Project no. 600663

PRELIDA

Preserving Linked Data
ICT-2011.4.3: Digital Preservation

D3.2 Consolidated State of the Art

Start Date of Project: 01 January 2013
Duration: 24 Months

Organisation Name of Lead Contractor for this Deliverable: APA

Version Final

Project co-funded by the European Commission within the Seventh Framework programme

Document Information

Deliverable number:	3.2
Deliverable title:	Consolidated state of the art
Due date of deliverable:	Dec 2014
Actual date of deliverable:	Dec 2014
Author(s):	David Giaretta
Participant(s):	APA
Workpackage:	3
Workpackage title:	State of the Art Assessment
Workpackage leader:	APA
Est. person months:	4.5
Dissemination Level:	PU
Version:	FINAL
Keywords:	digital preservation, linked data

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level
0.1	2014-11-17	Draft	David Giaretta (APA)	Initial draft
0.2	2014-12-07	Draft	David Giaretta (APA)	Revision with inputs from partners, especially Carlo Meghini (CNR). Added discussion of Use Cases
0.3	2014-12-09	Draft	David Giaretta (APA)	Further clarifications
1.0	2014-12-20	Candidate Final	David Giaretta (APA)	Clarifications added to address comments from Carlo Meghini (CNR) and others, and to make this document consistent with D4.3 Roadmap.
1.1	2014-12-23	Candidate Final	David Giaretta (APA)	Updated references
1.2	2014-12-28	Final	David Giaretta (APA)	Responded to comments from partners
Final	2014-12-30	Final	David Giaretta (APA)	Final responses and clarifications

Abstract

This document summarises the state of the art of the understanding of the fundamental techniques of digital preservation and then systematically discusses these techniques in the context of Linked Data current practices.

Table of Contents

Executive Summary

This document summarises the state of the art of the understanding of the fundamental techniques of digital preservation using the concepts introduced by OAIS [1], and then systematically discusses these techniques in the context of Linked Data current practices. The question addressed is whether these techniques are applicable to Linked Data, whether current Linked Data practices involve new techniques which might be more broadly applicable and finally whether there are improvements to current linked data practices.

The conclusions are that current Linked Data preservation practices can be significantly improved by systematic application of the general digital preservation techniques. Linked Data presents a number of specific challenges in terms of distribution and changeability; while these are not qualitatively unique challenges, nevertheless they are potentially quantitatively unique i.e. Linked Data may involve much more widely distribution and components which are less controlled in terms of persistence and variability.

The state of the art in digital preservation can provide a number of tools and services which could be applied to Linked Data but there is a need for solutions to the challenges of distribution, which is likely to be more widely beneficial, and specific tools and services which fit within the context of the general techniques.

This abstract approach is complemented by *D4.3 Roadmap* [2], which takes are more detailed look at the specific use cases.

1 Introduction

This document outlines the fundamental ideas about digital preservation and then describes in detail how these concepts may be applied to Linked Data.

1.1 Outline of document

In section 2 the threats which have been identified across disciplines are discussed and their applicability to Linked Data is described. Section 3 is the core of the document, outlining the key preservation concepts and techniques and describing how these apply to Linked Data, including the importance of identifying preservation aims where a number of possible options are discussed. These are compared to the existing practices to see whether lessons can be learned - these lessons will be useful input to the companion deliverable, D4.3 Consolidated roadmap, which discusses what remains to be done in order to create the technology required to deal with the preservation of Linked Data.

Section 4 examines several Use Cases which have been the subject of study of PRELIDA; the current practices are discussed and potential improvements identified.

Key issues in sustainability i.e. who pays or preservation, and why, are discussed in section 5. Examples of archives which claim to preserve linked data are described in section 6, together with ideas about how these repositories might be audited.

Section 7 looks at some options for international agreements which have implications for the digital preservation of all types of digital objects including linked data.

1.2 Methodology

Rather than starting from the position that Linked Data is special in terms of how to do digital preservation, we begin by looking at the general digital preservation ideas and applying these to Linked Data. We summarise the general concepts about and techniques for digital preservation and then apply these to Linked Data, comparing these to what is currently being proposed for the preservation of Linked Data. This will allow us to provide a context for these attempts at preservation and also to identify how these preservation efforts can be improved.

Having done this we then see whether there are other preservation needs needed for Linked Data which are not covered by the standard preservation techniques. In this way we fulfil the bi-directional aims of PRELIDA.

Existing services and tools used for digital preservation and their application to LD is discussed. Additional tools and services required for LD are identified.

1.3 Related Information

D3.1 State of the Art was the initial version of this document, which collected basic information. This document takes a more ordered approach, building on D3.1 and the other work undertaken by Prelida, including information gathered in the workshops reported in other deliverables.

The deliverables concerning the Roadmap (D4.1 Gap analysis [9], D4.2 First version of the roadmap [10] and D4.3 Consolidated Roadmap [2]) are related to the State of the Art, in that they complement (at different stages of development) the state of the art with an analysis of what is missing and an agenda about achieving it.

2 Threats to be countered

The PARSE.Insight project collected information through large scale surveys of researchers, data managers and publishers [2], which showed that there was a widespread view, consistent across disciplines and national boundaries, about the major threats to the preservation of digitally encoded information .

The threats and the responses identified and the way in which these apply to Linked Data is given in the table below.

	Threat	Impact on Linked Data
1	Users may be unable to understand or use the data e.g. the semantics, format, processes or algorithms involved	The semantics of the links between named objects and the semantics associated with the objects themselves may be lost. The specific relevance in terms of LD include: ● RDF/XML format, for example, may change over time. ● semantics of the tags may change over time ● ontologies and other related web resources may no longer be available - as happens with many web resources
2	Non-maintainability of essential hardware, software or support environment may make the information inaccessible	Changes may occur in technical infrastructure such as the internet including the DNS system and the host servers, for example ● the triple stores used to hold the LD information ● the SPARQL engines used to query LD datasets ● inference engines used to compute implicit data in LD datasets
3	The chain of evidence may be lost and there may be lack of certainty of provenance or authenticity	The authenticity (rather than the correctness) of the linkages depends upon the reliability of the name resolvers and other parts of the infrastructure. Authenticity of changes over time depends on the trustworthiness of the people responsible for the linked objects. Some aspects specific to LD include: ● re-writing of URIs, for example at ingest time to point to web archives or to other repositories that ensure the long-term access ● uncertainty that the URI resolution mechanisms get to the original resources
4	Access and use restrictions may make it difficult to reuse data, or alternatively may not be respected in	There may be access restrictions to the linked resources

	future	
5	Loss of ability to identify the location of data	The specific relevance in terms of LD is the issue of URL decay.
6	The current custodian of the data, whether an organisation or project, may cease to exist at some point in the future	This is a general issue which is certainly applicable to LD
7	The ones we trust to look after the digital holdings may let us down (e.g. do not maintain usability)	This is a general issue which is certainly applicable to LD

3 Digital preservation and Linked Data

OAIS defines the following fundamental concepts:

Long Term Preservation [5] as *The act of maintaining information, Independently Understandable by a Designated Community, and with evidence supporting its Authenticity, over the Long Term.*

Independently Understandable [5]: *A characteristic of information that is sufficiently complete to allow it to be interpreted, understood and used by the Designated Community without having to resort to special resources not widely available, including named individuals*.

However one can go a little further by looking in more detail at what “use” might mean.

Preserving a simple JPEG image is relatively straightforward. To preserve it means that it can be displayed or printed in the future - it seems reasonable to say that there are few other options.

More complex digital objects have many more options - perhaps too many to be able to state explicitly.

Preservation aims, while not an OAIS concept, enable one to refine the definition of understandability and usability.

3.1 Preservation Aims

Understandability and usability are very general concepts. Being able to use some digitally encoded information could cover almost anything, including printing the 0’s and 1’s as wallpaper. Normally the implication is that the Designated Community can do things other than render (print) them - especially if it is scientific data. Even scientific data could be used in various ways. As an extreme example given a scientific dataset containing measurements of ocean temperature across the Pacific Ocean on 1st Oct 2014, one can think of uses ranging from (1) extracting the value of the temperature measured at some specific location, to (2) using the information to contribute to a study of global warming. In principle, depending on the intelligence/skills of the user, being able to do the first may allow the second to be achieved.

The repository could relatively easily provide Representation Information to achieve (1) but would need to provide much more Representation Information in order to guarantee the ability to achieve (2) - again depending on the definition of the Designated Community [5].

Defining Preservation Aims helps to guide the repository in terms of clarifying the amount of Representation Information to provide, and may make the work of the Designated Community easier.

Examples of preservation aims for a dataset may include trying to enable the Designated Community to:

- process the dataset and generate the same data products as previously
- understand the dataset and use it in analysis tools
- combine the data with other data to calculate derived quantities
- ...

The Preservation Aims can also influence what data is selected for preservation. For example if the aim is to be able to combine two datasets then both datasets should be preserved - of course they may be regarded as a single digital object.

For Linked Data one can look at a number of possibilities depending on the Preservation Aims.

3.1.1 LD Preservation Aim: Underlying data usability

Linked Data is often a means of publishing the underlying data, e.g. held in a scientific format such as HDF or in an SQL database, rather than simply publishing it in its native form. For example DBpedia [8] is captured and stored in a database, not as RDF. The RDF is derived from the database using specific software which encodes some choices, for instance how to transform the original data values into URIs or literals, how to encode attributes as properties. Because of this the original data and their transformations into LD are not the same.

Therefore one could choose to preserve the data and the software if we wish to preserve DBpedia more or less in isolation, but this may not be the same as preserving the RDF version of DBpedia.

3.1.2 LD Preservation Aim: RDF usability

Alternatively one may decide that the important thing is the RDF itself and its usability.

3.1.3 LD Preservation Aim: Services

On the other hand it may be more important to preserve the usability of the services such as the inference capabilities.

3.1.3.1 SPARQL

One of the possible services is the provision of SPARQL endpoints however many public SPARQL access points are dysfunctional most of the time [6] and [11].

3.1.4 LD Preservation Aim: Time dependence

Another Preservation Aim may be to be able to see the Linked Data situation at some time in the past. As an example for DBpedia data (in RDF format, or Tables-CSV format) are archived and the user requests specific data (or the entire dataset) as it was at a specific date in the past, e.g., the RDF description of topic Olympic games at 1/1/2010.

Conceptually the archive would keep AIPs containing snapshots of the information at specific times.

In terms of practical implementations, we have two complementary approaches. One builds entirely on the web architecture, extending it in order to access past representations of a resource state. This avenue is pursued by the Memento project. The other relies on the development of ad hoc systems for the creation, maintenance and access to provenance information of any resource. Hyberlink and Diachron are two important projects following this approach.

3.1.4.1 Memento

The Memento protocol specified as RFC 7089, defines interoperability for access to resource versions based on a resource's generic URI and a desired date/time. Memento's paradigm is as distributed as the web itself, and hence can work in a hybrid environment of centralized and decentralized archives. Memento is fully aligned with the web architecture, REST, and "follow your nose" principles.

Over the past 5 years, the Memento protocol has been adopted by many major publicly accessible web archives. Currently, there is a focus on getting it adopted for versioning systems such as wikis, software control system, evolving technical specification, etc.

From the very outset of the Memento project, its relevance for Linked Data has been demonstrated [12].

3.1.4.2 Hiberlink

In the Hiberlink project (Edinburgh and LANL) and in the Internet Robustness project (Harvard), various ideas are explored that closely relate to the problem of interconnectedness of Linked Data with other Linked Data. Hiberlink focuses on web resources that are referenced in scholarly publications and Internet Robustness focuses on web-based legal literature and the blogosphere. The projects share the notion of a core collection that someone cares about (cf. a Linked Data set) and resources that are linked from it (cf resources interconnected with the Linked Data set) [13]. They share the notion of pro-actively archiving the linked resources at crucial moments in the lifecycle of the core collection. Both projects are exploring a variety of ways to achieve this. Harvard's amberlink approach is to cache linked resources along with the core collection [14]. Hiberlink mainly looked into pushing linked resources into web archives.

3.1.4.3 Diachron

Diachron [15] links data preservation to evolution management of "LOD datasets". It sees the main challenges as

- diverse data models
- dynamic datasets
- recoverable versions
- changes as first-class citizens
- cross-snapshot queries

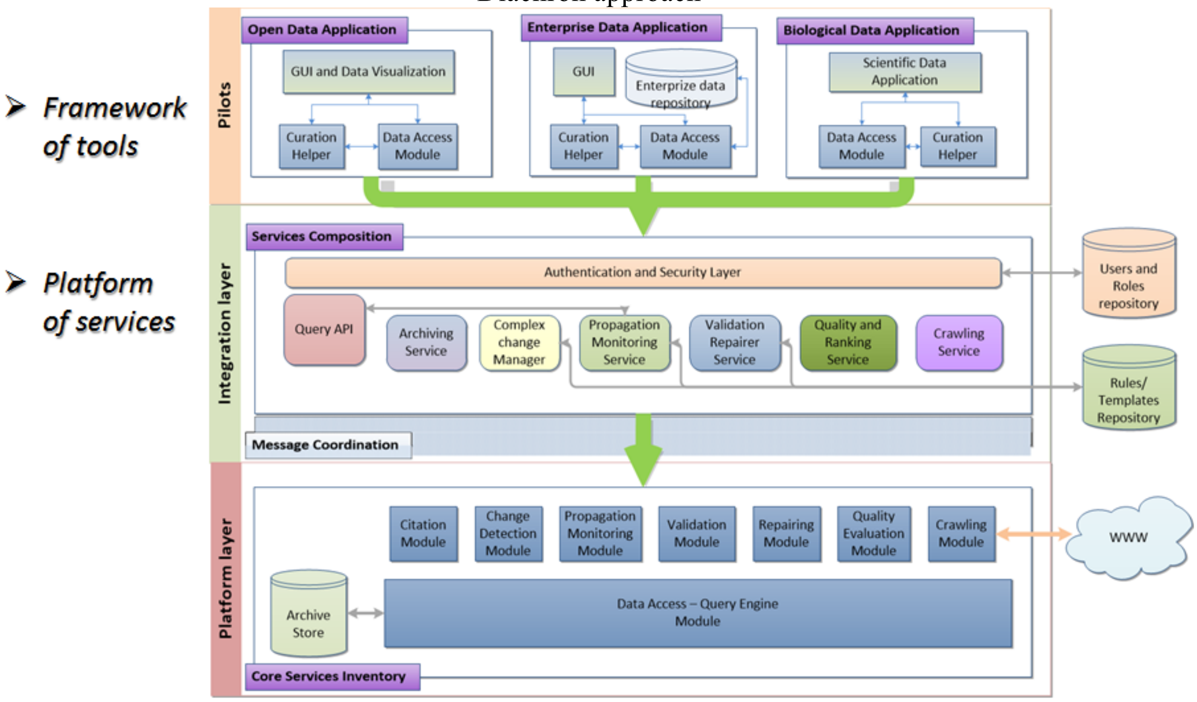
It is undertaking 3 pilot programmes involving:

- open government data
- open enterprise data
- open scientific data

These will

- Monitor the changes of LOD datasets (tracking the evolution)
- Identify the cause of the evolution of the datasets in respect with the real world evolution of the entities the datasets describe (provenance problem)
- Repair various data deficiencies (curation problem)
- Temporal and spatial quality assessment of the harvested LOD datasets and determination of the datasets versions that need to be preserved (appraisal)
- Archive multiple versions of data and cite them accordingly to make the reference of previous data feasible (archiving and citation)
- Retrieve and query previous versions (time traveling queries)

Diachron approach



3.2 Fundamental approaches

3.2.1 OAIS requirements

OAIS specifies what is required for conformance:

- support the OAIS Information Model and
- fulfil a number of mandatory responsibilities:
    a. Negotiate for and accept appropriate information from information Producers.
    b. Obtain sufficient control of the information provided to the level needed to ensure Long Term Preservation.
    c. Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and, therefore, should be able to understand the information provided, thereby defining its Knowledge Base.
    d. Ensure that the information to be preserved is Independently Understandable to the Designated Community. In particular, the Designated Community should be able to understand the information without needing special resources such as the assistance of the experts who produced the information.
    e. Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, including the demise of the Archive, ensuring that it is never deleted unless allowed as part of an approved strategy. There should be no ad-hoc deletions.
    f. Make the preserved information available to the Designated Community and enable the information to be disseminated as copies of, or as traceable to, the original submitted Data Objects with evidence supporting its Authenticity.

All the above mandatory responsibilities apply to the preservation of LD. Addressing (b) we see that the OAIS must gain sufficient control over the information in order to ensure that it can preserve that information. Where the information is encoded as a single file then this is relatively easy - it can be copied to the repository. However for a distributed system such as we find for Linked Data, then we have a more difficult issue.

For Linked Data one could (1) maintain the distributed sources of information; alternatively one could (2) copy the some or all of the various components to the repository. We consider situation (2) where the information is kept within the OAIS - although we note that the OAIS may be distributed. If the various distributed sources of LD become part of the OAIS then we have case (1).

As an intermediate case it would be possible to created a set of preserved LD datasets, each member of the cloud being an OAIS devoted to preserve a specific dataset to which the other members of the set somehow link.

In both cases the obvious difficulty is the potential open-ended nature of Linked Data as one piece is linked to one or more other pieces. This linkage cannot be infinite since there are only a finite number of links in the world, but recognising that even a small part of this finite number could be very difficult.

In the case that the linked objects are copied, and so the copies have new URIs, then an important part of the Provenance, and so important with respect to Authenticity, is the original URI and the time of collection.

Therefore in what follows we refer to “the” archive although this may be distributed and we assume there are adequate resources to deal with the preservation of the Linked Data.

3.2.2 Techniques

It used to be said that *the dominant approach to digital preservation has been that of migration. Migration is the process of transferring data from a platform that is in danger of becoming obsolete to a current platform* [16].

However while this has some truth to it for rendered digital objects such as images or simple documents, it is inadequate for data.

The next sections look at the three basic techniques, which include the “emulate or migrate” as a subset. The three fundamental techniques are:

- Add Representation Information - this is very much broader than, but includes, emulators
- Transform - a more accurate description of "migration"
- hand over to the next in the chain of preservation

Each of these is described in detail in the next subsections. In each section a general description of the technique is given followed by a discussion of the applicability to Linked Data.

3.2.3 Maintaining AIPs

It is important to bear in mind that the creation of the AIPs is fundamental to the preservation of digitally encoded information, irrespective of any changes that may occur.

Therefore in what follows it is essential to realise that all the information needed for an up to date AIP is maintained. This will not be repeated in the discussions in the following sections but is summarised here. For each of the techniques:

- Add Representation Information

    - In principle adding Representation Information should not require changes to the Preservation Description Information (PDI) [5] or other parts of the AIP, apart from the Representation Information
- Transform
    - This requires great changes in the AIP. The Provenance Information needs to cover
        - the links to the original AIP
        - the details of the Transformation including the reasons why this new object is worthy to be regarded as sufficiently authentic
        - details of the Packaging Information which will probably change
        - any relevant changes to the Package Description
        - details of Access Aids [5] which may be significantly changes
        - Representation Information for the new object
- Handover
    - The AIPs is maintained by the repository. If/when the AIPs are handed over to one or more other repository additional Provenance Information must be added about that handover from the first repository.

3.3 Fundamental techniques

3.3.1 Add Representation Information

Representation Information is defined by OAIS as: *The information that maps a Data Object into more meaningful concepts*. This is expanded as follows: *Since a key purpose of an OAIS is to preserve information for a Designated Community, the OAIS must understand the Knowledge Base of its Designated Community to understand the minimum Representation Information that must be maintained. The OAIS should then make a decision between maintaining the minimum Representation Information needed for its Designated Community, or maintaining a larger amount of Representation Information that may allow understanding by a larger Consumer community with a less specialized Knowledge Base, which would be the equivalent of extending the definition of the Designated Community. Over time, evolution of the Designated Community's Knowledge Base may require updates to the Representation Information to ensure continued understanding.*

Each piece of Representation Information is encoded, often as a digital object, and that digital object may require its own Representation Information in order for the Designated Community to be able to understand and use it. The same applies to each of those pieces of Representation Information, producing a network of connections, which is called a Representation (Information) Network (RIN).

As described in section 3, one of the key threats to preservation is that the digital objects will not be understandable and usable. Another threat is that software or hardware is no longer available.

One, perhaps the only, way to overcome this threat without changing the object is to add Representation Information. In the case of hardware being unavailable the Representation Information would be an emulator.

A repository would need to be clear on its definition of the Designated Community, which in turn determines the Representation Information required. As the Knowledge base of that Designated Community changes, additional Representation Information must be added.

There are services which can assist the efforts required to have adequate Representation Information including

- Network of Registries of Representation Information - to share such things as shemas and ontologies
- Orchestration service - to help to share information about changes

Examples of relevant efforts include the SCIDIP-ES project http://www.scidip-es.eu, the software from which can be tested at http://int-platform.digitalpreserve.info.

We can examine this in terms of several different potential decisions of Designated Communities and potential evolution of its Knowledge Base. Note of course that we cannot predict these future changes, and moreover OAIS does not demand that the all the possible Representation Information be collected at once. However these “thought experiments” should clarify what information we should be prepared to add to the collection of Representation Information required to ensure the Designated Community can understand and use the LD of interest.

3.3.1.1 LD Representation Information

Designated Community: users of Linked Data

One can think about the stack of information and processes which such a user currently has support for, noting at each point the potential threats/ changes which may need to be countered.

Let us assume that the user has a link - a (HTTP) URI - to some Linked Data and consider the various steps.

- the HTTP URI must be resolved so that the RDF can be accessed
    - this requires the ability to recognise the HTTP URI as a string which can be resolved, and then resolving it to a particular address; this address must be able to then provide the correct sequence of bytes in some way
        - currently the DNS system (perhaps after a Persistent Identifier look-up) resolves the HTTP URI to an IP address. The Internet infrastructure and TCP/IP directs the packets to that address. The recipient must recognise the byte sequence and send the appropriate response using an acceptable protocol via TCP/IP back to the requesting machine.
        - in the future, assuming the internet and DNS are available, the various hosts may no longer exist or the HTTP URIs may no longer be resolvable given the rate of decay of URLs.
- On receipt of the requested RDF, local software parses it and identifies schema, ontologies and other files that may be needed. These can be retrieved in a similar way to the initial file. Each retrieved file is parsed and may itself point to further files.
- Having parsed and gathered the information together, software, not necessarily the same as the parser, is used to satisfy queries from the user.

The required pieces of information, including files and software, are Representation Information; the lookup system and network infrastructure **could** also be regarded as Representation Information in the sense that in order to use the RDF files one needs to resolve the URIs. This could be the "normal" network infrastructure including DNS; an alternative would be to create a local version of the resolution system. The amount of Representation Information collected depends upon the Knowledge

Base of the Designated Community, which may change over time. As an extreme example if at some point in the far future the internet as we know it is about to replaced by something very different then those responsible for preserving LD could decide that additional Representation Information such as the definition of the network protocols used by LD such as TCP/IP, HTTP and various RFCs should be collected, and perhaps local implementations be kept available.

As a concrete example one could image that some time in the future when the current network infrastructure, such as the use of TCP/IP, is to be replaced, then the archive could add information about TCP/IP as additional Representation Information as one of the potential preservation techniques.

We can see here the Representation Information Network (RIN) is made up of

- distributed files each of which will have its own RIN. Note that each file could be generated on the fly by the server which receives the request.
  - schema
  - other ontologies
  - imported files
  - explanations of the meaning of the various symbols in the ontologies, supplementing the information provided by the ontologies i.e. the linkages between the symbols.
  - structural information such as Unicode definitions
- software written in various languages, and each of which relies on RINs including various libraries, operating systems and hardware
  - parsers
  - query resolvers

In general terms one could:

- collect the files, to whatever required depth, and use appropriate indirection to ensure that embedded locator strings (e.g. the HTTP URIs) still find the appropriate files; the Memento system is an example of this, adding in timestamping.
- collect the software, either source code or compiled files, together with the required libraries etc. or different underlying software - emulators.

3.3.2 Migration

As an alternative to keeping the bytes unchanged and adding Representation Information, it may be advantageous to change the bytes - and of course add a whole new set of Representation Information associated with those bytes. The advantages may be because of cost savings or for more straightforward perhaps wider, usability.

The issues which is arise include:

- the choice of the particular transformation to use
- whether the new object can be claimed to be an authentic representation of the original. Note that it becomes impossible to verify the authenticity of the “new” object simply by checking that the fixity (e.g. checksums or digests) match the original (unless the transformation is reversible in which case one can reverse the process before calculating the digest).

Experience indicates that, except in the rare cases that the transformation is reversible, there will be loss of some information and so a key question is: has enough information been retained?

OAIS introduced two related terms related to these questions:

- **Information Property Description:** The description of the Information Property. It is a description of a part of the information content of a Content Information object that is highlighted for a particular purpose.

- **Information Property** [5] is *that part of the Content Information as described by the Information Property Description. The detailed expression, or value, of that part of the information content is conveyed by the appropriate parts of the Content Data Object and its Representation Information.*

- **Transformational Information Property** [5] *as an Information Property the preservation of the value of which is regarded as being necessary but not sufficient to verify that any Non-Reversible Transformation has adequately preserved information content. This could be important as contributing to evidence about Authenticity. Such an Information Property is dependent upon specific Representation Information, including Semantic Information, to denote how it is encoded and what it means. (The term 'significant property', which has various definitions in the literature, is sometimes used in a way that is consistent with its being a Transformational Information Property).*

The choice of Information Properties and Transformational Information Properties (TIFs) are made according the judgement of people - perhaps the various stakeholders such as data creators, funders or repository managers. The judgement that the TIFs have acceptable values is the responsibility of a person, for example the repository manager. The reputation of that person is then an important factor in the consumers' judgement of the authenticity of the "new" object.

3.3.2.1 LD Migration

Assume the Linked Data we are interested in is in the form of RDF/XML serialised as a file on a server. There are several possible transformations - for example to various serialisations, as files or as byte sequences in a database:

- RDF/XML
- Turtle
- N-Triples
- JSON
- etc

In the future there may be other serialisations.

These transformations are not reversible in that going in a cycle e.g. from an RDF/XML file to a Turtle file and then back to an RDF/XML file again, the latter will be very similar, but not identical, to the original e.g. there may be differences in statement order. This lack of reversibility will not affect the "understandability and usability" of the RDF **but** means that authenticity cannot be verified by checking at the bit-level such as comparisons of checksums or digests, and hence Transformational Information Properties must be identified.

Transformational Information Properties might include:

- the ability to link between named things
- the ability to describe the meaning of those links at some level
- the ability to resolve queries about the things

- it may be required to provide information about the changes over time of each object/link

The parser would need to be changed to deal with the serialisation chosen. It might implement one of the proposed LD APIs [17]

The links within each object may be transformed or else the Representation Information for the links must specify how to find the location of the linked object.

3.3.3 Hand-over

If the repository is unable to preserve the information, perhaps because there are insufficient resources allocated, then it should hand over the information, together with any information needed to make up the appropriate Archival Information Package e.g. Representation Information, Provenance, Access Rights etc.

3.3.3.1 LD Hand-over

Handing over LD should present no additional challenges except that there may be a significant number of distributed objects. It is worth re-iterating the importance of updating and maintaining the Provenance, in particular the URIs - including the new URIs, the URIs used by the previous archives, and the original URIs.

3.4 Changes – discovering and responding

If nothing changed then preservation would be unnecessary. Actions are needed to ensure preservation of digitally encoded information *because* things change. In order to react to such changes a number of things need to happen and there are tools and services which can support those responsible for undertaking that preservation. Examples are shown in the following table (many of these services have been developed by SCIDIP-ES [7] and other projects).

When things change one needs to:	**Support**
know that something has changed e.g. software, hardware, knowledgebase of a Designated Community or environment	**Brokerage service** allows interchange of information about (potential) changes
understand the implications of that change	**Gap Identification service** helps curators understand the "gaps in understandability" that might arise – for preservation this is gap with respect to the Designated Community[i]
Decide on the best course of action for preservation, based on costs and risks involved	**Preservation strategy toolkit** helps curators to investigate, decide between, and document, alternative courses of action

The responsibility for discovering the changes relies on effort from many sources. The key source of information is people; technology can support the capture and sharing that information; in the future there may be a greater technological input.

Deciding on and putting the responses into action is certainly the responsibility of the archive, but the data producer may be required to play a role.

The changes of concern in earlier sections are related to a particular piece of information (at a particular time) and the Designated Community's ability to understand and use it.

However there are other changes which are particularly of importance for Linked Data, for example changes in the linked data of interest e.g. DBpedia arising from changes to Wikipedia. Initiating the preservation of such changed information would in OAIS terms be the responsibility of the Producer who would submit the appropriate information to the archive to allow the creation of new AIPs. These would logically be separate new AIPs but the implementation may involve more complex digital objects. For example the difference between the old and the new RDF files may be kept in order to reconstruct the new set of RDFs. Another common technique that might be used is to share Representation Information between AIPs, for example placing vocabularies in a Registry/Repository of Representation Information; the SCIDIP-ES implementation supports versioning. Note that in this case to use those ontologies fairly seamlessly within the Linked Data there would have to be a local redirection – the equivalent of a local cache. The implementation of such redirection would ultimately be the responsibility of the Consumer but the archive may implement it in order to provide a better service to users. The alternative would be to change the RDF of interest.

4 Use Cases

The use cases discussed in D3.1, the initial draft of the State of the Art, were CEDAR (an archive at DANS), DBpedia and Europeana. The preservation techniques used for these use cases is discussed in terms of the 3 fundamental digital preservation techniques described in section 3.

4.1 CEDAR

The CEDAR project works from the available deposited datasets as Excel files to produce LD.

If we assume that the preservation aims are to be able to continue to use LD then the preservation being undertaken seems minimal.

4.1.1 Improvements:

- Add Representation Information, as described above for RDF.

Issues arise in terms of the enrichment of the data, and these may be viewed as Adding Representation Information.

4.2 DBpedia

Wikipedia provides the "raw" data used by DBpedia, from which DBpedia creates RDF. DBpedia stores different versions of the entire dataset as RDF or CSV dumps, as a versioning mechanism.

The DBpedia contains more than 27 million links to other Linked Datasets.

The preservation strategy is to keep these RDF and/or CSV dump files. The external LD to which these are linked are not part of the preserved information, nor are the querying or other software. The evolution of terms is not tracked.

In terms of the basic preservation techniques described in section 3, this is a very minimal approach to preservation.

4.2.1 Improvements

- Add Representation Information, as described in section 3
- ...

4.3 Europeana

As described in D3.1, Europeana functions as a metadata aggregator: its partner institutions or projects send it (descriptive) metadata about their digitized objects to enable centralized search functions. The datasets include links to the websites of providers, where users can get access to the digitized objects themselves. As this metadata is stored by Europeana, Europeana has no specific requirement for specific metadata preservation policies on the provider's side. This is less true for the problem of link rot on providers' websites. Often providers do not use (or do not send) persistent web identifiers, which results in broken links between Europeana and provider's object pages, when these get different web addresses.

Also Europeana has embarked on enriching this data, linking for example to GEMET, Geonames and DBpedia. While sets GEMET are very stable, DBpedia is much more dynamic, and not monotonic (i.e., DBpedia facts may sometimes be retracted during updates, while others are added). Europeana download dumps of external sets to store a part of it in its main databases, so the Europeana services would not be disrupted, should the external datasets undergo massive changes.

There are several complexities in the Europeana case. There are several potential sets of preservation aims.

4.3.1 Preservation Aim: Preserve the RDF

Versions of the RDF can be stored by Europeana.

4.3.1.1 Potential Improvements

The Europeana RDF so that it is usable would require

- the schema - those specific to Europeana as well as any imported
- all the raw data on which it depends - see the next Preservation Aim
- associated software

4.3.2 Preservation Aim: Preserve the "raw" data

By "raw" data is meant the data about which the metadata is harvested - because the data that Europeana uses changes. This implies some level of link rot.

- Dumps of this remote data are being collected. They would presumably be used within a preserved European because if the published links may no longer meaningful in the context of updated third-party sets. Europeana could re-publish its "cached" version of the third-party data. But in a Linked Data setting it would be extremely confusing for users, if such re-publication shows statements that have become very different, or even incompatible with the original source.
- In any case Europeana generates its internal identifiers from the identifiers sent by its providers, which are not always persistent.

4.3.2.1 Potential improvements

- Add Representation Information (as described above) e.g.
    - information about schema involved
    - information about the link between the original URIs and the re-directed identifiers
    - Add versioning information (a special type of Representation Information)
        - Currently there is no versioning at all in the data that Europeana re-published. Europeana hopes to make progress at some point in the future, by providing information on incremental modification using the tested means of an OAI-PMH server for RDF/XML representation of the object records stored by Europeana.

4.3.3 Preservation Aim: Preserve the services

To keep the Europeana services running one could undertake the 2 preservation aims above

5 Sustainability

A critical issue that is recognised in digital preservation is the question "who pays and why?".

The APARSEN project has produced an integrated view of digital preservation which is built directly on business processes [18].

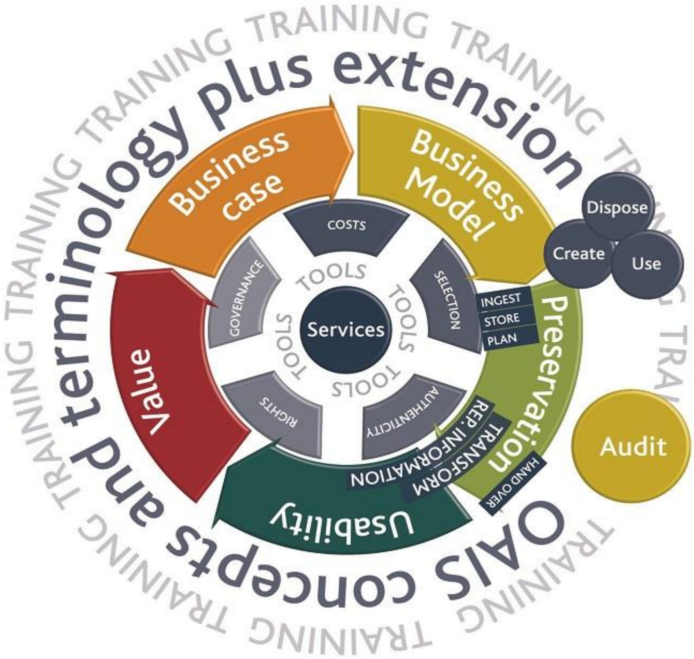


Figure 1 APARSEN Integrated View of digital preservation

The diagram above illustrates the basic sequence of activities to implement a sustainable business process centred in the preservation of digital objects, to be embedded in the overall business cycle of organisations responsible for securing the future usage of such assets.

Note that the focus here is on preservation. There is a large number of other models with which one may be tempted to compare; these tend to be focused on the creation of digital objects and the publication of results, or the academic lifecycle, but those models tend to ignore the business model aspects, i.e. how to implement the delivery of Digital Preservation value proposition over time.

It should be borne in mind that in reality there may be a number of iterations. For example to create a Business case, Value may be re-visited and revised as may be Usability; these iterations are omitted in the flow shown above for the sake of clarity.

The activities may be summarised as follows:

5.1 Preserve..

the object by a variety of sub-processes

- Ingest

- Store
- Plan preservation, including identifying the designated community (ideally this should be done at the earliest opportunity – certainly before the creation of the digital objects, if we want to secure the best conditions for future usage and we must secure a proper value justification to secure financial resources flows)
- The basic steps in preservation to counter changes are:
  - **create adequate Representation Information** for the Designated Community and/or
  - **transform** to another format if necessary or
  - if preservation cannot be carried on by the current organisation then **hand over** to the next organisation in the chain of preservation
- Evidence about the authenticity of the digital objects must also be maintained, especially when the objects are transformed or handed over.
- Confirmation of the quality of preservation can come from an Audit (with possible certification)

5.2 Usability

Usability is of course closely connected with, but is not limited to, preservation.

- Digital objects and digital collections should remain **usable**, i.e. one (human or artificial agent) should be able to understand and use the digital material. This is closely related to task performability. Various tasks can be identified and layered, e.g. rendering (for images), compiling and running (for software), getting the provenance and context (for datasets), etc. In every case task performability has various prerequisites, (e.g. operating system, tools, software libraries, parameters, representation information etc.). These prerequisites are termed Representation Information in OAIS and the minimum amount of Representation Information needed is determined by the definition of the Designated Community.
- **Additional Representation Information** may be created to enable a broader set of users to use and understand the digitally encoded information
  - Other communities may use different analysis tools and it may be convenient to transform the digital object to a more convenient format. This will itself require its own Representation Information; the semantic RepInfo may be unchanged but new structural RepInfo will certainly be needed.
- The digital objects should also be discoverable in some sensible way – bearing in mind that some information will be publicly available whereas other information will be restricted.

5.3 Value proposition

The portfolio of Value proposition/s will provide the core of the answers to “Why preserve a certain digital collection and who would be willing to pay for it?”

- Value propositions must be created by the identification, classification and quantification of the expected benefits which may be obtained by the targeted communities of customers and users from the continuous usage of the preserved objects, which in turn depends on the needs of the users and the usability conditions created for such preserved objects
- the objects will probably be more useful to one type of user community than to another, and this may change over time. These differences and changes must be addressed by a portfolio of Value propositions (as well as by the design and implementation of adequate business models)

- rights may be associated with the objects, perhaps arising from the value or potential value of the object. These rights can generate revenue, and the revenue generation in turn depends on the business model used.

The value need not be measured simply in monetary terms, although monetary resources will undoubtedly be required to undertake the preservation activities. Maintaining links may be one of the key reasons for assigning value - it has been noted that data becomes more valuable when it is combined.

Having understood what the value could be allows one to see how this potential value may be converted into actual resources through Business Cases.

5.4 Business case

- There is an increasing demand from decision makers to justify: the need for objects to be preserved, the benefits derived of their usage, the costs involved in the preservation, as well as other resources required for preservation
- Its implementation will be addressed by one or more business models
- There will almost certainly be options for trade-offs between costs, risks and capabilities

The Business Case(s) should then be converted into practical use through a Business Model.

5.5 Business model

- The business model lays out the business logic, i.e. how the value proposition is consistently delivered to the beneficiaries.
- Decisions about the mix of sources providing the financial resources required for implementing and operating the preservation business process will be based on the characteristics of the users and customers base (the target groups), the competition in the provision of the preserved assets as well as in the nature and dynamics of the formulated business case. Costs play an important in these considerations.
- The resources may be used at the very start to create new digital objects, which will presumably have been created for a specific purpose and which then may be either disposed of or be preserved.
- A selection process will be needed to decide what is to be preserved. This will presumably be based on business case and risk considerations. It may also depend on the interest of other possible curators of the information.
- The financial resourcing may be (perhaps should be) part of the budgets needed to create the digital objects. However some or all of the objects created may be disposed of rather than preserved.

Each of these steps will be assisted by the use of tools and/or services, such as the ones the VCOE should be able to offer.

Shown as a circle outside the main figure, because it would normally be done by a third party, the audit process will be important to provide assurance of the quality of the preservation, and will also take into account the financial viability of the archive (at least long enough for a handover of the digital holdings to take place successfully) .

The underpinning components are first the use of a consistent terminology, the OAIS terminology with extensions to cover those aspects outside the OAIS remit and second the training modules covering all aspects of the common vision.

6 Archives

6.1 Memento related

For over 3 years now, there has been a publicly accessible, Memento compliant DBpedia archive available, see http://mementoweb.org/depot/native/dbpedia/ . This archive is, as per the Memento protocol, integrated with DBpedia itself, in the sense that DBpedia URIs provide a "timegate" link to the archive (it suffices to look at DBpedia response headers to see them).

6.2 Audit and Certification of LD preservation

The audit of a repository which claims to be trustworthy in terms of preserving Linked Data would follow the metrics of ISO 16363 [19]. The broad grouping of metrics are:

- Organisational infrastructure
- Digital Object Management
- Infrastructure and security risk management

All the metrics are applicable to LD archives although special concerns arise if the LD collection is distributed.

7 Inputs to standards

The results from PRELIDA will be taken into account in the ongoing standardisation efforts in CCSDS/ISO [20] and the related RDA activities [21]. The aim is to develop terminology to describe and standardise the various stages of the data lifecycle from proposal, through collection, to preservation, adding value and re-use. Then each stage will be the subject of a more detailed description and standardisation process where applicable.

This work is at a very early stage of planning but the aim is to produce the first standard in 2 years.

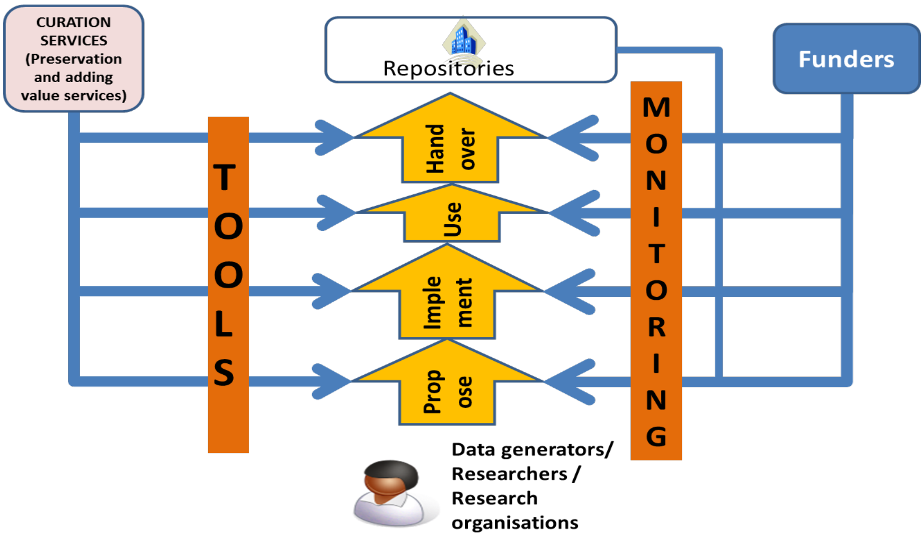
7.1 CCSDS/ISO

The CCSDS/ISO group on Data Archive and Ingest is starting work on a new project on standardising the Information Curation Process. This will be carried out in collaboration with the RDA groups described below.

7.2 RDA

RDA has 2 Interest Groups relevant to preservation which will work with CCSDS, providing inputs from a wide variety of disciplines and help to test ideas

- Preservation e-Infrastructure
  - This group is working on definitions of services which will help to improve preservation capabilities.
- Active Data Management Plans
  - This group has begun work on the various stages from the proposal stage to handover to the repository, shown in the following figure.


CURATION SERVICES (Preservation and adding value services)
Repositories
Funders
TOOLS
MONITORING
Hand over
Use
Imple ment
Prop ose
Data generators/ Researchers / Research organisations

Details are available from https://rd-alliance.org/groups/active-data-management-plans.html

8 Conclusions

Systematic study of the fundamentals of general digital preservation indicates that these are applicable to Linked Data, although of course there are aspects that are specific to Linked Data, for example the specific types of Representation Information that must be collected.

None of the cutting edge activities in preserving Linked Data seem to indicate any missing general concepts. On the contrary, the general preservation aspects do suggest a more systematic approach to preserving Linked Data are useful. On the hand there may be specific tools which are needed. Some exist already, including:

- a system of Registries of Representation Information (see the SCIDIP-ES services http://int-platform.digitalpreserve.info/dashboard/registry/)
- a system to collect information about changes e.g. changes in schema or software (see the SCIDIP-ES service http://int-platform.digitalpreserve.info/dashboard/orchestration-service/

Other services and tools are still to be developed. The document D4.3 Consolidated Roadmap discusses additional work that is needed.

9 References

[1] Reference Model for an Open Archival Information System (OAIS). ISO 14721:2012 or later, available from http://public.ccsds.org/publications/archive/650x0m2.pdf

[2] PRELIDA Deliverable D4.3. Consolidated Roadmap. Available from the PRELIDA web site: prelida.eu

[3] PRELIDA Deliverable D3.1. State of the Art. Available from the PRELIDA web site: prelida.eu

[4] PARSE.Insight Survey Report, available from http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

[5] See the Digital Preservation Glossary – SKOS based – available at http://www.alliancepermanentaccess.org/index.php/consultancy/dpglossary/. This provides HTTP URIs for terms e.g. http://www.alliancepermanentaccess.org/index.php/consultancy/dpglossary/#Long_Term_Preservation http://www.alliancepermanentaccess.org/index.php/consultancy/dpglossary/#Designated_Community

The terms are linked together using the SKOS ontology

[6] Herbert v de Sompel private communication

[7] SCIDIP-ES project tools and services available from http://int-platform.digitalpreserve.info/

[8] DBpedia – see http://dbpedia.org/About

[9] PRELIDA Deliverable D4.1. Report on the consolidation and dissemination workshop. Available from the PRELIDA web site: prelida.eu

[10] PRELIDA Deliverable D4.2. First version of the Roadmap. Available from the PRELIDA web site: prelida.eu

[11] See http://ruben.verborgh.org/blog/2013/09/30/can-i-sparql-your-endpoint/

[12] LDoW Workshop paper, which was most enthusiastically received, among others, by Tim Berners-Lee: *http://arxiv.org/abs/1003.3661*

[13] H van de Sompel et al, see http://www.slideshare.net/hvdsomp/creating-pockets-of-persistence

[14] Amberlink see http://amberlink.org/

[15] DIACHRON project – Managing the Evolution and Preservation of the Data Web – see http://www.diachron-fp7.eu/

[16] S Granger, Emulation as a Digital Preservation Strategy, 2000, http://www.dlib.org/dlib/october00/granger/10granger.html

[17]For example see https://code.google.com/p/linked-data-api/wiki/Specification

[18]APA/APARSEN integrated view of digital preservation, see http://www.alliancepermanentaccess.org/index.php/community/common-vision/

[19] PTAB ISO 16363 website see http://www.iso16363.org/

[20] CCSDS Data Archiving and Ingest Working Group – see http://cwe.ccsds.org/moims/default.aspx#_MOIMS-DAI

[21] RDA Active Data Management Plans Interest Group – see https://rd-alliance.org/groups/active-data-management-plans.html