Project no. 600663

PRELIDA

Preserving Linked Data
ICT-2011.4.3: Digital Preservation

D4.2 First version of roadmap

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

University of Huddersfield

Version [draft,1]

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

---

Document Information

Deliverable number:	D4.2
Deliverable title:	First version of roadmap
Due date of deliverable:	08|2014
Actual date of deliverable:	09|2014
Author(s):	Grigoris Antoniou, Sotiris Batsakis, Antoine Isaac, Andrea Scharnhorst, José María García, René van Horik, Carlo Meghini.
Participant(s):	HUD,CNR, APA, UIBK
Workpackage:	WP4
Workpackage title:	Roadmapping the future
Workpackage leader:	HUD
Est. person months:	5
Dissemination Level:	PU (Public)
Version:	1
Keywords:	Digital Preservation, Linked Data, Gap Analysis

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level

Abstract

The present document is a first version of the long term preservation of Linked Data roadmap deliverable. Based on current state of the art on digital preservation and Linked Data and a description of related use cases, corresponding challenges and limitations of existing approaches are identified. Then, based on these challenges, a draft roadmap is proposed for dealing with ingestion of Linked Datasets and changes into the dataset. Keeping track of changes and related technical and organizational challenges is also addressed. The present document is intended to form the basis for the scientific content of the final PRELIDA Workshop and is a stepping stone towards the preparation of a detailed final roadmap at the end of the project.

Table of Contents

Executive Summary

Rationales of the deliverable are to offer a description of use cases related to the long-term preservation and access to Linked Data, and then identify and analyse challenges, problems and limitations of existing preservation approaches when applied on Linked Data. Based on the analysis of these limitations, solutions for identified technical issues are proposed. Technical issues are related with organizational issues and best practices for Digital Preservation of Linked Data, and these are presented as well. This can be considered as the first step towards describing a detailed roadmap that will lead to efficient digital preservation of Linked Data. The present document is intended to form the basis for the scientific content of the final PRELIDA Workshop and is a stepping stone towards the preparation of a detailed final roadmap at the end of the project.

1 Introduction

1.1 Rationale

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. Following the gap analysis, the second objective is to propose a roadmap for dealing with Linked (Open) Data preservation. This deliverable can be considered as the first step towards describing this roadmap that will bridge this gap and will lead to efficient digital preservation of Linked Data.

1.2 Purpose of the roadmap

The aims of the roadmap are to:

- Enumerate all *peculiarities* of Linked Data compared documents and other type of data such as Web data, multimedia and software using use cases. Based on the Gap analysis report (PRELIDA deliverable D.4.1[1]), several issues were identified, mainly that Linked Data is dynamic and distributed, often depended on external datasets requiring coordination of several stakeholders. In addition reasoning capabilities and often querying capabilities (e.g., SPARQL endpoints) must also be preserved.
- Examine and propose possible solutions to technical problems related to *OAIS compliance*. Duties of stakeholders, ingestion of archived datasets and managing of changes are important issues here. Changes to a dataset can be direct (i.e., modification of data), or indirect (change in representation standards, external vocabularies, storage hardware and software such as reasoners). The archiving mechanism should deal with all the above issues.
- Examine and propose solutions related to best practices for digital preservation. Scope of preservation, stakeholders and their responsibilities are not strictly technical issues but they are highly relevant to digital preservation as well.

1.3 Structure of the report

This document consists of the following parts: Section 2 consists of a description of digital preservation standards, with particular emphasis on the OAIS (Reference Model for an Open Archival Information System) framework and Linked Data. Section 3 consists of an analysis of uses cases, which will be used to illustrate clearly the challenges that Linked Data and Digital Preservation communities will face when trying to achieve efficient preservation of Linked Data. Section 4 consists of a description of technical challenges and solutions related to Linked Data preservation. These challenges concern: (a) compliance with OAIS model and the corresponding responsibilities, (b) dataset ingestion and (c) managing changes. Summary and conclusions are the last parts of this deliverable.

[1] PRELIDA Deliverable D4.1 Analysis of the limitations of Digital Preservation solutions for reserving Linked Data. Available from the PRELIDA web site: prelida.eu

2 Background and related work

This report aims to identify challenges arising when digital preservation is applied on Linked Data and propose a roadmap for addressing them. 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. A more detailed description of the above topics is provided in the corresponding PRELIDA project deliverable "D3.1 State of the art"[2], but a short description is provided here to make the document more self-contained.

2.1 Digital Preservation

Digital preservation can be defined as activities ensuring access to digital objects (data and software) in the long term. In addition to that, preserved content must be authenticated and rendered properly upon request. 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 Trusted Digital Repositories (TDR) that are based on the OAIS reference model.

2.1.1 OAIS Reference model

Standardization requirements for Digital preservation led to the adoption of the OAIS reference model for the corresponding tasks. 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. The OAIS 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 Open Archival Information System reference model (OAIS) is an ISO standard (ISO 14721) that provides fundamental concepts for preservation and fundamental definitions so people can speak without confusion. The OAIS reference model has been developed under the direction of the "Consultative Committee for Space Data Systems" (CCSDS) and was 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

[2] http://www.prelida.eu/sites/default/files/D3.1%20State%20of%20the%20art.pdf

design of system architectures, but it does not ensure consistency or interoperability between implementations.

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.
- 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 which are fundamental to understand and authenticate a piece of digitally encoded information. The OAIS model is the basis against which procedures of certification are set up, which in turn determines if a digital archive can claim to be a Trusted Digital Repository. The key elements for preservation are: Trust, Authentication and Sustainability.

2.2 Linked (Open) Data

Data traditionally was considered to be a closed asset, but today it is considered to be a critical resource. 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. Opening data in addition to saving and processing it internally, can lead to the creation of businesses using this open data to create new value and services. This may create new revenues to states and corporations and is part of the developing of the so called *data economy*. On the other hand the loss of control enforced by the processing of requests comes at a cost: the data which that is made open can, and probably will, be used in unexpected ways. Furthermore it can be combined with other datasets and interpreted in a non-standard 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 datasets are made available in an open license and they are uploaded and/or referenced. What all these portals have in common is that they allow end users to download entire datasets or parts of datasets. 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. 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[3] and can be summarized as:

- Use the Web as a platform to publish and re-use identifiers that refer to data, and
- Use a standard data model for expressing the data (RDF).

The Resource Description Framework[4] (RDF) is a way to model data as a list of statements made between two resources identified by their unique identifiers (URI). 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[5] (TTL), TriG[6], RDF/XML[7], and RDFa[8] 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 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[9]. This is similar 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 is used for "small data" (e.g. social profile on a home-page) whereas the database approach is adopted for "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. 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 (Web-based and database-based) 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. 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. In order to preserve this, the LD 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.

[3] http://www.w3.org/DesignIssues/LinkedData.html

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

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

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

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

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

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

2.3 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 "online" Web of data faces the same problems as web archiving. But there are more challenges when the semantics and the overlap between these two facets of Linked Data are considered. These challenges will be studied in section 4.

3 Use cases and gap analysis

3.1 Use cases

Analyzing specific use cases is an important step towards identifying technical organizational and economic challenges on digital preservation of Linked (Open) Data. In the following two use cases, DBpedia and Europeana, will be presented in order to identify Linked Open Data preservation issues. DBpedia a crucial use case because it is a core part of the LOD cloud being its most referenced node. The Europeana project on cultural heritage preservation is also an important use case, as it involves the preservation of metadata from different, independently maintained sources such as museums and libraries. Both use cases were presented in detail in PRELIDA deliverable 4.1[10] and they were analyzed at PRELIDA midterm workshop (Deliverable 2.5[11]).

Examples of projects in which the Linked Data paradigm is put into practice 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. The DIACHRON project[12] is a highly relevant research effort towards this direction. Auer et al. [4] identifies main issues related to LOD preservation for different use case categories, namely Open Data Markets, Enterprise Data Intranets, and Scientific Information Systems. These use cases will be discussed at the final PRELIDA workshop and are expected to be included in the final roadmap delivered at the end of PRELIDA.

3.1.1 Digital Preservation for 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 can be retrieved directly, be it through a SPARQL end-point or as Web pages. Knowledge

[10] http://prelida.eu/sites/default/files/PRELIDA-D4.1.pdf

[11] http://www.prelida.eu/sites/default/files/D2.5.pdf

[12] http://www.diachron-fp7.eu/

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.

DBpedia archiving is currently handled by the DBpedia association[13] itself and not by an external organization. Since DBpedia data is extracted from Wikipedia data and is transformed to RDF format, these two organizations 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 License (GFDL) [14]. DBpedia content (data and metadata such as the DBpedia ontology) is available to end users under the same terms and licenses as the Wikipedia content.

DBpedia preserves different versions of the entire dataset by means of DBpedia dumps corresponding to a versioning mechanism[15]. Besides the archived versions of DBpedia, DBpedia live[16] keeps track of changes in Wikipedia, and extracts newly changed information from Wikipedia infoboxes and text into RDF format[17]. 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 archived[18].

DBpedia dataset contains links to other datasets containing both definitions and data (e.g., Geonames). 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.

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.

Use case 1: RDF Data archiving and retrieval

DBpedia data (in RDF format, or Tables-CSV format) are archived and the user requests specific data (or the entire dataset) as it was at a specific date in the past, e.g., the RDF description of topic Olympic games at 1/1/2010. The preservation mechanism must be able to provide the requested data in RDF (or Table) format. Retrieving data for a specific time interval, e.g., 2010-2014, 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.

---

[13] http://wiki.dbpedia.org/Association

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

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

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

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

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

Use case 2: Rendering data as Web page

The user requests the DBpedia data for a specific topic at a given temporal point or interval as in Use case 1, but rendered as web page. 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 valid for, 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?

Use case 3: SPARQL Endpoint functionality

The main requirement here is to reconstruct the functionality of the DBpedia SPARQL endpoint at a specific temporal point in the past. There are different kinds of queries that must be handled corresponding to different use cases:

a) Queries spanning across RDF data into DBpedia dataset only
b) Queries spanning across DBpedia dataset and datasets directly connected to the DBpedia RDF dataset (e.g., Geonames)
c) Queries spanning across 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) or (c) when the content of external links is requested.

3.1.2 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.

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. As this metadata is stored by Europeana, Europeana has no specific requirement for specific metadata preservation policies on the provider's side. 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.

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[19], Geonames[20] and

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

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.

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 up 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.

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[21], 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. 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. 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 processes 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 for this objective. It must be trustable by data consumers. However, as noted, Europeana

[20] http://geonames.org

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

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?

3.2 Gap analysis

This section provides a summary of the deliverable D4.1 "Analysis of the limitations of Digital Preservation solutions for preserving Linked Data" [12]. The first step in gap analysis was to identify the peculiarities of Linked Open Data when compared to digital objects and other forms of data that typically are handled by digital preservation systems. This is crucial for preparing a roadmap towards efficient solutions for Linked Data preservation. Classification of Linked Data was 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 [1] 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): Different statements may be made at any time and so the "boundary" of the object under consideration changes in time.
- 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 such as the truth value of a statement.
- Non-rendered: Non-rendered digital objects need to be processed to produce any number of possible outputs. Typically Linked Data is not rendered and adopts standards, such as RDF, that are open, widely adopted and well supported.
- Passive: The Linked Data is usually represented in the form of statements or objects (typically RDF triples) which are not applications. Also, besides preserving data, software that handles data should be preserved in some cases, such as a SPARQL endpoint.

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.
- Authenticity and provenance is a major issue in preservation, further complicated by the fact that LOD are distributed and typically not centrally controlled.
- 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 in general) and 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).

- 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. 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.
- Linked Data depends on the web infrastructure and in particular on the de-referenciation of HTTP URIs. With respect to this issue all projects addressing link rot and content rot are relevant.
- Linked Data is 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 serialized using different formats such as RDF/XML, N3, Turtle and JSON-LD. For each form its durability can be assessed.
- 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.
- 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 standard way to express metadata about RDF triples. Labelling, named RDF graphs and various forms of reification (e.g., N-ary approach[22]) have been proposed for addressing this issue.

Based on the questions raised in the previous section several issues and problems were identified in D4.1:

- *Selection*: Which LOD data should actively be preserved?
- *Responsibility*: Who is responsible for "community" data, such as DBpedia?
- *Durability of the format*: Which formats can we distinguish? RDF, Triple Store, Software, SPARQL, etc. Can we make a classification?
- *Rights / ownership / licenses*: LOD is by definition open (which is not always the case for LD in general), but how to preserve privacy then?
- *Storage*: Highest quality is storage in "Trusted Digital Repository". But which other models can be used?
- *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.

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

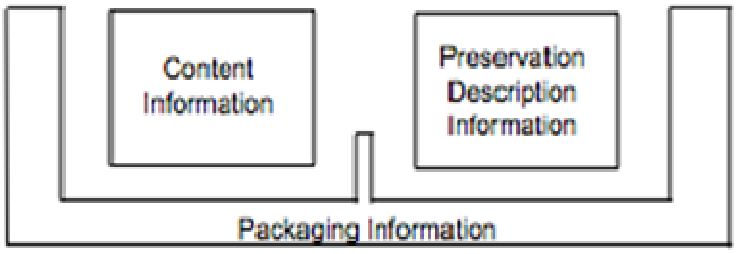
4 Technical challenges

This section deals with several issues related to digital preservation of Linked Data identified using the use cases of previous section. Specifically these technical issues are related to (a) ingesting a Linked Data dataset, and (b) managing the changes that can impact on a Linked Data dataset[23]. In what follows, by "Linked Data dataset" (LDD for short) we mean a 5 star dataset, that is one expressed in RDF with links to a significant number of other web resources, including datasets but also web pages, documents, and in general anything that can be identified by an HTTP IRI . Concept definitions (ontologies) expressed in OWL are also covered in this section. By making this choice we place ourselves in the most general and technically challenging case.

4.1 Oasis model compliance

According to the Reference Model, "an OAIS is an archive, consisting of an organization, which may be part of a larger organization, of people and systems that has accepted the responsibility to preserve information and make it available for a designated community." In light of this, one of the goals of PRELIDA is to discuss how the concepts and functions introduced by OAIS can be used for the preservation of Linked Data.

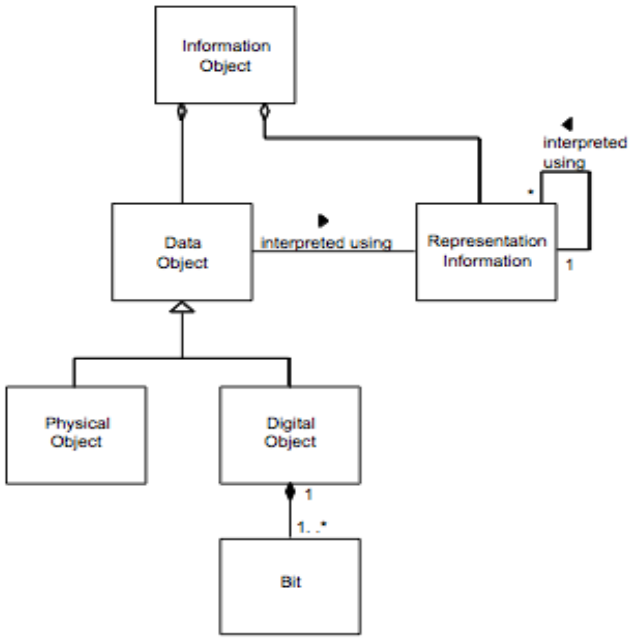
A brief description of OAIS model is presented in section 2.1.1 and a more detailed description is provided in deliverable D3.1 (State of the art) of PRELIDA project. An archived information package consists of content and preservation description information.



OAIS Information package

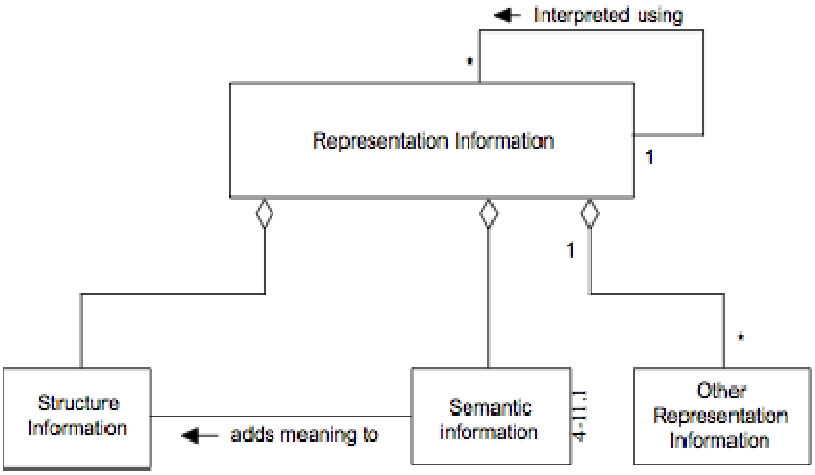
The Content Information contains the data to be preserved, whereas the Preservation Description Information includes various types of knowledge required (or rather, recommended) for the preservation of the content. The Content Information is in turn structured as an information object:

[23] https://webfoundation.org/2011/11/5-star-open-data-initiatives/



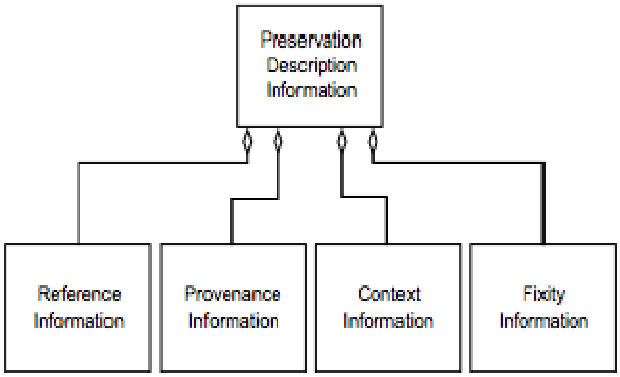
OAIS Content Information Package

where the Data Object is the data to be preserved and Representation Information is information "needed to make the Content Data Object understandable to the Designated Community". Representation Information is composed of various parts:



OAIS Representation Information

The other part of an OAIS Information Package is given by Preservation Description Information (PDI for short). PDI is structured in the OAIS Information Model as follows:



OAIS Preservation Description Information

One observation that was made from an archival point of view, was that the notion of designated communities only helps partially - because by default we cannot know what communities in the future might be interested in the archived material; an archive partly relies on current requests of communities, partly it relies on the gut feeling of archivists, and there will be always an arbitrary element in archiving. This may be even more sensitive in the case of Linked Data, which are typically created with an idea of sharing in mind, and as such tend to be less community - specific, typically by crossing community barriers and by linking to popular datasets, which concretely means by using popular URIs for identifying the resources they are about. Indeed, Linked Data shifts away from the paradigm of self -containedness often assumed by archives.

By its nature, Linked Data refer to other resources outside the graph they belong, so that when archiving one graph, one has to decide what to do with the links going out of the graph. The question then naturally arises, what to do with the links. This question is a special case of a more general question concerning how to make a Submission Information Package (SIP) out of an LDD. An LDD is not an OAIS, it is, at best, just the content part of an OAIS. What do we need to add to an LDD such that it can be accepted by an OAIS as a SIP is the issue covered in the following section.

4.2 Ingesting a LD dataset

This section deals with the problem of how to construct a Submission Information Package (SIP) that includes, as Content Data, an LD dataset.

4.2.1 Self-containedness

Ideally, an archive would like to ingest a self-contained SIP, thus avoiding any dependence on resources that are out of the archive's control. In the case of an LD dataset, this means to obtain, for instance by crawling the web, all the resources that the LD dataset links to (via IRIs).

- Let us consider RDF resources first. This ingest-all strategy seems to be feasible as far as vocabularies and ontologies (typically in OWL) are concerned, since their size is not prohibitive and there is a limited number of them[24]. However, it is doubtful whether it can be

[24] For example http://lov.okfn.org/dataset/lov/ stores many vocabularies, but the file gathering them all is only 8.4 megabytes, 64740 triples.

applied to any RDF resource, because it exposes the archive to the risk of archiving a large portion of the LOD cloud[25]. Here, a boundary has to be set in the context of the negotiation between the data producer and the archive. The notion of boundary, however, needs to be clearly defined, and the terms to describe it need to be established, so that a common practice can be created.

- For all other kinds of resources, there exist the same boundary problem, and in addition there is the problem of digital rights management (DRM), due to the fact that these resources may not be freely accessible. Again, then, the boundaries of the SIP need to be defined in the negotiation phase. One alternative solution to ingestion could be to rely on web archives (such as for instance the Internet Archive) for time-based access to non-RDF resources.

4.2.2 Serialization

Once the content data of the SIP are defined, a serialization format has to be chosen for the LD dataset being ingested and the related RDF resources. It has been suggested that RDF 1.1 n-quads[26] [5] are good candidates for the serialization of RDF for archiving purposes. The specifications of the chosen serialization format are also to be ingested (or referenced to) in the SIP as Structure Information, which is part of the Representation Information of the SIP.

Structure Information is given by (definition of) the serialization format. RDF is a data model, there are many serializations, all Unicode based (RDF/XML, RDFa in HTML, Turtle, etc). RDF serializations are mostly interchangeable (although named graphs in RDF/XML require tricks[27] and JSON-LD[28] may not cover everything), and there is no evidence that some serializations are better than others. PRELIDA will seek to establish a contact with the Data Best Practices W3C group[29], asking the group if they can make a recommendation on serialization for archive, or make all serializations kept fully compatible over time.

4.2.3 LD Dataset Description

The SIP needs to have a description of the content data, and for this VoID[30], DCAT[31] and PROV[32] have been proposed as suitable vocabularies for describing (respectively):

- general metadata based on Dublin Core, access metadata, structural metadata, and links between datasets

[25] State of the LOD Cloud in 2014 (Retrieved in September 2014): http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/

[26] http://www.w3.org/TR/n-quads/

[27] https://dvcs.w3.org/hg/rdf/raw-file/default/rdf-dataset/index.html

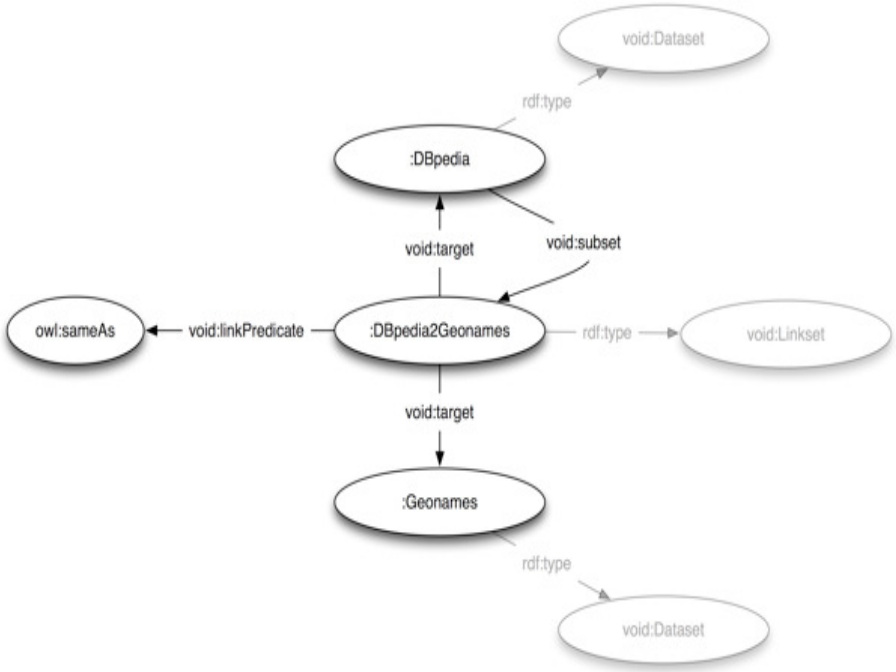
[28] http://www.w3.org/TR/json-ld/

[29] http://www.w3.org/2013/dwbp/wiki/Main_Page

[30] http://www.w3.org/TR/void/

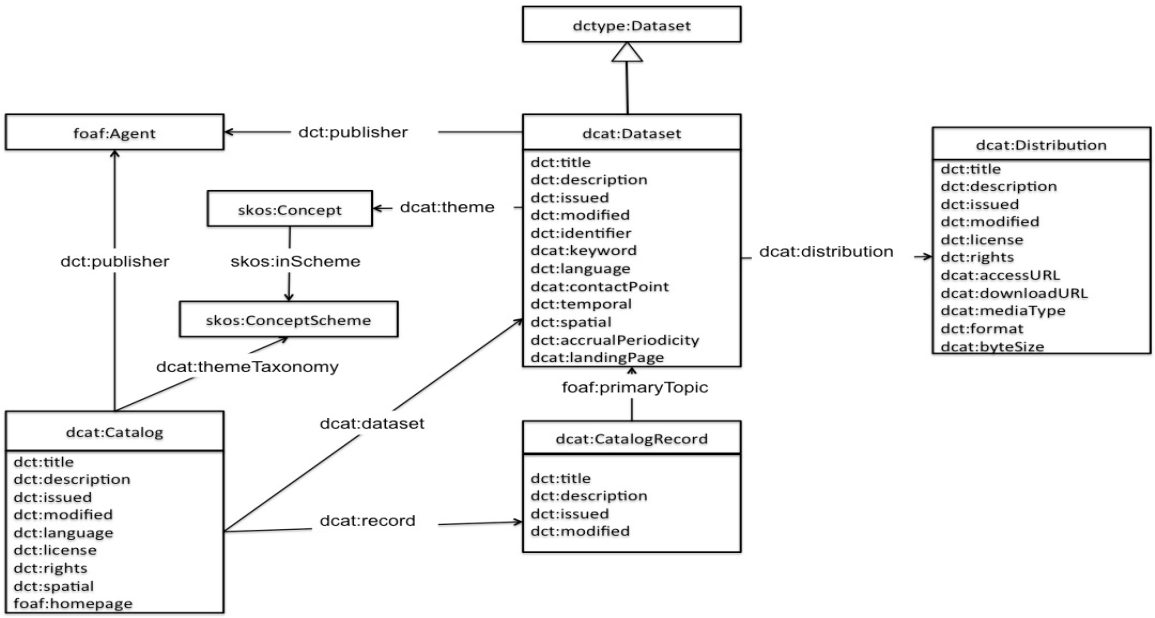
[31] http://www.w3.org/TR/vocab-dcat/

[32] http://www.w3.org/TR/prov-o/



Linked Dataset description example using VoiD (source:W3C[33])

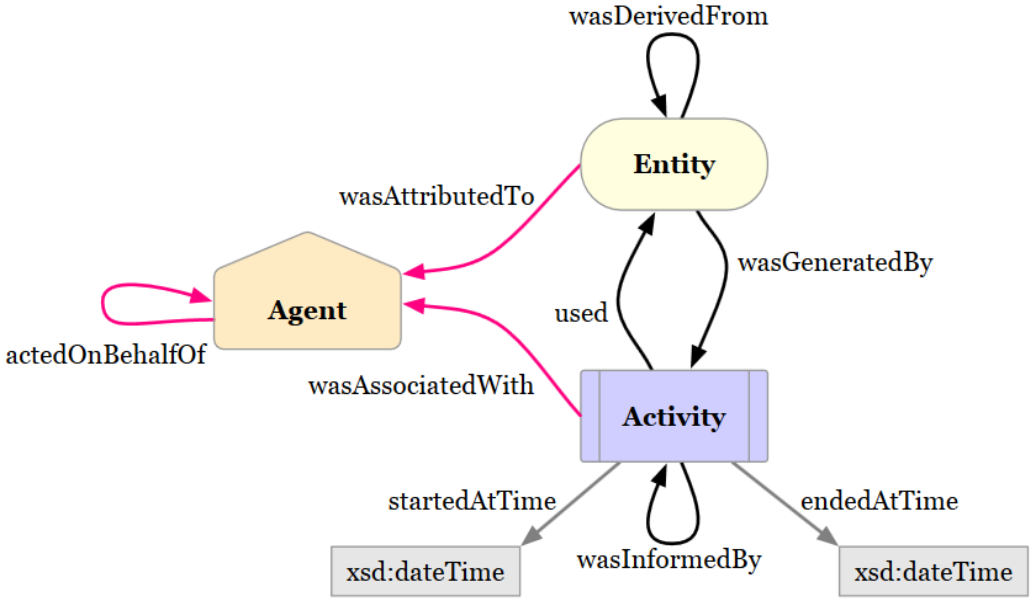
- the LD dataset in data catalogs



Data catalog vocabulary example using DCAT (source:W3C)

[33] http://www.w3.org/TR/void/

- provenance information



Data provenance description example using PROV vocabulary (source:W3C)

This information must be provided by the data producer and validated by the archive upon receiving the SIP. It belongs to Representation Information (VoID and DCAT) and to Preservation Description Information-PDI (PROV). The PROV vocabulary[34] is recommended by the W3C for expressing Provenance Information. PROV is designed precisely to represent how the RDF was made, what the history of this dataset is before and after ingest. PROV is about documenting the provenance of an object, not about offering a metamodelling mechanism.

4.2.4 Reasoner preservation

Semantic Information consists of two parts: the semantics of RDF[35] plus the semantics of the specific RDF vocabulary the graph is built on. The former is archived by the W3C in form of documents containing the various recommendations. The latter is given by the vocabularies (or ontologies, or terminologies) referred to by the Data Object, typically in OWL format. Based on these reasoning can be applied to the LD.

Part of the preservation strategy must consider the preservation of related application software. Although preservation of the software itself is not a strict requirement since the specific software (especially if it is widely used), may be preserved anyway elsewhere, information about the specific reasoners used along with the preserved data and their version must be part of the archived information. Related software can be OWL/RDF reasoners[36] or SPARQL endpoints. This is also very important for OWL ontologies, since reasoners are equally important to the ontologies themselves.

[34] Provenance Working Group. The PROV Namespace. W3C Document 19 May 2013. http://www.w3.org/ns/prov

[35] http://www.w3.org/TR/rdf11-mt/

[36] http://www.w3.org/2001/sw/wiki/OWL/Implementations

Since OWL vocabularies are also required for reasoning, semantics of OWL and metadata about the specific version used must be preserved as well. Specifically, description of OWL version or OWL profile used for concept description and reasoning are important here[37].

4.3 Managing changes

A crucial aspect of preservation is to keep the preserved data always accessible and usable by the Designated Community, as established by the OAIS Reference Model. In order to achieve this goal, an OAIS needs to take appropriate actions to contrast the changes that time brings to:

(1) the technological architecture that supports the archival and access to the data

(2) the ontological architecture that underlies the Representation Information and the Preservation Description Information associated with the preserved content.

In what follows we will review the types of changes that may affect the preservation of an LD dataset, discussing for each type what kind of actions is required.

4.3.1. Changes to the technology used by the archive to preserve the data

Description

An OAIS is based on a computerized information system, which is a complex technological artifact, supported by several hardware and software components. Any of these components may malfunction or may become obsolete and may therefore require to be replaced.

Example

The hard disks used by the archive go out of order, or a file format that was in use in the archive is no longer supported.

Responsibility

It is the responsibility of the archive to monitor such changes, and take actions (such as migration of the data to a new format or to a new medium) in order to make sure the data remain accessible. In case of an LDD, the selection of a new format to which the dataset must be migrated, must be based on the recommendations from the W3C.

Status

This is a core topic in digital preservation, and the results obtained so far provide an archive with solid methods and tools for dealing with this kind of problems [13]. The application of these methods and tools to LDDs does not pose any additional problem.

Required technology/standards

Standard preservation practices are adequate here. W3C archiving recommendations (regarding format, compatibility and data migration) will be required.

[37] http://www.w3.org/standards/techs/owl#w3c_all

4.3.2 Changes to the Content Data being preserved

Description

The preserved data are an image of an information system that is currently in use by the holding institution, and as a result of this usage, the data in the information system change, meaning that some element is deleted, or updated, or that a new data element is created.

Example

The DBPedia LDD is continuously updated by the addition of new triples.

Status

Existing mechanisms and policies are adequate for internal data, but links to external datasets are not handled by existing archives (e.g., DBpedia).

Responsibility

This type of change is rather uncontroversial from the preservation point of view: when the owner of the data decides that the changes are significant enough, a new snapshot of the data is taken by re-ingesting the Content Data to the archive. The archive's sole responsibility is to possibly keep track of the versioning relationships that exist between the different snapshots taken in time from the same Dataset.

Required technology/standards

Standard mechanisms and policies are adequate for internal data, but links to external datasets introduce a problem here [12]. Solutions based on crawling (and research on technical issues such as refresh rate, crawling frontier) can be put in place. Alternatively a mechanism for propagating changes and notifying corresponding archives may be deployed as well.

4.3.3 Changes to the Representation Information or to the Preservation Description Information

Description

Representation Information and Preservation Description Information are recommended by the OAIS Reference Model to be added to the Content Data for preservation purposes. This information may change, either because the holding archive updates them (see first example below) or because some event outside the holding archive requires a change to them (second example).

Examples

(1) The serialization format of the preserved LDD becomes obsolete, and the holding archive migrates the data to the newly recommended format. The new format must be recorded in the Representation Information, and the migration has to be recorded in the provenance section of the PDI.

(2) The organization producing the data goes out of business, and the responsibility of the data is transferred to a different organization. This change needs to be reflected in the Context Information section of the PDI of the preserved LDD.

Responsibility

This case is similar to the previous one, in that a new Submission Information Package is created which must be ingested and properly related to the one it is a new version of. But in case the change generates from inside the archive, there is the preliminary step by the OAIS to alert the producer of the data of the need to record the change and negotiate with them the new submission.

Status

Serialization formats are defined by W3C and are standardized. Detection of changes is an open problem.

Required technology/standards

Similar as the previous case, standard policies and recommendations for detecting changes (crawling strategies or notification mechanism) must be defined.

4.3.4 Changes to the vocabularies used in the LDD or to the additional information stored with it.

Description

In the preservation of LD, both the Content and the additional information stored for preservation purposes (Representation Information and Preservation Description Information) are expressed in terms of vocabularies that may change any time, due to the addition of new terms (and of the involved axioms) or to the deprecation of old terms. In this case, the data being preserved do not change directly, but the change to the vocabularies may have an influence on their semanitcs, making some statements obsolete or false.

Examples

(1) As a consequence of political evolution, East Germany and West Germany no longer exist because they are (re)united into Germany. In order to reflect this new situation, the gazetteer that was in use in the archive is updated by the authority maintaining it: a new term for Germany is introduced, whereas the old terms for East and West Germany are deprecated. The statement that the content data was generated in East Germany is part of context in the Preservation Description Information, and must be updated because the term “East Germany” used in it will soon be obsolete and the Designated Community will no longer understand it.

(2) As a consequence of scientific discovery, the definition of planet has changed and what was so far classified as a planet may no longer be so. In order to reflect this new situation, the ontology of astrophysics that was in use in the preserved LDD is updated by the authority maintaining it: a new term for planet is introduced and properly axiomatized, whereas the old term is deprecated. The statement that the Content Data is about a planet is part of the Representation Information and needs to be retracted because, according the new meaning of planet, it is no longer true.

Responsibility

This case is tackled by the joint action of the archive and the data holder. The archive should have in place a mechanism to monitor the vocabularies of the preserved LD and, whenever a change occurs to one such vocabulary, the archive should alert the data holder that some action is required. The action though rests with the data holder who is in the best position to decide how to change the data and when to re-submit it.

Status

Preserving external vocabularies is not handled by existing LOD archives.

Required technology/standards

Detecting changes or being notified as in previous cases is an option. Since vocabularies are usually small in size, a more aggressive crawling strategy (i.e., frequent ingestion of all related vocabularies) than in other cases can be the standard practice. Also the definition, perhaps by W3C, of a set of centrally preserved core vocabularies is also an option.

4.3.5 Changes to web resources other than those discussed here.

Description

In the preserved LDD there may be URIs referring web resources, that is information resources that have a representation on the web, accessible via the HTTP protocol, other than those discussed so far. These resources may disappear or change their state at any time, and, as a consequence, the reference in the preserved data may no longer reflect the creator's intention.

Examples

(1) The preserved LDD are astrophysical data that contain the URL of an image, and the image goes offline after a few years. As a consequence, the preserved LDD has a dangling reference.

(2) The Representation Information of the same astrophysical LDD refers to a PDF document describing some important characteristics of the preserved data. The PDF document was online at ingestion time, but after a few years the organization maintaining it changes their access right policy and the document is put behind a billing service. As a consequence, it is no longer accessible in the same modality.

Responsibility

This case is tackled by the joint action of the archive and the data holder, as in the case of external RDF datasets and vocabularies.

Status

This problem is similar to the Web archiving problem and a complete technical solution for all cases is not considered to be feasible. Nevertheless partial solutions are feasible.

Required technology/standards

For these, web archiving solutions have been indicated. As an alternative, the solutions proposed by projects such as Memento [3] dealing with archiving of different versions of Web resources (since similar mechanisms are required for archiving different versions of LD), can be adopted.

4.3.6 Changes to the knowledge base of the designated community.

Description

In an OAIS, the role of the Designated Community is central. In particular, a piece of information is considered by the OAIS as usable if it can be understood based on the knowledge base of the Designated Community. In fact, the knowledge base of the Designated Community forms the basis on which the whole knowledge structure of the preserved information relies. This knowledge base is hardly expressed in a formal way in a single structure. In most of the cases it is distributed amongst textbooks and papers, and its language may vary from entirely informal to formal, and may include pictures and diagrams. As any kind of knowledge, also the knowledge base of the Designated Community is subject to change, due to changes in the domain of discourse, or to change in the knowledge of the domain of discourse.

Examples

The term planet has acquired a new meaning as described above, but in this case there is no formal ontology defining it in a formal way; the term is only defined in the textbooks of the designated community and directly used, e.g. in some Representation Information. This case is similar to the previous one with the difference that there is no ontology to be updated: this fact simplifies one aspect, but leaves the same propagation problem as in the previous case. Additionally, the detection problem becomes somehow harder: the change in the knowledge base may go unnoticed for some time, since there is no digital representation of it.

Responsibility

This case is tackled by the joint action of the archive and the data holder, as in the previous case.

Status

This problem is a case of the Web archiving problem and a complete technical solution for all cases is not considered to be feasible.

Required technology/standards

Crawling strategies are more complex here since resources to be crawled are more (and size can be considerably bigger, i.e., videos instead of RDF files). This case is similar to Web archiving problem and some recommendations and best practices can be defined but a complete technical solution for all cases is not considered to be feasible.

4.4 Dealing with changes

The change management problem that poses some challenges concerns the propagation of ontology change to the archived descriptions (Representation Information or Preservation Description Information) that contain it. Of course there are lightweight approaches to coping with these changes. For instance, an archive may just add to the Representation Information or to the Preservation Description Information a reference to some source explaining the difference between the current and the previous notions. Or, it may just indicate that there has been a change in the context (vocabulary) that may matter for the designated community. However, if an algorithmic (automated) approach is

required, the way of tackling this problem strictly depends on the requirements of the designated community. In particular:

- If the designated community requires accessing and using the preserved data on the basis of the new term, then the occurrences of the old term have to be replaced, and this implies a re-rewriting of parts of the LDD. Techniques for doing so have been researched in the context of RDF [6,7,8,9]. The re-writing operations can be distinguished in basic (e.g., insert, update or delete) and complex changes, the latter being sets of basic changes that form logical units (such as merge, split, or change of graphs). Algorithms for computing the differences between ontology versions and for translating them in re-writing operations are, amongst others, PROMPTdiff [7] or COntoDiff/CODEX [8][9]. A more general approach to concept evolution can be found in [6]. The modified data has to be re-ingested, and it is the responsibility of the archive to maintain the proper connection between the previous and the updated data.
- If the designated community requires accessing and using the preserved data on the basis of both the old and the new term, then mappings have to be created and used in the access function of the archive. This problem reduces to mapping the new vocabulary (i.e., the language including the new term(s)) to the old one, and for doing this a number of techniques developed in the last decade in the context of data integration on the web, can be employed.

Status

Technical solutions have been proposed, but standard policies and recommendations for detecting and dealing with changes must be defined.

Required technology/standards

Besides the application of existing tools mentioned above, a centrally controlled mechanism for preservation and notification of changes of core vocabularies and datasets can be defined. Alternatively a standard for LDDs can be put in place ensuring that the dataset is providing required metadata (e.g., modification information) and/or notification policies for all related organizations.

5 Recommendations

A list of desirable actions and features for a preservable LDD will be presented in the following. This list is just a starting point that is used to identify all the features an LDD archive should consider. In practice, recommendations would probably not have all these.

- Selection and appraisal of data: identify the boundaries of the LDD that has to be preserved, perhaps using a Concise Bounded Description as defined in [10]
- Gather every RDF datasets (using quads to identify RDF graphs) that are relevant for the LDD to be preserved. Default strategy is complete closure. Both for vocabularies (ontologies) and instances. There are vocabularies that describe the provenance of crawl/imports/ingests of Linked Data[38].
- Whenever an LDD is collected into a SIP, the owner of the LDD should be alerted that any change in that LDD is relevant for the collector, who is made part of a list of subscribers that

[38] See http://ldif.wbsg.de/#provenance

have to be notified of any change. For notifications when a dataset changes: ResourceSync[39] can be used (it is used for instance for DBPedia synchronization). A lightweight alternative for LDD (usingVOID) is offered by dady[40].

- Submit data in a standard serialization (such as N-quads). Consider conversion between formats .
- Include VoID/DCAT/PROV description in Representation Information. Also Resource Shape -like [11] data validation instructions. And the corresponding ontologies (DCAT and VoiD ontologies, etc.).
- Specification documents should be also preserved, (i.e., RDFS, OWL, serialization specs).
- Time-stamps for the crawls of the collected datasets. Perhaps several ones: the snapshot time, the date of last modification, etc (the snapshot time is most important).
- Reasoners and SPARQL engines (triple store) are also to be preserved for accessing purposes.
- Submit every representation (HTML+RDFa, JSON) served in content negotiation. It has to be negotiated between producer and archive, in the light of what is wanted in the Dissemination Information Package. For the HTML part we could rely on existing web archiving (e.g. Korea national library has done work on this).

Finally, it has been observed that the distributed nature of Linked Data suggests that a distributed approach may be more appropriate than a traditional one for preservation of LDD. In such an approach, an OAIS can be spread over several archives, each storing a part of the Content Data. This distributed structure would be more suitable to archive a LDD, whose references to external entities can be managed as references to other OAIS managing those entities.

Overall besides the recommendations above a formal set of recommendations by organizations such as W3C may be the outcome of PRELIDA project. This draft report can be considered as a step towards this direction. Also centrally controlled mechanisms for core vocabularies and datasets and/or a set of standards and best practices should be defined and (hopefully) adopted.

6 Conclusion

This report examines and proposes solutions to issues related to the long term preservation of Linked (Open) Data. In order to achieve the project's objective it combines 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, that 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.

[39] http://www.niso.org/workrooms/resourcesync/ (see also http://www.openarchives.org/rs/toc and http://www.openarchives.org/pmh/)

[40] https://code.google.com/p/dady/wiki/Demo

The Linked Data paradigm concerns the technology to publish, share and connect data on the web, data that has formal semantics and is 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 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 [3] dealing with archiving of different versions of Web resources are highly relevant to PRELIDA.

This deliverable describes existing standards and provides some directions towards the creation of solutions to prevent the loss of Linked Open Data by means of analyzing specific use cases. The information in this draft report will be used for defining the research agenda of the final PRELIDA workshop in October 2014. Based on the workshop outcomes a detailed final the report will be created in order to arrive at concrete solutions and approaches at the final version of the roadmap document, which will be the final deliverable of the PRELIDA project.

In order to provide deal with changes and provide solutions for the long term preservation of Linked Data the following three issues: version, fixity and responsibility should be addressed. In each of these, technological challenges not currently addressed are identified. 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. A related set of recommendations, which will be further refined after the final workshop, is presented in this document.

Versioning concerns the temporal aspect of Linked Data that requires attention as in the course of time data is enhanced, adjusted and deleted. Timestamping all ingested data and keeping track of changes, direct and indirect as suggested in section 4, is the recommendation for dealing with the dynamic aspect of LOD. 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.

Trust is a key concern 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 research cycle. Although, LOD, as any digital object can be recorded, it remains to be negotiated which ensemble of digital objects should be archived [2].

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 assumes that the meaning of a resource can be given in a local description, which is not usually the case in Linked Data. 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 labor. This issue and proposed strategies is presented in Section 4.

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 since they are actually text (typically Unicode), representing RDF triples and storing and preserving text is a known problem. As explained in detail above, it is the differentiation between LOD living on the web, of which the 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. 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.

As use cases of section 3 illustrated, current Linked Data archiving is not adequate for long term preservation, mainly because of not dealing with the archiving of linked resources and vocabularies which is crucial for preserving Linked Data and reasoning capability over them. An additional issue is the preservation of querying, rendering and reasoning software along with the data.

Initial analysis of dealing with technical challenges of Section 4 will form and PRELIDA final workshop agenda and subsequently the content of the consolidated roadmap. An initial recommendation for related best practices is presented in section 5. This may form the basis for an official set of recommendation for Linked Data preservation by W3C and related organizations. This in turn can lead to efficient long term preservation of Linked Data such as DBpedia, which is a great source of information regarding human knowledge and contemporary culture and which is not currently preserved with a view of long term preservation. The final roadmap (towards which this document can be considered as a first step) will help establishing best practices and policies and adopting technical solutions for this problem, which in turn will be of great help to future researchers among others.

Bibliography

[1] Giaretta, . Advanced digital preservation. Berlin [etc] pp 31-39: Springer, 2001.

[2]Treloar, A. Van de Sompel, H. 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

[3]Ainsworth, Scott G., et al. How much of the web is archived? Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries. ACM, 2011.

[4] 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.

[5] Gavin Carothers. RDF 1.1 N-Quads. A line-based syntax for RDF datasets. W3C Recommendation 25 February 2014. http://www.w3.org/TR/n-quads/

[6] Siarhei Bykau, John Mylopoulos, Flavio Rizzolo, Yannis Velegrakis. On Modeling and Querying Concept Evolution. Journal on Data Semantics, (1), pp. 31-55, 2012.

[7] N . F. Noy and M. A. Musen. Promptdiff: A fixed -point algorithm for comparing ontology versions. In R. Dechter and R. S. Sutton, editors, AAAI/IAAI, pages 744–750. AAAI Press / The MIT Press, 2002.

[8] M. Hartung, A. Gross, and E. Rahm. Codex: exploration of semantic changes between ontology versions. Bioinformatics, vol.28, pages 895–896, 2012.

[9] M. Hartung, A. Groß, and E. Rahm. Conto –diff: Generation of complex evolution mappings for life science ontologies. Journal of Biomedical Informatics, vol. 46, pages 15–32, 2013.

[10] Patrick Stickler. CBD-Concise Bounded Description. W3C Member Submission 3 June 2005. http://www.w3.org/Submission/CBD/

[11] Arthur Ryman. Resource Shape 2.0. W3C Member Submission 11 February 2014. http://www.w3.org/Submission/shapes/

[12] PRELIDA Deliverable D4.1 Analysis of the limitations of Digital Preservation solutions for preserving Linked Data. Available from the PRELIDA web site: prelida.eu.

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