Project no. 600663

PRELIDA

Preserving Linked Data
ICT-2011.4.3: Digital Preservation

D4.1 Analysis of the limitations of Digital Preservation solutions for preserving Linked Data

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

University of Huddersfield

Version [draft]

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

Document Information

Deliverable number:	[D4.1]
Deliverable title:	[Analysis of the limitations of Digital Preservation solutions for preserving LD]
Due date of deliverable:	[March 2014]
Actual date of deliverable:	[March 2014]
Author(s):	[Grigoris Antoniou, Sotiris Batsakis,Antoine Isaac, Andrea Scharnhorst, José María García, René van Horik, David Giaretta, Carlo Meghini]
Participant(s):	[HUD, CNR, APA, UIBK]
Workpackage:	[4]
Workpackage title:	[Roadmapping the future]
Workpackage leader:	[HUD]
Est. person months:	[6]
Dissemination Level:	[PU (Public]
Version:	[1.0]
Keywords:	[Digital Preservation, Linked Data, Gap Analysis]

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level

Abstract

Rationale of the deliverable: Collect, organise and offer a detailed description of use cases related to the long-term preservation and access to Linked Data. Identify challenges, problems and limitations of existing Preservation approaches when applied on Linked Data. Analyse these limitations.

Table of Contents

Executive Summary

PRELIDA project objectives include the identification of differences, and the analysis of the gap existing between two communities: **Linked Data** or Linked Open Data as part of the semantic web technology and **Digital Preservation** as discussed in the context of archives and digital libraries. While both communities deal with data, focus, methods and objectives differ. This deliverable deals with the task of identifying technical issues where this gap exists and the implication of this for efficient digital preservation of linked data. This can be considered as the first step towards describing a roadmap that will bridge this gap and will lead to efficient digital preservation of linked data.

1 Introduction - Digital Preservation and Link Data Gap analysis

1.1 Problem statement

PRELIDA project objectives include the identification of differences, and the analysis of the gap existing between two communities: **Linked Data** or Linked Open Data as part of the semantic web technology and **Digital Preservation** as discussed in the context of archives and digital libraries. While both communities deal with data, focus, methods and objectives differ. This deliverable deals with the task of identifying technical issues where this gap exists and the implication of this for efficient digital preservation of linked data. This can be considered as the first step towards describing a roadmap that will bridge this gap and will lead to efficient digital preservation of linked data.

The first part of this document consists of a description of digital preservation standards and framework, with particular emphasis on the OAIS (Reference Model for an Open Archival Information System) framework. Then practices and solutions of the digital preservation community will be examined with respect to specific use cases involving Linked Data. Linked data on the other hand have characteristics that make their preservation a particular challenging task for the digital preservation community. Detailed analysis of uses cases will be used in order to illustrate clearly the challenges that Linked Data and Digital Preservation communities will face when trying to achieve efficient preservation of Linked data, which are dynamic, heterogeneous, distributed and interconnected. This process will help both communities, since Linked Data community will be aware of practices, solutions and discussions going on in the Digital Preservation community, and Digital Preservation community will be able to understand and prepare to meet challenges posed by the peculiarities of Linked Data.

1.2 Purpose of the gap analysis report

The purpose of the gap analysis report is to provide insights into issues such as:

- Identify and list differences between Linked Data and other types of data with respect to digital preservation requirements. This report will help identify if Linked Data preservation can be reduced to reliable storing of RDF data or if additional issues arise and must be taken into account. These issues may require special treatment of Linked Data.
- Determine to what extent Linked Data preservation can be viewed as a special case of Web archiving, or these two tasks (Linked Data preservation and Web preservation) still have different archiving requirements?
- Identify differences between Linked Data preservation and other special types of data (e.g. multimedia content, software) that are also challenging in terms of digital preservation requirements.
- Identify through specific use cases the importance of digital preservation for Linked Data. Also identify the most important tasks and needs. For example is the full functionality related with an RDF dataset (e.g., SPARQL endpoints) that must be also preserved or just

the data in order to keep track of changes in to the dataset without needing to restore functionality corresponding to each specific time in the past?

- Do previous versions of Linked Data need to be preserved at all?
- If there is a serious reason for preserving Linked Data does this reason apply directly to datasets related to the specific Link Dataset that must be preserved? Notice that in case of Linked Data other related datasets often contain information and definitions required for representing and reasoning over the dataset that must be preserved.
- Since Linked Data are interconnected, does preservation implies keeping track of changes on other datasets directly or indirectly connected to a specific RDF dataset?
- Linked Data complicate preservation requirements in terms of stakeholders, rights over data, ownership of an interlinked dataset and ownership of the archived version. These are issues to investigate.

Answering the above questions will offer important insights into the issue of Linked Data preservation for both involved communities i.e., Digital Preservation and Linked Data. This in turn is important for data providers, service providers, technology providers and end user communities associated with archiving or Linked Data.

1.3 Structure of the report

This document consists of the following parts: Background and Related Work section summarizes the current state of the art in both Digital Preservation and Linked Data communities. Digital preservation and related issues are presented in this section. Particular emphasis is given to the OAIS reference model and a subsection of this report is dedicated to the presentation of OAIS. Then state of the art in Linked Data, in terms of current standards, main applications, trends and best practises are presented. Background section also contains an overview of existing examples of Linked Data archiving.

The third section presents two specific use cases of Linked Data and issues related to their digital preservation. The two use cases are DBpedia and Europeana. Both use cases will be analyzed with respect to the OAIS reference model by identifying stakeholders, functional entities and information for each use case. Detailed descriptions of the use case, identified challenges and suggestions for best practices to deal with these challenges will be presented.

From the identified challenges in each of the use cases we expect to specify a list of required actions which will form the basis for designing a roadmap addressing related issues. This will be the content of the gap analysis section of this report. Finally all the above will be summarized and presented in a compact form in the conclusions section.

2 Background and related work

This report aims to identify challenges arising when digital preservation is applied on Linked Data. In the following background and state of the art of both digital preservation and Linked Data will be presented. A separate subsection consists of the description of the OAIS reference model.

2.1 Digital Preservation

Digital preservation can be defined as activities ensuring access to digital objects (data and software) as long as it is required. In addition to that, preserved content must be authenticated and rendered properly upon request. Potential problems related 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 even the simple loss of the digital objects. Presenting examples cases of the impact of problems appearing in practice related to preservation of digital objects is an issue that has not been brought into attention until recently[1]. Such work can increase awareness regarding digital preservation in a way similar for example to the *Atlas of Damage* for off-line material (Most et al., 2010). Combining visual documentation with a typology, and good practices determined in a large number of digital preservation projects can lead to a tool each digital librarian and archivist can use as a handbook.

Digital preservation as a concern is around since the upcoming of computer technology 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. Related work for example includes (Borgman, 2007) and the report "Preserving Digital Information" (Waters, Garrett 1996). An important issue is that digital information documents have a rather short life (Rothenberg 1995 p.42). Since the 1990s 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.

Several initiatives have been started internationally and nationally and a number of solutions and recommendations were formulated to address the issues mentioned above. Example recommendations are:

- Using file formats based on open standards
- Using the services of digital archives to store the objects for the long-term
- Creating and maintaining high quality documentation (for example the PREMIS[2] standard), specifically developed to create preservation metadata so in the future the digital objects can be reused
- Making 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).

Ross (Ross, 2000) distinguishes three classes of digital materials that scholars may need - retroconversion, new digital content and by-products of contemporary life - that will form the digital record of the future. Scholars must be aware that active involvement in documentation issues of

[1] See for example: http://digitalpreservation.nl/seeds/where-is-our-atlas-of-digital-damages/

[2] See: http://www.loc.gov/standards/premis/

digital materials is essential for long-term access to them. The classification as described by Ross, made almost 15 years ago, and clarifies why initially digital archiving was directed towards **digitized analogue objects** (e.g documents and photographs) and **"digital born" objects**, such as digital texts and databases. Linked Open Data objects obviously belong to the digital born objects for which an analogy in the analogue and off-line world cannot be easily found. The **digital record of the future** mentioned as the third class by Ross still had to emerge. Initially the digital preservation infrastructure was aimed at digital preservation of the first two classes of digital materials: digitized analogue objects and digital born objects which mimicked form and function of analogue objects.

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. Key components of the digital preservation infrastructure are the so-called Trusted Digital Repositories (TDR) that are based on the OAIS reference model, which will be presented in detail in the following. The characteristics that a TDR should adhere to are currently under debate and development. But there is agreement on the fact 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. The framework contains a number of levels ranging from basic self-certification to extended certification carried out by external auditors.

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

1. The **technology preservation strategy,** preservation of the original software and hardware that was used to create and access the information,
2. The **technology emulation strategy**, future computer systems emulate older, obsolete computer platforms as required, and
3. 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 paradigm 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.

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 (Giaretta 2011, pp 31-39) proposed different possible classifications of digital objects.

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. 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 their own. Secondary analysis, replication, enrichment and combining of digital objects are important functions a TDR must support, thus extending its rationale.

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 concrete use cases. The conclusions summarize insights and challenges.

2.2 OAIS Reference model

Standardization requirements for Digital preservation led to the adoption of the OAIS reference model for the corresponding tasks. OAIS reference model will be presented in detail in the following.

The OAIS reference model (Reference Model for an Open Archival Information System) establishes a common framework of terms and concepts relevant for the long term archiving of digital data. It is entirely based from the perspective of an archive. This means, that the model details the processes around and inside of the archive, including the interaction with the user. But, it does not make any statements about which data would need to be preserved.

The OAIS reference model has been developed under the direction of the "Consultative Committee for Space Data Systems" (CCSDS) and adopted as ISO standard 14721. An OAIS is defined as an archive and 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 Open Archival Information System reference model (OAIS) is an ISO standard (ISO 14721) that provides:

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

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

1. Negotiate for and accept appropriate information from information Producers.
2. Obtain sufficient control of the information provided to the level needed to ensure Long Term Preservation.
3. 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.
4. 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.
5. 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.

6. 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 which 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 more fine grained set of terms - much more detailed than simply using the term “metadata”.

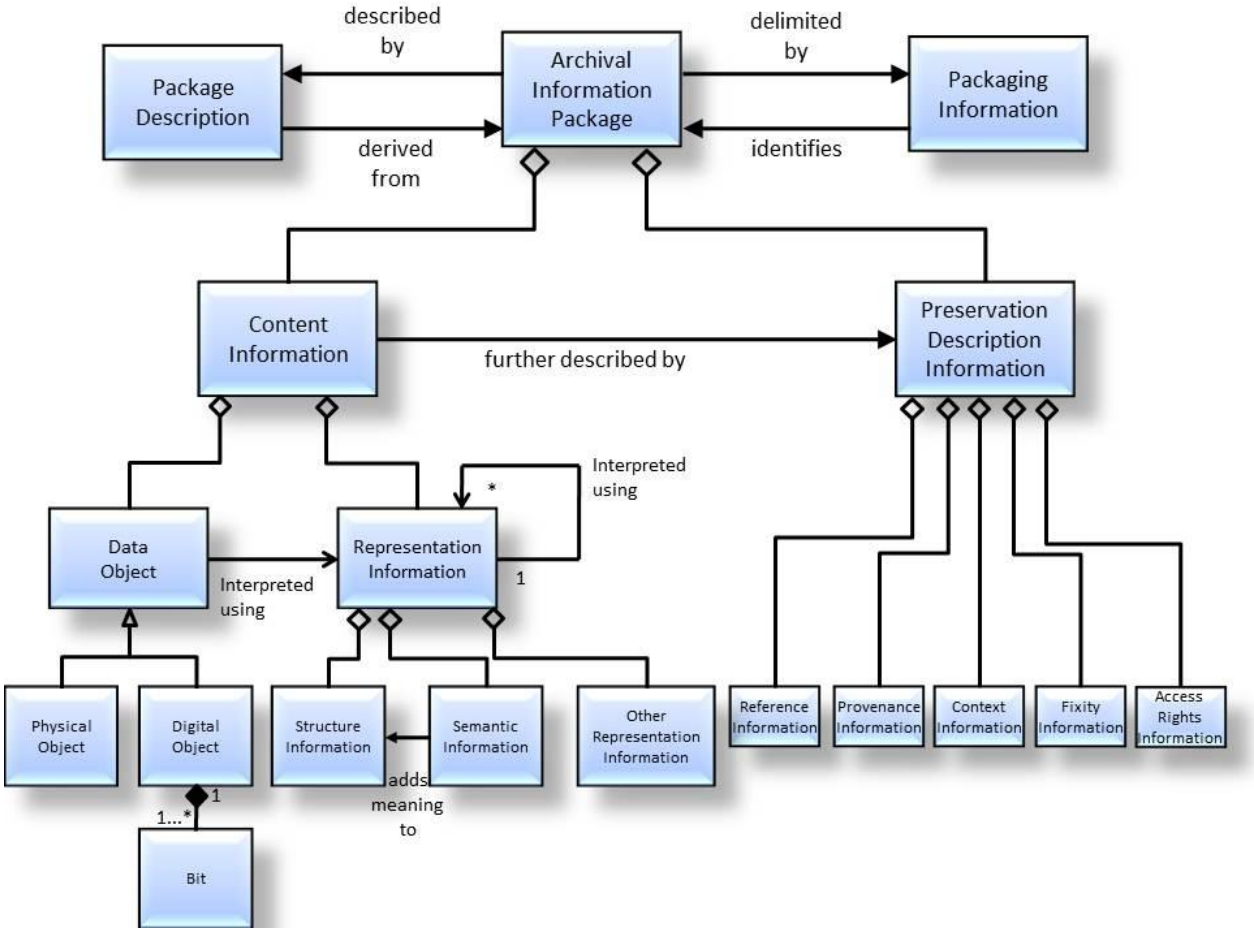

Figure 1 : OAIS Archival Information Package

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 (5) indicates that even if the repository itself fails, there should have been made arrangements to hand over the digital objects, and the AIP construct must ensure that the appropriate information has been captured in advance.

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.**

2.3 Linked (Open) Data

Data traditionally was considered to be a closed asset, but today it is considered to be a critical resource equally significant to resources such as oil (Rotella, 2012). The value of data comes with the usage of data after appropriate processing. Processing data by other parties implies that data must be shared by allowing access to third parties. Typically data sharing among companies is based on pair agreements made on a case by case basis: a company with an interest in the data of another company will contact it to make an agreement to exchange data against a monetary compensation. This can also be the case between governments and citizens when personal data and the Public Sector Information (PSI) EU directive apply[3]. This directive essentially states that every data collected with public funds shall be made accessible to the public when public requests it. Data falling under this directive have been long considered to be closed data that can be shared only on-demand. These demands should be formally approved and processing such demands is a costly administrative process which pushed the institutions into making the data freely accessible to everyone directly. 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[4]. In addition to saving on administrative processes spending, opening up public datasets can lead to the creation of businesses using this open data and bringing back indirect revenues to the state. On the other hand the loss of the government control enforced by the processing of requests comes at a cost: the data which are made open can, and probably will, be used in unexpected ways. Furthermore they can be combined with other datasets and interpreted in a wrong way or yield more information than intended, thus raising for example privacy issues.

Open data portals demonstrate the effects of opening access to data. A data portal is a place where data sets are made available in an open license and they are uploaded and/or referenced. There are more than 150 of such data portals in Europe aiming at providing access to a wide range of data sets both in the public, scientific and cultural heritage domains[5]. What all these portals have in common is that they allow end users to download entire data sets or parts of data sets. A user can get a file containing data in a particular serialization format and conceptual model.

After downloading open data, the following task is data integration and data analysis. The objective is to combine all the heterogeneous data acquired from different sources into one consistent dataset that can be used by a given application. An important issue is to create unambiguous terms. “Portsmouth”, for example, may refer to a city in the UK, several cities in the US or a football team. The main idea behind Linked Open Data (LOD), but also behind Linked Data in general, is to use unique identifiers instead of ambiguous words for both the concepts referred to in the dataset and the data model, and definitions applying to the data. The design principles of LOD are defined by Tim Berners Lee and can be summarized as

- Use the Web as a platform to publish and re-use identifiers that refer to data,

[3] See: http://ec.europa.eu/digital-agenda/en/open-data-0

[4] See report at: http://policyexchange.org.uk/images/publications/the big data opportunity.pdf

[5] See: http://ec.europa.eu/digital-agenda/en/open-data-portals

- Use a standard data model for expressing the data (RDF).

The Resource Description Framework[6] (RDF) is a way to model data as a list of statements made between two resources identified with their unique identifiers (URI). For example, data telling that "Athens is the capital of Greece and is called Atene in Italian" can be expressed as two statements in RDF (See Figure 2).

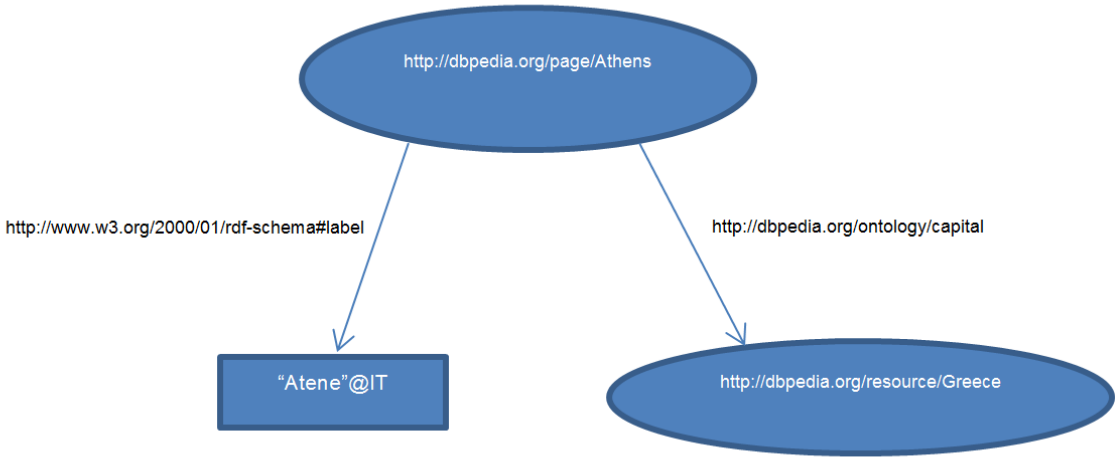

Figure 2 : An example RDF representation of "Athens is the capital of Greece and is called Atene in Italian"

The drawing of Figure 2 follows the common representation convention of using ellipses for resources and squares for literal values. It can be observed that by using a resource instead of a literal for "Athens" the two statements can be connected. Following the same principle across several datasets leads to the creation of a "Web of Data".

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[7] (TTL), TriG[8], RDF/XML[9], and RDFa[10] are such 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

[6]See: http://www.w3.org/standards/techs/rdf#w3c_all

[7]See: http://www.w3.org/TR/turtle/

[8]See: http://www.w3.org/TR/trig/

[9] See: http://www.w3.org/TR/REC-rdf-syntax/

[10]See: http://www.w3.org/TR/rdfa-syntax/

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 request specific type of documents. Typically, HTML for human consumption and RDF for machine consumption
- As a database: RDF can be stored in optimised graph databases ("triple store") and queried using the SPARQL query language[11]. This is similar in spirit to storing relational data in a relational database and query it using SQL.

There are several considerations that must be taken into account when deciding between the three approaches. One of them is the size of the dataset; typically the annotation approach used for "small data" (*e.g.* social profile on a home-page) whereas the database approach rules "big 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 3).

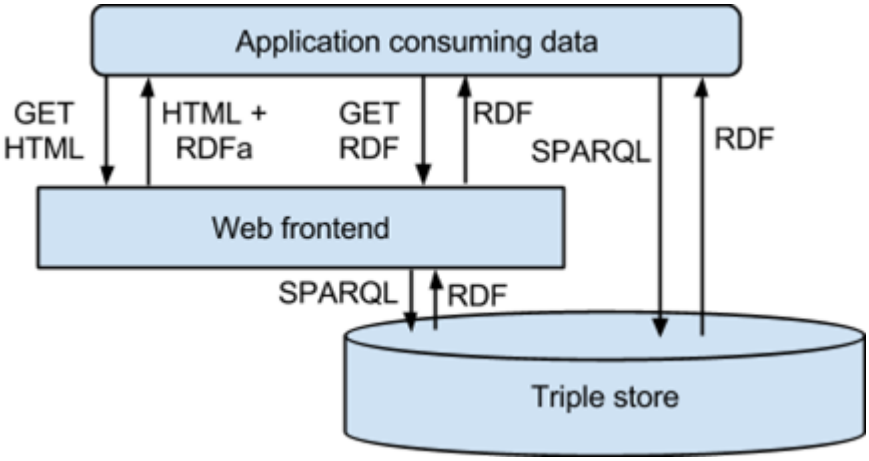

Figure 3: A common publication architecture for RDF data

The architecture depicted on Figure 3 is the one in place for DBpedia, 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 [12] http://dbpedia.org/page/Amsterdam

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

[12] 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[13]: http://dbpedia.org/resource/Amsterdam

- With a SPARQL query sent to the end point[14]: http://dbpedia.org/sparql

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. 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 categories of Web of Data out there, for which different preservation strategies can be proposed.

The differentiation between the two categories of Web 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 the users that use Linked Data without requiring online access (**offline 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 and other forms of Web data. The archived content can be considered as non-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 **(online 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 dereferenceable". 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.

Web of Data: Offline and Online use

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

---

13 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)

13

14 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)

exclusively accessible on the Web and cannot be queried using SPARQL, a query language for RDF.

- 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 which 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. Other approaches concern the harvesting of Web-only data to make it accessible in a triple store or the extraction of structured information from Web pages as RDF dumps. The majority of Linked Data consumption is performed "off-line", using statements stored in a triple-store.

2.4 Digital Preservation and Linked Data

The presence of these two different forms of Web data is very important for the goal of preserving them. In fact, two preservation strategies can be employed 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, Microdata …). 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 N Triples 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 dereferenceable or non dereferenceable 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 (i.e., broken links), 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. 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. 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.

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 “European Union”, which in turns links to the concept “Europe”. Whereas Paris has now a conceptual definition that can be considered stable, this is not the case for European Union (which will evolve with changing members) or even Europe. 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.

- **Overlap**: RDF data dumps are easy to preserve, share, load and consume. These RDF data dumps are already largely used on data portals as a way to publish/share/consume Linked Open Data. DBpedia archiving is such an example. 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 “European Union”, 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 15-20 years. Furthermore, the matching Web resource may have gone missing by that time. Another issue is that two data set preserved from two different time-frames may refer to the same concept “European Union” 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.

3 Use cases gap analysis

Analysing specific use cases is an important step towards identifying technical challenges on digital preservation of Linked (Open) Data. In the following two use cases, DBpedia and Europeana, will be analysed in order to identify Linked Open Data preservation issues. DBpedia is a core part of the LOD cloud, thus it is a very important use case. The Europeana project for cultural heritage preservation is also an important use case, involving preservation of metadata from different sources such as museums and libraries.

3.1 Digital Preservation for DBpedia

3.1.1 Description of DBpedia

DBpedia's 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 is 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. DBpedia is selected as a use case because it is one of the core parts of the Linked Open Data cloud and it is interlinked with numerous LOD sets.

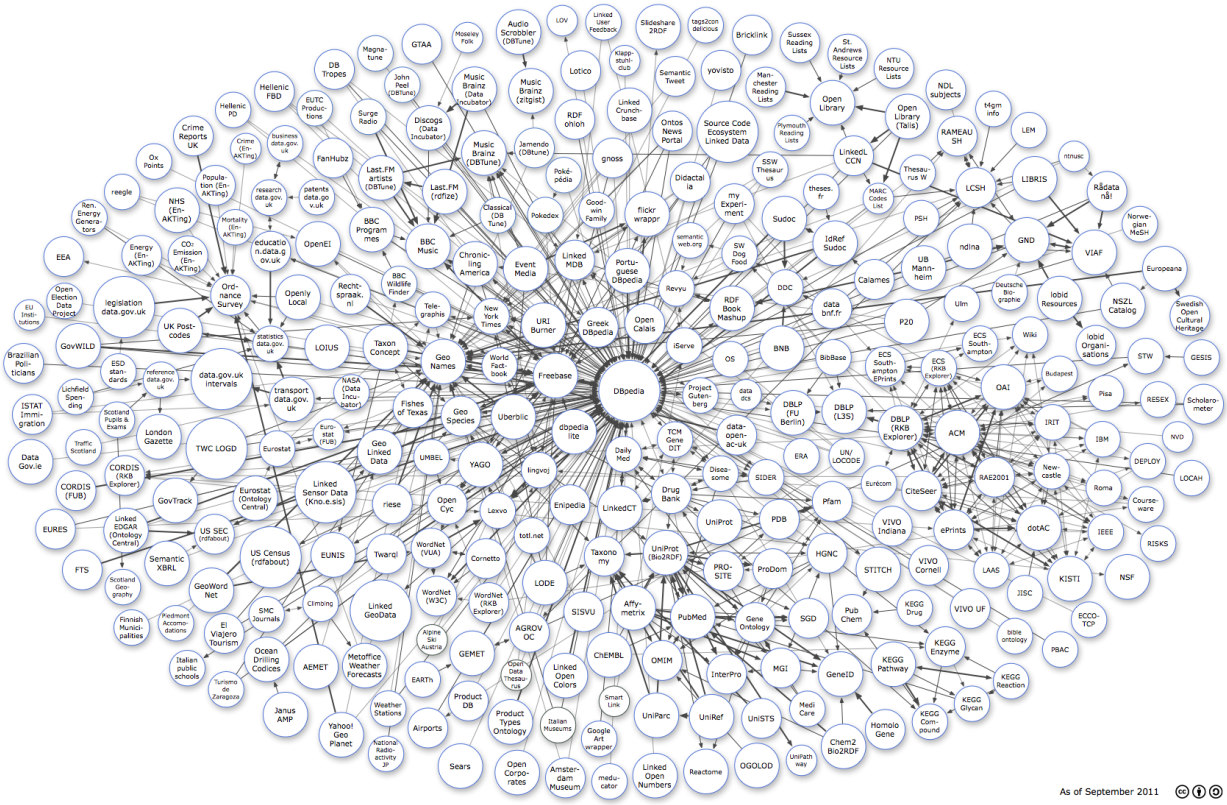

Figure 4: Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch[15]. DBpedia is the core node in this diagram.

The archiving mechanism of DBpedia is presented in the following.

3.1.2 DBpedia archiving

DBpedia archiving is currently handled by DBpedia association and not by an external organisation. Since DBpedia data is extracted from Wikipedia data and is 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)16. DBpedia content (data and metadata such as the DBpedia ontology) is available to end users under the same terms and licences as the Wikipedia content.

[15]See: http://lod-cloud.net/

[16]See: http://en.wikipedia.org/wiki/Wikipedia:Copyrights

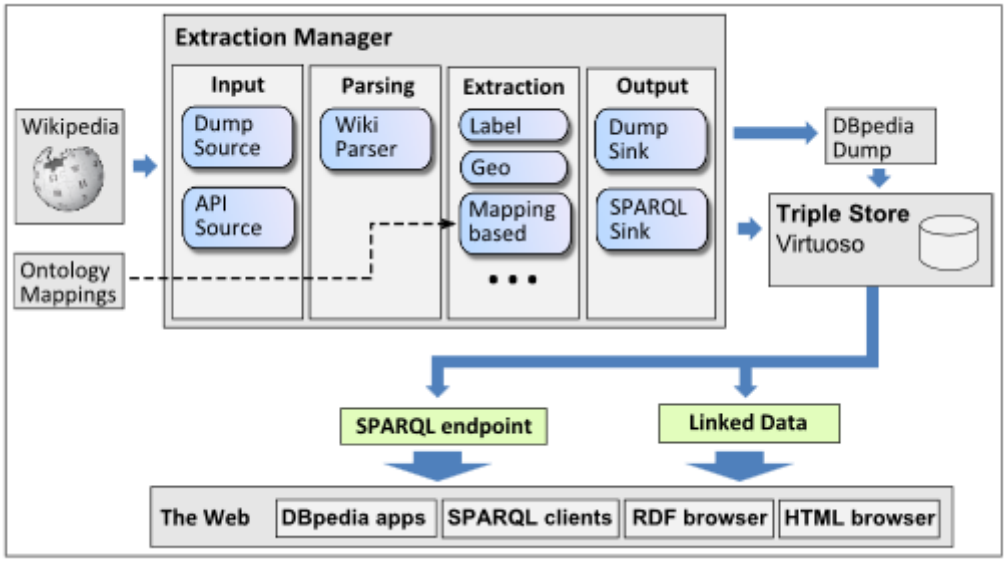

Figure 5: DBpedia data extraction mechanism[17]

DBpedia preserves different versions of the entire dataset by means of DBpedia dumps corresponding to an versioning mechanism18. Besides the versioned versions of DBpedia, DBpedia live19 keeps track of changes in Wikipedia and extracts newly changed information from Wikipedia infoboxes and text into RDF format20. DBpedia live contains also metadata about the part of Wikipedia text that the information was extracted, the user created or modified corresponding data and the date of creation or last modification. Incremental modifications of DBpedia live are also archived21. DBpedia dataset contains links to other datasets containing both definitions and data (e.g., Geonames). There are currently (February 2014) more than 27 million links from DBpedia to other datasets. DBpedia archiving mechanisms also preserve links to these datasets but not their content. Preserved data is DBpedia content in RDF or tables (CSV) format. Rendering and querying software is not part of the archive although extraction software from Wikipedia infoboxes and text used for the creation of DBpedia dataset is preserved at GitHub.

The above description indicates that currently DBpedia preservation stakeholders are the DBpedia association22 and end users seeking access to previous versions of the DBpedia dataset, either in RDF format of as a Web page or through a SPARQL endpoint. Cooperating organizations such as

[17] See: http://svn.aksw.org/papers/2013/SWJ_DBpedia/public.pdf

[18] See for example: http://downloads.dbpedia.org/3.9/en/

[19] See http://live.dbpedia.org/

[20] See for example the entry for Berlin at: http://live.dbpedia.org/page/Berlin

[21] See for example: http://live.dbpedia.org/changesets/2014/

[21]

[22] See: http://dbpedia.org/association

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. Currently supported data formats are RDF and CSV. Adopting open standards such as RDF and following W3C specifications clearly 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 functionality are not directly preserved and supported. Also answering complex requests about the evolution of specific data over a temporal interval is 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 not yet supported.

In the following specific use cases based on possible interactions and user requests are presented. Use cases are:

- Request of archived data in RDF of CSV format
- Request of rendered data in Web format
- Submitting SPARQL queries on the archived versions of the data

The above three use cases can be further refined with respect to the format of the request i.e., if it corresponds to a specific time point or interval. Also they can be refined with respect to the requirement of getting data from external sources.

3.1.3 DBpedia preservation list of use cases

Use case 1: RDF Data archiving and retrieval

DBpedia data (in RDF format, or Tables-CSV format) is archived and the user requests specific data (or the entire dataset) as it was at a specific date in the past, e.g., the RDF description of topic Semantic Web at 1/1/2010. The preservation mechanism must be able to provide the requested data in RDF (or Table) format. Timestamps specifying the interval that the data was valid (i.e., dates that the data was created or modified for the last time before the date specified into the user request, and first modification or deletion date after that date) are a desirable feature of the mechanism. Retrieving data for a specific time interval, e.g., 2010-2012, instead of a specific date is a more complex case since all versions of the data and their corresponding validity intervals with respect to the request interval must be returned.
Currently complex requests involving intervals are not handled by the DBpedia archiving mechanism. RDF data containing links to other LOD for specific time points can be retrieved, but the content of links to external LOD is not preserved.

Use case 2: Rendering data as Web page

User requests the DBpedia data for a specific topic at a given temporal point or interval: e.g., description of the topic Semantic Web at 2010, or between 2008-2012 presented in a format such as the following:
http://dbpedia.org/page/Semantic_Web
The preservation mechanism should be able to return the data in RDF format and in case description is modified during the given interval all corresponding descriptions, the intervals that each one distinct description was true, modification history, differences between versions and editors should be returned as in the first use case. Rendering requested data as a Web page will introduce the following problem: can the functionality of external links be preserved and supported as well or not? Currently rendering software is not part of the preservation mechanism, although using a standard representation (RDF) minimizes the risk of not been able to render the data in the future.

Use case 3: SPARQL Endpoint functionality

Reconstruct the functionality of the DBpedia SPARQL endpoint at a specific temporal point in the past. Specifically the user will be able to issue queries through a Web interface such as:
http://dbpedia.org/sparql
specifying both the query and the time point (this use case can be extended by supporting interval queries as in use case 1 and 2 above) . There are different kinds of queries that must be handled corresponding to different use cases:

a) Queries spanning into RDF data into DBpedia dataset only

b) Queries spanning into DBpedia dataset and datasets directly connected to the DBpedia RDF dataset (e.g., Geonames)

c) Queries spanning into DBpedia data and to external datasets connected indirectly with DBpedia (i.e., through links to datasets of case b).

Currently SPARQL end-point functionality is not directly preserved, i.e., the users must retrieve the data and use their own SPARQL end-point to query them. Then, they will be able to issue queries of type (a) above, but not queries of type (b) when the content of external links is requested. Also requests of type (c) cannot be handled by the current preservation mechanism.

3.2 Linked Data Preservation for Europeana

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), now mainly by means of an API usable by everyone.

Basic Europeana sources

The main source of data for Europeana are its cultural data providers—museums, libraries, and archives. 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.

Dependence on third-parties linked datasets

Cultural Heritage providers are not Europeana's only source of data. 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[23], Geonames[24] 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.

[23] GEneral Multilingual Environmental Thesaurus, http://www.eionet.europa.eu/gemet/

[24] http://geonames.org

24

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, even incompatible with the original source.

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 constructs, 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[25].

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[26].

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).

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[27], either by http content negotiation or in the form of RDF dumps.

[25] Cf. case studies at http://pro.europeana.eu/carare-edm and http://pro.europeana.eu/mimo-edm

[26] http://www.w3.org/2004/02/skos/

[27] http://data.europeana.eu

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. Europeana try to address these issues by implementing redirection mechanisms between old and new identifiers. In addition Europeana tries to convince providers to send 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-)publishes. Progress on this issue is expected soon (as of February 2014), 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 data on behalf of its providers, who often have their own policies in place. This will raise issues if one day Europeana has to provide preservation data to its own consumers, which should reflect the preservation 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.

Use Case 1: aligning different time-versions of data for linked data consumption.

For Europeana it is important to be get a seamless access to data for resources, even when that data change. It could be that a description of an object in Europeana, given by a provider, uses a third-party URI that is now deprecated in the most updated version of that third party linked dataset. Best practices on how to represent updates or deprecation of URIs and accompanying data would be needed, for data providers to inform properly the data consumers. Rules for consuming the published information should also be defined, so that the entire community proceeds the same way the versioning data.

Use Case 2: preserving data that aggregates other datasets

Europeana aims to be a reference point for accessing cultural objects. The metadata it aggregates plays the key role in this objective. It must be trustable by data consumers. However, as noted, Europeana has no mandate to preserve its providers' metadata. In fact

the metadata it receives from them is only a derivative, a reformatted version of it. Sometimes with less data, sometimes with more (e.g. for controlled rights statement that applies to the content representing a cultural heritage object). European's problem becomes the one of preserving an interconnected set of dataset views. What should be the best practices for doing this?

4 Gap analysis-Different characteristics of Linked (Open) Data compared to traditional archives

The first step in gap analysis is to identify the peculiarities of Linked Open Data when compared to other forms of data that typically are handled by digital preservation systems. This will help identifying gaps between the two communities (Linked Data and Digital Preservation) and prepare a roadmap towards efficient solutions for Linked data preservation. Classification of Linked Data will be based on classification schemes for digital objects in general. There are different possible classifications of digital objects, for example the following classification was proposed for the APARSEN project (Giaretta 2011, pp 31-39) according to whether the digital object under consideration is

- static vs dynamic
- complex vs simple
- active vs passive
- rendered vs non-rendered

Applying this classification to Linked Data yields the following: linked data are dynamic, complex, passive and typically non-rendered. While this is not an exhaustive 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? This is the first question to be answered when digital preservation is applied over Linked Data. A common answer cannot be given for each case in advance and it depends on a specific dataset. But since Linked Data are usually dynamic, which isn't the case in many preservation applications for other types of objects, older versions should be preserved. This is the case for example of DBpedia where both older versions and incremental changes are archived. Also Europeana consists of metadata that are dynamic, although changes are not that often as DBpedia.
  - Different statements may be made at any time and so the “boundary” of the object under consideration changes in time. This is typically the case in Linked Data, thus for example new Links can be added to other datasets

containing for example definitions to concepts of the preserved dataset. This can introduce indirect changes to the preserved dataset by means of changes in the definition of a concept, causing in turn the need to preserve the linked dataset containing the definitions as well. This is an issue not commonly addressed in digital preservation systems and in existing Linked Data archives such as for example the DBpedia and Europeana.

- Complex
  - Linked Data is typically about expressing statements (facts) whose truth or falsity is grounded to the context provided by all the other statements available at that particular moment. Related information possibly contained in other linked datasets may be part of the data needed to specify properties of statements such as the truth value of a statement. On the other hand the preservation community is concerned with preserving what has been expressed and not the truth or falsity of the information. It is important to note that the notion "truth" here is applied with two different meanings: the truth of the content to be preserved (Linked Data) versus truly preserving, meaning to ensure the authenticity of the object independent from its content. The second meaning is usually addressed in the Digital preservation systems while the first (which is closely related to the fact that Linked Data are interconnected and dynamic) is not.
- Non-rendered
  - Non-rendered digital objects need to be processed to produce any number of possible outputs. Is what is done with LD very varied? This gives rise to a great number of possible preservation objectives, compared to rendering an image. The answer to the above question also contributes to specifying requirements of the preservation mechanism. Typically Linked data are not rendered and adopt standards such as RDF that are open, widely adopted and well supported. This simplifies that preservation mechanism if the only preservation objects are the data themselves. This is the case in the current DBpedia archiving mechanism. On the other hand if rendering is a requirement then appropriate software and perhaps platform emulators must be preserved in addition to data themselves. For example DBpedia RDF data can be rendered in both HTML format and through a SPARQL end-point. While this is part of the functionality of the current version of DBpedia, rendering is not supported for the archived versions, by the existing archiving mechanism. This is also the case for Europeana.
- Passive
  - The linked data is usually represented in the form of statements or objects (typically RDF triples) which are not applications. The statements are normally handled by applications and, right now, applications are not perfect or indefinitely scalable. The previous statement illustrates the fact that

besides preserving data, software that handles data should be preserved in some cases. Rendering functionality or access method, such as a SPARQL endpoint for archived DBpedia data is an example case not handled by the existing archiving mechanism.

In addition to the above Linked data is distributed and this fact complicates authenticity of preserved data and increases uncertainty.

- Linked Data is typically distributed and the persistence the preserved objects depend on all the individual parts and the ontologies/vocabularies with which the data is expressed. A lot of data is essentially dependent on OWL ontologies that are created/maintained/hosted by others. LOD is based on the Web and as such they suffer from 404 errors, but their effect may be stronger for machine readable data based on formal semantics than it is for Web documents. Would mirroring/redundancy be the main goal of "preservation" for LD rather than the long term access? The above question also is very important. If the answer is positive then emphasis must be given to preserving related datasets as well. The next important issue raised then is if this has to be repeated for datasets related to the datasets related to the preserved dataset etc. Identifying the minimal set of data needed to be preserved in such a scenario is a hard problem. The scope of the archiving process is critical for determining these boundaries. The three different cases in DBpedia preservation use case 3 illustrate this fact. Also preserving related data leads to similar problems to Web archiving which is also a very complex topic. Current archiving mechanism such as the one used in DBpedia don't address this issue. Europeana uses a cached copy of Linked Data (e.g., data from DBpedia) in order to avoid broken link errors, but this introduces the danger of inconsistency of stored and live data on Linked Datasets, in case Linked Data has changed.
- Authenticity is a major issue in preservation. There are efforts for dealing with the provenance of digitally encoded information over time. Can these techniques simply be applied individually to the various parts of each triple? The answer to this question in case of interlinked datasets is also complex because authenticity may have to be ensured for external datasets too.
- LOD are **uncertain**: LOD quality may be compromised by various data imperfections due to limitations of the underlying data acquisition infrastructures (which is a problem of Web data also) the ambiguity in the domain of interest since various definitions and natural language terms used are ambiguous (and formal semantics may not solve this problem if definitions are not accurate). Similarly, when LOD is produced by automatically extracting structured information from text the results are approximate and uncertain. In this respect, representing uncertainly and answering queries over corresponding RDF graphs is a challenging problem not yet related to long-term LOD interpretability.
- Linked Data is 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. This is similar to recording and archiving time series of measurements of temperature. In this case preserving the mere numerical value is

not useful 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 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. The OAIS model differentiates between the data object and its representation information. This is related to the question of data and metadata - a discourse which can easily fill books. This can be expressed as drawing boundaries between the object itself, its description, and its meaning. In case of LOD, an object's meaning is often defined on external linked datasets, thus keeping track of changes in external datasets is critical. Inconsistencies between archived versions of external datasets and live versions are an important issue, as illustrated for example in Europeana use case.

- Linked Data depend on the **web infrastructure** and in particular on the dereferenciation of HTTP URIs. With respect to this issue 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 key, as well as the interoperability between different identifiers. 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**[28] demonstrated the urgency of the problem.
- Linked Data are **accessible in many ways**: through SPARQL end-points, as RDF dumps, as RDF dumps plus a sequence of incremental updates, as RDFa, as microdata and others as demonstrated in the DBpedia use case. Linked Data descriptions are modelled using RDF and can be serialised using different formats such as RDF/XML, N3, Turtle and JSON-LD. For each form its durability can be assessed. For example XML is both machine and human readable; this is also the case for formats mentioned above such as Turtle and JSON. But, more important here is the dichotomy between a database representation (the data base web of data) and the web representation (web version of web of data). Ideally both cases (e.g., Web rendering and SPARQL end-point in DBPedia use case) must be handled by a preservation mechanism.
- 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 the literature. 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. Currently, vocabulary has been developed to address the problem of versioning which in the language of LOD is covered by the concern to have good provenance (Groth &Frew, 2012). But the fact that part of the LOD cloud might miss provenance in space (versions) and time can lead to an increase 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**

[28] See: http://hiberlink.org/

triples, and at the moment there is no standard ways way to express metadata about RDF triples. Labelling, named RDF graphs and various forms of reification (e.g., N-ary approach[29]) have been proposed for addressing this issue.

4.1 Technical gap-identified issues and problems

Based on the questions raised in the previous section several issues and problems are identified. These are:

- **Selection:** Which LOD data should actively be preserved?
- Who is responsible for "community" data, such as DBpedia? Although in some cases the answer to this question is clear (for DBpedia for example DBpedia association is responsible for preservation, another case is the Dublin Core ontology which is maintained by DCMI[30] which is also maintaining FOAF[31]) the question becomes then, which sustainability model for such associations. Increasingly government data is made available as LOD. Agencies publishing the data should be aware of their role to create durable, trustworthy, authentic LOD. In addition to that other LOD may depend on a dataset e.g., for providing definitions. Should such data preserved and if yes can that reasoning apply iteratively?
- **Durability of the format**: Which formats can we distinguish? RDF, Triple Store, Software, SPARQL, etc. Can we make a classification? Can we create guidelines to create durable LOD formats? E.g. RDF is based on open standards so they can be considered as very durable. What about Triple Stores? Also use of Persistent Identifiers contributes to durability of LOD. Standardization efforts in Linked Open data community and the compliance to W3C standards greatly reduce risks related to durability of the adopted format.
- **Rights / ownership / licenses**. LOD are by definition open (which is not always the case for LD in general), but how to preserve privacy then? This issue is raised for example in the PILOD project reports, a Dutch project to create LOD from governmental data (Gueret, 2013). Additional issues are which licenses to use, which Creative Commons code? Answers to such questions must be provided with all involved stakeholders such as Wikipedia in case of DBpedia and content providers such as museums in case of Europeana. For Europeana for example the preserved metadata is available under the CC0[32] licence. Rights and licenses issue is the main difference between open and non-open LD since different licences are typically required for each case. Other than that Linked Open Data and other forms of LD do not differ with respect to preservation requirements.
- **Storage**. Highest quality is storage in "Trusted Digital Repository". But which other models

[29] http://www.w3.org/TR/swbp-n-aryRelations/

[30] See: http://dublincore.org/

[31] See: http://www.foaf-project.org/

[32] See: http://creativecommons.org/about/cc0

can be used? One example is providing multiple copies/mirrors (CLOCKS). Also there is the issue of scale, similar to problems encountered when dealing with Web archiving. The issue of scale is very important in specific use cases involving large scale, interconnected and highly dynamic datasets when linked datasets must also be preserved. DBpedia use case 3 is such an example where scale problems similar to Web archiving appear.

- **Metadata and Definitions**. Documentation is required to enable the designated community to understand the meaning of LOD objects. Are LOD objects "self-descriptive"? That depends on where to put the boundary of LD objects. If a LD object doesn't include the ontology(ies) that provides the classes and properties used to express the data, than the object is not self-descriptive, and there's an additional preservation risk. Do they need any additional data elements to facilitate long term access? Are these metadata adequate and how to prove that?

Based on the above list of issues and the description of DBpedia and Europeana Use Cases, the following comments can be made:

DBpedia use case:

For the DBpedia use case stakeholder is the DBpedia association, but in case of external data sources (use cases 2 & 3) additional organizations must participate in the archiving process directly, which is not addressed in the current archiving process. Since DBpedia data is retrieved by Wikipedia and is open, licensing issues are not currently a problem. Also the selection of open standards such as RDF ensures to a large degree the durability of the data, the same argument applies to the rendering of data in HTML or the reconstruction of the SPARQL endpoint functionality. On the other hand since corresponding software is not preserved there is no guarantee that the rendering functionality and the SPARQL endpoint functionality will be preserved in the long term, even if the DBpedia data is available. Part of the data preserved as of February 2014 is metadata corresponding to the source of information (i.e., the Wikipedia page that data were extracted from, the user and the data of the last modification). This was not the case of previous versions of the DBpedia archive, were only the time that the data was archived was recorded.

Europeana use case:

For the Europeana use case stakeholder is the Europeana foundation, external data providers and organizations such as museums and libraries, and data consumers. Cultural organizations usually have their own archiving process besides Europeana. Yet Europeana harvests metadata (under Creative Commons CC0 1.0 Universal Public Domain Dedication licence[33]) which must be provided further in a trustable manner. Rendering functionality is not preserved into the archive, although corresponding software is preserved separately. Metadata from external Linked Data is also cached in order to deal with the issue of changing data, and for performance reasons. On the other hand this practise may introduce inconsistencies between cached and live data especially when this practise is used on dynamic Linked Open Data such as DBpedia.

[33] See: http://pro.europeana.eu/web/guest/data-exchange-agreement

5 Conclusions

This report examines 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 consisting 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, which is 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 allowing people to 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.

The linked data paradigm concerns the technology to publish, share and connect data on the web, data that have formal semantics and are machine readable. 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. On the other hand this is similar to the challenges and criticisms raised for the Web. This is exactly the reason why projects such as **Memento** (Ainsworth, Scott G., et al. 2011) dealing with archiving of different versions of Web resources are highly relevant to PRELIDA, since similar mechanisms are required for archiving different versions of L(O)D.

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 the loss of linked open data by means of analysing specific use cases. 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 DBpedia, 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. **DIACHRON** project is a highly relevant research effort towards this direction. Auer et al. (Auer Sören, et al. 2012) identifies main issues related to LOD preservation for different use case categories, namely Open Data Markets, Enterprise Data Intranets, and Scientific Information Systems. These issues are: ranking datasets, crawling datasets, diachronic citations (i.e., data from different sources for the same fact), temporal annotations, cleaning and repairing uncertain data, propagation of changes over Linked Data, Archiving multiple versions and **longitudinal querying** (i.e., querying over different versions of data spanning over a temporal interval as in the DBpedia use case of this report). All these issues are

highly relevant to PRELIDA project, and this report is a first step towards creating a roadmap for dealing with such issues. Since this is also the case for DIACHRON project close cooperation between the two projects is expected.

In order to provide solutions for the long term preservation of linked data the focus should be on the following three issues: **version, fixity and responsibility**. In each of those aspects we find technological questions not yet solved. But the main lesson to be learned from Digital Preservation is that the essence 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 it 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 their research cycle. Although, LOD, as any digital object can be recorded, it remains to be negotiated which ensemble of digital objects should be archived. 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)

The question of 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 lookup mechanism is important to establish shared understanding of that a URI identifies”. This assumes 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, which is the most common case for LOD as illustrated at DBpedia use 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.

Linked Data or Linked Open Data are a specific form of digital objects. The problem for LOD lies not with the notation of the data model. On contrary, LOD are expressed in ASCII, they are actually text, representing RDF triples. Storing and preserving text is a known problem. As explained in detail above, it is the differentiation between LOD living on the web, and which main part are URIs pointing to web resources; and LD living in a database like environment, which creates most problems for archiving. Any attempt to archive LOD as part of the living web shares problems to archive web

resources.

Thus 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" or to keep them active, so that the links can remain intact. Mirroring or LOCKSS[34] (Lots of Copies Keep Stuff Safe) approach can be also applied for LOD preservation. 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). Overall Preservation of Linked Data is a complex issue involving dynamic interconnected data, combining characteristics of databases and the Web and also both data and applications for rendering and processing them.

Bibliography

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

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

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.

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.

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.

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.

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

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]

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

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

[34] See http://www.lockss.org/

Ainsworth, Scott G., et al. (2011). How much of the web is archived? *Proceedings of the 11th annual international*

ACM/IEEE joint conference on Digital libraries. ACM.

Auer, Sören, et al. (2012). Diachronic linked data: towards long-term preservation of structured interrelated information. *Proceedings of the First International Workshop on Open Data*. ACM.