Project no. 600663

PRELIDA

Preserving Linked Data

ICT-2011.4.3: Digital Preservation

D3.1 State of the art assessment on Linked Data and Digital Preservation

Start Date of Project: 01 January 2013

Duration: 24 Months

APA

Version : final

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

Document Information

Deliverable number:	3.1
Deliverable title:	State of the art assessment on Linked Data and Digital Preservation
Due date of deliverable:	Feb 2014
Actual date of deliverable:	Feb 2014
Author(s):	Sotiris Batsakis, David Giaretta, Christophe Gueret, Rene van Horik, Maarten Hogerwerf, Antoine Isaac, Carlo Meghini, and Andrea Scharnhorst. Comments and text contributions from Albert Moreno-Penuela, Peter Doorn, Marat Charlaganov and Menzo Windhower.
Participant(s):	CNR, APA, HUD, UIBK
Workpackage:	3
Workpackage title:	State of the Art Assessment
Workpackage leader:	APA
Est. person months:	6
Dissemination Level:	PU
Version:	1.0
Keywords:	Digital preservation, linked data

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level

Abstract

The state of the art of Linked Data technologies and standards and of Digital Preservation solutions, standards and technologies is presented, along with an analysis of the characteristics of Linked Data that make their preservation different from that of other digital resources (A consolidated version of the report will be published at the end of the project).

Table of Contents

Executive Summary

This report is the result of bringing together representatives of the Linked (Open) Data and the Digital Preservation communities in a workshop, supplementing desk research. It presents an overview of the fundamental concepts and current capabilities of Digital Preservation and Linked Data. This is followed by our initial ideas of where Digital Preservation seems to have answers and where there seem to be no answers – yet.

In M24 a consolidated version of this document will be published.

1 Introduction - contextualizing PRELIDA

1.1 Problem statement

PRELIDA's point of departure is the identification of a gap between two communities: **Linked Data**[1] (LD) or Linked Open Data (LOD) as part of communities in the computer sciences who develop semantic web technologies and **Digital Preservation**[2] as discussed in the context of archives, libraries, and museums. Actually scanning through the report about the first workshop[3] there seem to be three notions around which the debate circulates: PRESERVATION, DATA, and WEB (SEMANTIC WEB TECHNOLOGIES). One could also see DATA as a notion, which concerns, and this way also links both communities.

According to the PRELIDA Description of Work (from now on "DoW" for short) the motivation for PRELIDA grows from the statement that the Linked Data community might not be aware of the discourse and developed solutions in the Digital Preservation community. Accordingly, information transfer is one goal of this report. On the other hand the DoW states that Linked Data have characteristics, which form a challenge to the Digital Preservation community.

From the workshop, both these initial statements are true.

To identify and describe those issues in a language comprehensible to both communities, is the second goal of this report.

1.2 Questions to be answered

Currently among researchers and information providers one can find quite different ideas and opinions about preserving linked data, different "preservation objectives" as well as different "preservation strategies". One frequently used statement says "just store the RDF[4]", implying that we do not need to do anything special when comparing linked data to other data. This report addresses primarily those differences in perception in the *academic discourse*, and the different open questions the academic communities involved have identified. However, as the report will detail, solutions to those questions cannot be developed properly without further limiting the scope of the problems addressed. For this, higher level responsibilities, such as scientific integrity, governmental openness to public, transparency in governmental decision-making, etc. need to be articulated to frame an otherwise open and unlimited academic search process.

As said above, one could think "preserving linked data" will in fact be exactly the same as "preserving a relational database" to which one would reply "just store the SQL dump". Intuitively, one might think that one would need to have special things to do with respect to the links and networked aspect (web) of the data, but it turns out in the actual debate that such a differentiation is rather more harmful

---

[1] Wikipedia defines Linked Data as "a term used to describe a recommended best practice for exposing, sharing, and

[2] The reference model for an Open Archival Information System (OAIS) defines Long Term Preservation as "The act of maintaining information, Independently Understandable by a Designated Community, and with evidence supporting its Authenticity, over the Long Term." The Wikipedia entry on Digital Preservation states "Digital preservation can be understood as the series of managed activities necessary to ensure continued access to digital information for as long as necessary,..." http://en.wikipedia.org/wiki/Digital_preservation For a taxonomy of terms as defined by the experts see http://www.dpconline.org/advice/preservationhandbook/introduction/definitions-and-concepts

[3] PRELIDA Deliverable D2.2, available on http://www.prelida.eu/results/deliverables

[4] RDF stands for Resource Description Framework and will be further explained in section 2.

than useful. (see a recent debate on the Internet[5] and section 4) What really is a burning issue, which relates to the Web aspect of the data, is the long-term stability of the URIs of the resources (e.g. by applying mirroring techniques). To ensure they always return the intended result even if the original source shuts down is what many, non-archivist, people expect when they speak about "preservation". Related to this is the question: How to settle license issues related to linked open data in relation to its long-term availability? One model to look at is the Perma CC service[6], that allows users to create citation links that will never break, and has an "opt in" function enabling the user to decide whether a source should be managed by the service or not[7].

This report aims to respond to these questions by (1) introducing the basic concepts of LD and LOD on the one side and Digital Preservation on the other side. One important point, still controversially discussed inside of the LD community, is if the technology and the underlying data should be maintained on the web under the authority of a network of data providers, or if it is appropriate and maybe even needed, to re-create resources and information at one specific place at the web, under one authority. We will also show, that LD can also be created locally, without, or outside of the web. As stated above, if references are made to web resources, issues of web archiving, or stability of both location and content of web resources (link rot versus content rot) become relevant.

In a second step we revisit standards, typologies, and classifications developed in a decade of research projects in Digital Preservation. We interrogate how they can be used when reflecting if and how to preserve Linked Data.

Thirdly, we shortly describe three use cases from current practices, where the problem how to preserve LOD or LD takes a very concrete shape. This case description is set up as in a format, and meant to start off a collection of case studies that will be extended in the final version of this document.

1.3 Purpose of the report

A number of stakeholders are listed: data providers, service providers, technology providers and end user communities. They are actually stakeholders for both the Linked Data community and the Digital Preservation community. The results of this Coordination and Support Action have been designed to be of multiple uses:

- an inventory - a knowledge base - of material (reports, publications, codes, projects, discursive reflections in blogs, …) around the issues how to preserve Linked Data -in the form of an overview sections together with a bibliography
- to create a *market place*[8] where meetings and collaborative writing takes place to organize a trans-community discourse, to clarify and align notions, to reach a shared view and vocabulary
- to provide user communities access to forefront solutions - and to feed back immediate needs from those user communities (archives, research communities, ….) into the debate with the aim to push further the problem definition and solution process among the ICT experts.[9]

[5] http://krr.cs.vu.nl/2013/10/on-the-use-of-http-uris-and-the-archiving-of-linked-data/ [cited 11 January 2014]
[6] http://perma.cc/

[7] See: http://www.perma.cc/about The service developed by Harvard Law School and aimed at curating legal sources. [cited 12 January 2014]
[8] This is the function of PRELIDA meetings.
[9] This will be supported by the nature of this text, which is both an introduction and general review as well as

1.4 Method of the report

We think it therefore most appropriate and most important to use a language as general as possible. Following Galison's model of a *trading zone* (Galison, 1997) we are aware that this includes 'translations' of technical terms into a more mundane, broadly comprehensible language. Moreover we apply principles of science mapping (as mind maps) to illustrate the landscape of traditions, conceptual models, notions, projects, persons and institutions. Departing from such a more comprehensive level, we introduce the OAIS Reference Model as a framework developed by the space data community to enable mutual understanding on digital preservation and at the same time as a blueprint against which measures to raise awareness of archival concepts needed for long term digital information preservation and access can be taken.

We start with a short contextualization of the explorative task of the state of the art report. Hereby we reference back to pre-web situations as often as appropriate. The motivation for such a ***historically informed sketch of the current research landscape*** is to widen the mutual understanding between communities which so far have operated in quite distinct areas of the large and scattered science landscape. By reference to the history of the domain specific discourse we increase the overlap in mutual understanding. In other words, we refer to situations known in science history that we all can relate to, independently of the community we belong to nowadays. This way we create a ground on which experts can 'locate' their area of expertise in a wider framework and non-experts also can engage with the discussion. The latter group might this way be able to identify which parts, which notions, which techniques and which experts user communities they need to approach for their specific needs.

The next section (Section 2) ***summarizes the current state of the art*** in both communities. The following sections give a description of the exchange of ideas between the LD and DP communities.

The third section presents the state of art in Digital Preservation on the basis of results of DP projects in the last decade. Building on this, Section 4 discusses general aspects of preserving Linked Data. Section 5 starts an envisioned **collection of use cases** by presenting a first use case. By use cases we mean occasions, projects where questions of preserving linked data occur. More particular we will describe one project, CEDAR, DANS is involved in a format that can be used for the description of other use cases in other partner institutions.

From the identified challenges by each of the communities separately, and their manifestation in the practice (use cases) we draw a list of preliminary **conclusions**. Those will be turned into a list of action in course of PRELIDA, in other workpackages, leading eventually to the design of a roadmap.

In all sections we refer to existing literature, listed in a bibliography at the end of the document.

---

listing very specific needs. One could also imagine follow up 'implementation' projects from PRELIDA.

2 Definitions and terminology

2.1 Preservation - Linked - Data. Describing the context of the discourse

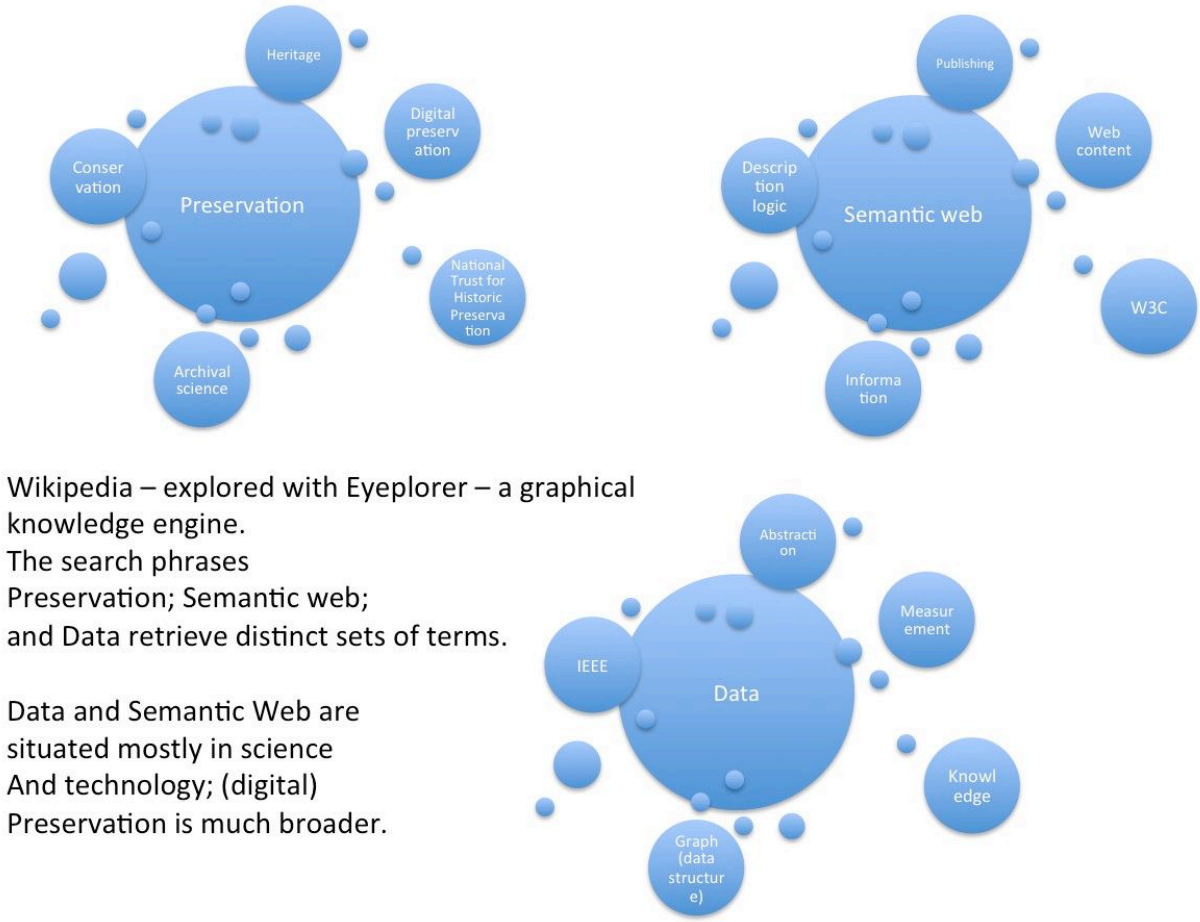


Figure 1 Preservation - Web - Data - main issues addressed in Wikipedia

From all three (**preservation, linked or web or semantic web,** and **data**) clearly preservation is the oldest, and one with a large use across scientific disciplines. Wikipedia lists no less than 26 different reference points for this notion.[10] In the area of cultural heritage and education - the area we focus on - preservation has been a concern of administrations through thousands of years which build archives to preserve bills, laws, land titles and so on. One could say from the cuneiform scripts on clay table to digital records[11]. Given role of archives, museums and libraries for this task since the early modernity it cannot be a surprise that the discourse about **Digital Preservation** is led by scientific communities

[10] http://en.wikipedia.org/wiki/Preservation

[11] For a short summary why digital preservation of records is a concern for the *Memory of the world* see http://www.unesco.org/new/en/media-services/single-view/news/digital_preservation_preserving_heritage_and_protecting_civil_rights/#.UnTXNCSE4-I

as information sciences, archival sciences, library and information sciences. The workflows at archives and libraries including steps of selection of material to be preserved, standards to index and document holdings, ways to ensure long-term preservation and policies of access all flavour the actual discussion about digital preservation. At the same time libraries and archives have been shaken profoundly since the emergence of computers, digitization and more lately the web. From having reached ironclad authority as public institutions accompanying industrialization and modernization, they suddenly found themselves threatened from being closed down, and are continuously trying to re-define their role and position in society at large and more particularly for academics. The societal status of being a librarian or an archivist needs still to be regained. Their question *What to do with Linked Data?* is the newest variant among questions around *What to do with digital material of various kinds?* This on-going transformation institutionally with new players emerging almost hourly is accompanied with a struggle for new identity among actors which can show mood swings from almost unbearable institutional proudness and stubbornness towards feeling helpless. This is not meant to serve as a characteristic of involved concrete parties and persons, but rather to sketch the overall situation. [12]

On the other side, the web as a technology has only been around for about 25 years, and yet has penetrated every corner of society (Slevin 2000; Webster, 2002). The representatives of web based technology come quite rightly with the attitude of explorers of new territory, they are engineers and makers from the attitude of their hearts - the new masters of our information and knowledge management. The Web is the greatest digital resource of our era, and thus Web archiving has emerged as the process of collecting portions of the World Wide Web to ensure the information is preserved in an archive for future researchers, historians, and the public. Multiple efforts have been devoted to the purpose of preserving the contents of the Web, such as the Internet Archive[13], archiving various types of web content such as HTML pages, style sheets, Javascript, images and video. However, there is an emerging part of the Web, called the Semantic Web (Berners-Lee et al, 2001), that consists of different content as the traditional Web. As envisioned in 2001 the Semantic Web was conceived as an evolution of the existing Web, built essentially on the paradigm of the document, into a "semantic" Web, built on the paradigm of meaning and structured data. By that time, most of the contents of the Web were designed for humans to read, but not for computer programs to process meaningfully. Computer programs could parse the source code of Web pages to extract layout information and text, but they had no mechanism to process their semantics. In other words, the Semantic Web "is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation" (Berners-Lee et al. 2001). The community movement motivated by this vision has been developing technologies such as RDF and OWL to make the Web evolve from a Linked Document space into a Linked Data space, where every structured data[14] portion is given a URI and is linked to other structured data portions. (see section 2.3) This way, a big graph of linked data is being created on the Web in parallel to the big graph of linked documents. RDF is the language in which Linked Data is expressed (just like HTML is the language in

[12] Andrew Prescott's blog Digital Riffs gives ample evidence for this in each of his paper-lenght blog entries. http://digitalriffs.blogspot.nl/ Andrew states about himself "I am Professor of Digital Humanities at King's College London. I was formerly a Curator of Manuscripts at the British Library, and have worked in digital humanities units and libraries in Sheffield, Lampeter and Glasgow." Andrew is also the one who called for a return of Academic Librarianship in his paper given at the Digital Humanities Congress in Sheffield 2012 (see http://digitalriffs.blogspot.nl/2012/09/made-in-sheffield-industrial.html)

[13] http://archive.org/web/

[14] http://structured-data.org/

which Linked Documents are expressed). Linked Data are a form of formal knowledge. This explains the urgency felt to discuss its preservation. At the same time this formal knowledge lives on the web, is by nature distributed, can be accessed in different ways and is in constant flux. These are all features that make preservation to a big challenge.[15]

Still the web since born has one problem that sets it quite opposite not to talk outside any consideration of preservation. Designed to organize knowledge **flows** it comes intrinsically without a memory. (Berners-Lee and Fischetti, 1999; Chung and Keenan, 2006) Each website we browse through is actually a graph of web resources. (Huberman, 2001) An often made mistake is to think about a website as a document or a record, while it is a network of resources - as if you would have a text document and all figures, tables and references would actually come as extra bundles. In this network the resources are dynamic, they can change over time content-wise, and they can come from different locations. Returning to our analogy: the tables and figures might be provided by different parties and change asynchronously.[16] This makes it even harder to think up any scheme to archive parts of the web. Still from the very beginning web archiving has been a growing concern, articulate by communities of scholars that care about the web and the internet as home for wealth of cultural artifacts.[17] The Internet Archive with its Wayback machine and international debates about web-archiving are the most visible expressions. (Masanes, 2006; Niu, 2012).[18]

In parallel to the development of web technology we find debates on the persistence or ephemerality of URLs (Koehler, 1999). URIs, persistent identifier can be understood as countermeasures against *link rot* in web resources. Increasingly it was understood that persistence of URLs is not a technological issue only, but - as always in technological innovation - an issue of social dynamics - negotiations, agreements and maybe institutionalization. **Memento**[19] - an invention (and project) to 'dig out' existing earlier versions of websites bridges between an institutional approach to web-archiving - very similar to traditional archives and libraries; the mundane need of a web browser to see earlier versions; and the use of web technology in a way that embeds archiving into the distributed character of web-based information and the dynamic way views from web resources are presented to the viewer as "a document". Memento functions as a browser plugin. It uses the fact that historic versions of web resources are kept, either by the Content Management System behind a website or by internet archives. Installed at a server it allows institutions to archive their websites internally and the visitor of the websites to see earlier versions. (Sanderson et al. 2011) **Transactional archiving** - another principle proposed by Herbert van de Sompel and his team equally responds to the question what to archive, when and in which way. Urged by the increase of references in scholarly publications which cannot be retrieved anymore - called reference rot - Herbert van de Sompel and others set up a new project **Hiberlink**, which analyses the problem systematically. (Sompel et al., 2013) One can

[15] The attributes: web-based infrastructure; distributed by nature; accessible in different ways; and changeable leading to the problem of versioning have been identified at the first PRELIDA workshop.

[16] The consequence of this network dynamics for the integrity of the scholarly record has been illustrated lately in a presentation of Andrew Treloar and Herbert van de Sompel, where on slide 40, for an archived (!) website the timestamp of different resources belonging to this particular site have been indicated. (Treloar, Van de Sompel 2014).

[17] See the Association of Internet Researchers as an example http://aoir.org/

[18] Web archiving played a role in different projects. A search in the CORDIS database on the exact term "web archiving" delivered the FP7 project: Living web archives. http://www.liwa-project.eu/

[19] http://www.mementoweb.org/news/

predict similar, and probably even more complex problems when extending the web of scholarly publications with the web of research data.

The emergence of Linked Data, or the Web of Data as the newest technology on the HTTP level of the Internet architecture gives new impulse to the discussion of web-archiving. Linked Data can also be understood as a way to create indices to knowledge resources. If we compare just for a moment, web resources with books, URLs can be compared to call numbers; and the use of controlled vocabulary to characterise them with the classification systems - or Knowledge Organization Systems used for books in libraries. Those web-based kind of indices can in principle target any kind of information. They can also be directed to objects or resources from our cultural heritage that have been archived traditionally. Think in terms of objects from musea, information about places-events-persons, or statistical information about welfare. Applied to this kind of information, the possibility emerges to create an all-encompassing catalogue to cultural heritage. The emergent web of knowledge resembles dreams of Paul Otlet with the foundation of the *Institut International de Bibliographie.* (Rayward, 1996; 2013) What makes such an enterprise much more complex than in times of Otlet, is that the objects to be bibliographically described are moving targets, as well as the means to bibliographically describe them. To return to our analogy: imagine a library in which the books after being stacked by a librarian at a shelf would start behind her or his back to re-locate themselves and maybe so disappear. And if this were not enough: while the librarian writes index cards to them to later go into subject or author catalogues, somebody also would play cards with those already being put into the drawers.

Aside of these technologically-inherent problems, one also has to be aware that Linked Data as a (research) technology is for a large part situated in the area of fundamental research. New concepts, concepts of proofs and pushing the technological boundary are at the core of the discourse. The transfer of knowledge and technique from the research cycle into information services usually requires the innovation to be mature and consolidated. This is a lesson from innovation studies - at least what concerns large-scale adoption of new technologies. (Rogers, 2003) The knowledge exchange in PRELIDA and possible envisioned implementation (use cases) represents an encounter between fundamental and applied research at a very early phase of the innovation diffusion curve. It aims at a co-evolution of developing - or at least reflecting about - new technologies and new services. It is important to keep this in mind when it comes to expectations of what PRELIDA can deliver. It is - given the available time and resources - not feasible to expect new standards, generic solutions, or massive implementations. What can be expected is to deliver a fairly comprehensive state-of-the-art snapshot in a very volatile research and service environment.

2.2 What do we mean by digital preservation?

2.2.1 Digital preservation – initial considerations

Digital preservation has been a concern from the rise of computer technology after WW II. But, it took momentum with the emergence of the Internet, the scaling up of digitization and the changes in scholarly practices in the digital age. (Borgman, 2007) Some sources (Documentation Abstracts, 2002) point to "Preserving Digital Information" (Waters, Garrett 1996) as a 'landmark report' in the 1990s. In 1995 Rothenberg raised general awareness of the problem that digital documents have a rather short life. Digital media "*will last forever - or five years. Whichever comes first*" (Rothenberg 1995 p.42). From 1995 onwards several digital preservation projects and studies were carried out on a wide range of subjects. They consisted of inventories and assessments of digital resources, tools and methods to preserve digital material and standards, and guidelines to support digital preservation.

Barbara Sierman from the Royal Library in the Netherlands published a blog entry September 14, 2012 asking "Where is our Atlas of Digital Damages?" (Sierman, 2012). As Bill LeFurgy responded in another blog "Her argument spurred action, and the "Atlas of Digital Damages" now is up and running on Flickr[20]. This is a crowdsourced effort, and anyone can upload pictorial evidence of bits gone bad. There are currently a few dozen images available, but it is easy to imagine building quite a large collection of compelling images."[21] One could imagine that similar to the *Atlas of Damage* for off-line material (Most et al., 2010) the visual documentation is lined up with a typology, and good practices determined in a large number of digital preservation projects, leading to a tool each digital librarian and archivist can use as a handbook.

There have been many rather informal definitions of digital preservation over the years. For example:

"Digital preservation refers to the series of managed activities necessary to ensure continued access to digital objects for as long as necessary"[22]. *"The goal of digital preservation is, hence, the accurate rendering of authenticated content over time"*[23].

However we should note that definitions like these refer to "access" and "rendering", which are useful for certain kinds of digital content, but not, for example, for Linked Data.

2.2.2 The role of OAIS

As noted in the introduction the OAIS Reference Model now plays a fundamental role in digital preservation activities. This is in part because it takes a more logically complete view of preservation, and moreover one which can be tested. To this end OAIS defines Long Term Preservation as

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

The various components of this are further defined as follows.

---

[20] See: http://www.atlasofdigitaldamages.info/v1/ The Flickr site or group by January 2014 contains 99 examples of "digital damages". [cited 13 January 2014]

[21] http://blogs.loc.gov/digitalpreservation/2012/10/bits-breaking-bad-the-atlas-of-digital-damages/

[22] Definition taken from: Neil Beagrie and Maggie Jones "Preservation management of digital materials: The Handbook" (Digital Preservation Coalition), The printed handbook was published in 2001. The online version contains updates until November 2008. See: <http://www.dpconline.org/publications/digital-preservation-handbook> [cited 13 January 2014]

[23] Quotation from Wikipedia entry http://en.wikipedia.org/wiki/Digital_preservation Accessed February 1, 2014.

Long Term:

A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing Designated Community, on the information being held in an OAIS. This period extends into the indefinite future.

Independently Understandable:

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.

Designated Community:

An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities. A Designated Community is defined by the Archive and this definition may change over time.

Authenticity:

The degree to which a person (or system) regards an object as what it is purported to be. Authenticity is judged on the basis of evidence.

Information:

Any type of knowledge that can be exchanged. In an exchange, it is represented by data. An example is a string of bits (the data) accompanied by a description of how to interpret the string of bits as numbers representing temperature observations measured in degrees Celsius (the Representation Information).

Data:

A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing. Examples of data include a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen.

Representation Information:

The information that maps a Data Object into more meaningful concepts. An example of Representation Information for a bit sequence which is a FITS file might consist of the FITS standard which defines the format plus a dictionary which defines the meaning in the file of keywords which are not part of the standard. Another example is JPEG software which is used to render a JPEG file; rendering the JPEG file as bits is not very meaningful to humans but the software, which embodies an understanding of the JPEG standard, maps the bits into pixels which can then be rendered as an image for human viewing.

From these definitions and supporting concepts a solid basis for digital preservation is presented in OAIS. These can be used in the first approach to preservation of LD. We come back to this model in section 3.

2.2.3 Threats to digital preservation

Moreover threats to digital object preservation have been discussed as follows:

> *The main threats to long-term access to digital objects are file format obsolescence, storage medium failure, the fact that value and function of the digital object cannot be determined anymore (often due to the lack of appropriate documentation) and simple loss of the digital objects.*

While these considerations are important for Linked Data nevertheless there are clearly other threats such as the way in which things are linked, and the authenticity of the data.

A more thoroughgoing discussion of threats to digitally encoded information has been provided by the PARSE.Insight project (2008-2010). Through a number of large surveys of researchers, data managers and publishers, with several thousand responses world-wide it became clear that the following threats were very widely recognized, cross national boundaries, and across disciplinary boundaries. Note that this, consistent with the idea of *trading zone* (Galison, 1997), involved 'translations' of technical terms into a more mundane, broadly comprehensible language.

Table 1 Threats to digital preservation identified by PARSE.Insight

	Threat
1	Users may be unable to understand or use the data e.g. the semantics, format, processes or algorithms involved
2	Non-maintainability of essential hardware, software or support environment may make the information inaccessible
3	The chain of evidence may be lost and there may be lack of certainty of provenance or authenticity
4	Access and use restrictions may make it difficult to reuse data, or alternatively may not be respected in future
5	Loss of ability to identify the location of data
6	The current custodian of the data, whether an organisation or project, may cease to exist at some point in the future
7	The ones we trust to look after the digital holdings may let us down

2.2.4 Remedies

By the year 2000 three main strategies towards digital preservation have been described. [Beagrie and Jones, 2001, p 26]. These are

- The **technology preservation strategy,** preservation of the original software and hardware that was used to create and access the information,
- The **technology emulation strategy**, future computer systems emulate older, obsolete computer platforms as required, and
- The **digital information migration strategy**, digital information is re-encoded in new formats before the old format becomes obsolete.

The three digital preservation strategies were applied for different purposes and user groups and to a wide range of digital materials, such as computer programs, digital images, electronic texts and web

pages. For a number of years the digital preservation paradigms described above dominated the research direction, debate and focus of the digital preservation community. In the course of time this strategy discussion moved to the background and new insights emerged on what should be done to preserve digital objects. Since initiatives have been started internationally and nationally and a number of solutions and recommendations were formulated to cope with the issues mentioned above. Examples for recommendations are: use file formats based on open standards, use the services of digital archives to store the objects for the long-term, create and maintain high quality documentation (e.g. the PREMIS standard, specifically developed to create preservation metadata[24] so in the future the digital objects can be reused, or make use of multiple storage facilities to reduce the risk that the objects get lost (e.g. by applying the LOCKSS (Lots Of Copies Keeps Stuff Safe) method[25]).

However based on its collected information, PARSE.Insight (2008-2010) proposed a number of approaches to counter the broader collection of threats referred to in Table 1.

Table 2 Threats and resolutions

	Threat	Requirements for solution
1	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
2	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
3	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
4	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
5	Loss of ability to identify the location of data	An ID resolver which is really persistent
6	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
7	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

[24] PREMIS Data Dictionary for Preservation Metadata, see: <http://www.loc.gov/standards/premis/>

[25] See: <http://www.lockss.org/ >

In the course of time consensus has been reached on the features of digital preservation services that are required to guarantee long-term access to them. A key component of the digital preservation infrastructure are so-called Trusted Digital Repositories (TDR) that are based on the OAIS reference model[26]. Which exact characteristics a TDR should adhere to is currently under debate and development. There is agreement that a TDR should meet criteria that are formally checked by an audit and certification process. A number of certification initiatives do exist and they collaborate in a European framework for audit and certification[27]. The framework contains a number of levels from basic self-certification to extended certification carried out by external auditors[28].

In the APARSEN project the ISO standard for the Audit and Certification of Trustworthy Digital Repositories (ISO 19363) has been used as a landscape for checking the coverage of various aspects of preservation. It can be used in this way because the standard contains metrics covering, in principle, all the things which need to be done by a trustworthy repository – see Figure 2

[26] ISO 14721: 2012 *Space data and information transfer systems - Open Archival Information System - Reference Model*, International Organisation for Standardisation. Also published as: Reference model for an Open Archival Information System (OAIS). Online available at <http://public.ccsds.org/publications/archive/650x0m2.pdf> [Cited 15 January 2014)

[27] See: <http://www.trusteddigitalrepository.eu> [Cited 16 January 2014]

[28] The "Dataseal of Approval" contains 16 guidelines for a trusted digital repository can be applied and checked in online self-assessment process. See <www.datasealofapproval.org>. A more extended and formal audit and certification is described in the technical Recommendation "Audit and Certification of Trustworthy digital repositories", see: <http://public.ccsds.org/publications/archive/652x0m1.pdf> It provides a detailed specification of criteria by which digital repositories shall be audited, based on the OAIS reference model.

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	

Figure 2 Examples from SCIDIP-ES, mapping services to the ISO 16363 metrics

An important issue in this development is the emergence of new types of digital objects that cannot be classified according to the traditional "document" oriented approach and for which the traditional metaphor of storing objects in archives, and retrieving them with inventories and catalogues is not valid any more. The APARSEN project [29] (see also Giaretta 2011, pp 31-39) proposed different possible classifications of digital objects - an issue we come back to later.

A new term emerged for the activities that are required to manage digital objects for the long term: **digital curation**. Digital curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle.[30] The notion grew that digital archiving is not the last phase of a linear process in which objects are stored and kept for future generations, but that digital objects have a life cycle of its own. Secondary analysis, replication, enrichment and combining of digital objects are important functions a TDR must support, thus extending its rationale.

In section 3 below we present an overview about achievements in Digital Preservation (DP) in greater detail, presenting projects and dimensions along which the discourse of DP unfolds. Implications for archiving Linked (Open) Data (LD) are discussed at various places in this document: from the perspective of LD as a developing technology; from the perspective and experiences of DP; and from the perspective of a very concrete use case. The conclusions summarize insights and challenges.

[29] APARSEN is a FP7 project, a Network of Excellence: see http://www.alliancepermanentaccess.org/index.php/aparsen/

[30] Definition from the UK Digital Curation Centre <http://www.dcc.ac.uk/digital-curation/what-digital-curation>

2.3 What do we mean by Linked Data?

2.3.1 From Data, to Open Data, to Linked Open Data

Once considered to be a closed asset, data is now considered to be the "new oil" (Rotella, 2012). Its true value comes with its usage at its gets processed. But in order for this data to be processed by other parties it has to be shared. Data sharing among companies has a long history paved with pair agreements made on a case by case basis: company A with an interest in the data of company B will contact it to make an agreement to exchange data against a monetary compensation. Governments and citizens display a similar pattern when personal data and the **P**ublic **S**ector **I**nformation (PSI) directive come into play.[31] This directive essentially states that every data collected with public funds shall be made accessible to the public when this public requests it. Data falling under this directive have been long considered to be closed data to be shared only on-demand. Processing such demands is a costly administrative process which pushed the institutions into making the data widely accessible to everyone right away, fully open. As an example, the UK - a pioneer in the open data landscape - can save Between £16bn and £33bn a year by opening up its data according to a report by the Policy Exchange think tank[32]. Besides saving on administrative processes spendings, opening up public datasets yields the expectations of the creation of businesses using this data and bringing back indirect revenues to the state. But the loss of the control point that represented the processing of requests comes at a cost: the data made open can, and will, be used in unexpected way; combined with other datasets and interpreted in a wrong way.

The first manifestation of this liberation of data is the open data portals. A data portal is a place where data sets are made available in an open license are uploaded and/or referenced. There are more than 150 of such data portals in Europe[33] aiming at providing access to a wide range of data sets both in the public, scientific and cultural heritage domain. What all these portals have in common is the possibility to download data sets or parts of data sets: a user is invited to get a file containing data in a particular serialization format and conceptual model.

For a data consumer, the task that comes after downloading open data is data integration and data analysis. The objective is to combine all the heterogeneous data acquired from different sources into one coherent dataset that can be used by a given application. The main challenge is to create unambiguous terms. "Boston", for example, may refer to a city in the US, several cities in the UK, a baseball team or even a music band [MetaWeb]. The main idea behind Linked Open Data (LOD) is to use unique identifiers instead of ambiguous words for everything, from the concepts referred to in the dataset to the model used to express the data. The design principles of LOD are defined by Tim Berners Lee[34] and can be summarized by (1) use the Web as a platform to publish and re-use identifiers that refer to data, and (2) use a single data model for expressing the data (RDF).

---

[31] Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information See: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:02003L0098-20130717:EN:NOT [accessed February 25, 2014]

[32] See: http://policyexchange.org.uk/images/publications/the big data opportunity.pdf [cited 29 January 2014]

[33] See: http://www.slideshare.net/OpenDataSupport/open-data-support-service-description Examples of data portals ar: http://open-data.europa.eu and http://publicdata.eu.[cited 9 January 2014].

[34] See: http://www.w3.org/DesignIssues/LinkedData.html. The design principles for Linked Open Data were defined in 2006. [cited 15 January 2014]

The Resource Description Framework (RDF)[35] is a way to model data as a list of statements made between two resources identified with their unique identifier. For example, the sentence "Lille is in France and called 'Rijse' in Dutch" can be expressed as two statements in RDF (see Figure 3).

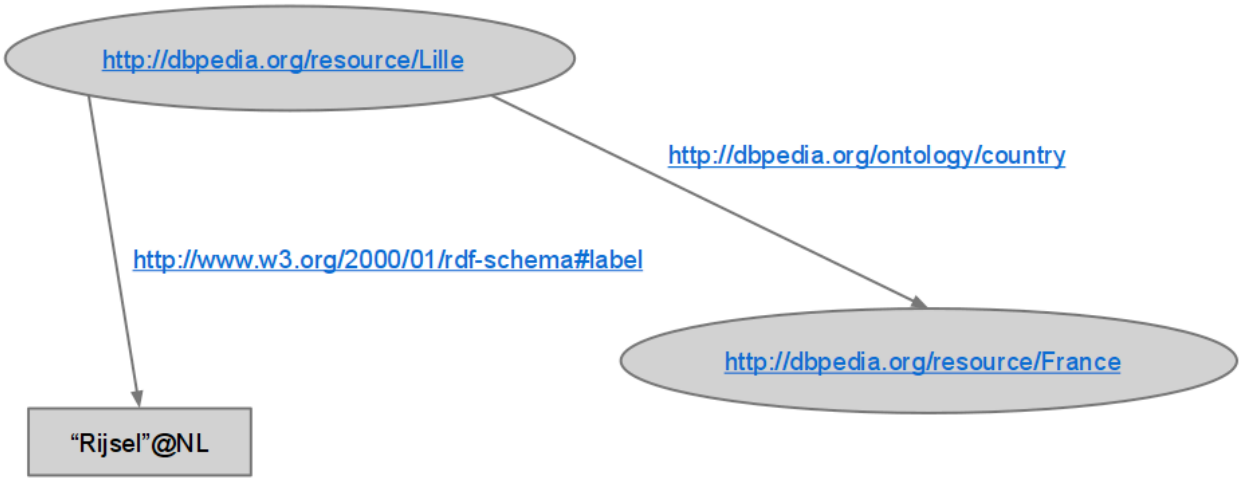


Figure 3 An example RDF representation of "Lille is in France and called Rijsel in Dutch"

The drawing of Figure 3 follows the common representation convention of using ellipses for resources and squares for literals. It can be observed that by using a resource instead of a literal for "Lille" the two statements are connected. Following the same principle across several datasets leads to the creation of a "Web of Data", a pre-integrated dataset.

2.3.2 Publishing and consuming the Web of Data

RDF is a modelling language that let users express their data along, with the schema describing it, as a graph. There exists then several serialisation formats for this RDF data. Turtle[36] (TTL), TriG[37], RDF/XML[38], RDFa[39] are only but a few examples. In fact, one can distinguish 3 ways to publish RDF data:

- As annotation to Web documents: the RDF data is included within the HTML code of Web pages. Software with suitable parsers can then extract the RDF content for the pages instead of having to scrape the text.
- As Web documents: RDF data is serialized and stored on the Web. RDF documents are served next to HTML documents and a machine can ask which type of document it needs. Typically, HTML for human consumption and RDF for machine consumption

[35] RDF, Resource Description Framework is a standard model for data interchange on the web. See: http://www.w3.org/RDF/ [cited 5 January 2014]

[36] Turtle is a textual syntax for RDF. See: http://www.w3.org/TR/turtle/ [cited 24 January 2014]

[37] TriG is an extension of the Turtle RDF syntax. See: http://www.w3.org/TR/2014/PR-trig-20140109/ [cited 24 January 2014]

[38] See: http://www.w3.org/TR/REC-rdf-syntax/. RDF/XML syntax specification was defined in 2004 [cited 24 January 2014]

[39] RDFa is a syntax for embedding RDF in Web pages through HTML attributes. See: http://www.w3.org/TR/rdfa-syntax/ [Cited 24 January 2014]

- As a database: RDF can be stored in optimised graph databases (called "triple stores") and queried using the SPARQL query language. This is similar in spirit to storing relational data in a relational database and query it using SQL.

There is a variety of considerations that come into play when deciding between the three approaches. One of them is the size of the dataset, the annotation approach is commonly used for "small data" (*e.g.* social profile on a home-page) whereas the database approach rules "large data" (*e.g.* the content of Wikipedia expressed as RDF). Most often what is put in place is a combination of all three approaches (see Figure X).

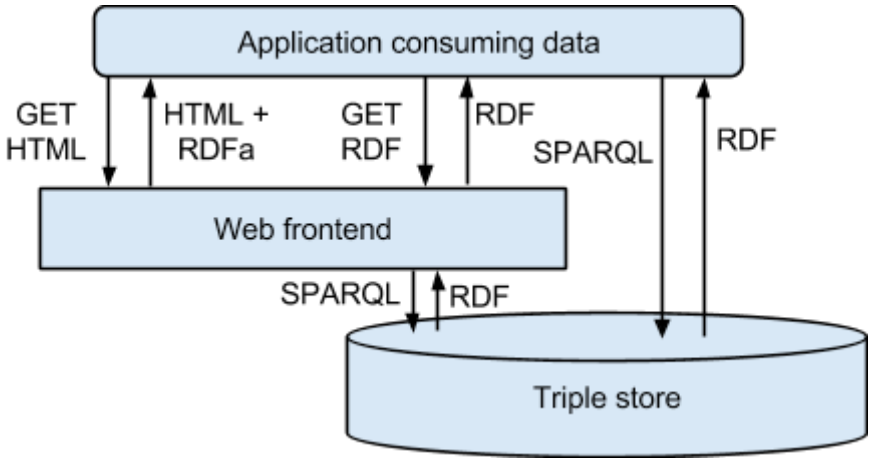


Figure 4 A common publication architecture for RDF data

The architecture depicted on Figure 4 is the one in place for DBpedia[40], an RDF version of the structured content available in Wikipedia. The description of "Amsterdam", the city in the Netherlands, can be queried from the three different ways as introduced above (all links valid on January 16 2014):

- As annotations through the RDFa markup present in the HTML page http://dbpedia.org/page/Amsterdam (see http://www.w3.org/2012/pyRdfa/extract?uri=http%3A%2F%2Fdbpedia.org%2Fresource%2FAmsterdam&rdfa_lite=false&vocab_expansion=false&embedded_rdf=true&validate=yes&space_preserve=true&vocab_cache_report=false&vocab_cache_bypass=false for the output)
- As RDF content via content-negotiation with the resource http://dbpedia.org/resource/Amsterdam (see http://www.w3.org/RDF/Validator/rdfval?URI=http%3A%2F%2Fdbpedia.org%2Fresource%2FAmsterdam&PARSE=Parse+URI%3A+&TRIPLES_AND_GRAPH=PRINT_TRIPLES&FORMAT=PNG_EMBED for the output)
- With a SPARQL query sent to the end point http://dbpedia.org/sparql (see http://dbpedia.org/sparql?default-graph-uri=http%3A%2F%2Fdbpedia.org&query=DESCRIBE+%3Chttp%3A%2F%2Fdbpedia.org%2Fresource%2FAmsterdam%3E&format=text%2Fplain&timeout=30000&debug=on for the output)

[40] See: http://wiki.dbpedia.org/Architecture [Cited February 3 2014]

The three outputs are expected to contain the same RDF data. Several formats can be queried for it, from RDF/XML to CSV to JSON[41]. But whereas DBpedia shows the example in terms of flexibility for the user, not all RDF datasets are published that way. There are in fact pretty much two Web of Data out there, for which different preservation strategies can be proposed.

The differentiation between two Webs of Data comes back if we take the perspective of a user, consuming Linked Data. We need to distinguish between two different types of users of Linked Data. First, some users of Linked Data do not care about keeping them functionable online. (**off-line use)** They typically store local replicas of the RDF data they need to use, just as copying locally a traditional database, but don't use it to follow links online from one piece of data to the other. In such a case, a hypothetical Linked Data Archive (LDA) would only need to store RDF data dumps just as it stores HTML, Javascript, and the rest of the Web. The archived content can be considered "dead" (i.e. not actively used), and the original URI authority (i.e. the owner of the original domain) could be replaced by some meta-data describing it. Second, some other users use Linked Data on the Web **(on-line use)**, and thus they care about being able of jumping from the URI of one piece of data to the other. The technical notion for this is "making URIs de-referenceable". In order to preserve this, the LDA would need to implement a de-referencing service that could fetch out of the archive the description of a particular URI and return it as requested. Ideally a redirect would be established from the original domain name, and the LDA could then return different historical versions of the resource.

2.3.3 The two Webs of Data

Above we describe different ways to publish data according to Linked Data principles. But the publication of data in this form is not an activity standing for itself. It is connected to a further use of these data.

Resulting from the different options for publishing data according to the Linked Data principles, one can observe two versions of the Web of Data:

- The "Web" Web of Data: a network of semantically linked resources published exclusively on the Web. This is for example the case for most personal web pages, annotations added to pages to support the Open Graph protocol from Facebook or annotations added to enhance the indexing of Web pages by the major search engines (see Schema.org). This content is exclusively accessible on the Web and can not be queried using SPARQL, a query language for RDF[42].
- The "Data-base" Web of Data: a set of RDF statements stored in an optimised database and made queryable using SPARQL. This set of resources uses URIs that are not expected, and most of the time are not, dereferencable. As such this Web of Data is a graph disconnected from the Web.

These two Webs are closely related. Most often, a Web front-end with dereferenceable URIs will be supported by a database Web via the usage of a Linked Data frontend[43]. Other approaches concern the harvesting of Web-only data to make it accessible in a triple store[44] or the extraction of structured

[41] Comment of Rene vH to be taken up in the final version "How durable are these formats? This can be elaborated on in the consolidated version of the report. Also a good topic for a workshop"

[42] See: http://www.w3.org/TR/rdf-sparql-query/ [cited 11 January 2014]

[43] An example of a Linked Data frontend for SPARQL endpoints is Pubby, see: http://wifo5-03.informatik.uni-mannheim.de/pubby/ [cited 19 January 2014]

[44] See e.g. http://www.sindice.com [cited 15 January 2014]

information from Web pages as RDF dumps[45]. The majority of Linked Data consumption is performed "off-line", using statements stored in a triple-store.

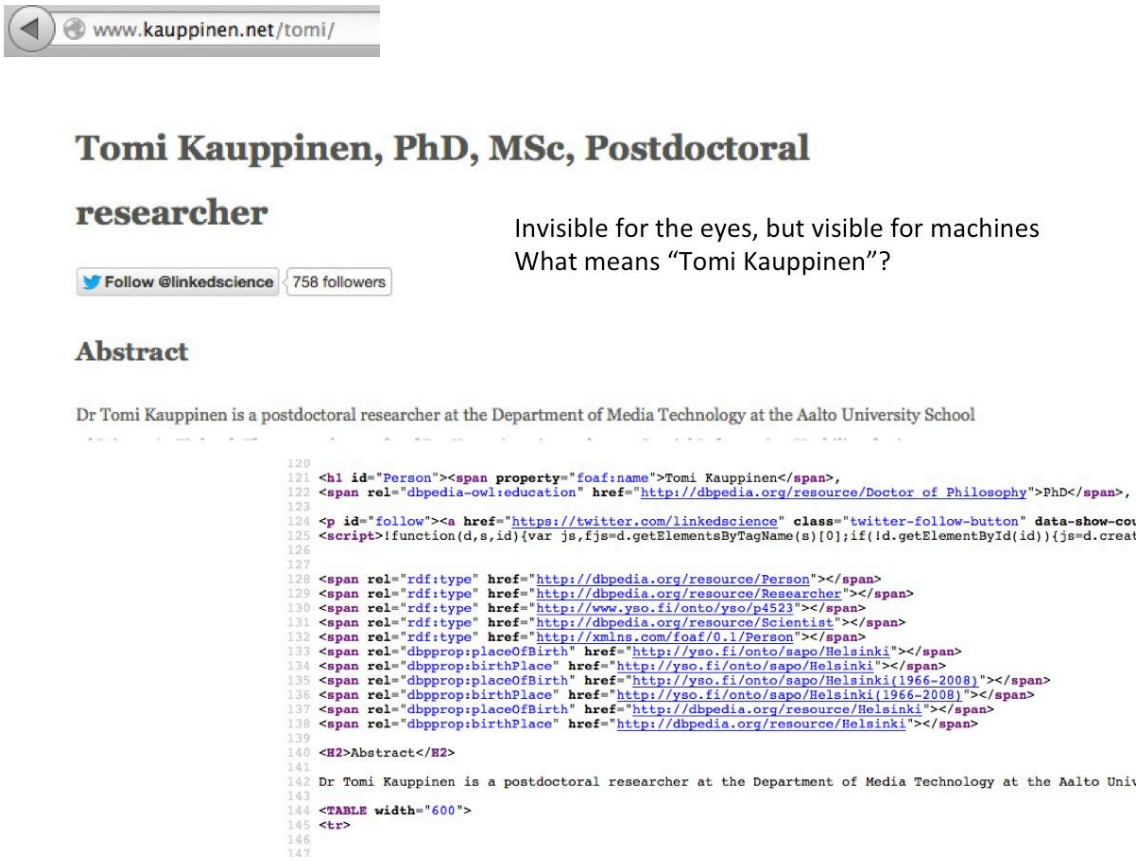


Figure 5 Example of annotations on the home-page of Tomi Kauppinen. These are only accessible on the Web and can not be queried from a triple store.

[45] See e.g. http://any23.apache.org [cited 15 January 2014]

3 Relevant dimensions addressed by digital preservation projects

As noted in section 2.2 there are several views of what digital preservation is, and despite the logical set of concepts provided by OAIS, the research in digital preservation is very fragmented. In part this arises from the funding mechanisms that are in place. These tend to force the projects to promote distinct solutions and, in a very real sense, to "oversell" their solutions. The effect is to cause uncertainty when selecting solutions.

There are however significant motivations for moving into a new regime.

- The EC has decided not to prioritise DP for funding research within the first stage of H2020 – although this may be reversed later stages of H2020. Thus an appraisal of the research that has been undertaken is necessary and an integrated view developed.
- Society expects to benefit from the resources invested in creating research, and other, data. For example Commissioner Kroes has stated several times "Data is the new Gold". Thus the data must continue to be re-used over the long-term, and production level services be created to enable this.
- The audit and certification of the ability of repositories to undertake digital preservation are being put in place. Thus claims of digital preservation will be tested.

In the rest of this chapter expands on these motivations and gives an overview about digital preservation research projects so far.

3.1 Digital preservation research projects

The current state of art concerning the long-term access to digital objects to a large extend is based on the outcomes of a number of EU projects. The report *Research on Digital Preservation within projects co-funded by the European Union in the ICT programme* (Strodl et al. 2011) gives an overview of the research on digital preservation of initiatives co-funded by the ICT program of the EU. The first projects aimed at raising awareness and the creation of a scientific community addressing the topic of digital preservation. The first activities were influenced by the archive and library community and the research was mainly focussed on office documents and images. In the next phase a number of technical projects were carried out that resulted in tools and services, such as file format registries and metadata management systems. This led to the availability of concrete solutions. The projects also have influence on international standardization initiatives in wide range of digital preservation fields, such as the OAIS reference model, audit and certification standards and metadata guidelines.

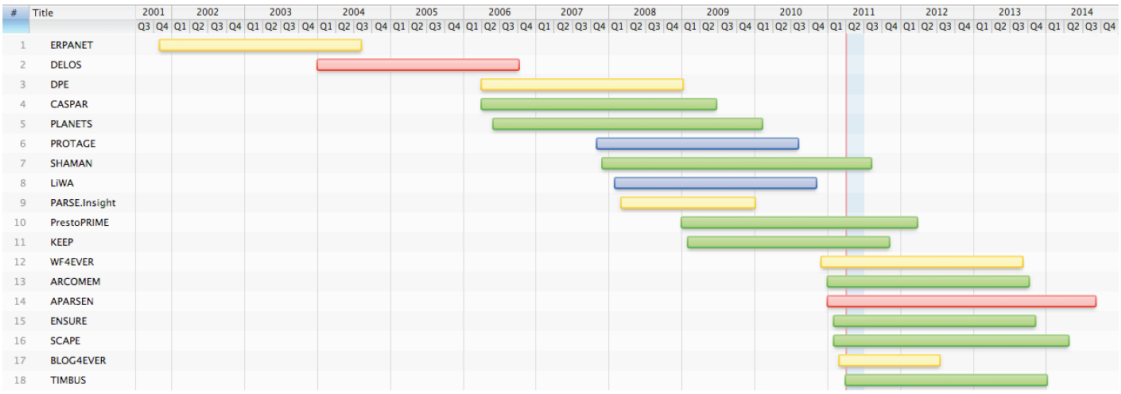


Figure 6 Reproduction of an overview about projects funded by the EC between 2001 and now devoted to Digital Preservation, Courtesy of Strodl et al. 2011, page 13. Color coding: blue = 'Specific Targeted Research Project', red = 'Network of Excellence', yellow =

The 7th EU framework program, started in 2007, provided means to start fundamental research concerning the digital preservation of complex digital objects, such as ontologies, interactive objects and embedded objects. Examples are the LIWA project[46] addressing web archiving and the TIMBUS project[47] aiming at the preservation of business processes. Another field of fundamental research concerns the validation of objects according to format specifications and policies. The PLANETS[48] and SCAPE[49] projects are examples of this. Outreach and networking is another important topic of the research activities on digital preservation. An example is the APARSEN project aimed at the establishment of a Network of Excellence on digital preservation[50].

The report of Strodl et al. (2011) classifies twenty of those projects along dimensions of **content type** (e.g., office documents, audio-visual material, research data etc.) (p. 15, Fig. 3), **targeted audiences** (memory institutions, scientific institutions, government, enterprises and private) and **key institutions** involved in DP projects and featuring in multiple (>2) projects.

Among the results of those projects three topics can be identified that potentially might be relevant for the preservation of linked open data objects.

- Object classification and validation
- Persistent identifiers
- Audit & Certification / Trusted Digital repositories

More recently, in the framework of DASISH (Data Service Infrastructure for the Social Sciences and Humanities) - an FP7 project several reports address digital preservation as part of evolving research infrastructures in the social sciences and humanities (Kvalheim et al., 2012; Anonymous, 2012).

[46] See: <http://liwa-project.eu/> [cited 12 January 2014]

[47] See: <http://timbusproject.net/> [cited 12 January 2014]

[48] See: http://www.planets-project.eu/ The Open Planets Foundation has been established to provide practical solutions and and expertise in digital preservation, building on the on the research and development outputs of the Planets project. See: http://www.openplanetsfoundation.org/[cited 12 January 2014]

[49] See: http://www.scape-project.eu/ [cited 12 January 2014]

[50] See: http://www.aparsen.eu [cited 12 January 2014]

Compared with the overview of earlier projects presented in Strodl et. al's report, one can state that currently focus has moved to services and service providers (commercial and public ones), standards negotiated and awaiting implementation, and best practices around user communities.

Between the partial exploration of early 2000's projects, and projects such as DASISH and APARSEN lies a struggle for conceptual clarification which cornerstones are summarized in the next section.

3.1.1 Digital preservation: standards, strategies, tools and services - fragmentation

3.1.1.1 OAIS-Reference model

The Open Archival Information System (OAIS) reference model has been developed under the direction of the "Consultative Committee for Space Data Systems" (CCSDS) and adopted as ISO standard 14721[51]. The OAIS reference model establishes a common framework of terms and concepts relevant for the long term archiving of digital data. It is entirely formulated from out the perspective of an archive. This means, that the model details the processes around and inside of the archive, inclusive of the interaction with user. But, it does not make any statements about which data would need to be preserved.

In the reference model, an OAIS is defined as an archive, consisting of an organisation of people and systems that has accepted the responsibility to preserve information and make it available for a "Designated Community". A Designated Community is defined as "an identified group of potential consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities". The OAIS model is widely used as a foundation stone for a wide range of digital preservation initiatives. The model can be considered as a conceptual framework informing the design of system architectures, but it does not ensure consistency or interoperability between implementations.

The OAIS reference model provides:

- fundamental concepts for preservation
- fundamental definitions so people can speak without confusion

Cyberinfrastructure Vision for 21st Century Discovery[52] states that this is "now adopted as the de facto standard for building digital archives". A short summary has been produced by Lavoie[53].

A conformant repository must support the OAIS Information Model and fulfil the following responsibilities:

- Negotiate for and accept appropriate information from information Producers.
- Obtain sufficient control of the information provided to the level needed to ensure Long Term Preservation.
- 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.

[51] Available free from http://public.ccsds.org/publications/archive/650x0m2.pdf
[52] http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf
[53] Introduction to OAIS http://www.dpconline.org/component/docman/doc_download/91-introduction-to-oais-introduction-to-oais

- 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.
- 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.
- 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.

The OAIS Information Model introduces a number of concepts that are fundamental to the understandability and authenticity of a piece of digitally encoded information. The diagram of the Archival Information Package (AIP) shows the various components. These components provide a much finer grained set of terms - much more detailed than simply using the term "metadata".

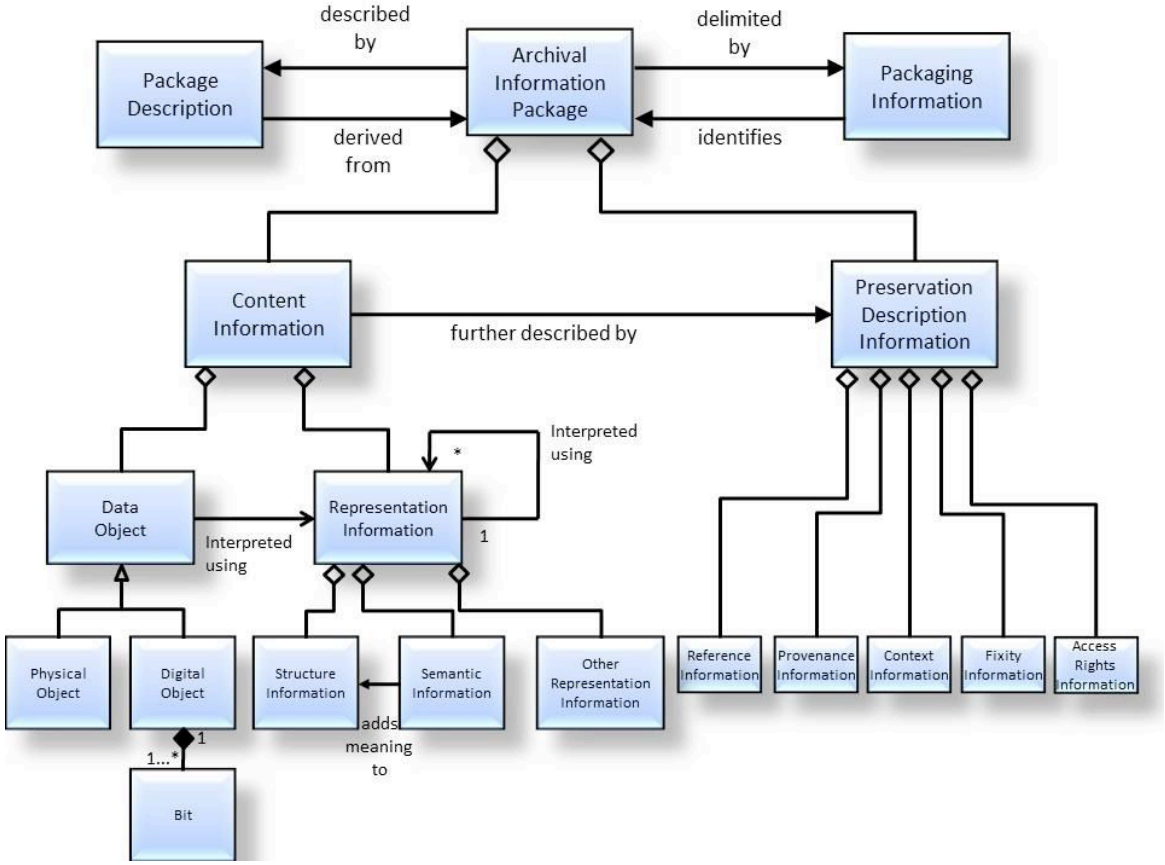


Figure 7 OAIS Archival Information Package

Note that the AIP consists of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an OAIS; it contains all the information needed for the preservation of the digital object of interest.

Mandatory responsibility indicates that even if the repository itself fails, it should have made arrangements to hand over the digital objects, and the AIP construct ensures that the appropriate information has been captured in advance.

Those in the library world often use the mantra "emulate or migrate" but a better mantra would be "add Representation Information or Transform or hand over".

The OAIS model is the conceptual basis against which procedures of certification are set up, which determines if a digital archive can claim to be a so-called Trusted Digital Repository. The key elements hereby are: **Trust, Authentication and Sustainability.**

3.1.2 De-fragmentation of research efforts

The APARSEN[54] project has been investigating various silos of research into different aspects of digital preservation. The general approach for each silo is illustrated Figure 8.

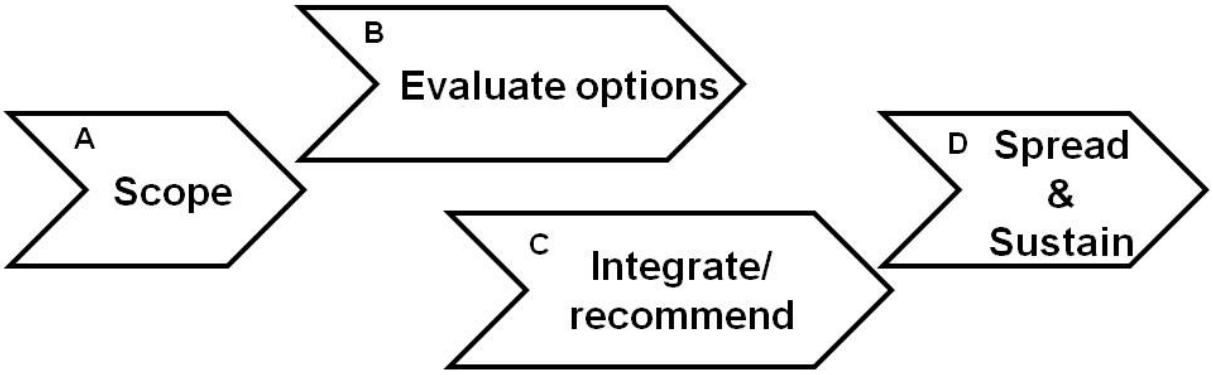


Figure 8 APARSEN approach to silos in DP research

Each area is scoped, the various relevant pieces of research are evaluated, with additional research undertaken for clarification where necessary, then an integrated view, with recommendations, are produced. The results are available of the public website at http://www.alliancepermanentaccess.org/index.php/aparsen/aparsen-deliverables/ and so will not be discussed here. The silos in which APARSEN divided the digital preservation world are shown below, grouped into 4 topics, trust, sustainability, usability and access.

Table 3 APARSEN topics and separate areas

Trust	• Certification of repositories • Reputation and trustability of datasets, publications and people • Authenticity
Sustainability	• Business cases • Preservation services • Cost/benefit analysis • Storage solutions • Scalability
Usability	• Intelligibility • Use by common tools

[54] http://www.aparsen.eu

	• Cross domain usability • Interoperability
Access	• Identification of datasets, publication, people • Rights and responsibilities • Policies and governance

However it is important to note that APARSEN is producing an integrated overall view which embeds digital preservation into the overall business cycle of organisations responsible for securing the future usage of such assets. The current view is shown below.

The aim is to create a unified view which brings together a consistent, coherent view of digital preservation and which forms the basis for the APA's advice, consultancy, services, tools and training.

Figure 9 The common vision representing the digital preservation lifecycle

Figure 9 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 ([35],[36],[37]) with which one may be tempted to compare; these tend to be focussed 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:

- 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)
- Usability
    - 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.

- 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.
- 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
- 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.
    - 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.
    - This 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.

3.2 Digital preservation e-Infrastructure projects

As a natural progression, one would expect (some) research to evolve into usable products or services. It appears that the EC has the same expectation of digital preservation. Thus although the EC has

decided not to prioritise DP for funding research within the first stage of H2020 – although this may be reversed later stages of H2020 – there is funding for preservation in e-Infrastructures.

There is a major project in preservation e-Infrastructures underway namely SCIDIP-ES[55].

SCIDIP-ES builds on the CASPAR research project. CASPAR developed a number of tools and services to support digital preservation for a wide variety (potentially any) types of digitally encoded information. Evidence was collected to support these claims[56]. The approach is firmly based on the OAIS concepts including tools and services to deal with Representation Information and Authenticity, and also address many of the threats identified by PARSE.Insight (see Table 2).

The aim is to help repositories to respond things changes:

Table 4: SCIDIP-ES services and tools to counter changes

Requirement	Action	Icon
One must know something has changed.	A person gives information about a change to the Orchestration service	
Identify the implications of that change	The Orchestration service informs the Gap Identification Service	
Decide on the best course of action for preservation	The data curator uses the Preservation Strategy Toolkit to decide on a course of action.	
One may need extra RepInfo to fill the gaps	The RepInfo may be either already existing (created by someone else) – obtained from a Registry/Repository of RepInfo OR the data curator may create it using the RepInfo Toolkit	

[55] http://www.scidip-es.eu

[56] CASPAR: Validation-Evaluation Report (D4104) http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=CASPAR%3A+Validation-Evaluation+Report+%28D4104%29

		REPINFO TOOLKIT 
Alternatively the curator may decide that the digital object must be transformed, this must be done as a separate activity. If transformed the question arises as to how to maintain data authenticity	The Authenticity toolkit guides the curator in creating adequate evidence.	AUTHENTICITY 
Alternatively: hand it over to another repository	The Orchestration service can be used as a broker to help identify a repository to which to hand over.	
Make sure data continues to be usable	The RepInfo will ensure the digitally encoded information is understandable/ usable by the designated community. In particular the Virtualisation aspects should help to make the information in an automated way, for example in different software tools.	VIRTUALIZATION 

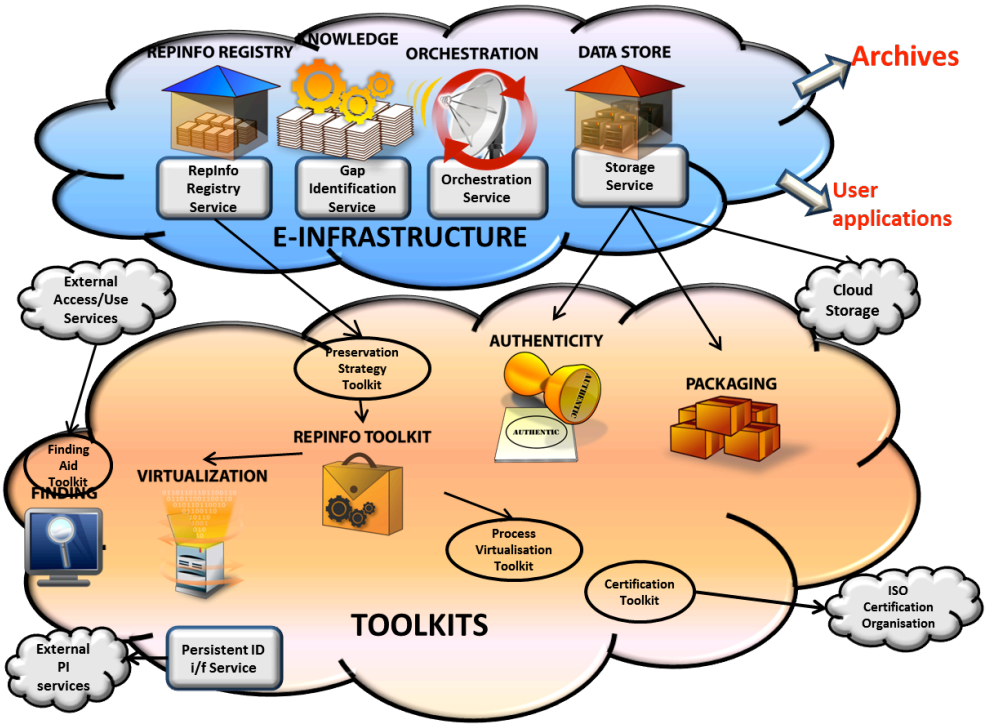


Figure 10 SCIDIP-ES services and toolkits

4 Initial ideas on preserving Linked (Open) Data

4.1 First thoughts from the Linked Open Data perspective

The presence of these two facets of Web data matters for the goal of preserving them. In fact, two preservation strategies can be observed depending on the data at hand:

- Web Data can be preserved just like any web page, especially if there is structured data embedded in it (RDFa, Micro-data, …). It is possible to extract structured data from any Web page that contains annotations in order to expose it to the user via various serialisation formats.
- Database Data can be preserved just like any database. RDF is to be considered as the raw bits of information which are serialised in RDF/XML, Trig, HDT, Turtle or Ntriples files (to name just but a few). The preservation of such files is similar to what would be done for relational databases with the goal of providing data consumers with a serialisation format that can be consumed with current software.

An envisioned Linked Data Archive taking care of the "Web" Web of data faces the same problems as web archiving. Related to the split between the need of de-referenceable or non de-referenceable URIs is what we call the *reference rot* problem, a combination of the well-known link rot problem and the less discussed content decay problem. Link rot is about links that stop functioning, whereas content decay is about the linked content changing over time, possibly to the extent that it stops being representative of the content that was initially referenced[57]. While some specialists claim that the traditional Web never really suffered from 404's (the error users typically get when retrieving a non existing URI), it may be harder for machine agents than for human agents to recover from link rot and content decay[58]. Solving reference rot in the Linked Data case may be feasible by ways of attaching timestamps to different versions of URIs; this would allow historical versions of a resource to be reachable by archived Linked Data browsers. While this is a rather minimalistic solution, one could also imagine a workflow in which (1) the owner of the original LOD namespace have this namespace redirects to the archive; and (2) that the archive accepts to handle the possible traffic.

But there are more challenges when the *semantics* and the *overlap* between these two facets of Linked Data is considered. For example:

- **Semantics**: the archiving of a Web document consists of its own text and other Web resources that are embedded in it. This provides a complete set of resources that can be used to re-create the visual representation of the page. This view differs for a Web of resources where the links between the resources matter and evolve in time. For instance, a Web resource for the city "Paris" may have a link to the concept "Europe", which in turns links to the concept "Eurasia". Whereas Paris has now a conceptual definition that can be considered stable, this is not the case for Europe (which will evolve with changing members) or even Eurasia (which depends on continental drift). A preserved version of "Paris" will have to be preserved with its context in order to remain meaningful in the years to come. On a global graph interconnecting several data sources through shared conceptualization, this context is *infinite*. The only way to preserve the Web of Data in a meaningful way would be to snapshot it entirely, a scenario that is intractable from an architectural point of view.

[57] http://mementoweb.org/missing-link/

[58] http://krr.cs.vu.nl/2013/10/on-the-use-of-http-uris-and-the-archiving-of-linked-data/

- **Overlap** : RDF data dumps are easy to preserve, share, load and consume. These are already largely used on data portals as a way to publish/share/consume Linked Open Data. As long as the URIs in these files are considered not to have any Web existence one willing to grasp the meaning of the resources at the time of preservation will have to load the relevant snapshots dated from the same preservation time. If an archived dataset from 1998 contains references to a resource "Europe", the matching definition as of 1998 will have to be downloaded from an archive and loaded in the same knowledge base. Unfortunately, the Web-link for the resource "Europe" will not be trustable as this concept has evolved over the last 20 years. Furthermore, the matching Web resource may have gone missing by that time.
- **Overlap** : another issue is that two data set preserved from two different time-frames may refer to the same concept "Europe" while implicitly using two different versions of it. The two will point to the same URIs but because of the difference of context at the time of preservation use two different descriptions associated to that very same entity.

Since Web data are in fact not simply rendered web pages, other more basic considerations include the evolution, and perhaps replacement of, RDF itself. Clearly we expect to deal with RDF over the long term because if we expected the encoding to always be kept up to date then that would be the preservation mechanism.

In a related way the meaning of the relationships is encoded in the RDF. At the moment these are mostly relatively simple and well documented but there is in principle no limit to the complexity which may be introduced. The semantics of these relationships would probably be embedded in software used contemporaneously with the data.

The evidence about authenticity of the LD also should be maintained. In a distributed environment this may be an increasingly difficult issue.

The fundamental concepts of digital preservation suggest that we need to be able to (logically) construct Archival Information Packages. These would be needed for the various, for example, RDF files, or the source of the LD, for example the databases and associated software.

Currently, discussion how to best archive Linked Data takes place in the semantic web community, and the blog entry of Wouter Beek of October 7, 2013 documents this.[59] In a response to this blog entry Herbert van de Sompel points to the importance of Memento when it comes to archiving Linked Data. [60]

4.2 First thoughts from the Digital Preservation perspective – applying the concept of a digital object[61]

Compared to tangible objects such as books or archival sources, digital objects are available in a wide range of appearances from simple stand-alone files to specialised software programs. A classification of the digital objects to be managed can help to apply the most suitable method to provide long-term access to them. A classification brings things together and can be based on several principles. Examples are classifications based on the senses used to experience them, classification by medium or classification by subject. The classification and documentation of resources is a significant aspect of their longevity. The importance of the formulation of digital object classification as part of solution to

[59] http://krr.cs.vu.nl/2013/10/on-the-use-of-http-uris-and-the-archiving-of-linked-data/

[60] http://krr.cs.vu.nl/2013/10/on-the-use-of-http-uris-and-the-archiving-of-linked-data/#comment-4317

[61] Classification provided by David Giaretta

provide long term access to them is very well illustrated by the quote "What is difficult to identify, is difficult to describe and therefore difficult to organize" (Svevonius, 2000, p 13)".

The set of metadata elements that are drawn from a number of metadata schemas combined and optimised for a particular local application is called an 'application profile'. By definition, an application profile cannot introduce new metadata elements. Each metadata element has to come from an existing metadata schema. Thus, application profiles reuse existing metadata elements. The difference between a metadata schema and an application profile is that a metadata schema only declares metadata elements whereas an application profile reuses existing metadata elements. The Dublin Core Metadata Element Set (DCMES[62]) [ISO15836:2003] is an example of a metadata schema. (Horik, 2005, p 69). Based on classification schemes a number of application profiles are developed that can be used to document objects. Obviously the Dublin Core Data Element Set plays an important role in this. DCMES, consisting of a set of 15 data elements aimed at "resource discovery", has a huge user community and facilitates interoperability. Local interpretations of DCMES (called "qualified DCMES") and application profiles using DCMES elements are developed to document a wide range of objects for a wide range of purposes. Also for long-term access.

The perceived value and importance of digital objects is not fixed over time and within an interest group. Not all digital objects are valuable enough to justify the efforts to guarantee its long-term accessibility. A number of selection methods are developed, such as the "Decision tree for selection of digital materials for long term retention" (http://www.dpconline.org/advice/preservationhandbook/decision-tree). The selection of digital objects to be preserved should be addressed by the policy of the organisation that takes responsibility to provide long-term access to the digital objects.

The features and characteristics of the digital objects as well as its (future) value and importance determine which standards, strategies, tools and services should be applied. The knowledge base and expertise to provide durable access to digital objects is a moving target, but in the course of time consensus is reached on a number of principles relevant for digital preservation. Central to this is the **OAIS reference model.**

There are different possible classifications of digital objects. The APARSEN project (see also Giaretta 2011, pp 31-39) proposed[63] the following partial classifications. The purpose of this has been to provide a **partial** view of the variety of types of digital objects which exist "in the wild" and which one might be required to preserve. The reason has been to ensure that one can at least recognise the possibilities when confronted with the challenge of preserving a digital object.

- Dynamic vs static
    - Dynamic: The basic idea is that the various changes are important and there is a desire to make queries about such changes. If the information changes but one is not interested in the older versions then either we are not in the domain of digital preservation. Alternatively, in terms of the model in Figure 9 one is choosing to preserve something more valuable until, at some future time, may lose its value when the next version comes along. In this latter case the Provenance should reflect the change but might not give exact details of the changes/additions.
    - Static: the information is unchanging

[62] See: <http://www.dublincore.org>

[63] The description of these is available at http://www.alliancepermanentaccess.org/index.php/knowledge-base/tools/tools-for-preservation/rough-classification-of-digital-objects/

- simple vs complex
    - simple: normally are treated as a whole – for example an image – or
    - complex: normally treated as a collection of simpler parts,
- non-rendered vs rendered
    - Rendered: digital objects which are usually processed by some software to produce a rendering which is presented to a human user who can then interpret what he/she sees/hears/feels/tastes. This can include simple documents, pictures, videos and sounds. These we will refer to as Rendered Digital Objects because is in 100 years' time as long as one could display/print/render the digital object then a reasonable person would agree that a good job had been done in terms of preservation..
    - Non-rendered: digital object for which it is not enough to simply render it but for which one needs to know what the contents mean in order to be able to further process it Examples include scientific data, where just being able to display/print the numbers is not normally regarded as useful. For example one would not think it adequate simple to be able to print out in 100 years, say, 10 PB of data produced by the LHC – one must be able to perform further computations with it. The end result ***may*** be displayed for a human to view e.g. as an image, but normally there will be many views, all of which may be useful but none of which would be regarded as adequate for further processing.
- passive vs active
    - Passive: something which is used by other applications (software) to do something. For example a document file is used by a word processing programme to print the document or display it on the screen.
    - Active: an object which does something. For example the word processing application or the astronomical analysis software mentioned in the previous paragraph might be the digital objects to be preserved.

Of course a particular digital object may be regarded by different repositories in different ways in terms of preservation – in other words different repositories may have different preservation objectives. One repository may actually want to print LHC data as an artistic installation, and be able to do this in the future; another, scientific, repository may want to regard LHC data as something to be further processed. When we think about LHC data, or any digital object, we could think about a digital object from all these above points of view but we should make sure we at least cover the most likely preservation objective. In other words we should be sure that we do not limit ourselves to thinking of Linked Data as being rendered when considering preservation.

Applied to Linked Data as a new data model, one might, as a first attempt, say that linked data encoded in some way, for example as an RDF triple, is dynamic/complex/non-rendered/passive. Based on this classification a number of questions can be raised:

- Dynamic i.e. changes over time:
    - Do people need archived version of LOD datasets or are the most up to date version only what is needed?
    - Different statements may be made at any time and so the "boundary" of the object one is considering changes in time.

- Complex
    - LD is all about expressing statements whose truth or falsity is very much grounded to the context provided by all the other statements available at that particular moment. The preservation community is, according to David Giaretta, "concerned with preserving what has been expressed - whether true or not". It is important to note that the notion "truth" is here applied with two different meanings: the truth of the content to be preserved (LD) versus the truth of the preservation, the latter meaning to ensure the authenticity of the object independent from its content.
- Non-rendered
    - Non-rendered digital objects need to be processed to produce any number of possible outputs. What is done with LD is very varied. This gives rise to a great number of possible preservation objectives, compared with rendering an image. For example one piece of LD may be combined with several others to produce a new result which may be presented to a human in many different ways, or may be used by other applications. At any point new statements by be made in new pieces of LD and so new inferences may be created. It is of course these possible inferences rather than the display of a particular encoding (N3, RDF etc) which is of interest.
- Passive
    - The linked data is usually in the form of statements or objects which are not applications. The statements are normally themselves operated on by applications and, right now, applications are not perfect nor indefinitely scalable. Plus, LD is supposed to be about raw data. Rendering matters less for preservation then. Unless one wants to preserve the applications, but that's quite a different story-one that is orthogonal to L(O)D.

Other questions include:

The LD is normally distributed and the persistence the "object" depends on all the individual parts. LOD is based on the Web and as such suffers from 404 errors. But their effect is stronger for data than it is for documents because of the linkages between them. The persistence of the identifiers, including domain names, as well as the actual files or databases are both concerns here.

Authenticity is a major issue in preservation. There are beginning to be ways of dealing with the provenance of digitally encoded information over time. Can these techniques simply be applied individually to the various parts of each triple?

4.3 Technological challenges around L(O)D as specific digital objects[64]

Linked Data are a form of **formal knowledge**.

> As for any kind of data or information, the problem for long-term preservation is not the preservation of an object as such, but the preservation of the meaning of the object. Think in terms of archiving longitudinal time series of measurements of temperature. The essence of here is not to preserve the mere numerical values. They are useless without measurement units, location and time information and preferably the circumstances of the recording. To the same extent as the meaning of the numerical value is not automatically attached to its symbol (an

[64] Information provided by Carlo Meghni

integer or a rational number), a URI which is basically a string of symbols does not carry the semantics it has in principle when embedded in a larger data graph and when living there.

The OAIS model differentiates between the data object and its representation information. Sometimes also shorten to the question of data and metadata - a discourse that can easily fill books. In essence we ask for drawing boundaries between the object itself, its description, and its meaning. From the projects reviewed in the Strobl et al. report (2001), CASPAR is most near to those questions[65] while SCIDIP-ES puts the CASPAR research into production.

Linked Data depend on the **web infrastructure**, and in particular on the dereferenciation of HTTP URIs.

As discussed in section 2.1 all projects addressing link rot and content rot are relevant. But not all of the discussed solutions target long-term preservation. From the perspective of an archive Persistent identifiers are a key and the interoperability between different identifiers. This is currently addressed in WP22 in APARSEN ("Identifiers and Citability"). But, despite of efforts to define good practices when URI are given and to develop web archiving strategies to counter loss of information, projects as **Hiberlink** show the urgency of the problem.[66]

Linked Data are **distributed** in nature, since it is not only possible, but indeed strongly recommended that Linked Data datasets reference each other.

Referencing has to do with (persistent) identifiers, again. But, when we talk about preservation of distributed information it is important to determine the boundaries of the object to be preserved. With this aspect we enter the field of the object format definition (and format registries).

Linked Data are **accessible in many ways**: through SPARQL end-points, as RDF dumps, as RDFa, as microdata and others.

Linked Data descriptions are modelled using RDF and can be serialised using different formats (RDF/XML, N3, Turtle, JSON-LD, and others). For each form its durability can be assessed. But, more important here is the dichotomy between a data-base representation (the "data base web of data") and the living on the web representation ("web-web of data") (section 2.3.3)

In order to cope with change, Linked Data datasets and vocabulary should be **versioned**, and any reference to a versioned dataset should also mention a specific version.

The existence of versions for formal knowledge is nothing new in scholarship. Books for instance appear in different editions, and going even further back, middle-age collections of sheet music provide a good example for objects with very changeable boundaries. Different sheets of music can appear in different editions of music.[67] Currently, vocabulary has been developed to address the problem of versioning which in the language of L(O)D is covered by the concern to have good provenance (Groth &Frew, 2012). It does not take away that part of the L(O)D cloud might miss provenance in space (versions) and time. This leads to an increase

[65] http://www.casparpreserves.eu/publications/deliverables.html The work continued in the APARSEN project: WP 25 "interoperability and intelligibility". The main person in this field is Yannis Tzitizikas of FORTH. See also: http://ercim-news.ercim.eu/en80/special/knowledge-management-for-digital-preservation

[66] http://edina.ac.uk/projects/hiberlink_summary.html

[67] Personal communication of Marnix van Berchum; see also http://www.cmme.org/

of ambiguity. One example is the publishing of UDC[68] numbers (a decimal classification system for concepts - applied by libraries) with reference to the edition of the UDC. The UDC consortium cares for a classification scheme for helping library catalogue their content. Being a reflection of the world view, this classification evolves over the years and the UDC produces new releases of its master reference file (MRF) every year. Library using this MRF currently overwrite all the data from the previous years with the data from the new year (or part of it). This leads to inconsistencies that would benefit from using Linked Data technologies to trace the changes in the MRF and eventually refer to specific releases for particular classes found in it. On the other side if library catalogues are expressed in LD formats publishing a UDC number without reference to time and edition can be quite misleading.[69] While LOD strives to reduce ambiguity, this is an example for amplification of ambiguity.

Preservation requires the expression and recording of several kinds of metadata about the preserved object. For preserving Linked Data such **metadata should be associated with triples**, and at the moment there is no obvious way (apart from reification) to express metadata about RDF triples[70].

This aspect has a strong overlap with the provenance issue. (see above)

4.4 Implementation of DP principles for preserving LOD

Selection: Which LOD data should actively be preserved? (See the example of the Dutch Historic Census below) Who is responsible for "community" data, such as DBpedia? Increasingly government data is made available as LOD. Agencies publishing the data should be aware of their role to create durable, trustworthy, authentic LOD. The business process described with Figure 9 should be able to help with this.

Creation of the Archival Information Package: There are several aspects to this.

Representation Information (RepInfo):

Structural RepInfo: including the XML and RDF standards, as well as the definition of Unicode. Which formats can we distinguish? RDF, Triple Store, Software, SPARQL, etc. What about Triple Stores? Also use of Persistent Identifiers contributes to durability of LOD

Semantic RepInfo: the semantics, as discussed above, as well as the basic semantics of RDF.

Other RepInfo: ranging from the software in which the semantics may be embedded to the de-referencing software used in the web.

The Registry of RepInfo which SCIDIP-ES has implemented, described in section 3, should provide the ability to share RepInfo which has been created/collected. In fact the basic standards such as RDF, XML, Unicode etc. should also already the in that Registry. Claims that RDF is a "self-describing" format imply that no external associations are needed – this in general is not something that can be depended on since, at the very least the semantics is external. The key issue is to associate the RepInfo with, for example, the RDF. The AIP provides a way to do this.

[68] UDC stands for Universal Decimal Classification. Developed by Paul Otlet, in this Knowledge Organization System concepts can be expressed by complex strings of symbols. (Smiraglia et al., 2013)

[69] Personal communication with Aida Slavic, editor in chief of the UDC.

[70] "You can use quads. Then at least one can add metadata to the (named) graph. Sometimes there is talk about having an identifier/URI per triple, maybe such a quintuple approach would be better suited for preservation." Comment of Menzo Windhouwer.

Authenticity: the evidence needed to support claims.

Authenticity tools are be produced by SCIDIP-ES – supplementing whatever the host system of the RDF has

Packaging: The overall association of the AIP components together is referred to in OASI as packaging. SCIDIP-ES uses a variety of techniques including OAIS-ORE, SAFE and XFDU as packaging techniques.

In addition as things change SCIDIP-ES (see Table 4) provides a number of tools and services that should be applicable to the LD world.

Additionally

Rights / ownership / licenses. LOD are by definition open, but how to preserve privacy than (see for a discussion of these issues reports on the PILOD project, a Dutch project to create LOD from governmental data (Gueret, 2013) Which licenses to use, which Creative Commons code? APARSEN has investigated preservation of DRM.

Storage. Highest quality is storage in "Trusted Digital Repository". But which other models can be used: one example is providing multiple copies/mirrors (CLOCKS). APARSEN has a number of studies on Storage and on Scalability of solutions.

4.5 Summary

In section 3 we examined DP projects for their relevance for LD preservation. In this section 4 we tried to identify those features of LD that at first glance present a challenge to long-term digital preservation. We counter the requirements from the LD perspective with what epistemic frameworks DP might offer to solve this. What is striking at first glance is the variety of typologies, dimensions, and dichotomy used to build a valid reference framework to approach the problems. Figure 11 summarizes the schemas. Clockwise we start with the dimensions along which projects are allocated in the Strodl et al. report; the challenges from the perspective of digital objects in general; the challenges emerging from the nature of LD; and the lessons to be learned from current DP practices. Strikingly, there is little correspondence between those schemes. Some of the overlap is captured by using similar pictograms and colour codes. But, even more striking: communities - be it in the role of data producers, data consumers, or data providers seem not to be the main concern of our discourse in PRELIDA so far. We will inspect specific use cases in the following section, and see how there the community and institution aspect come to the foreground again.

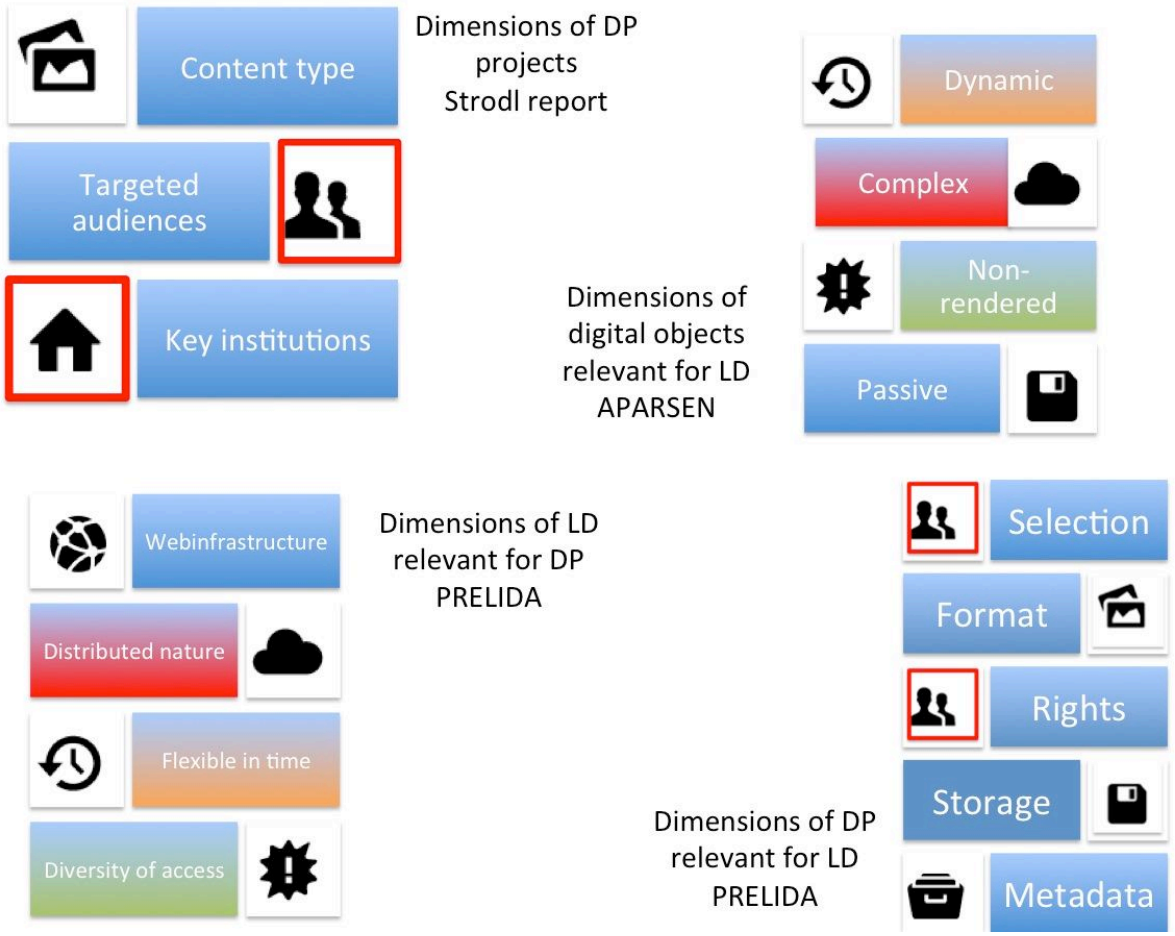


Figure 11 Different dimensions used in DP (Entypo pictograms by Daniel Bruce — www.entypo.com)

5 Use cases

This section presents three important use cases from projects or organizations thoroughly involved in the creation or archiving of Linked Data and therefore highly interested in informing PRELIDA. The use cases are presented in a discursive, informal way. It is expected that they will take a more software development orientation in the consolidated version of this report.

5.1 CEDAR - From research explorations to archiving services - the case of the Dutch Historic Census Collection

5.1.1 Description of the project

CEDAR is a project in the Computational Humanities programme of the KNAW, and the abbreviation stands for "Census data open linked – From fragment to fabric – Dutch census data in a web of global cultural and historic information.[71]" The projects runs from 2010 to 2015, and two PhD students and one postdoc are the appointed staff on it. The PhD supervisors come from social history for one PhD, and from semantic web for the other.

5.1.2 Context of the project[72]

CEDAR is an important project for DANS. As an archive, CEDAR offers possibilities to check data quality; represents an experiment with Linked Data and Linked Open Data and semantic web technologies; and is suppose to help the archive to improve the access and reuse of data. To judge the importance of this specific project, one can best describe CEDAR as one step in a sequence of projects around the digitization of census material in the Netherlands, and the creation of user interfaces to it.

At the beginning of the census project stand book publications. This is how the micro-level information collected from visiting houses can be recorded and preserved long-term. In the census data set CEDAR deals with - short labelled as Historic census - information has been aggregated from 17 instances of measurement (meaning collecting the census information) - 1795, 1830, 1840, 1849, 1859, 1869, 1879, 1889, 1899, 1909, 1919, 1920, 1930, 1947, 1956, 1960, 1971. They are represented in tables, which layout changes with the different information collected. Also the tables from different years are not always been published in books sequentially. Some books contain information from different census years.

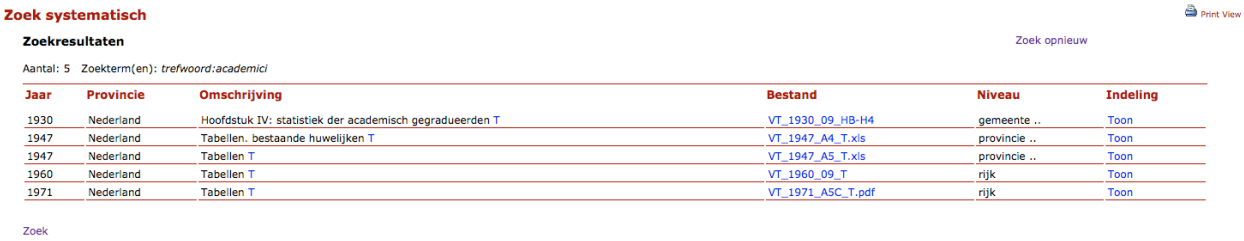
To enable access to the book publications for the wider public, the books have been scanned[73]. Later, in another project the tables have been digitized by manual data entry - OCR proved to be not feasible. Both tables and images have been published on-line using a Content-Management Systems. Provided at the website www.volkstellingen.nl this web resource was quite popular. The CMS also contains an information retrieval part and allows to search over indexed resources in a quite sophistic way. For instance, a so-called systematic search for a keyword such as "academici" retrieves all five sources which contains information on this occupational category.

---

[71] http://www.cedar-project.nl/

[72] In the final version all use cases need to be aligned according to the content and need to be re-edited to focus on the preservation aspect only. For CEDAR we document a lot of information not relevant for preservation.

[73] A description of the digitization process can be found in (in Dutch), Doorn, P.K. and van Maarseveen, J.G.S.J., Twee eeuwen volkstellingen gedigitaliseerd. In: *Twee eeuwen Nederland geteld. Onderzoek met de digitale Volks-, Beroeps- en Woningtellingen* 1795-2001. Den Haag (2007)



Zoek systematisch

Print View

Zoekresultaten

Zoek opnieuw

Aantal: 5 Zoekterm(en): *trefwoord:academici*

Jaar	Provincie	Omschrijving	Bestand	Niveau	Indeling
1930	Nederland	Hoofdstuk IV: statistiek der academisch gegradueerden T	VT_1930_09_HB-H4	gemeente ..	Toon
1947	Nederland	Tabellen. bestaande huwelijken T	VT_1947_A4_T.xls	provincie ..	Toon
1947	Nederland	Tabellen T	VT_1947_A5_T.xls	provincie ..	Toon
1960	Nederland	Tabellen T	VT_1960_09_T	rijk	Toon
1971	Nederland	Tabellen T	VT_1971_A5C_T.pdf	rijk	Toon

Zoek

The original CMS would even have allowed a full text search. But, this functionality has been deactivated due to the aging of the underlying architecture and related security issues. A problem of normal digital preservation, so to say. However, from the point of view of an archive, the website volkstellingen.nl represents an access point to the resources - the table and images. They themselves - the source material - are part of a Trusted Digital Repository, EASY, for which long-term preservation strategy is in place[74].

The CEDAR project has been set up to fill the gap between digital preservation and optimal access as near as possible to the original source material (on paper), a service required by the community of experts from social history. The standard of data representation in their field is a database, which can be queried. A problem emerges in the transition from the raw data to a database representation, because any database representation requires harmonization, meaning mapping between different expressions of variables over time. One easy to be grasped case is the census on occupations. In each decade we find different occupations and different named occupations. Some of them can be easily matched, some of them are new - because new industrial sectors emerge, some of them vanish, and others again merge or split. But, also other information is not easy to be matched. How age is recorded varies, the same hold for the household situations. Other kind of questions, reflecting other views of what a society defines as being worth to be recorded, leads to variables that cannot be automatically related to each other. Usually harmonization is part of the social history research process and a certain harmonization is authorized by an author (or a group of them). To create a specific harmonization is not the task of an archive. To store a created one, in contrast, is a task. The decision to go for an RDF data representation from the side of the archive was, that this new data model allows to keep the authenticity of the original information.

Another aspect that pops up whenever the data in the census are inspected is to clean up the data. Sometimes the original data contain errors, some of them can be logically detected, e.g., if the total number of inhabitants in an area does not match the sum of inhabitants in all subareas. DANS has set up an ongoing curation process concerning the digitized files. Excel tables are checked against the original printed statistics, obvious errors are corrected, annotations are added. This process is ongoing. Equally on-going is to add census data from later years. Both curation processes are handled according to standards for TDR: the original sources (images) are maintained, corrected digital expressions (EXCEL tables) are recorded as versions of a data set. All versions have its own individual persistent identifier.

In order to be able enhance the quality of the data by using LOD principles, CEDAR has decided to work from a copy of the available excel files the deposited datasets in the version of October 2010. In order to have a clear overview of the contents of the dataset, a script was written, TabExtractor. TabExtractor offers a summary of a collection of Excel spreadsheets at the data and metadata levels.

[74] The digitized census tables are stored in the EASY archive repository and can be found at: https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:44159 [cited 20 January 2014].

The CEDAR dump entails 2288 tables with 33283 annotations in 507 Excel files. (Ashkpour, Moreno Penuela, 2013)

5.1.3 Arguments to use a LD or LOD data representation

- The "raw data" structure can be preserved, all changes to it are additions to the original datagraph, which provenance can be recorded.
- Extracted information on entities (geographical locations, occupations, time, household situation) can be enriched linking it to other semantic referenceable resources. For instance, one can imagine that when searching for "accademici" one get's additional information about this specific occupation in a certain time, such as typical images from other collections. This way the context of the dataset can be enriched (semi)automatically.
- Alternative harmonization schemes can be developed and applied next to each other, allowing to judge the resulting variance in terms of numerical values. The latter one can be seen as an error margin on the data value in longitudinal studies resulting from ambiguity in the interpretation of the data.
- Visualization of the data, e.g. by using GIS systems.

5.1.4 Problems addressed in CEDAR

- How to move from Excel expressions of tables to RDF? This problem has been solved by a combination of TABLINKER and manual styling of the excel tables.
- How to support with the RDF expression the/an harmonization process? CEDAR is currently working on this problem, and one element relevant for the preservation discussion is the use of authorized vocabularies. As much as possible, the interpretation of variables will be coupled to already existing semantic descriptions of the variables. The definition of own vocabulary will be restricted. This Dutch Census Specific vocabulary will be published with trustworthy parties.
- How to trace provenance for any additions/changes in the RDF graph?
- How to design GUI's which allow a database-like querying of the new data representation specific enough for experts; and intuitive and broad enough for the interested lay audience?
- How setting up a workflow of enriching the data model which can be implemented and used by third parties?
- How and where to preserve the results of the project?

5.1.5 Problems concerning preservation resulting from the LOD

- How to (re-)import the new data representation into the current archival system? How to ensure that a link is kept to the original sources? This seems to be the easiest problem, the RDF graph, any new RDF graph can be handled as a new version to a table or a completely new dataset. EASY can ingest these RDF files. Depending on the structure, these will result in one complete dataset, a dataset per RDF-file, or somewhere in between (a dataset per census year). EASY will currently assign a persistent identifier (URN:NBN) and metadata to each dataset. It is under investigation how DataCite DOI identifiers can be assigned per dataset, and how individual files can also be identified persistently. EASY as an archive is tailored towards

preserving the data. What is the impact on an OAIS-like system as EASY if it should also preserve the service to resolve requests for parts of these data?

- In order to enrich the dataset we would like to use vocabulary from other parties. They exist as web resources, and end up as URIs in the to be archived RDF graph(s). How, to ensure the stability of those URIs?

- If the Dutch census data are published as LOD, they in principle can be referenced to and re-used in other data models and data stores. How do we keep the boundaries around the original object, authorized changes and additions (authorized by whom, experts and/or archivists), and experimental use?

- Which metadata we need to use for an RDF graph? Need we transfer all provenance properties into a Dublin core format, as URIs, in another form?

In summary our questions resemble the issues Wouter Beek raised in his blog entry (Beek, 2013) to which we pointed in 4.1 already.

5.2 DBpedia use case

5.2.1 Description of DBpedia

DBpedia objective is to extract structured knowledge from Wikipedia and make it freely available on the Web using Semantic Web and Linked Data technologies. Specifically data are extracted in RDF format and they can be retrieved directly, through a SPARQL end-point or as Web pages. Knowledge from different language editions of Wikipedia is extracted along with links to other Linked Open Data datasets. The archiving mechanism of DBpedia is presented in the following.

5.2.2 DBpedia archiving

DBpedia archiving is currently handled by DBpedia and not by an external organisation. Since DBpedia data are extracted from Wikipedia data and are transformed in RDF format these two organisations are closely cooperating for the dataset creation in the first place and the ability of the dataset to evolve, besides the archiving. Wikipedia content is available using Creative Commons Attribution-Sharealike 3.0 Unported License (CC-BY-SA) and the GNU Free Documentation Licence (GFDL). DBpedia content (ontology as metadata and data)[75] is available to end users under the same terms and licences as the Wikipedia content.

[75] As the following exchange of comments shows is the definition of what is count as data and what as metadata is still debated among different experts. [C1 AI] “Isn't it (meta)data, not content?” [C2 SB] “Both DBpedia ontology (metadata) and data are available to users.” [C3 AI] “Quite confusing still. For me (and I believe many people in the preservation community) metadata is data about something, structured according to a schema/ontology. So an ontology doesn't really count as metadata.” [C4 SB They also preserve metadata (that was a recent addition) for example extraction date from Wikipedia, see for example: http://live.dbpedia.org/page/Berlin]

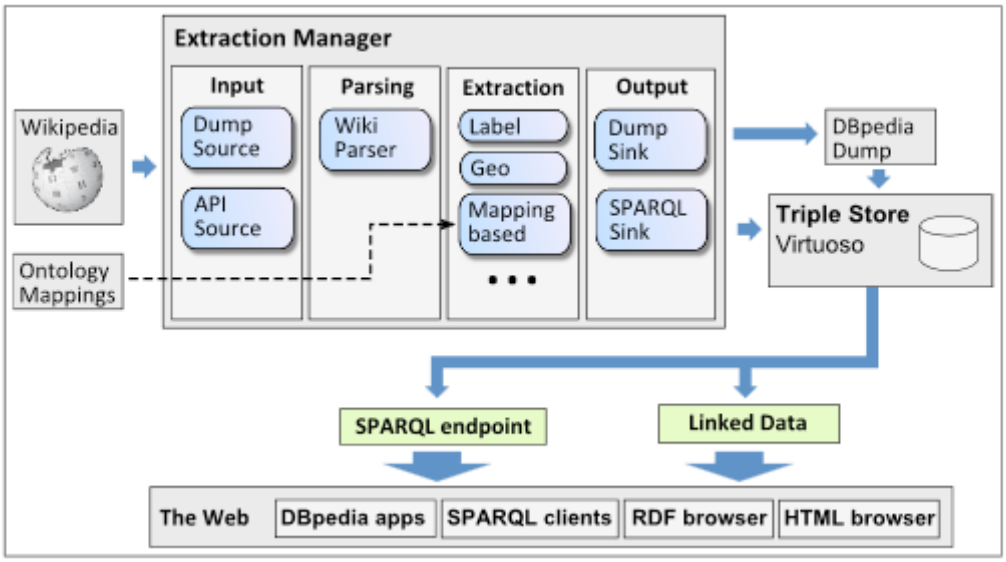


Figure 12 DBpedia extraction mechanism[76]

DBpedia preserves different versions of the entire dataset by means of DBpedia RDF (or CSV) dumps corresponding to a versioning mechanism[77]. Besides the archived versions of DBpedia, DBpedia live[78] keeps track of changes in Wikipedia and extracts newly changed information from Wikipedia infoboxes and text, into RDF format. DBpedia live contains also metadata about the part of Wikipedia text that the information was extracted, the user that has created or modified corresponding data and the date of creation or last modification. Incremental modifications of DBpedia live are also archived[79].

DBpedia dataset contains links to other Linked Open Data datasets containing definitions and information (e.g., Geonames). There are currently (February 2014) more than 27 million links from DBpedia to other datasets. DBpedia archiving mechanism is used for the preservation of links to these datasets but not their content. Preserved data are DBpedia content in RDF or tables (CSV) format. Rendering and querying software are not part of the archive although extraction software from Wikipedia infoboxes and text used for the creation of DBpedia dataset is preserved.

5.2.3 DBpedia archiving problems

The above description indicates that currently DBpedia preservation stakeholders are the DBpedia organization and end users seeking access to older versions of the DBpedia dataset either in RDF format or as a Web page or through a SPARQL endpoint. Cooperating organizations such as Wikipedia and linked datasets creators provide data and access for the creation of the DBpedia dataset but they are not involved in the archiving process.

[76] See: http://svn.aksw.org/papers/2013/SWJ_DBpedia/public.pdf
[77] See for example: http://downloads.dbpedia.org/3.9/en/
[78] See http://live.dbpedia.org/
[79] See for example: http://live.dbpedia.org/changesets/2014/

Currently supported data formats are RDF and CSV. Adopting open standards such as RDF and following W3C specifications reduces the risk of not been able to reproduce the data in the future upon request. This argument also applies for the Web rendering and SPARQL endpoint functionality. On the other hand, since corresponding software and hardware platforms required for preserving SPARQL-endpoint and Web rendering functionality are not part of the preservation mechanism this risk is not eliminated.

Summarizing, using the DBpedia archive users can retrieve valid versions of data for specific time points in the past but rendering and SPARQL end-point functionality are not directly preserved and supported. Also answering complex requests about the evolution of specific data over a temporal interval are not directly supported. Specifically a version of DBpedia for a specific time point can be retrieved, but a more complex query requesting all valid versions of data during a temporal interval and the modifications that have happened during the interval is a functionality that is not yet supported.

5.3 Europeana

5.3.1 Description of the project

Europeana.eu is a platform for providing access to digitized cultural heritage objects from Europe's museums, libraries and archives. It currently provides access to over 30M such objects.

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. Europeana re-publishes this data openly (CC0), mainly by means of an API usable by everyone.

5.3.2 Basic Europeana sources

The main source of data for Europeana are its cultural data providers—museums, libraries, archives, mostly. These are often taking great care of their data, including metadata and digital content, with appropriate preservation policies. However for most of them the metadata is sent as batches in a discrete way, with infrequent updates. 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. This is however rather a traditional issue of preserving access to web pages, not one of linked data preservation.

5.3.3 Dependence on third-parties linked datasets

Cultural Heritage providers are not Europeana's only source of data, however. To compensate for certain quality lacks in the providers' data, especially considering multilingualism or semantic linking, Europeana has embarked on enriching this data. This is mostly done by trying to connect the cultural objects in Europeana with a small set of "important" (especially, large, semantically structured and multilingual) reference linked datasets. At the time of writing, Europeana connects to GEMET, Geonames and DBpedia. Once the links to contextual resources (places, persons) from these datasets, have been created, the data on these resources is added to Europeana's own database, to later be exploited to provider better services. This introduces a dependency towards external linked datasets, which Europeana has to take into account. 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. Yet the use of Europeana data outside of Europeana itself could be impacted, if the published links that are 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.

5.3.4 On the way to more linked data dependencies

As the experiments on re-using third-party linked data proved quite successful, Europeana started to encourage its providers to proceed with some linking by themselves. Since they know the data better, they are in better position to come with the best data enrichment processes. At the same time, Europeana was updating its data model to include a richer set of construct, enabling the provision by providers of local authority files, thesauri and other knowledge organization systems.

The conjunction of both efforts has already led to some projects sending data that includes:

- links to the same external linked data sources, that Europeana already uses for its own enrichment;
- links to projects' and institutions' own thesauri, classification expressed themselves as linked data.

Two illustrative projects are CARARE and MIMO.

In a first phase, Europeana has encouraged such providers to send data on the new contextual linked data resources embedded in their "traditional" metadata. It is now starting to harvest this linked data on the web, using the standard linked data de-referencing techniques, on the condition that this linked data is made available using the vocabularies recommended by the Europeana data model, such as SKOS .

Of course this can have drastic consequence regarding our own requirements on preservation of such datasets. The entire cultural sector would then become more sensitive to some reference datasets becoming unavailable, be they references to one institution, a group thereof or an entire sector (e.g., libraries).

5.3.5 Europeana as data publisher

As said, Europeana re-distributes the metadata it aggregates from its partners, in a fully open way. This is done via its API, mainly. But there have been experiments using semantic mark-up on object pages (RDFa, notably with the schema.org vocabulary) and in the form of "real" linked data , either by http content negotiation or in the form of RDF dumps.

However, the data that Europeana gathers changes. This implies some level of link rot. Europeana generates its internal identifiers from the identifiers sent by its providers, which are not always persistent. When there are updates, this can result in an object being provided a new identifier, and eventually a new HTML page and (linked data) URI, while the old identifiers die. We try to address

issues by implementing redirection mechanisms between old and new identifiers. And convince our providers to send us more stable identifiers to start with, which is relatively well-engaged, as the need of persistent identifiers is being accepted in more circles besides Europeana.

There is also (less dramatic) content decay, as the metadata statements sent by providers, or Europeana's own enrichments, change. Currently there is no versioning at all in the data that Europeana re-published. We hope to make progress soon, 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. This will however constitute only a first step, as this will only reflect changes in the data as harvested by Europeana, not reflecting the more granular updates that could happen on the providers' side (e.g. when a specific library updates a record in its catalogue).

One must note however, that Europeana has no mandate to preserve its providers' data, who often have their own policies in place. This will raise issues if one day Europeana has to provide preservation-level information to its own consumers, which should reflect the preservation-level information of its providers. Europeana should aim at being as transparent as possible, yet a new layer should be added, to reflect that the data made available by Europeana is more than the basic sum of what has been directly provided by providers: it's been massaged to a common data model, while some values were normalized and enriched.

6 Conclusions

This report concerns issues related to the long term preservation of linked (open) data. It brings together the research results of two communities, working respectively on solutions to curate digital objects and on solutions to create a semantic web of linked data objects.

The main approach in the digital preservation community is to document fixed digital objects and store them in a Trusted Digital Repository, a repository that meets specific requirements based on standardized audit and certification procedures. The OAIS reference model is an important standard that provides fundamental concepts for digital preservation activities. It also provides definitions so people can speak without confusion. The research activities in the digital preservation community can be summarized as working towards testable and provable approaches to guarantee that digital objects are usable for a designated community in the future. For this a number of tools and services are developed and are part of the developing e-Infrastructures. With the emergence of L(O)D new problems emerge for the DP community. As described above the problems are related to the specific features of the LOD data model, and possible new communities getting involved into DP, namely those who publish and (re-)use LOD.

The linked data paradigm concerns the technology to publish, share and connect data on the web. This web of data is created with the help of a number of standards and protocols, such as RDF, triple stores and SPARQL endpoints. The linked open data paradigm currently is rapidly gaining ground as it offers a great potential for building innovative products and services by creating new value from existing data. The dynamic character of linked open data objects and the absence of a central administration to manage the objects are the main factors that threaten the long-term availability and usability. Prior to any service beyond the research cycle, the volatility of the data (model) has implication for the expert community in LOD itself, and for any inner-academic use. What is the impact on scientific integrity when researchers base their conclusions on drifting concepts, on data that disappears, or data that isn't owned by one owner? Which measures the community itself has develop so far to ensure scientific integrity on which the current on-going exploration of LOD is based?

The linked data paradigm emerged recently and we now can observe a growing attention for digital preservation solutions to guarantee long term access to this type of data. What can both communities learn from each other? This report describes the state of art in general terms and provides some directions towards the creation of solutions to prevent that linked open data objects get lost. The information in the report will be updated in order to arrive at some concrete solutions and approaches towards the end of the PRELIDA project. Examples of projects in which the linked data paradigm is put into practice, such as the CEDAR project, deliver important use case information that can be used to find out how and to what extent approaches from the digital preservation community can be used to curate the data.

In order to provide solutions for the long term preservation of linked data we have identified, in section 3, a number of technical solutions, digital preservation tools and services, which should be directly applicable to LD. In section 4 we laid out different high-level classifications and frameworks relevant for archiving LD. More detailed investigations are needed about the practicalities involved.

The following three, non-technical, issues: **version, fixity and responsibility,** merit special discussion. In each of those aspects we find technological questions yet not solved. But the main lesson to be learned from Digital Preservation is that an important aspect of Digital Preservation are social interactions which lead to norms, best practices, and standards followed by communities and implemented in institutions.

Versioning concerns the temporal aspect of linked data that requires attention as in the course of time data is enhanced, adjusted and deleted. How to preserve these changes and how to keep track of different versions of a data object - is a technical aspect? But, at which frequency versions should be archived; how they should be described for re-use is a question only to be solved by the involved communities. The second issue concerns the actual characteristics of linked open data objects and the selection and implementation of dedicated tools and services to preserve these **fixed objects**. By definition linked data objects are related with each other raising issues concerning the boundaries and format of the objects. The common agreement and understanding of the features of linked data object is an important building block for data curation activities. Trust is a keyword in digital preservation and requires that **key stakeholders** in the linked open data arena have the authority and take the responsibility to develop and maintain an infrastructure in which linked data can be curated. In this infrastructure legal aspects concerning the creation and use of data objects are settled as well as the quality of the data objects. Responsibility is taken by the communities producing and curating LD data as part of the their research cycle. Although, L(O)D, as any digital object can be recorded, it remains to be negotiated which ensemble of digital objects should be archiving. The dichotomy between **recording and archiving** - recently introduced by Andrew Treloar and Herbert van de Sompel is a useful framework against which the issue of preserving Linked Data should be discussed. (Treloar, van de Sompel, 2014)

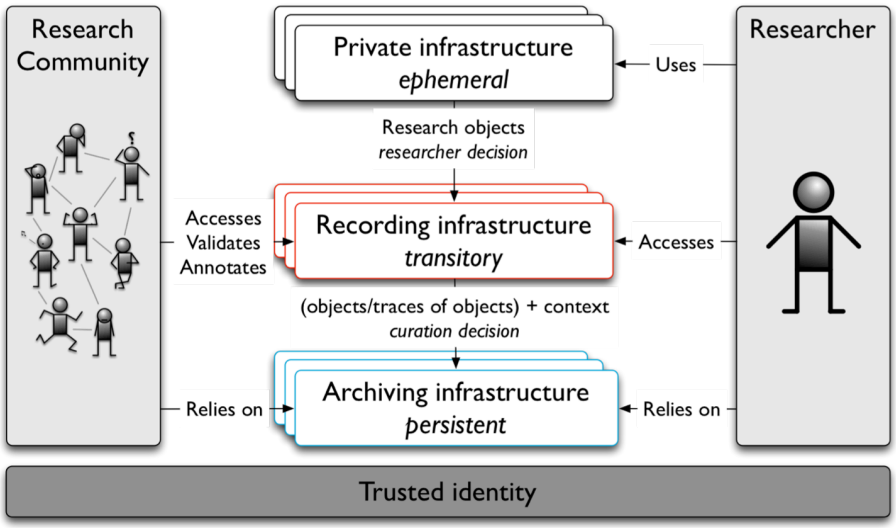


Figure 13 Andrew Treloar, Herbert van de Sompel, Slide 47, CC-BY-SA

The question, how much Linked Data context needs to be archived so that it retains its original meaning can be approached on a technical level. There, two approaches can be envisioned. The first is the one the COOL URI Interest Group of the W3C and Memento adhere to: "A look-up mechanism is important to establish shared understanding of that a URI identifies"[80]. This assumes, hence, that the meaning of a resource can be given in a local description. On the other hand, others may argue that the meaning of a resource can only be understood by looking-up the contents of all its surrounding resources. In such a case, all Linked Data from the archived Linked Data must be archived too. At the end, the communities of LD producers, LD users and the archivist need to negotiate a division of labour.

[80] http://www.w3.org/TR/2008/NOTE-cooluris-20081203/#semweb

Linked Data or Linked Open Data are encoded as digital objects and so have many of the same issues in terms of preservation. Linked Data stands for a specific data representation, and the characteristics of this data model make Linked Data different to other data models. The problem for LOD lies not with the notation of the data model. On contrary, LOD are expressed in Unicode, they are actually text, which can best be understood using the analogy to a large index[81] made for machines to be consumed. Storing and preserving the Unicode text is a known problem. As explained in detail above, there are two aspects of L(O)D which present a challenge to preservation: 1) semantic information which is often only implicit (by dereferencing URIs) documented; and 2) the distributed nature of LOD. Concerning the latter the differentiation between LOD living on the web (main part are URIs pointing to web resources); and LD living in a database like environment is important. Preserving the LOD creates most problems. Any attempt to archive LOD as part of the living web shares problems to archive web resources.

As Peter Doorn put it: "it will be important to distinguish between the straightforward preservation of the linked data in an archive on the one hand, and keeping linked data „serviceable". [In other words] to keep them active, alive, so that the links can remain intact. Mirroring, as [mentioned in some discussions], is an element, but perhaps it makes sense to look here at the (Controlled)LOCKSS approach.[82] Perhaps even we could even think of a variant especially dedicated to linked data: LOCKLODS (Lots of Copies Keep Linked Open Data Safe). Alternatively, one could also draw the parallel with the difference between data archiving and sustaining software (or a service). Data should be archived in a stable state to retain its usefulness; whereas software needs to be maintained and developed (both need a proper version control).

Eventually, solutions to the preservation of Linked Open Data cannot be developed properly without identifying actual needs of stakeholders, such as research communities, libraries and archives, but also governmental information services in the broadest sense. For this, higher level responsibilities, such as scientific integrity, governmental openness to public, transparency in governmental decision-making, etc. need to be articulated to frame an otherwise open and unlimited academic search process. The question to start with is: who is in need to preserve LOD, and why, for which purpose? For this question all others unfold. If it is a research community, which needs to preserve LOD as part of the integrity of the scholarly record, there are in principal two options? That the research community itself takes care of the preservation or negotiates a division of labour with information service providers. The same holds if the stakeholder is a governmental organization publishing statistical or other information in form of LOD. In both cases it is part of the negotiation to determine the goals of the digital preservation and its form.

[81] Personal communication Dan Brickley

[82] See http://www.lockss.org/ and http://www.clockss.org/clockss/Home

7 Bibliography

[1] Anonymous. (2012). *Data Service Infrastructure for the Social Sciences and Humanities - D4.1 Roadmap for Preservation and Curation in the SSH* (pp. 1–40).

[2] Ashkpour, A., & Moreno Penuela, A. (2013). CEDAR project, Technical report: Miniproject advances, Iteration 2. Amsterdam, The Hague. Retrieved from http://cedar-project.nl/wp-content/uploads/Miniproject-Technicalreport.pdf

[3] Beek, W. (2013). On the use of HTTP URIs and the archiving of Linked Data. *KRR blog*. Retrieved December 09, 2013, from http://krr.cs.vu.nl/2013/10/on-the-use-of-http-uris-and-the-archiving-of-linked-data/

[4] Berners-Lee, T., & Fischetti, M. (1999). *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor*. San Francisco: HarperSanFrancisco.

[5] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, *284*(5), 34–43. doi:10.1038/scientificamerican0501-34

[6] Borgman, C. L. (2007). *Scholarship in the digital age: Information, infrastructure, and the Internet*. Cambridge, Mass: MIT Press.

[7] Chun, W. H. K., & Keenan, T. (2006). *New media, old media: A history and theory reader*. New York: Routledge.

[8] Documentation Abstracts, Inc., & Council on Library and Information Resources. (2002). *The state of digital preservation: An international perspective : conference proceedings : Documentation Abstracts, Inc., Institutes for Information Science, Washington, D.C. April 24-25, 2002*. Washington, D.C: Council on Library and Information Resources.

[9] Galison, P. (1997). *Image and logic: A material culture of microphysics*. Chicago: University of Chicago Press.

[10] Giaretta, D. (2011). *Advanced digital preservation*. Berlin [etc].: Springer.

[11] Groth, Paul, and James Frew. *Provenance and Annotation of Data and Processes: 4th International Provenance and Annotation Workshop, Ipaw 2012, Santa Barbara, Ca, Usa, June 19-21, 2012 : Revised Selected Papers*. Berlin [etc.: SpringerLink, 2012.

[12] Guéret, C. (2013). How to publish Open Data on the Web. In E. Folmer, M. Reuvers, & W. Quak (Eds.), *Linked Open Data - Pilot Linked Open Data Nederland* (pp. 115–120). Amersfort: remwerk Amersfort.

[13] Horik, R. . (2005). *Permanent Pixels: Building blocks for the longevity of digital surrogates of historical photographs*. The Hague: DANS.

[14] Huberman, B. A. (2001). *The laws of the Web: Patterns in the ecology of information*. Cambridge, Mass: MIT Press.

[15] Jones, M., Beagrie, N., Resource: The Council for Museums, Archives and Libraries., & British Library. (2001). *Preservation management of digital materials: A handbook*. London: The British Library for Resource, the Council for Museums, Archives and Libraries.

[16] Koehler, W. (1999). An analysis of web page and web site constancy and permanence. *Journal of the American Society for Information Science*, *50*(2), 162–180. doi:10.1002/(SICI)1097-4571(1999)50:2<162::AID-ASI7>3.0.CO;2-B

[17] Kvalheim, V., Kiberg, D., Vestrheim, E., Kvamme, T., De Smedt, K., Mueller, A., … Wloka, B. (2012). *Data Service Infrastructure for the Social Sciences and Humanities. D4.2 Report about Preservation Service Offers* (p. 179). Retrieved from http://dasish.eu/publications/projectreports/D4.2_-_Report_about_Preservation_Service_Offers.pdf

[18] Masanès, J. (2006). *Web archiving*. Berlin: Springer.

[19] Most, P. van der, Defize, P., & Havermans, J. (2010). *Archives Damage Atlas - a tool of assessing damage*. (E. van der Doe, Ed.) (p. 143). The Hague: Metamorfoze.

[20] Niu, J. (2012). An Overview of Web Archiving. *D-Lib Magazine*, *18*(3/4). doi:10.1045/march2012-niu1

[21] Rotella, P. (2012). Is Data The New Oil? *Forbes, Tech, 4/02/2012*, http://www.forbes.com/sites/perryrotella/2012/04/0.

[22] Rayward, W. B. (1997). The origins of information science and the International Institute of Bibliography/International Federation for Information and Documentation (FID). *Journal of the American Society for Information Science*, *48*(4), 289–300. doi:10.1002/(SICI)1097-4571(199704)48:4<289::AID-ASI2>3.0.CO;2-S

[23] Rayward, W.B. (2013) From the index card to the World City: knowledge organization and visualization in the work and ideas of Paul Otlet. In: *Classification and visualization: interfaces to knowledge: proceedings of the International UDC Seminar, 24-25 October 2013, The Hague, The Netherlands* . Edited by Aida Slavic, Almila Akdag Salah & Sylvie Davies. Würzburg: Ergon Verlag, 2013. Pp. 1-42.

[24] Rogers, E. M. (2003). *Diffusion of innovations*. New York: Free Press.

[25] Ross S., *Changing trains at Wigan: Digital preservation and the future of digital scholarship.* London (National Preservation office), 2000. Online available at: <http://eprints.erpanet.org/45/> [cited 16 January 2014]

[26] Rothenberg, J. (1995). Ensuring the Longevity of Digital Documents. *Scientific American*, *272*, 42–47. doi:10.1038/scientificamerican0195-42

[27] Sanderson, R., Phillips, M., & Van de Sompel, H. (2011). Analyzing the Persistence of Referenced Web Resources with Memento, 4. Digital Libraries. Retrieved from http://arxiv.org/abs/1105.3459

[28] Sierman, B. (2012). Where is our Atlas of Digital Damages? *Post at the blog "Digital Preservation Seeds."* Retrieved from http://digitalpreservation.nl/seeds/where-is-our-atlas-of-digital-damages/

[29] Slevin, J. (2000). *The internet and society*. Malden, MA: Polity

[30] Smiraglia, R. P., Scharnhorst, A., Akdag Salah, A., & Gao, C. (2013). UDC in Action. In A. Slavic, A. Akdag Salah, & S. Davies (Eds.), *Classification and visualization: interfaces to knowledge* (pp. 259–270). Würzburg: Ergon Verlag. Retrieved from

http://www.udcc.org/index.php/site/page?view=visualization

[31] Sompel, H. Van de, Klein, M., Sanderson, R., & Nelson, M. (2013). Thoughts on Referencing, Linking, Reference Rot. *MementoWeb.org*. Retrieved from http://mementoweb.org/missing-link/

[32] Strodl, S., Petrov, P., & Rauber, A. (2011). Research on Digital Preservation within projects co-funded by the European Union in the ICT programme. *Vienna University of* .... Retrieved from http://www.ifs.tuwien.ac.at/~strodl/paper/Report - Research on Digital Preservation.pdf Also available at http://cordis.europa.eu/fp7/ict/telearn-digicult/report-research-digital-preservation_en.pdf

[33] Svenonius, E. (2000). *The intellectual foundation of information organization*. Cambridge, Mass: MIT Press.

[34] Treloar, A. Van de Sompel, H. (2014) Riding the Wave and the Scholarly Archive of the Future. Presentation at DANS, The Hague, January 20, 2014. Slides available http://www.slideshare.net/atreloar/scholarly-archiveofthefuture

[35] Webster, F. (2002). *Theories of the information society*. London: Routledge.

[36] Life Cycle Models for Digital Stewardship, by Bill LeFurgy, 2012, see http://blogs.loc.gov/digitalpreservation/2012/02/life-cycle-models-for-digital-stewardship/

[37] Review of Data Management Lifecycle Model by Alex Ball Univ Bath, 2012, see http://opus.bath.ac.uk/28587/1/redm1rep120110ab10.pdf

[38] Data Lifecycle Models and Concepts by CEOS, 2012, see http://www.ceos.org/images/DSIG/Data%20Lifecycle%20Models%20and%20Concepts%20v13.docx