Data Pronom is configured to point to the National Archive for the document formats, but it has been established that a large Aerospace Manufacturer has a better RIN store for the CAD data, so the system is configured to reference this data source for the CAD formats, as it is trusted to have a more up to date preservation policy for the CAD formats than the National Archive.
Selected implementations	**CASPAR prototype**

	PRONOM Linked data[17]**.** The PRONOM registry contains (Representation) information about file formats, compression techniques and encoding types. Linked data is about linking up related data on the web, to help expose, share and connect data, information, and knowledge through using URIs and RDF. (Resource Description Framework, aka Triples) Initially this will concentrate on modelling and publishing file format data already stored in PRONOM, using linked data standards. This is the largest core of data within PRONOM, and its first step to transform the data will be to convert existing data to RDF to describe features of each format. The new version of PRONOM will be extensible, so at a later stage the data model will be enhanced to improve other areas of information in the database. Eventually it is hoped linked data from other external data sources will populate PRONOM, transparently showing where the information came from and in doing so develop a more comprehensive technical registry. It is expected that the new version of PRONOM will be an open source system with a completely open code base. **SCIDIP-ES** sustained service for representation information registry (available from 2014 onwards)
Other relevant initiatives	UDFR
Next steps	Work with SCIDIP-ES: • Representation Information Registry holding copies of Representation Information of all types which can be shared and enhanced by contributions from many people. • Virtualisation techniques to facilitate easier integration into contemporary tools • Preservation features should be embedded in the 'creation' environment, automating/facilitating the generation of necessary representation information (data, models, configurations, ...). • Knowledge Gap Manager which provides a semi-automated way of identifying where additional Representation Information needs to be created, based on information collected by the Orchestrator/Broker • Processing Context which helps to maintain information about the processing history of a dataset Extend PRONOM Linked Data: Looking at extending the prototype developed to date so that it can clearly demonstrate the capability of accessing Policy information (RIN) from multiple suppliers in an orchestrated way.
Final destination	World-wide loosely federated Registries A set of services, supported over the long term, which make it easier to maintain adequate Representation Information, particularly after active work on the dataset has ceased or slowed. Automated capturing of the creation and processing context. An interconnected list of RDF / Linked Data providers of Representation Information Network data. Allows users to tailor their preservation policies based on their choices of who to Trust, and extend their registries with their own RIN as required.
Date of update	March 2013

[17] http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

3.5 AUTOMATED METADATA CREATION/MAINTENANCE

Name of service	Automated metadata creation/maintenance
Service description	Tool for automating the creation of metadata repositories by scanning file systems, identifying file types, extracting the embedded metadata.
Scenario	Described in the web page and also in the paper[18]: Yannis Marketakis, Makis Tzanakis and Yannis Tzitzikas, 'PreScan: Towards Automating the Preservation of Digital Objects', *International ACM Conference on Management of Emergent Digital EcoSystems (MEDES'09)*, Lyon, France, Oct. 2009.
Selected implementations	**PreScan:** currently desktop tool **Metadata extraction tool** developed by the National Library of New Zealand [19] **EMET**[20] is a file processing tool for extracting metadata embedded in JPEG and TIFF files. **Extractor**[21], which came from the PLANETS project, can extract metadata from a wide range of image files. It is part of the XCL suite (Extensible Characterization Language). **FITS**[22] wraps several tools to provide a toolset for identifying, validating and extracting technical metadata for various file formats. The tools include Jhove, DROID, Exittool, FFIdent, meta-extractor. **Jpylyzer**[23] was developed by the SCAPE project and is hosted by the Open Planets Foundation. It is a validator and property extractor for JPEG 2000 files.
Other relevant initiatives	Creation of metadata from the web is also possible: **ContextMiner**[24] provides tools to collect data and metadata by automated crawls. It supports automated crawls from blogs, YouTube, Flickr, Twitter, and open Web. It is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 United States License.
Next steps	Testing of scalability and comparison of functionalities of implementations to allow recommendations for particular situations.
Final destination	Service for repositories, archives, individuals that have to curate collections stored in plain file systems
Date of update	April 2014

[18] http://www.ics.forth.gr/isl/PreScan/docs/PreScan_Medes09.pdf
[19] http://meta-extractor.sourceforge.net/
[20] http://www.artstor.org/global/g-html/download-emet-public.html
[21] http://planetarium.hki.uni-koeln.de/planets_cms/extractor
[22] http://code.google.com/p/fits/
[23] http://www.openplanetsfoundation.org/software/jpylyzer
[24] http://www.contextminer.org

3.6 METADATA MIGRATION

Name of service	Metadata migration
Service description	Tool for migrating metadata to new schemas (expressed in RDF/S) enriched with steps for curating the quality (specificity) of the migrated descriptions.
Scenario	One description is at: http://www.ics.forth.gr/isl/RIMQA/; other scenarios are described in the paper: Y. Tzitzikas, A. Analyti and M. Kampouraki, 'Curating the Specificity of Metadata while World Models Evolve', *Proceedings of the 9th Annual International Conference on Digital Preservation (iPres2012)*, Oct. 2012, Toronto.[25]
Selected implementations	**RIQMA:** currently desktop tool
Other relevant initiatives	—
Next steps	Test in real environments
Final destination	Application to repositories, digital libraries, Linked Open Data
Date of update	March 2013

[25] http://www.ics.forth.gr/~tzitzik/publications/Tzitzikas_2012_iPres_Specificity.pdf

3.7 ENVIRONMENT MONITORING (PRESERVATION WATCH)

Name of service	Environment monitoring (preservation watch)
Service description	A service that monitors changes to the external environment that might have an impact on planning and implementation of preservation activities in the repository. Such changes include impending obsolescence of hardware, software and formats, and changes in the knowledge base of the designated community. The preservation watch service should be customisable to the needs of the particular repository and should feed into preservation planning.
Scenario	A repository of experimental data from a scientific facility contains the results of measurements taken on many sample materials over a period of years. The older data is in a specialised format developed by the facility itself, while the newer data is in a widely used community standard format. In order to reanalyse the data, a particular suite of software tools, developed by the scientific community, is required. The watch service detects that a major new release of the software is forthcoming, which will no longer be able to read the older format. This triggers preservation planning resulting in migration of all the old files into the new format. Digital archives have external dependencies, e.g., to operating systems necessary for the data stored in the archive. Thus a preservation watch service has to monitor if these dependencies change. It has to check for instance if the content fails to conform to defined policies, if the formats archived are still supported, or if there are new tools available to render the archived content.
Selected implementations	Initial development in the **PLANETS** project, being taken further in **SCAPE**
Other relevant initiatives	The reconfiguration of the **ENSURE** project[26] is responsible for catching events that indicate that initial assumptions are violated (costs of storage service, fragility of data, etc.)
Next steps	Integrate repository events and mechanisms for monitoring these (i.e. ingest , access, plan execution, ...) to be fed back into the watch component. Monitoring websites and react to specific events as well as rendering analysis of online components.
Final destination	An adaptable and generalisable service framework for monitoring many kinds of changes in the external environment, using customised plug-ins to allow extension and adaptation to complex types of monitoring. The service should be able to handle uncertainty (for example in timescales) and offer human-understandable explanations of its reasoning.
Date of update	April 2014

[26] http://ensure-fp7-plone.fe.up.pt/site/

3.8 KNOWLEDGE MODEL COMPARISON

Name of service	Knowledge model comparison
Service description	Ability to compare knowledge models expressed in RDF/S. Important for identifying changes in community knowledge (e.g. terminology). It is related to WP25.
Scenario	Described in the papers: Y. Tzitzikas, C. Lantzaki and D. Zeginis, 'Blank Node Matching and RDF/S Comparison Functions', *Proceedings of the 11th International Semantic Web Conference (ISWC'12)*, Nov 2012, Boston, USA.[27] C. Lantzaki, Y. Tzitzikas and D. Zeginis, 'Demonstrating Blank Node Matching and RDF/S Comparison Functions' , Demo at the 11th International Semantic Web Conference (ISWC'12), Nov 2012, Boston, USA.[28]
Selected implementations	Both desktop tool (command line) and web application offering this as a service
Other relevant initiatives	—
Next steps	Testing and validation in a wider range of domains
Final destination	General purpose tool that can be exploited for various purposes. For aiding humans to understand the evolution, for aiding the managers of metadata repositories, for building versioning, synchronization services, etc.
Date of update	April 2014

[27] http://www.ics.forth.gr/~tzitzik/publications/Tzitzikas_2012_ISWCpaper.pdf
[28] http://www.ics.forth.gr/~tzitzik/publications/Tzitzikas_2012_ISWCdemo.pdf

3.9 PRESERVATION PLAN FORMULATION

Name of service	Preservation plan formulation
Service description	Supports the user in planning preservation actions ahead by constructing a preservation plan. A preservation plan defines a series of preservation actions to be taken by a responsible institution due to an identified risk for a given set of digital objects or records. Such risks can be for instance file format obsolescence, problematic migration paths but also organizational risks.
Scenario	**Scenario 1:** A repository is considering a file format migration from TIFF to JPEG2000 (JP2) but is uncertain about the risks. It would like to know which file characteristics are important and should remain untouched when the conversion is taking place. **Scenario 2:** A repository is considering the migration of a large amount (2m) of image files consuming 80TB of tape space from TIFF to a newer format. Such a migration project is a critical task as the original images might be put under risk, costs can be high and the success of the overall project depends on many factors. Hence thorough planning is required that tackles questions regarding general feasibility, target image format, storage as well as the costs. Should the storage system be changed as tape storage with its linear access has become inefficient and the costs for disk storage has decreased? What are the significant properties of the images? How can we identify them? Can loss of quality be accepted? What image file formats are there available and what makes them more suitable for long term storage? Should we keep the originals or can we replace them completely? A service is needed that provides answers to these questions in a reproducible, trustworthy way. Based on a solid source of information, decisions can be made. **Scenario 3:** Video games (console games, computer games) are part of the contemporary culture and hence need to be preserved. Video game preservation is a non-trivial task as games are interactive (e.g. player input and output), use non-deterministic features (e.g. random events) and are often available only on highly specialized platforms. Being part of modern art and culture, video games need to be preserved. Hence the essential characteristics such as video, audio, gameplay, timing and many other properties need to be identified, described and preserved.
Selected implementations	**PLATO**[29] **SCIDIP-ES Preservation Strategy Toolkit**
Other relevant initiatives	**PLATTER**[30] (Planning Tool for Trusted Electronic Repositories) provides a basis for a digital repository to plan the development of its goals, objectives and performance targets over the course of its lifetime in a manner which will contribute to the repository establishing trusted status amongst its stakeholders. However this is more of a service used once when setting up the

[29] www.ifs.tuwien.ac.at/dp/plato/
[30] http://www.digitalpreservationeurope.eu/platter/

	repository, and as such complements the service Audit and certification of repositories (3.27).
Next steps	The growing complexity of digital objects will increase the complexity of preservation planning and consequently more sophisticated tools and services will be required in the near future. Preservation planning services will be designed to address not only technical issues but also to support planning decisions based on preservation goals, institutional policies, user requirements, economical issues and other constraints. A comprehensive knowledge base that collects, maintains and provides the insights gained from similar scenarios. This should also be fed into a decision support system that facilitates automated judgement.
Final destination	One or more services that enable organisations to get advice on their preservation strategies based on automated decision support systems that ensure guaranteed long term accessibility.
Date of update	April 2014

3.10 OBSOLESCENCE SUBSTITUTION

Name of service	Obsolescence substitution
Service description	Service to enable the sharing of information about the availability of hardware and software and their replacements/substitutes. Note that there is a close connection with the brokerage service.
Scenario	A performing artist finds a masterpiece of (formerly) modern music which requires a signal processing system which used to run on an Apple Mac to add a special type of reverberation to the sound. The artist has a number of options including finding the signal processing software together with a working Apple Mac, or an emulator running on his/her computer. A way to sharing information about hardware and software would facilitate the re-performance of this masterpiece.
Selected implementations	**SCIDIP-ES** Brokerage service (at least some of this) **EF Emulator archive service / EF Software archive service** — see 'Emulation' service entry for more information.
Other relevant initiatives	—
Next steps	Development and sharing of information about emulation and migration strategies Development of orchestrator/broker to share available substitutes Bringing together services with several functions: • acting as a clearing house for demands for Representation Information • collecting information about changes in availability of hardware, software, environment and changes in the knowledge bases of Designated Communities • brokering agreements about datasets between the current custodian, which is unable to continue in this role, and an appropriate successor. Improve the emulator and software archive services by containing more content and extended metadata structure.
Final destination	A set of services which make it easier to exchange information about obsolescence of hardware and software and techniques for overcoming these.
Date of update	May 2013

3.11 DEPENDENCY MANAGEMENT

Name of service	Dependency management
Service description	Ability to express the dependencies of digital objects, regarding performability of various tasks, and services for checking perfomability and possible gaps. It is related to WP25.
Scenario	Some in CASPAR deliverables. Also in the publication: Y. Marketakis and Y. Tzitzikas, 'Dependency Management for Digital Preservation using Semantic Web technologies' , *International Journal on Digital Libraries (IJDL)* , 10(4), 2009 (accepted for publication on May 2010[31] Recent advances, in the context of SCIDIP-ES and APARSEN/WP25 are described at: Y. Tzitzikas, Y. Marketakis and Y. Kargakis, 'Conversion and Emulation-aware Dependency Reasoning for Curation Services' , *Proceedings of the 9th Annual International Conference on Digital Preservation (iPres2012)*, Oct. 2012, Toronto.[32]
Selected implementations	Prototype implementations in **CASPAR** and **APARSEN**
Other relevant initiatives	See also knowledge model comparison (3.8). Both services reason about the consequences of changes in the environment, whether changes in community knowledge base or changes in ability to perform tasks.
Next steps	The prototype model for managing migration and emulation using a dependency management approach will be developed as part of the WP25 activities (Task 2520).
Final destination	To provide services that help archivists in checking whether the archived digital artefacts remain intelligible and functional and identifying risks and the consequences of probable losses or obsolescence of digital resources.
Date of update	March 2013

[31] http://www.springerlink.com/content/969066n487501h06/

[32] http://www.ics.forth.gr/~tzitzik/publications/Tzitzikas_2012_iPres_DepMgmtForCovertersEmulators.pdf

3.12 AUTHENTICITY EVIDENCE MANAGEMENT

Name of service	Authenticity evidence management
Service description	Ability to bring together evidence from diverse sources about the Authenticity of a digital object: Authenticity is not a Boolean concept. It is in general not possible to state that an object is authentic. Instead one can provide evidence on which a judgment may be made about the degree to which a person (or system) may regard an object as what it is purported to be. This evidence will be technical, for example details of what has happened to the object (Provenance) as well as social, for example does one trust the person who was in charge of the system under which the object has been held. In general the provenance information associated with various objects will be encoded according to one of a multitude of different system e.g. CIDOC-CRM[33], OPM[34]. There is at minimum a need to be able to interpret and present provenance evidence in a uniform way so that users can make an informed judgment about the degree of belief that a data object is what it is claimed to be. These tools would also facilitate the collection of appropriate evidence. The problem has been thoroughly addressed by APARSEN within WP24, and a detailed methodology has been proposed, which is based on collecting authenticity evidence along the whole digital resource lifecycle. The model and the guidelines have been successfully put to test on experimental environments provided by the project partners.
Scenario	**Scenario 1:** A virtual reconstruction of the Taj Mahal created at the start of the 21st century shows that there has been, 50 years later, subtle damage caused by a local development. The developer disputes this and argues that the digital data on which the virtual reconstruction has been made is not what is claimed. What evidence can and should be provided to support the claims of authenticity and hence save the Taj Mahal? **Scenario 2:** A digital record is produced as relevant evidence in a legal case. The record is digitally signed and has been preserved in a repository for dozens of years. During that time its format has been converted to PDF/A. What evidence should be preserved and provided to prove the authenticity and the integrity of the record? For instance it should be proved that the digital signature is authentic by showing that a reliable person had checked it, that the preservation process, which may have involved several repositories, and specifically the format conversion did not affect the content of the record. **Scenario 3:** Two scientists dispute, each claiming to be the first one to have discovered a new subatomic particle. The claim is based on some experimental datasets and on preliminary reports where data are analysed and crucial conclusions are drawn. What evidence is necessary to collect and preserve, in order to prove the authenticity and integrity of the experimental datasets and the reports, and the exact time they were produced? Regardless of the specific scenario, strong techniques and support tools are

[33] http://cidoc.ics.forth.gr/
[34] http://openprovenance.org/

	needed to allow curators to support claims of authenticity. Standardization and interoperability are crucial issues, since a digital record may potentially undergo several changes of custody and be preserved in several different repositories during its lifecycle. All changes of custody should be conveniently documented and all the evidence should be strictly associated to the digital record. According to the OAIS model this is to be a relevant part of the PDI (Preservation Description Information).
Selected implementations	The basic idea behind an authenticity evidence service is to provide designers and managers of preservation repositories a way to: a) reduce the effort needed to implement in their repository the management of the authenticity evidence, by providing them a methodology and a set of tools, i.e. pre-implemented basic modules; b) guarantee the interoperability among different repositories, by providing convenient guidelines, and defining a standard structure according to which authenticity evidence should be collected and preserved. **APARSEN** has already provided convenient guidelines as part of the activity in WP24 Authenticity and Provenance. A further important step has been to implement the methodology, and has been performed as part of a cooperation between the APARSEN and **SCIDIP-ES** projects. As a matter of fact, one of the main goals of SCIDIP-ES is to upgrade a set of prototypal preservation services into scalable, robust e-infrastructure components to support digital preservation of all types of digital objects. More specifically, the joint APARSEN-SCIDIP-ES team has designed interoperable structures to preserve the authenticity evidence, and has implemented a prototype version of authenticity management service, to be used by preservation repositories, as well as by other repositories involved in the digital resource lifecycle, to cooperate and implement their own authenticity management policies. The main results of this cooperation are[35]: • the proposal of a set of standardized XML-based structures, to represent both the provenance graph and the authenticity evidence gathered and preserved in connection with digital resource transformations and changes of custody, that are formally called Authenticity Evidence Records; • the support of interoperability among different repositories in managing the authenticity evidence, through the definition and the reference to a common dictionary, based on PREMIS Data Dictionary for Preservation Metadata, a widely acknowledged reference within the digital preservation community; • the implementation, based on REST principles and Java technology, of the Authenticity Toolkit , i.e. of a prototype version of the authenticity management service that provides a set of basic functions for the management of the authenticity evidence according to our

[35] Briguglio, L., Salza, S., Guercio, M.: Preserving Authenticity Evidence to Assess Provenance and Integrity of Digital Resources. In: P. Nesi and R. Santucci (Eds.): ECLAP 2013, LNCS 7990, pp. 66–77, 2013. Springer, Heidelberg 2013

	model, that can be accessed through an API interface. Therefore, at this level of the implementation, the Authenticity Toolkit can be exploited, in the design and implementation of a preservation repository, by using, as basic modules of the architecture, the set of methods provided by the toolkit, that include a complete set of basic functions to manage, according to our model, the authenticity evidence.
Other relevant initiatives	SCIDIP-ES, CASPAR, SHAMAN, nestor
Next steps	Propose an authenticity formalism and get consensus on it as a crucial preliminary step towards interoperability. Develop international standards and common policies on authenticity and provenance. Creation of tools to capture evidence relevant to authenticity, e.g. by consolidating the SCIDIP-ES authenticity toolkit. Develop tools to map provenance to authenticity tools. Maintain the chain of evidence through (automated) digital audit (provenance) trails by embedding support for capturing knowledge about the actual operations performed.
Final destination	A set of standards and tools through which a user in the future can be provided with evidence on which he/she may judge the degree of Authenticity which may be attributed to a digital object.
Date of update	June 2013

3.13 APPRAISAL OF COLLECTIONS

Name of service	Appraisal of collections
Service description	A service which supports the process of evaluating records to determine which are to be retained as archives, which are to be kept for specified periods and which are to be destroyed.
Scenario	The library of a research centre must take decisions on what digital material to preserve and for how long. Of particular concern is 'grey literature' which does not represent the formal output of the organisation but is nonetheless a part of its intellectual record. It may include working drafts of documents, preliminary reports, emails, etc. There is a clear link to legal issues of records management and data protection (data about individuals). But equally there is the question of what is worth preserving: how does it contribute to the 'memory' of the organisation?
Selected implementations	An approach in the context of social Web archiving has been experimented in **ARCOMEM** project[36]. ARCOMEM introduces semantic technologies for the appraisal and selection of content that is to be archived to improve the quality of Web Archives.
Other relevant initiatives	Policy-based assessment of SIPs (3.2) is related.
Next steps	Elucidation of the basis in policies: what levels of policies influence appraisal and how.
Final destination	A coherent approach to the selection of material for preservation that takes account of the legal and regulatory framework while also allowing intelligent decisions based on assessment of long-term worth of digital assets.
Date of update	April 2014

[36] http://www.arcomem.eu/about/

3.14 DRM CLEARINGHOUSE

Name of service	DRM clearinghouse
Service description	Ability to deal with Digital Rights correctly in a changing and evolving environment: Allow the digital rights associated with an object to be presented in a consistent way, taking into account the changes in legislation. There are several digital rights expression languages in the academic community and commercial world - some are being standardised – the infrastructure must be able to cope with this variety and their evolution and possibly of the underlying rights. An associated problem is the circumstance in which the licence to access the object (or without which the required software is unusable) expires and the originating company no longer exists.
Scenario	A piece of software was produced by an inventor and is protected by a user key which must be renewed every year. Several years after the death of the inventor the software is needed by a researcher in another country with a different legal system. What restrictions on usage are there under this rather different system? Even if the software could legally be used, how can the appropriate software key be created? A way is needed to be able to handle the link between the rights and restrictions originally associated with the digital object and the legal system under which it is eventually used.
Selected implementations	**Didigo** as an implementation of the described examination service below - allows the detection of protected publications and report about the kind of the protection measure. **PECAN2[37]:** A central task for academic and research libraries is to have assurance of continuity of access to e-journal content: both long-term digital preservation and access to back content in the event of cancellation of subscription. PECAN is a JISC-funded project to investigate how best to support libraries and their patrons through access to e-journal content post-cancellation. In this project, the first step towards developing a potential service for use by HEIs and publishers alike will be taken by implementing the registry software needed to store, manage and access the entitlement information in a prototype. This will involve the collection of subscription data and the introduction of infrastructure in which it can be recorded.
Other relevant initiatives	JISC recently formed the working group JARVIG[38] to further extend the clearing house construction of PECAN in combination with KEEPERS registry which hold information which archive contains which publications.
Next steps	Share information on how constraints, imposed by DRM systems on preservation planning and preservation actions can be handled under different and changing legal systems. Develop a dark archive for holding tools to generate licences, which would only be used if the commercial supplier is unable to provide this capability.
Final destination	Registry/clearinghouse for rights information; dark archive of licensing tools
Date of update	March 2013

[37] http://edina.ac.uk/projects/pecan/
[38] http://www.jisc.ac.uk/whatwedo/programmes/preservation/jarvig.aspx

3.15 BROKERAGE BETWEEN REPOSITORIES

Name of service	Brokerage between repositories
Service description	Brokering of organisations to hold data and the ability to package together the information needed to transfer information between organisations ready for long term preservation: Projects and organisations can and do run out of funding for preserving digital holdings, for example projects from Earth Observation (EO) projects are often only funded for 10 years after the closure of the satellite from which the data is derived. There are in the EO case some more or less formal mechanisms for finding a host who could take over responsibility. A brokering/orchestration system is needed to formalise the finding of new hosts. However even if agreement is reached there is the issue of collecting all the information related to a set of digital objects held, perhaps in a variety of systems, by the original host, and transferring this to the new host, itself with a variety of systems. OAIS defines in very general terms an Archival Information Package which (logically) contains all the information needed for the long term preservation of a digital object. In addition to the Brokering/Orchestration mentioned above we need to be able to create the AIP so that these can be handed over to the new host.
Scenario	An archive finds that its funding agency has been wound-up and the archive must close in six months time. Moreover the data holdings are currently in a set of inter-related database tables with embedded binary large objects, and a sophisticated access system with much embedded business logic. How can the archive find someone willing to look after its holdings and how can they be handed over in practice? Although individual repositories tend to have specialised access systems tailored to help their users, attention must also be paid to ensuring that the holdings can be handed over if/when necessary, and appropriate tools and techniques are needed to help do this.
Selected implementations	**SCIDIP-ES**
Other relevant initiatives	This issue seems to have attracted little attention in general.
Next steps	Create tools for collecting and (logically) packaging information into AIPs using information from a variety of underlying information systems Investigate the options for mapping systems from one major system to another.
Final destination	A system which will allow organisations which are no longer able to fund the preservation of a particular dataset is able to find an organisation willing and able to take over the responsibility.
Date of update	June 2013

3.16 LONG-TERM ARCHIVING

Name of service	Long-term archiving
Service description	A long-term archiving service offers a secure, trustworthy repository for digital material, carries out checks on its integrity and assures its availability. It will operate a back-up service. Operations such as ingest and access are not part of long-term archiving, though some services might include thm as well. The motivation for such a service is that bespoke solutions are expensive and require an in-house infrastructure to run on. Many smaller organisations dealing with data preservation do not have such an infrastructure, money and/or expertise to setup and maintain an archiving solution. As a result, many collections are not preserved under the right conditions, leading to data loss and unwanted risks (see also PARSE.insight survey results).
Scenario	**Scenario 1:** An organisation does not have the capacity to invest in and host its own large scale infrastructure systems and servers, but does have a commercial need to ensure the long term accessibility of its digital assets. What it needs is a preservation service where there is no need for the up-front purchase of expensive storage and processing systems or specialist system administration skills. The organisation is not sure how much capacity is needed to provision the system, so instead is looking for a system where it only pays for the storage consumed, and that can increase the computer resources as and when required. There is no need to be involved in having to patch and manage the platform, and it must be maintained within a fully backed-up, highly-secure data centre in a territory of the client's choice. **Scenario 2:** An institution aims to preserve its digital assets but does not have resources to upgrade its institutional repository to a trusted and long term preservation oriented repository. Since the content has a strong social and historical value, the institution signs an agreement with the National Library for depositing contents in the Digital Stacks. Once the contents are deposited they receive an NBN that will be included in the content metadata.
Selected implementations	**Preservica**[39] Cloud based version of Tessella's SDB - Preservation as a Service. Preservica is a new cloud-based service that provides organisations with a world class, secure, and affordable solution to safeguard their digital assets. Built on the Tessella Safety Deposit Box (SDB) platform, used by some of the world's leading archives and libraries, it offers easy upload and access to digital artefacts, as well as providing world-leading Digital Preservation tools to make sure they are accessible long into the future. The **Digital Stacks service** is a public service of trusted digital repositories for the long term preservation of deposited digital resources through a scalable infrastructure (that can expand following the requirements and the available economic resources) and flexible (that can be used for the preservation of every type of digital resources). The Digital Stacks infrastructure is suitable not only for the library domain, but also for all the areas where long term access to digital resources must be ensured. The Digital Stacks guarantees for the deposited resources: viability: the quality

[39] http://www.preservica.com/

	of being readable from media. The bit sequences of each file are intact and readable.
Other relevant initiatives	The project Digital Preservation for libraries Libraries (DP4lib)[40] project is funded by the Deutsche Forschungsgemeinschaft (DFG). The goal of this project is to evaluate the feasibility of all options for establishing and running a ready-to-operate service for long-term preservation (LTP). In addition, the preceding conceptual work will be implemented in a piece of prototype software.
Next steps	SDB (and Preservica) is a fully supported and maintained product that has an active development roadmap, whereby SDB User Group members get the opportunity to vote on and influence the what new features and capabilities are introduced into the system over time with each six monthly release of the software. DP4lib: The next step, that is currently in progress, is to set up and establish a business and cost model based on the developed prototype. The resources and funds required to set up and operate service-based LTP have to be determined for this.
Final destination	Development of an actively used preservation service that incorporates best practices in its processes and functions. Since Digital Stacks is supported the Italian Ministry of Cultural Heritage (MIBAC) it aims to become one of the most important preservation service for Universities, Publishers and Cultural Heritage institutions in Italy.
Date of update	June 2013

[40] http://dp4lib.langzeitarchivierung.de/downloads/DP4lib-One-Pager-08-eng.pdf

3.17 INTEGRITY CHECKING

Name of service	Integrity checking
Service description	It is not only individual digital objects held in a repository whose validity must be checked. In many cases, the integrity of the contents of a whole is also an important factor. At one level, this can mean the continuing integrity of files (that they have not been changed or corrupted). It can also entail, for example, detection of the presence of duplicates or near-duplicates, and the verification of the completeness of the collection (no unexpected gaps) will be a part of overall quality assurance.
Scenario	A repository receives digitised material (say historic newspapers) to ingest and add to its collection. These Submission Information Packages are received from multiple sources over a period of time. How can the repository managers know whether duplicated material is arriving (different scans of the same original images). If that is the case, how can they decide which to retain and which to dispose of?
Selected implementations	**ACE**[41] (Audit Control Environment) is a set of tools to help archives monitor the integrity of collections. ACE continuously audits the contents of the various objects according to the policy set by the archive, and provides mechanisms for an independent third-party auditor to certify the integrity of any object. **Matchbox**[42] originates from the SCAPE project and is a duplicate detection tool for digital document collections. Ii identifies duplicated content, even where files are different, e.g. in format, size, rotated, cropped, colour-enhanced. It assists in assembling collections from multiple sources, and identifying missing files.
Other relevant initiatives	There is some overlap with the services related to ingest of material, characterizing and assessing the Submission Information Packages as they arrive. This service has a broader scope, though, being also concerned with the contents of a repository after ingest.
Next steps	Expected next steps towards full implementation and deployment of the service
Final destination	A configurable, policy-driven automated integrity checker, working in the background to monitor collection integrity at different levels and to alert the repository managers as well as provide input to preservation planning.
Date of update	April 2014

[41] https://wiki.umiacs.umd.edu/adapt/index.php/Ace:Main
[42] https://github.com/openplanets/scape/tree/master/pc-qa-matchbox

3.18 CLOUD STORAGE FOR PRESERVATION

Name of service	Cloud storage for preservation
Service description	It is a tempting solution to the needs of an organisation responsible for digital preservation to take advantage of cloud storage to take care of at least the 'bit storage' of its digital assets. Of course, to be considered a trustworthy repository, much more is expected, and at least some of this (such as validation of ingested material and ongoing integrity checking of the collection) could be performed by a cloud service set up specifically for the needs of long-term digital preservation.
Scenario	A small archive lacks the infrastructure and personnel to run its own digital repository. Nonetheless it receives material in digital form. A cloud service could be an attractive way of outsourcing responsibility and allowing long-term budgeting for preservation of these materials.
Selected implementations	**DuraCloud**[43] is a service that makes it easy to move copies of content into the cloud and store them with several different providers. It provides automated storage to multiple cloud providers with integrity checks. **Preservica**[44] is Tessella's cloud-based solution for digital preservation. An overview of a range of implementations of the service; there might of course be multiple implementations, particularly where a service has a relatively broad scope
Other relevant initiatives	Commercial cloud storage providers offer services that may be suitable for some aspects of long-term preservation. The APARSEN deliverable D27.1 'Recommendations about scalability' includes in its recommendations that 'the use of cloud storage should be considered', listing the advantages but noting that 'most clouds today do not have support of the full preservation process'.
Next steps	Inclusion of additional functionalities into cloud services, such as representation information dependencies and updating.
Final destination	Ideally, a cloud service that performs almost all the functions of a trustworthy digital repository. This vision will be limited by the domain/sector dependencies of any repository, and by the limitations on what must be performed locally.
Date of update	April 2014

[43] http://duracloud.org
[44] http://preservica.com

3.19 PRESERVATION POLICY CONSTRUCTION

Name of service	Preservation policy construction
Service description	A service for assisting in constructing complete and consistent preservation policies at different levels. The ISO/CCSDS standard 'Audit and Certification of Trustworthy Digital Repositories' defines Preservation Policy and declares that it must be consistent with the Preservation Strategic Plan; there is also a Preservation Implementation Plan.
Scenario	A repository of scientific data originating from a major international facility is reconsidering its preservation policies. Many types of policy statement have been put forward and are being applied, but the question is: how do they motivate each other? How does a lower-level policy contribute to a higher-level one? A simple policy statement such as 'All data files should be associated with an experiment' seems reasonable, but how is it derived from higher-level requirements? What policies are appropriate concerning the frequency of integrity checks? The policy construction service will assist with assembling a set of policies in a coherent way.
Selected implementations	Ongoing work in **SCAPE** project on modelling levels of policies and their relationships, developing a multi-level model of policies from the general to the actionable.
Other relevant initiatives	There are of course many resources on developing policies for research data management; the DCC has links at http://www.dcc.ac.uk/resources/policy-and-legal, including a template to help define preservation policies.
Next steps	Finalisation and validation of the multi-level policy model of SCAPE in a variety of domains/repositories.
Final destination	Integrated set of policy levels with clear relationships between them allowing clear traceability from higher-level to lower-level policies.
Date of update	April 2014

3.20 ANALYSIS OF AUTHENTICITY MANAGEMENT POLICIES

Name of service	Analysis of authenticity management policies
Service description	The experience collected and the case studies carried out for the development of WP 24 (Authenticity and provenance) have clearly illustrated that the digital preservation stakeholders are not prepared to properly understand the issues related to the authenticity concepts and principles and, consequently, to design and build adequate authenticity management policies, even if the need for documenting and governing these aspects is increasingly recognized for its role, as the training effort developed by APARSEN[45] [46] and the recent standards on digital repository auditing testify. The first step to support such requirements can be implemented in the form of a consultancy activity integrated by training courses with the aim of • analyzing the problems according to the standards, the legislation, the national and international recommendations and the specific needs of the designated community, • identifying appropriate authenticity evidence to be collected in connection with the definition of accurate and complete Preservation Description Information both in the Submission Information Package and in the Archival Information Package, • designing formal procedures for the gathering and the preservation of that information, • supporting the future use of these procedures with specific training events, designed according to the internal requirements of each organization. The service we propose includes the following phases: • analyzing the needs of the Designated Community and the specific organization; • identifying the lifecycle of the content information and the relevant events which can affect the authenticity evidence; • defining in detail the authenticity evidence to collect and the main steps to control; • defining the formal implementation of the authenticity management in the form of authenticity protocols.
Scenario	**Scenario 1:** A priori authenticity management A repository is managing on a continuing basis one or more workflows related to the preservation of predefined types of content information. The efficient management of the repository implies the definition of adequate authenticity management policies and specific authenticity evidence records consistent with the content information types.

[45] APARSEN D24.1. Report on Authenticity and Plan for Interoperable Authenticity Evaluation System (2012), http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/04/APARSEN-REP-D24_1-01-2_5.pdf

[46] APARSEN D24.2. Implementation and testing of an Authenticity Protocol on a Specific Domain (2012), http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/04/APARSEN-REP-D24_2-01-2_3.pdf

	Scenario 2: A posteriori authenticity management A repository, which until now has not deployed any systematic authenticity management policy, is interested in providing the consumer with appropriate evidence to assess the authenticity of its preserved content information. In a posteriori scenario the effort has to be dedicated to the collection of existing documentation and information relevant for building an authenticity evidence record to make available for users.
Selected implementations	The basic reference to be used in the analysis is given the guidelines and recommendations developed within **APARSEN** and documented. These guidelines have been implemented in cooperation with **SCIDIP-ES**[47] to provide a set of elementary authenticity services as modules of the SCIDIP-ES Authenticity Toolkit and the approach has also been successfully tested in a series of case studies[48]. Regulations under development in many countries and standards approved by ISO (ISO 16363 on auditing digital repositories) provide further references for evaluating and improving the structure of authenticity evidence records and authenticity protocols, referred to specific types of content information[49].
Other relevant initiatives	InterPARES, CASPAR, SCIDIP-ES, SHAMAN, PRELIDA
Next steps	Promote a standard structure for the AER which may guarantee the interoperability among different repositories, specifically in case of similar designated communities environments. Define a set of criteria for a formal certification of the authenticity management policies in long term digital repositories to provide accurate and complete documentation as part of an audit and consistently with ISO 16363.
Final destination	Provide an adequate formal framework, detailed methodologies and tools to develop both consulting services for the repository management and to provide certification services. Support training courses and formal education thanks to the development of formal guidelines.
Date of update	June 2013

[47] Briguglio, L., Salza, S., Guercio, M.: Preserving Authenticity Evidence to Assess Provenance and Integrity of Digital Resources. In: P. Nesi and R. Santucci (Eds.): ECLAP 2013, LNCS 7990, pp. 66–77, 2013. Springer, Heidelberg 2013

[48] Salza, S., Guercio, M.: Authenticity Management in Long Term Digital Preservation of Medical Records. In: Moore, R., et al. (eds.) iPRES 2012, 9th International Conference on Preservation of Digital Objects, Toronto, Canada, October 1-5, 2012, pp. 172–179 (2013), https://ipres.ischool.utoronto.ca/sites/ipres.ischool.utoronto.ca/files/iPres%202012%20Conference%20Proceedings%20Final.pdf

[49] Guercio, M.: Conservare il digitale. Principi, metodi e procedure per la conservazione a lungo termine di documenti digitali, Laterza 2013, Roma-Bari, pp. 23-34, 101-156. www.laterza.it/index.php?option=com_laterza&Itemid=97&task=schedalibro&isbn=9788858107959

3.21 FORMAT TRANSFORMATION

Name of service	Format transformation
Service description	For any digital collection there will be a certain point in time that action is required to retain it accessible for users. The most common actions are transformation (also known as migration), when the file formats need to be converted, and emulation, when the originally used computer platform is recreated. There are a considerable number of commercial products on the market providing format transformation, e.g., Adobe Acrobat which can be used to migrate for instance a Microsoft Word document into a PDF/A ISO-standardized document.
Scenario	In [50] the migration of the content of a digital preservation service for electronic journals, books, and other content is described in detail.
Selected implementations	In **SCAPE** different workflows implemented preservation action tools with Taverna , e.g., for a migration action with imagemagick convert or a MP3 to WAV migration using a ffmpeg web service . In PLANETS, RODA, and CriB different commercial and off-the-shelf tools were used for preservation action services, such as ImageMagick for handling bitmap images. RODA also included an own library for the migration of databases. **MIXED**[51] (Migration to Intermediate XML for Electronic Data) is a service that can convert files with tabular data (spreadsheets and databases) to supplier independent XML for long term preservation. This is the so called Standard Data Format for Preservation (SDFP). MIXED does not store the converted files, but returns them to you for download and keep 'for ever'. A SDFP file can also be converted back by MIXED to the original supplier dependent format, or to formats of other suppliers, or - in the future - to not yet existing formats.
Other relevant initiatives	There are numerous tools/services for performing particular format transformations for graphics files, sound files, office documents, etc.
Next steps	Build sustainable collections or registries for the preservation action services. These have to be comprehensive and might also be set up specifically for different domains.
Final destination	One or more operational preservation action services (migration, emulation) which can be used for access to digital collections and for validation of digital objects. Preservation action as a service: A set of preservation actions, deployed in a framework for integration, offered as a service.
Date of update	May 2013

[50] Sheila M. Morrissey, Vinay Cheruku, John Meyer, Matthew Stoeffler, William Howard, and Suresh Kadirvel. Migration at Scale: A Case Study. IPRES 2012, October 2012 , Toronto, Ontario, Canada .

[51] See http://mixed.dans.knaw.nl

3.22 FINDING AIDS

Name of service	Finding aids
Service description	Finding aids refer to any mechanisms by which content of a repository or multiple repositories may be searched, identified and located. Federated search over multiple repositories (3.23) is one specialised aspect of this; persistent identifiers (3.24) are an enabling mechanism. Because the scope of this service is very broad, and the functionality in general is not really specific to the context of preservation, this service is not completed as the others are, but is left as a placeholder for the sake of completeness.
Scenario	
Selected implementations	
Other relevant initiatives	
Next steps	
Final destination	
Date of update	December 2013

3.23 FEDERATED SEARCH

Name of service	Federated search
Service description	A service that allows a single point of entry to search across a large number of separate collections of digital material, bringing together resources of a similar nature that are otherwise geographically or organisationally separate. Federated search is also known as simultaneous searching, cross-database searching, parallel searching, broadcast searching, and integrated searching.
Scenario	Federated Search services provide a single search environment allowing users to search multiple information sources using a single, integrated interface. The user submits a single search query to distributed and heterogeneous sources and receives in return a unified list of results possibly with minimal duplication. Information sources can be heterogeneous in many respects, for example their location, the format of information they offer or the technologies they are based on.
Selected implementations	**ZPORTAL**[52] is a federated search engine. **Ex Libris MetaLib**[53] is an information portal to library collections. **Endeavor EnCompass**[54] is a digital management and search service for integrating information from heterogeneous sources. **Summon**[55] service is a Web-scale discovery service to library resources.
Other relevant initiatives	Examples of high-profile services that make use of fedetared search include: **Europeana**[56] provides a service that links archives, libraries, museums, and audio-visual material across Europe. It aggregates metadata from various cultural heritage providers. It provides a unified way to search various object collections using that metadata through a Web portal or an API. (formerly ISI) **Web of Knowledge**[57] is a research platform for information in science, social sciences, humanities and arts.
Next steps	Federated search for multilingual sources and multilingual access features
Final destination	Single federated search portal using persistent identifiers.
Date of update	March 2013

[52] http://oclc.org/zportal.en.html
[53] http://www.exlibrisgroup.com/category/MetaLibOverview
[54] http://www.infotoday.com/it/oct01/news19.htm
[55] http://www.serialssolutions.com/en/services/summon
[56] http://www.europeana.eu/
[57] http://wokinfo.com/

3.24 PID RESOLVER

Name of service	PID resolver
Service description	There is no shortage of things which are claimed to be Persistent Identifier systems. The issues associated with these are the scalability of the solutions and the longevity of the underlying organisational structure. A name resolving system whose persistence is guaranteed by an international, government based organisation is needed. This could build on one or more existing name resolving systems, strengthening the organisational structures underpinning the resolver.
Scenario	**Scenario 1:** A researcher reads a paper in a journal which refers to a dataset which he realises can be re-analysed and combined with some new data he has recently obtained. The paper has an identifier string for the dataset which after some investigation he sees is some sort of a "persistent identifier". Unfortunately the originator of that system is long gone, the DNS entry for the identifier name resolver system host has lapsed and the database system which was used is not available. A more permanent persistent identifier system is needed which itself has the appropriate longevity with committed long-term financial and social support. **Scenario 2:** A Project Officer who works for the European Commission is filling the Impact Assessment Form, which is part of the final evaluation report of a project he has been assigned. The project adopted the new EC policy on persistent identifiers (PIs), which means that any relevant entity connected to the project has a unique identifier. This means that the project itself has an identifier, together with every contributor (scientists, PhD students, developers, experts, ...) and every involved organization (universities, research institutes, companies, public bodies, ...). The Project Officer interrogates the ENS to get the PI of the project assigned to him. Once accessing the Project Management System of the EC, he uses the project ID to collect all he needs to know to fill the form: the complete list of papers published or submitted for publication by the project team, a rich collection of metadata about them (type of publication, degree of inter-organizational collaboration, impact factors of journals, statistics on external and self-citations, and the like), a list of patents filed in the course of the project. He can browse the graph of EC funded projects and explore the history of past collaborations. He can retrieve (and visualize on a map) other data, which prove the impact of the project in terms of jobs created within the participating partners, number and location of start-up companies and compute an overall impact indicator. This scenario shows how the integration of data through unique global IDs may allow policy makers to make much better informed decisions on the assignment of funds and provide a much more complete account to member states governments and European citizens.
Selected implementations	An example for an implementation of a PID resolver service is the URN:NBN Resolver for Germany & Switzerland[58] . The German National Library assigns and administrates URNs from the namespace ‘urn:nbn:de’ and offers an URN resolving service for Germany and Switzerland.

[58] http://nbn-resolving.org/

	URN:NBN Resolver for Italy - The Florence National Library assigns and administrates the 'urn:nbn:it' for the resources that are legally deposited in the National Library such as PhD Dissertation. OKKAM Entity Name System (ENS) is a large scale infrastructure for supporting the systematic creation and reuse of identifiers for entities which are named on digital documents and resources. The vision behind the ENS is that of enabling the Web of Entities, namely a virtual space where any collection of digital data and information about any type of entities (e.g. people, locations, organizations, events, products, ...) can be integrated across the boundaries of local systems, like the Web did for hypertexts. The idea is to cut to the root the proliferation of unnecessary new identifiers for naming the entities which already have a public identifier and support the interoperability between these identifiers. To this purpose, the system provides a set of core functionality, i.e. searching for entities, adding new entities, creating new IDs, adding/searching for alternative IDs. Through these basic services, the system provides content creators, editors and developers with a global infrastructure and a collection of new tools and plugins which support them to easily find public identifiers for the entities named in their contents/services, use them for creating annotations, build new network-based services which make essential use of these identifiers in an open environment (like the Web or large Intranets). EPIC: the European Persistent Identifier Consortium Service for the European Research Community is a PID service based on Handle for the allocation and resolution of persistent identifiers to the European Research community. The service is based on Handle and uses as a prefix the number 11858.
Other relevant initiatives	ORCID initiative for identification of academic authors, providing persistent identifiers for humans.
Next steps	The OKKAM team is now working on the development of the ENS 2.0 which aims to implement a new identification schema by integrating the Persistent Identifiers approach with that of Cool URIs. The German National Library currently cooperates with the National Library of Sweden to make the architecture of the resolving service redundant by using clustering technologies. As an observer DANS (NL) is also taking part in this project and might be a further node to improve the fail-safe stability in the future. The objective is to build up and establish a trustworthy and stable PID service for libraries in Europe (this could be also mentioned in the column final destination). Furthermore it is planned to implement a REST-interface for using common features like registering and updating URNs via a standardized machine interface.
Final destination	An identifier system for locating and cross-referencing digital objects which has adequate organisational, financial and social backing for the very long term which can be used with confidence. A large scale distributed service which permanently stores identifiers for entities and provides a collection of core services (e.g. entity matching, ID mapping and resolution) needed to support their pervasive reuse, on top of which several value added services can be built (e.g. global resolution service, citability and provenance services).

	A PID service, based on the handle system[59] for the allocation and resolution of persistent identifiers.
Date of update	June 2013

[59] http://www.handle.net/

3.25 EMULATION FACILITIES

Name of service	Emulation facilities
Service description	For any digital collection there will be a certain point in time that action is required to retain it accessible for users. The most common actions are migration, when the file formats need to be converted, or emulation, when the originally used computer platform is recreated.
Scenario	A performing artist finds a masterpiece of (formerly) modern music which requires a signal processing system which used to run on an Apple Mac to add a special type of reverberation to the sound. The artist has a number of options including finding the signal processing software together with a working Apple Mac, or an emulator running on his/her computer.
Selected implementations	**Emulation Framework (EF)**[60] supports the user in recreating an authentic computer environment using emulation. The EF automates almost all steps from object identification until compiling and configuring the emulated environment. It uses existing well known emulators such as QEMU and Dioscuri. The basic set of supported environments consist of x86, C64, Amiga, BBC Micro, Thomson and Amstrad, but this set can be extended. The EF comes with a Software archive web service and an Emulator archive web service to preserve emulators and software centrally in an organisation or across the internet. The **bwFLA** project provides emulation as a service. Their goal is to define and provide a practical implementation of archival workflows for the rendering of digital objects in its original environment. The purpose of the developed workflows is to use them in library and archival systems. Their distributed knowledge-base on different operating systems regarding emulator compatibility and hardware dependencies is part of their preservation action.
Other relevant initiatives	The concept of the 'Universal Virtual Computer', though this is still a subject of research[61].
Next steps	Further extend the EF with broader support of emulated platforms and software Lobby at EU for a change in the legal framework to allow reuse of old software and emulated hardware without charge Lobby at software/hardware vendors to allow the reuse of their software/hardware for preservation purposes.
Final destination	One or more operational preservation action services (migration, emulation) which can be used for access to digital collections and for validation of digital objects.
Date of update	April 2014

[60] http://emuframework.sf.net/

[61] van Diessen R. J., van der Hoeven J. R., van der Meer K., 2005. Development of a Universal Virtual Computer (UVC) for long-term preservation of digital objects, 31(3), pp. 196-208 doi:10.1177/0165551505052347

3.26 FULL REPOSITORY SERVICE

Name of service	Full repository service
Service description	This service is really an aggregate of other services, but is included because it could be offered as a full-scale service (a slightly different sense of the word 'service'). It encompasses any offering that outsources a wide range of repository functions in a coherent framework. Evidently it is closely connected with the services Long-term archiving and Cloud storage for preservation (though intended to be wider in scope than both of those).
Scenario	A small organisation with responsibility for archiving and preservation of material might lack the expertise and infrastructure to perform this task in-house. In such a case, a full repository service offered by another party would be an attractive option.
Selected implementations	**Archive-It**[62] is a web archiving service to capture, build, and manage collections of web content. It is a subscription web archiving service from the Internet Archive that helps organizations to harvest, build, and preserve collections of digital content. **RODA**[63] is a full-featured open-source digital preservation repository providing all the functionality prescribed by the OAIS reference model. It is capable of ingesting, managing and providing access to the various types of digital objects produced by large corporations or public bodies.
Other relevant initiatives	Cloud-based offerings such as **Preservica** (see 3.18) fulfil at least some of the functions required.
Next steps	Study of how completely the requirements for a trustworthy digital repository (for example, as defined in the ISO/CCSDS standard 'Audit and certification of trustworthy digital repositories'[64]) can be fulfilled by such a service.
Final destination	A range of repository services adaptable to organisations of different sizes and in different domains, covering all the functions of a trustworthy repository, with appropriate service agreements.
Date of update	April 2014

[62] http://www.archive-it.org
[63] www.roda-community.org
[64] http://public.ccsds.org/publications/archive/652x0m1.pdf

3.27 AUDIT AND CERTIFICATION OF REPOSITORIES

Name of service	Audit and certification of repositories
Service description	Certification process so that one can have confidence about whom to trust to preserve data holdings over the long term: Although one cannot guarantee anything into the indefinite future there has, for more than a decade, been a demand for an international process for accreditation, auditing and certification of digital repositories, based on an ISO standard.
Scenario	A funding agency wishes to instruct its researchers to deposit their data into one or other of the long term archives it will support. This will involve a large and continuing commitment of resources. How can the funder be sure that the archives it wishes to support are up to the job? An internationally recognised certification system would give funders and depositors a way to distinguish and evaluate archives.
Selected implementations	**SCIDIP-ES** will create a toolkit which will help. **Nestor** - Certification working group: Based on the DIN 31644: "Criteria for trusted digital repositories", the WG designed two certification processes with different levels. The first stage will include an extended self-evaluation based on the DIN with an assessment and the award of a nestor seal and registration, in order to help implement the 2010 agreed 'Memorandum of Understanding to create a European framework for audit and certification of digital repositories'. **DRAMBORA**[65] is intended to facilitate internal audit by providing repository administrators with a means to assess their capabilities, identify their weaknesses, and recognise their strengths.
Other relevant initiatives	Data Seal of Approval[66], TRAC[67]
Next steps	Support the development of a set of ISO standards about digital repository audit and certification Help set up the organisation and processes to provide accreditation and certification services
Final destination	An internationally recognised accreditation, audit and certification process with a well-defined and long-lived support organisation, with appropriate tools and best practice guides.
Date of update	March 2013

[65] http://www.repositoryaudit.eu/
[66] http://datasealofapproval.org/
[67] http://www.crl.edu/PDF/trac.pdf

4 CLASSIFYING SERVICES AND MAPPING THE LANDSCAPE

4.1 SERVICES AGAINST THE STANDARD FOR AUDIT AND CERTIFICATION OF REPOSITORIES

In order to obtain a view of the landscape of preservation services, two different approaches have been taken. One is to position the services within the structure of metrics of the standard on audit and certification of digital repositories[68]. The rationale is that the metrics evaluate functions that a trustworthy repository must carry out; thus, in principle at least, they could be offered as services by a third party (whether automated services or human consultancy/advice services).

It might be thought that this is not really an independent analysis from the OAIS functional model on which the list of high-level services was based, since the audit and certification standard derives ultimately from OAIS. However, the standard has been the subject of lengthy and intensive development aimed at making precise what a repository must do in many areas to be considered trustworthy, and it is this that gives value to taking it as a basis for the landscape of services.

The ISO standard 16363, which is the same as the CCSDS standard CCSDS 652.0-M-1 'Audit and certification of trustworthy digital repositories' [5], follows from the OIAS standard to provide reliable and comprehensive assessment of digital archives or repositories, measuring the degree to which they can be trusted to preserve their contents into the future and maintain access and usability. The standard offers a detailed specification of criteria by which digital repositories may be audited. As with OAIS, its scope is the entire range of digital repositories, despite its origins in the space science field.

The standard presents a series of metrics under the following main headings:

- Organizational Infrastructure
- Digital Object Management
- Infrastructure and Security Risk Management

Each metric is accompanied by discussion and examples of how a repository can show it is meeting the requirement expressed in the metric.

The point is that the activities and operations specified in this standard as being expected of a trustworthy repository imply the possibility of services to assist with them.

In many cases it was necessary to interpret the metrics in terms of a corresponding service. As a typical example, metric 3.1.2 reads:

> *The repository shall have a Preservation Strategic Plan that defines the approach the repository will take in the long-term support of its mission.*

The corresponding service is taken to be 'Production of Preservation Strategic Plan'—that is, a service that will assist with that task which, if performed satisfactorily, will mean that the metric is satisfied.

Another example is metric 4.2.2, which reads:

> *The repository shall have a description of how AIPs are constructed from SIPs.*

Here the corresponding service is titled 'Set-up of description of how AIPs are constructed from SIPs'—so the service assists with the creation of the description that is required to satisfy the metric. One might argue that such 'one-off' tasks are unlikely to be the subject of services, as opposed to those tasks that are routinely and repeatedly carried out as part of the repository's operations. Nonetheless there is a potential, and the purpose of the exercise is to understand the landscape and perhaps to identify opportunities that had not been considered before.

A final example, where less interpretation is needed, is metrics 4.2.8, which reads:

[68] 'Audit and certification of trustworthy repositories', CCSDS, 2011, available at http://public.ccsds.org/publications/archive/652x0m1.pdf.

The repository shall verify each AIP for completeness and correctness at the point it is created.

Here the corresponding service is simply 'Verification of each AIP for completeness and correctness at the point it is created'—the service is a direct translation of the requirement of the metric.

The criterion for a matching a service to a metric was that the service contributes in some way to the fulfilment of the conditions expressed by the metric.

Metric no	Metric interpreted as service	Correspondence in list of services in WP21	Notes
3	ORGANIZATIONAL INFRASTRUCTURE		
3.1	GOVERNANCE AND ORGANIZATIONAL VIABILITY		
3.1.1	Production of mission statement		
3.1.2	Production of Preservation Strategic Plan	Preservation policy construction Brokerage between repositories	Includes succession plan, contingency plan, escrow arrangements, and monitoring of organizational environment
3.1.3	Production of collection policy	Preservation policy construction	
3.2	ORGANIZATIONAL STRUCTURE AND STAFFING		
3.2.1	Identification of duties the repository needs to perform	Preservation policy construction	Includes staff with adequate skills and experience (recruitment, number, professional development)
3.3	PROCEDURAL ACCOUNTABILITY AND PRESERVATION POLICY FRAMEWORK		
3.3.1	Definition of Designated Community and associated knowledge bases		
3.3.2	Production of Preservation Policies	Preservation policy construction Analysis of authenticity management policies	
3.3.3	Documentation of history of changes to operations		
3.3.4	Transparency and accountability (could this be a service?)		
3.3.5	Definition/collection/tracking of information integrity measurements	Analysis of authenticity management policies	
3.3.6	Conduct of self-assessment and external certification	Audit and certification of repositories	
3.4	FINANCIAL SUSTAINABILITY		
3.4.1	Production and implementation of short- and long-term business planning processes		
3.4.2	Definition and audit of financial practices and procedures		
3.4.3	Commitment to analyze and report on financial risk, benefit, investment, and expenditure (service?)		
3.5	CONTRACTS, LICENSES, AND LIABILITIES		
3.5.1	Set-up and checking of contracts or deposit agreements	DRM clearinghouse	
3.5.2	Tracking and managing IPR and restrictions	DRM clearinghouse	

4 DIGITAL OBJECT MANAGEMENT																
4.1	**INGEST: ACQUISITION OF CONTENT**															
4.1.1	Identification of Content Information and Info Properties to preserve	Preservation policy construction														
4.1.2	Specification of information associated with Content Info at deposit	Acquisition and maintenance of Rep Info														
4.1.3	Specification for recognition and parsing of SIPs	Characterization of SIPs Policy-based assessment of SIPs Automated metadata creation/maintenance Appraisal of collections														
4.1.4	Identification of Producer	Appraisal of collections														
4.1.5	Ingest process which verifies each SIP for completeness and correctness	Characterization of SIPs Policy-based assessment of SIPs Automated metadata creation/maintenance Appraisal of collections														
4.1.6	Defining and obtaining sufficient control over Digital Objects															
4.1.7	Provision of responses to producer/depositor															
4.1.8	Keeping contemporaneous records of actions and administration processes that are relevant to content acquisition	Characterization of SIPs? Automated metadata creation/maintenance?														
4.2	**INGEST: CREATION OF THE AIP**															
4.2.1	Creation of definition that is adequate for parsing the AIP and fit for long term preservation needs		Includes components of AIP													
4.2.2	Set-up of description of how AIPs are constructed from SIPs															
4.2.3	Documentation of the final disposition of all SIPs															
4.2.4	Specification and implementation of persistent, unique identifiers for all AIPs	PID resolver														
4.2.5	Resources for provision of authoritative Representation Information for all of the digital objects	Automated metadata creation/maintenance Knowledge model comparison Acquisition and maintenance of Rep Info	Includes file type, RI to make understandable, assoc with Content Info													
4.2.6	Definition and documentation of processes for acquiring Preservation Description Information for its associated Content Information	Authenticity evidence management														
4.2.7	Ensuring that Content Information of the AIPs is understandable for their Designated Community at the time of creation of the AIP	Policy-based assessment of SIPs Automated metadata creation/maintenance Knowledge model comparison Dependency management? Acquisition and maintenance of Rep Info Format transformation Emulation facilities														
4.2.8	Verification of each AIP for completeness and correctness at the point it is created	Policy-based assessment of SIPs Automated metadata creation/maintenance Knowledge model comparison Dependency management? Acquisition and maintenance of Rep Info Emulation facilities														
4.2.9	Provision of an independent mechanism for verifying the integrity of the repository collection/content	Integrity checking														
4.2.10	Keeping contemporaneous records of actions and administration processes that are relevant to AIP creation	Integrity checking														

4.3	PRESERVATION PLANNING		
4.3.1	Production of documented preservation strategies relevant to its holdings	Preservation plan formulation	
4.3.2	Set-up of mechanisms for monitoring its preservation environment	Acquisiton and maintenance of Rep Info Dependency management Knowledge model comparison Automated metadata creation/maintenance Environment monitoring (preservation watch)	Includes adequacy of RepInfo
4.3.3	Set-up of mechanisms to change its preservation plans as a result of its monitoring activities	Automated metadata creation/maintenance Environment monitoring (preservation watch) Knowledge model comparison Obsolescence substitution Dependency management Metadata migration Acquisition and maintenance of Rep Info	Includes gathering extra RepInfo
4.3.4	Provision of evidence of the effectiveness of its preservation activities	Authenticity evidence management Integrity checking	
4.4	AIP PRESERVATION		
4.4.1	Production of specifications for how the AIPs are stored down to the bit lev	Preservation action? Long-term archiving?	
4.4.2	Keeping contemporaneous records of actions and administration processes that are relevant to storage and preservation of the AIPs	Authenticity evidence management Preservation action? Integrity checking	
4.5	INFORMATION MANAGEMENT		
4.5.1	Specification of minimum information requirements to enable the Designated Community to discover and identify material of interest	Finding aids Federated search	
4.5.2	Capture/creation of minimum descriptive information and ensure that it is associated with the AIP		
4.5.3	Maintenance of bi-directional linkage between each AIP and its descriptive information		
4.6	ACCESS MANAGEMENT		
4.6.1	Compliance with Access Policies (incl log access mgmt failures)		
4.6.2	Set-up and implementation of policies and procedures that enable the dissemination of digital objects that are traceable to the originals, with evidence supporting their authenticity	Metadata migration? Authenticity evidence management Analysis of authenticity management policies Emulation	Includes acting on problem reports

5	**INFRASTRUCTURE AND SECURITY RISK MANAGEMENT**		
5.1	**TECHNICAL INFRASTRUCTURE RISK MANAGEMENT**		
5.1.1	Identification and management of risks to its preservation operations and goals associated with system infrastructure	Environment monitoring (preservation watch)	Includes technol watch, backup, corruption detection, security updates, critical processes
5.1.2	Management of the number and location of copies of all digital objects	Long-term archiving	Includes synchronisation
5.2	**SECURITY RISK MANAGEMENT**		
5.2.1	Systematic analysis of security risk factors associated with data, systems, personnel, and physical plant		
5.2.2	Implementation of controls to adequately address each of the defined security risks		
5.2.3	Enforcement of delineated roles, responsibilities, and authorizations related to implementing changes within the system		This relates to staff
5.2.4	Preparation of suitable written disaster preparedness and recovery plan(s)	Long-term archiving?	Includes at least one off-site backup of all preserved information together with an offsite copy of the recovery plan(s)

4.2 SERVICES AND PRESERVATION SCENARIOS

The discussions within the APARSEN community and especially within the work package partners have shown that there are some similarities between the consideration of preservation services and preservation scenarios as developed in WP14 on common testing environments. Preservation scenarios were collected in WP14 as use cases of the involved work package partners. The intention was to get a summary of representative uses cases in the context of long term preservation and to contrast them with the identified test beds and tools for performing long term preservation actions.

The following table shows the mapping of the relationship between services and scenarios and uses the service as the starting point. Note that not all services are present, as some (such as 'Brokerage between repositories') do not correspond to the needs of any of the scenarios.

Service	Scenario
Characterization of SIPs	• File format validation of static online publications (newspapers in PDF/A-b) (PS-003) • Migration of a sample of digitization master files from TIFF to Jpeg 2000 (PS-004) • Detection of protection measures in online publications (PS-006) • Format identification and format validation of online publications (PS-007) • Add Structural & Semantic Representation Information to Scientific Data (PS-008) • Preservation of Born-Digital Government Departmental Records (PS-015) • Digitisation of Paper based Government Records for future access (PS-016) • Preservation of Scientific Datasets (PS-017)
Quality assurance of SIPs	• File format validation of static online publications (newspapers in PDF/A-b) (PS-003) • Detection of protection measures in online publications (PS-006) • Format identification and format validation of online publications (PS-007) • Preservation of Born-Digital Government Departmental Records (PS-015) • Digitisation of Paper based Government Records for future access (PS-016)
Policy-based assessment of SIPs	• File format validation of static online publications (newspapers in PDF/A-b) (PS-003) • Format identification and format validation of online publications (PS-007) • Evaluation of various ‘Office’ type migration paths (PS-014)
Acquisition and maintenance of representation information	• Format identification and format validation of online publications (PS-007) • Add Structural & Semantic Representation Information to Scientific Data (PS-008) • Migration to a more open format and adding more Representation Information (PS-010)
Automated metadata creation/maintenance	• File format validation of static online publications (newspapers in PDF/A-b) (PS-003) • Detection of protection measures in online publications (PS-006) • Format identification and format validation of online publications (PS-007) • Add Structural & Semantic Representation Information to Scientific Data (PS-008) • Preservation of Born-Digital Government Departmental Records (PS-015)

	• Digitisation of Paper based Government Records for future access (PS-016) • Preservation of Scientific Datasets (PS-017)
Metadata migration	• Migration of binary data to a more commonly used & accessible format (PS-009)
Environment monitoring (preservation watch)	• Migration to a more open format and adding more Representation Information (PS-010) • (Note: many of the 'migration' scenarios could also involve this service, with the preservation watch triggering planning and consequent migration in response to possible obsolescence of formats) • Evaluation of various 'Office' type migration paths (PS-014)
Knowledge model comparison	• *Link to work package 25 – Interoperability and Intelligibility*
Preservation plan formulation	• Evaluation of various 'Office' type migration paths (PS-014) • Digitisation of Paper based Government Records for future access (PS-016)
Obsolescence substitution	• Migration of binary data to a more commonly used & accessible format (PS-009) • Migration to a more open format and adding more Representation Information (PS-010) • Migrate binary data format to a standardised data description language (EAST) (PS-011) • Migration of Contemporary Performing Arts data in to alternative file formats (PS-012) • Migrate various World Heritage Site data into composite open formats (PS-013) • Evaluation of various 'Office' type migration paths (PS-014) • Preservation of Scientific Datasets (PS-017) • Migration of Social Science and Humanities Data between statistical formats (PS-018)
Dependency management	• *Link to work package 25 – Interoperability and Intelligibility*

Authenticity evidence management	• Preservation of Born-Digital Government Departmental Records (PS-015) • Digitisation of Paper based Government Records for future access (PS-016)
Appraisal of collections	• Preservation of Born-Digital Government Departmental Records (PS-015) • Digitisation of Paper based Government Records for future access (PS-016)
DRM clearinghouse	• *Link to work package 31 – Digital Rights and Access Management* • Detection of protection measures in online publications (PS-006)
Long-term archiving	• Preservation of Born-Digital Government Departmental Records (PS-015) • Digitisation of Paper based Government Records for future access (PS-016) • Preservation of Scientific Datasets (PS-017)
Integrity checking	• (Note: this service is in the background of many scenarios concerned with migration, ensuring that after the migration operations are carried out the contents of the repository maintain their integrity as a whole)
Preservation policy construction	• (Note: this service is in the background of many scenarios, linking high-level policies to more operational policies and thus motivating the scenarios)
Analysis of authenticity management policies	• *Link to work package 24 – Authenticity and Provenance*
Format transformation	• Migration of Datasets via an "intermediate XML format" (PS-001) • Migration of static online publications (theses) from PDF to PDF/A (PS-002) • Migration of a sample of digitization master files from TIFF to Jpeg 2000 (PS-004) • Migrations of audio files from BWF to FLAC and WavPack (PS-005) • Migration of binary data to a more commonly used & accessible format (PS-009) • Migration to a more open format and adding more Representation Information (PS-010) • Migrate binary data format to a standardised data description language (EAST) (PS-011) • Migration of Contemporary Performing Arts data in to alternative file formats (PS-012) • Migrate various World Heritage Site data into composite open formats (PS-013) • Evaluation of various 'Office' type migration paths (PS-014) • Preservation of Born-Digital Government Departmental Records (PS-015) • Digitisation of Paper based Government Records for future access (PS-016) • Preservation of Scientific Datasets (PS-017)

	• Migration of Social Science and Humanities Data between statistical formats (PS-018)
Federated search	• Preservation of Born-Digital Government Departmental Records (PS-015) • Digitisation of Paper based Government Records for future access (PS-016)
PID resolver	• *Link to work package 22 – Identifiers and citability* • *Link to work package 25 – Interoperability and Intelligibility*
Emulation facilities	• Migration of Contemporary Performing Arts data in to alternative file formats (PS-012)
Audit and certification of repositories	• Preservation of Born-Digital Government Departmental Records (PS-015) • Digitisation of Paper based Government Records for future access (PS-016)

The next examples illustrate the reasons why preservation services and preservation scenarios of WP14 are linked.

Obsolescence substitution <> Migration to a more open format and adding more Representation Information (PS-010)

The main objective of an obsolescence substitution service is to sharing information about the availability of hardware and software and their replacements/substitutes. The linked scenario "migration to a more open format" represents a use case for the conversion of astronomical data that is often stored in form of tables. In the case that an institution has no idea which format would be the best target format the obsolescence substitution service could give valuable advices by using its wide knowledge base and experiences. It is also conceivable that the obsolescence substitution service actively informs institutions in that domain about a possible decline of the availability of typical viewer software for astronomical data and gives recommendations to migrate their data stock.

Format transformation <> Migrations of audio files from BWF to FLAC and WavPack (PS-005)

The linked preservation scenario of the DNB needs a format conversion on a single-file level. The preservation action service on the other side should provide a service (on site of remotely) to facilitate batch conversion of data or on demand conversion or emulation. The service should provide possibilities to configure the target format and the used conversion tool. Thus the minimum requirement for using this service is the batch conversion a collection of many thousand audio files from BWF into FLAC and WavPack in a quality that meets the requirements of the customer.

Characterization of SIPs <> Format identification and format validation of online publications (PS-007)

The linked preservation scenario is also linked to the automated metadata creation/maintenance service. The scenario describes one of the important steps during a typical ingest process into a long term archive. Without file format identification the archiving institution has no information what kinds of objects are archived. This information is essential to establish an efficient migration strategy. This is the only way to address the objects which are concerned by a format migration at all. Moreover, format validation helps to sort out erroneous objects which may not be presented in the future. The identification and validation of the file format can be performed by the characterization service. If the service provides web-service protocols like REST / SOAP it is possible to integrate it in existing ingest workflows.

Careful readers surely noted that there are also links to other work packages. For example the work package 31 (Digital Rights and Access Management) has the objective to identify existing approaches for dealing with DRM protected material and work out advices (best practices) how institutions will be able to make sure that Digital Rights of digital objects are guaranteed and protected in the future. So the results of work package 31 are very important for building up a DRM clearing house service.

It should be one of the main objectives of the Virtual Centre of Excellence (VCoE) to be aware of these relations. Thus the given table can be considered as a first step for a tool that brings long term preservation use cases of customers/institutions and possible solutions together.

5 CONCLUSIONS FOR PRESERVATION SERVICES

On the basis of the above analyses, it is possible to draw some conclusions, first about the landscape of preservation services—which will form part of the common vision that APARSEN is constructing—then specifically relating to the Virtual Centre of Excellence. The conclusions about the landscape fall into three classes:

1. coverage;
2. limitations;
3. deeper description.

These will be discussed in turn.

1. Areas of the landscape where there seems to be less coverage of services

The mapping of the collected services against the metrics of for repository audit and certification, against the functional entities of OAIS, and against the scenarios of Work Package 14, reveals that there are some areas less covered by services than others. One might ask if this perception is illusory, as it is based on an admittedly incomplete collection of services. Nonetheless the criteria applied to choose services include generality, with a deliberate emphasis on breadth of coverage. The origin of many of the services in countering threats to long-term preservation such as those identified by PARSE.Insight also offers an assurance of breadth.

Looking at the mapping to the metrics for audit and certification of repositories, it is very evident that digital object management is well covered, while organisational infrastructure and infrastructure and security risk management are not. Perhaps these latter two areas offer less obvious opportunities for preservation services—after all, some of the functions of security risk management are not very specific to digital preservation, while organisational infrastructure is more oriented towards high-level statements, commitments and practices than to well-defined, repeatable operations. However, the fact that there are some services associated with these areas suggests that at least it is worth looking to see if there are other opportunities. Examples that immediately suggest themselves are services for assisting with ‘Definition of Designated Community and associated knowledge bases’ and ‘Management of the number and location of copies of all digital objects’ (independently of the long-term archiving service that is already mentioned under that item).

2. Limitations, or areas that have not been deeply explored

The preservation services discussed in this report have been motivated by a variety of factors, including systematic attempts to counter threats to long-term preservation, particular research interests, and commercial reasons. The services are presented in their own terms, relating to the problem they purport to solve and the approaches taken. However, it is desirable to go beyond this self-presentation of services to ask whether there are any limitations or assumptions not being made explicit, that will affect the applicability of the services.

For example, most of the collected services are purportedly domain-independent, but is this really so? In some cases at least the nature of the service will differ according to the domain or type of data, and there might be issues of scalability. It seems desirable to better understand the domain specificities and to construct a two-dimensional map of services—which would also be a useful tool for the VCoE.

There are some services that apparently contribute to several metrics in the audit/certification mapping. This suggests that the service is really fulfilling slightly different functions that could be better differentiated from each other. For example, the authenticity evidence management service contributes to the metric ‘Definition and documentation of processes for acquiring Preservation Description Information for its associated Content Information’ and to ‘Provision of evidence of the effectiveness of its preservation activities’. These are two rather different tasks of a repository: the former concerned with how to gather evidence of authenticity (among other things), the latter providing evidence as a form of quality assurance.

3. Deeper description of services

One form of deeper description has already been mentioned above: a two-dimensional map of services by function and by domain or type of data. There might be other approaches to analysing services that would be illuminating. For example, building an ontology of services that formalizes the relationships between the main entities linked to services. Certainly it seems desirable to develop some more comprehensive model that would allow services to be understood as a potential collaboration, rather than just a disparate set of non-interacting offerings.

As far as the Virtual Centre of Excellence is concerned, the work done on preservation services raises the following questions that should be considered when setting up the organisation.

- What services should the VCoE, or its members, offer to clients outside the organisation?
- Given the very disparate nature of preservation services, what is an appropriate charging model, if the VCoE does wish to make a commercial offering?
- What advice could the VCoE give on the availability and applicability of services, in response to a query from outside?
- What support is needed to provide such advice: for example, the two-dimensional map mentioned above, or a base of evidence about the effectiveness of particular services in practice?