



**Project no. 600663**

## **PRELIDA**

Preserving Linked Data  
ICT-2011.4.3: Digital Preservation

### **D4.3 Consolidated 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: D4.3  
Deliverable title: Consolidated roadmap  
Due date of deliverable: 12|2014  
Actual date of deliverable: 12|2014  
Author(s): Grigoris Antoniou, Sotiris Batsakis  
Contributors (internal): Antoine Isaac, Andrea Scharnhorst, David Giaretta  
José María García, René van Horik, Carlo Meghini  
Contributors (external): Mariano P. Consens, Yannis Stavrakas, Giorgos Flouris, Albert Meroño-Peñuela, Peter Burnhill, Mark Williams, Sławek Staworko, Ashkan Ashkpour, Christophe Gueret, Mariella Guercio

Participant(s): HUD,CNR, APA, UIBK, EUROPEANA  
Workpackage: WP4  
Workpackage title: Roadmapping the future  
Workpackage leader: HUD  
Est. person months: 6  
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 the final consolidated version of a roadmap on the long term preservation of Linked Data. 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 detailed roadmap is proposed for dealing with ingestion of Linked Datasets and changes into the dataset. Keeping track of changes and related technical, challenges are also addressed. Organizational and financial aspects of Linked Data preservation are presented as well. Another issue analysed in this report is how Linked Data can be used for digital preservation. A critical assessment of existing problems, approaches and proposals for Linked Data preservation is included, leading to a set of recommendations and a proposed research agenda.

## Table of Contents

Document Information .....	1
Abstract .....	1
Executive Summary .....	4
1 Introduction .....	5
1.1 Rationale .....	5
1.2 Purpose of the roadmap.....	5
1.3 Structure of the report .....	5
2 Background and related work .....	6
2.1 Digital Preservation.....	6
2.1.1 OAIS Reference model .....	6
2.2 Linked (Open) Data.....	8
2.3 Digital Preservation and Linked Data .....	10
3 Use cases and gap analysis.....	11
3.1 Use cases .....	11
3.1.1 Digital Preservation for DBpedia.....	11
3.1.2 Linked Data Preservation for Europeana .....	14
3.1.3 Linked Data Preservation for DIACHRON Project.....	17
3.2 Gap analysis .....	19
4 Technical challenges .....	21
4.1 OAIS model compliance .....	22
4.2 Ingesting a LD dataset.....	25
4.2.1 Self-containedness.....	25
4.2.2 Serialization.....	27
4.2.3 LD Dataset Description.....	27
4.2.4 Reasoner preservation .....	29
4.3 LD dataset changes .....	30
4.3.1 Changes to the technology used by the archive to preserve the data .....	30
4.3.2 Changes to the Content Data being preserved .....	31
4.3.3 Changes to the Representation Information or to the Preservation Description Information .....	31
4.3.4 Changes to the vocabularies used in the LDD or to the additional information stored with it. ....	32
4.3.5 Changes to web resources other than RDF/OWL .....	33
4.3.6 Changes to the knowledge base of the designated community.....	34
4.4 Dealing with changes .....	35
4.5 Dataset Evolution and Preservation .....	36
5 Organizational and Financial aspects of Linked Data preservation.....	38



6 Using Linked Data for Digital Preservation.....	41
6.1 The CEDAR use case.....	41
6.2 Privacy Aware Preservation and Linked Data .....	43
6.2.1 Privacy awareness in OAIS using ontologies .....	44
7 Assessment & Recommendations .....	48
8 Research Agenda.....	49
8.1 Defining the boundaries of LD archives .....	50
8.2 Change detection .....	50
8.3 Web archiving and LD preservation .....	51
8.3.1 Web archiving using Memento .....	52
8.3.2 Web archiving to preserve results for linked data queries .....	54
8.4 The need for refreshing OAIS in a web environment? .....	54
8.5 LD archiving and storage .....	55
9 Summary and conclusions.....	57
Bibliography.....	58



## **Executive Summary**

This deliverable aims 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 to Linked Data. In addition, usage of Linked data for preservation is examined. Based on the analysis of limitations of existing Linked Data preservation approaches, 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 report offers a detailed roadmap that will lead to effective and efficient digital preservation of Linked Data. Identified issues are analysed and a critical assessment is followed by a proposed research agenda and concluding remarks.

# 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, for example, archives, digital libraries and scientific data repositories. Following the gap analysis, the second objective is to propose a roadmap for dealing with Linked (Open) Data preservation.

## 1.2 Purpose of the roadmap

The aims of the roadmap are to:

- Enumerate all *peculiarities* of Linked Data compared to documents and other type of data such as Web data, multimedia and software using use cases. The Gap analysis report (PRELIDA deliverable D.4.1<sup>1</sup>) identifies several issues, mainly related to the dynamic and distributed nature of Linked Data, often dependent 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 or methodological problems related to *OAIS compliance* (i.e., make LOD preservation fit the OAIS framework). 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 preservation 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.
- Present a research agenda for addressing complex issues not covered by existing methodologies and tools.
- Study the use of Linked Data in support of general digital preservation solutions. Although not originally amongst the objectives of the project, this aspect has emerged during the execution of the project, especially during the workshop discussions.

## 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, ISO 14721:2012) framework and Linked Data. *The reader can skip section 2 if already familiar with the consolidated state of the art Deliverable (PRELIDA Deliverable D3.2).*

---

<sup>1</sup> PRELIDA Deliverable D4.1 Analysis of the limitations of Digital Preservation solutions for reserving Linked Data. Available from the PRELIDA web site: [prelida.eu](http://prelida.eu)

- 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 effective and efficient preservation of Linked Data.
- Section 4 consists of a description of technical challenges and possible solutions related to Linked Data preservation. These challenges concern: compliance with OAIS model and the corresponding responsibilities, dataset ingestion, managing and dealing with changes, and dataset evolution and preservation.
- Organizational issues are described in Section 5.
- Section 6 contains a detailed description of how LD can be applied to Digital preservation.
- LD preservation recommendations are presented in Section 7.
- A research agenda is proposed in Section 8.
- Summary and conclusions are the last parts of this deliverable.

## 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 are outlined. 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 “D3.2 Consolidated State of the art”<sup>2</sup>, but a short description is provided here to make the present document self-contained.

### 2.1 Digital Preservation

Digital preservation can be defined as a set of activities ensuring usability of 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 usability of digital objects. A key component of the digital preservation infrastructure is the Trusted Digital Repository (TDR) that is based on the OAIS reference model.

#### 2.1.1 OAIS Reference model

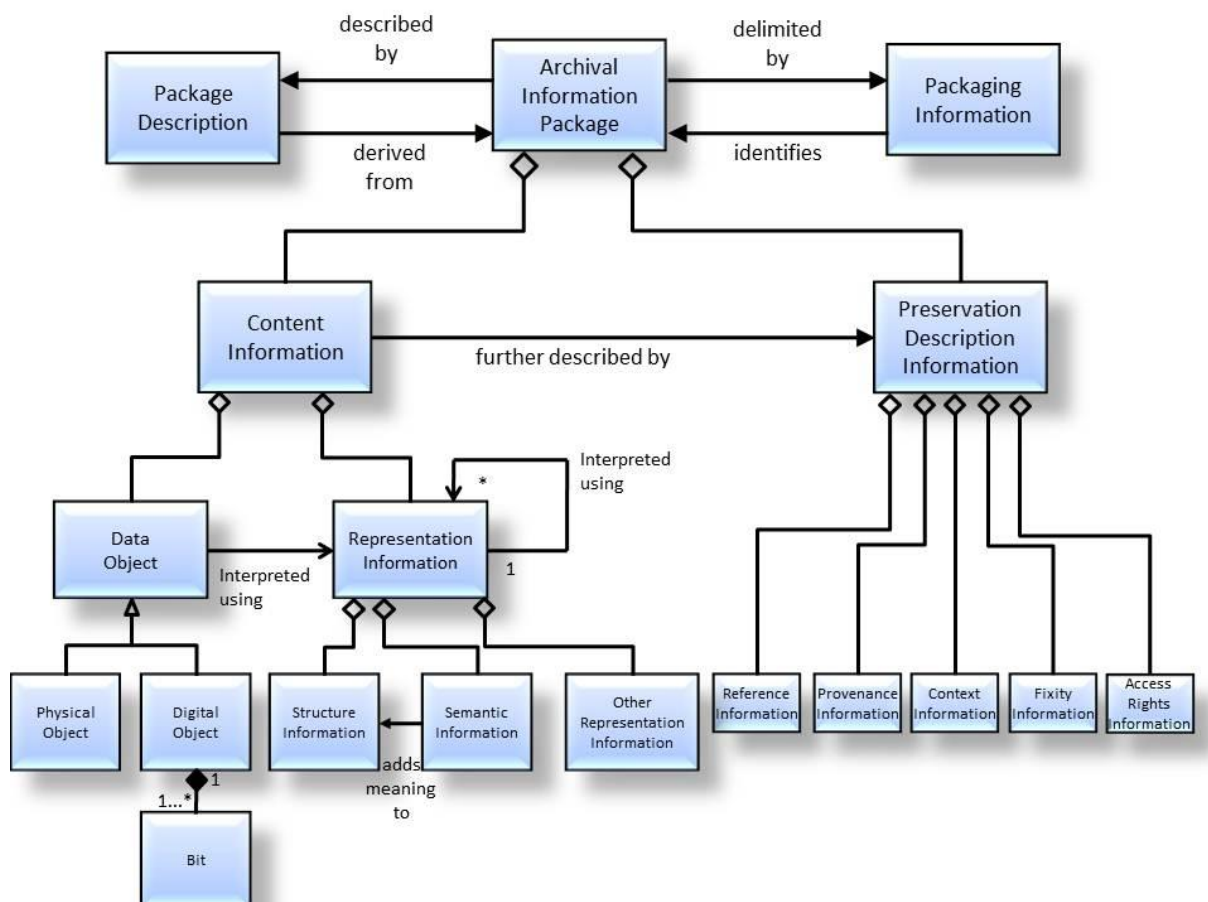
Standardization requirements for Digital preservation led to the adoption of the OAIS reference model for the corresponding tasks. The OAIS reference model (Reference Model for an Open Archival Information System) establishes a common framework of terms and concepts relevant to the long term archiving of digital data. The OAIS model details the processes around and inside 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

---

<sup>2</sup> Available on the project web site at page <http://www.prelida.eu/results/deliverables>

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



OAIS Archival information package

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 a 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 associated with 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 originally intended, thus raising for example privacy issues.



RDF example (source: W3C)

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<sup>3</sup> 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<sup>4</sup> (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<sup>5</sup> (TTL), TriG<sup>6</sup>, RDF/XML<sup>7</sup>, and RDFa<sup>8</sup> 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<sup>9</sup>. This is similar to storing relational data in a relational database and querying it using SQL.

---

<sup>3</sup> <http://www.w3.org/DesignIssues/LinkedData.html>

<sup>4</sup> See: [http://www.w3.org/standards/techs/rdf#w3c\\_all](http://www.w3.org/standards/techs/rdf#w3c_all)

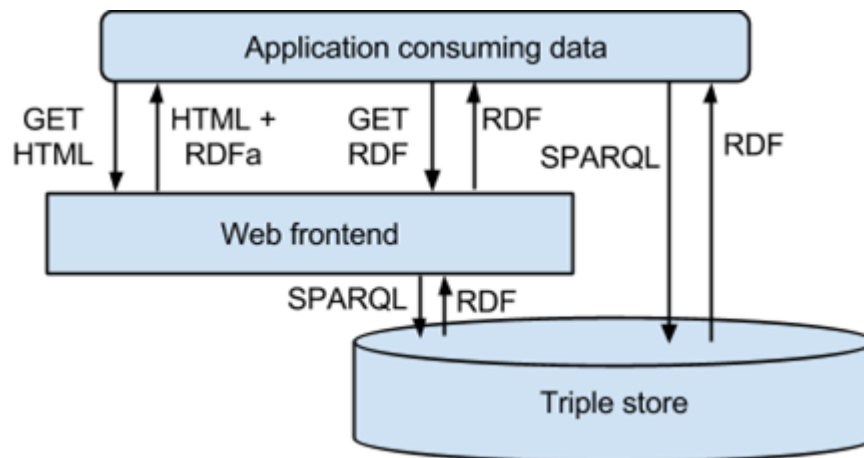
<sup>5</sup> See: <http://www.w3.org/TR/turtle/>

<sup>6</sup> See: <http://www.w3.org/TR/trig/>

<sup>7</sup> See: <http://www.w3.org/TR/REC-rdf-syntax/>

<sup>8</sup> See: <http://www.w3.org/TR/rdfa-syntax/>

<sup>9</sup> See: <http://www.w3.org/TR/rdf-sparql-query/>



Publication of RDF data

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.

## 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 it is published as structured mark-up in a web page. (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, 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.

In both cases, since, as stated above, “HTML for human consumption and RDF for machine consumption”, we are not relying on human capabilities to look at symbols on a screen, maintaining usability requires appropriate Representation Information.

An envisioned Linked Data Archive taking care of the “online” Web of data faces some of 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. Two use cases, DBpedia and Europeana, will be presented in order to identify Linked Open Data preservation issues. DBpedia provides a crucial use case because it is the core of the LOD cloud, being its most referenced node. The Europeana project on cultural heritage is also an important use case, as it involves the aggregation of metadata taken from different, independently maintained sources such as museums and libraries and subsequently processed by Europeana for normalization and enrichment. Both use cases were presented in detail in PRELIDA Deliverable D4.1<sup>10</sup> and they were analyzed at PRELIDA midterm workshop (Deliverable D2.5<sup>11</sup>).

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<sup>12</sup> is a highly relevant research effort towards this direction. Auer et al. [4] identify 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 were analyzed at the final PRELIDA workshop and are therefore included in this roadmap deliverable of PRELIDA.

#### 3.1.1 Digital Preservation for DBpedia

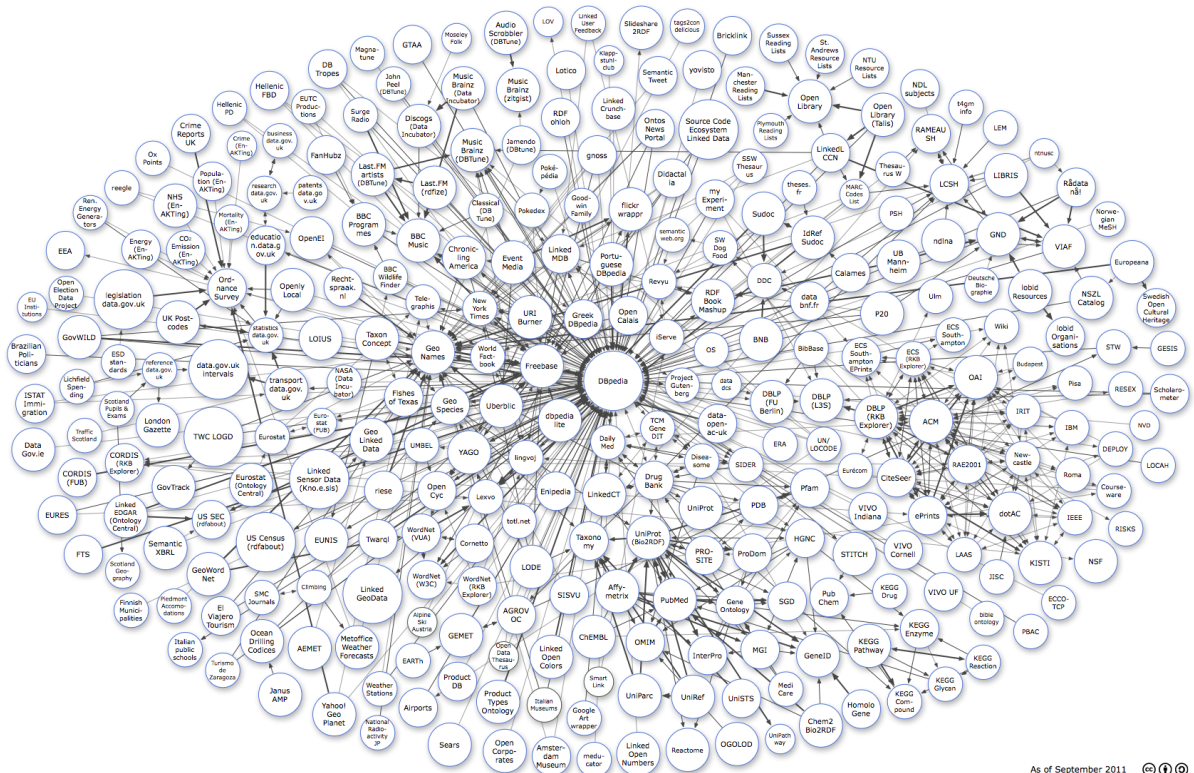
DBpedia's objective is to extract structured knowledge from Wikipedia and make it freely available on the Web using Semantic Web and Linked Data technologies. Specifically, data is extracted in RDF format and can be retrieved directly, be it through a SPARQL end-point or as Web pages. Knowledge from 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.


---

<sup>10</sup> <http://prelida.eu/sites/default/files/PRELIDA-D4.1.pdf>

<sup>11</sup> <http://www.prelida.eu/sites/default/files/D2.5.pdf>

<sup>12</sup> <http://www.diachron-fp7.eu/>



As of September 2011 

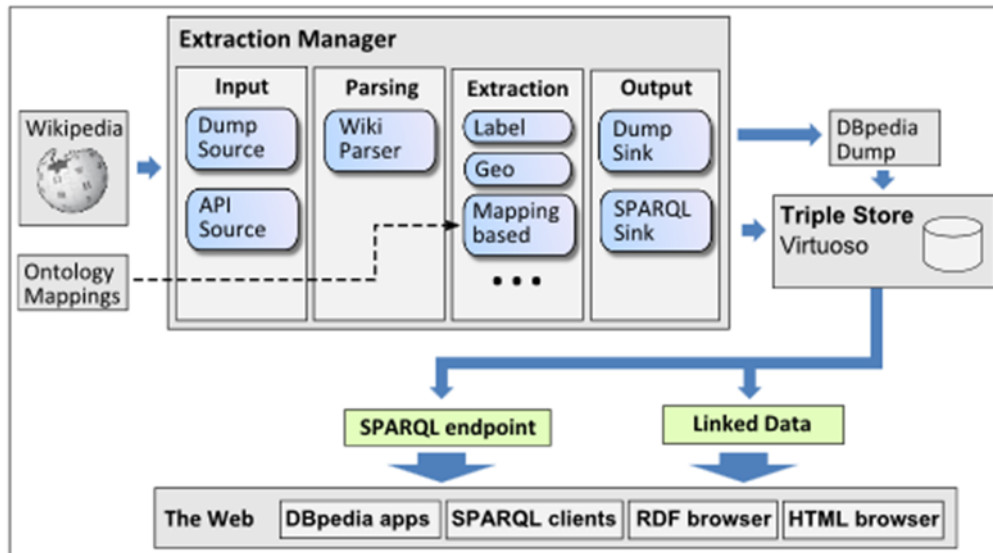
Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch<sup>13</sup>. DBpedia is the core node in this diagram.

DBpedia archiving is currently handled by the DBpedia association<sup>14</sup> 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)<sup>15</sup>. 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.

<sup>13</sup> See: <http://lod-cloud.net/>

<sup>14</sup> <http://wiki.dbpedia.org/Association>

<sup>15</sup> See: <http://en.wikipedia.org/wiki/Wikipedia:Copyrights>



DBpedia data extraction mechanism<sup>16</sup>

DBpedia preserves different versions of the entire dataset by means of DBpedia dumps corresponding to a versioning mechanism<sup>17</sup>. Besides the archived versions of DBpedia, DBpedia live<sup>18</sup> keeps track of changes in Wikipedia, and extracts newly changed information from Wikipedia infoboxes and text into RDF format<sup>19</sup>. 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<sup>20</sup>.

DBpedia dataset contains links to other datasets containing both definitions (typically Ontologies) 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 or 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.

<sup>16</sup> See: [http://svn.aksw.org/papers/2013/SWJ\\_DBpedia/public.pdf](http://svn.aksw.org/papers/2013/SWJ_DBpedia/public.pdf)

<sup>17</sup> See for example: <http://downloads.dbpedia.org/3.9/en/>

<sup>18</sup> See <http://live.dbpedia.org/>

<sup>19</sup> See for example the entry for Berlin at: <http://live.dbpedia.org/page/Berlin>

<sup>20</sup> See for example: <http://live.dbpedia.org/changesets/2014/>

#### *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<sup>21</sup> 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 an additional case, where all versions of the data and their corresponding validity intervals with respect to the request interval must be returned.

#### *Use case 2: Rendering data as a 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 a web page. The preservation mechanism should be able to return the data in RDF format, and in case the 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, since in this case DBpedia archive would have to keep archives of datasets that it doesn't import in DBpedia live SPARQL endpoint.

### **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 35M such objects.

Europeana functions as a metadata aggregator: its partner institutions or projects submit (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.

---

<sup>21</sup> [http://dbpedia.org/resource/Olympic\\_Games](http://dbpedia.org/resource/Olympic_Games)

The main source of data for Europeana is 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 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 deficiencies 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<sup>22</sup>, Geonames<sup>23</sup> 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 provide 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 mentioned before, Europeana re-distributes the metadata it aggregates from its partners in a fully open way. This is mainly done via its API, 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<sup>24</sup>, 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 tries 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 over time. 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.

---

<sup>22</sup> General Multilingual Environmental Thesaurus, <http://www.eionet.europa.eu/gemet/>

<sup>23</sup> <http://geonames.org>

<sup>24</sup> <http://data.europeana.eu>



*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, so that data providers can properly inform the data consumers. Rules for consuming the published information should also be defined, so that the entire community processes the versioning data in the same way.

*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 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 one of preserving an interconnected set of dataset views. What should be the best practices for doing this?

A project similar to Europeana related to digital preservation of cultural heritage is the Media Ecology Project<sup>25</sup> (MEP). The MEP project aims to preservation of audiovisual media content and RDF is used for allowing users to annotate and add metadata and descriptions to archived media files. RDF information can be related to provenance:

Users

\*Kemp Niver: Motion Pictures from the Library of Congress Paper Print Collection 1894-1912, 1967

---

```
<rdf:Description rdf:about="http://mep.dartmouth.edu/user/ur1617546/">
  <foaf:name>Kemp Niver</foaf:name>
  <foaf:mbox_sha1sum>b56d0fa3ea6bc6760d50c08918ffbcd623cf716</foaf:mbox_sha1sum>
  <foaf:homepage rdf:resource="http://mep.dartmouth.edu/user/ur1617546/">
  <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
</rdf:Description>
```

Person info example for MEP (source: MEP presentation, PRELIDA workshop)

---

<sup>25</sup> <http://sites.dartmouth.edu/mediaecology/>

## Commentary

A woman can be seen rolling out a pie crust in the center of a table on a set of a kitchen. A pie is cooling on the sideboard. As the woman continues her work, the door opens and a big man dressed as a tattered tramp enters and gestures a request for food. The woman replies by hitting him with a rolling pin, knocking him to the floor.

```
<rdf:Description rdf:about="http://mep.dartmouth.edu/users/john-bell/annotation">
  <rdf:type rdf:resource="http://scalar.usc.edu/2012/01/scalar-ns#Composite"/>
  <dcterms:title>The Tramp and the Muscular Cook</dcterms:title>
  <dcterms:description></dcterms:description>
  <sioc:content> A woman can be seen rolling out a pie crust in the center of a table on a set of a kitchen. A pie is cooling on the sideboard. As the woman continues her work, the door opens and a big man dressed as a tattered tramp enters and gestures a request for food. The woman replies by hitting him with a rolling pin, knocking him to the floor.
</sioc:content>
  <dcterms:isVersionOf rdf:resource="http://mep.dartmouth.edu/users/john-bell/annotation.0"/>
</rdf:Description>
```

Annotation & comments using RDF for MEP (source: MEP presentation, PRELIDA workshop)

Since MEP and Europeana have similar issues (media files and annotations), MEP is not analyzed as a separate use case but is rather mentioned as an additional example of issues related to European use case.

### 3.1.3 Linked Data Preservation for DIACHRON Project

The DIACHRON project<sup>26</sup> deals with the issue of evolution management for preservation. Three use cases are examined for this project:

- Open Governmental data, where “Data matchmaker” companies collecting data from Public sector, social networks, private sources and the web are involved. These data is multidimensional data in various formats (e.g., csv, relational, dsp)
- Open Enterprise data involving a large enterprise (Daimler Group). In this case Enterprise Structured Data - using Relational data models are preserved and LOD is employed for building rich Data Intranets
- Open Scientific data from the EMBL-EBI Bioinformatics Institute, supporting numerous biology research groups and activities. This use case involves heavily curated Biological Data (ontologies, experimental data and annotations).

#### *Open Government Data Use Case*

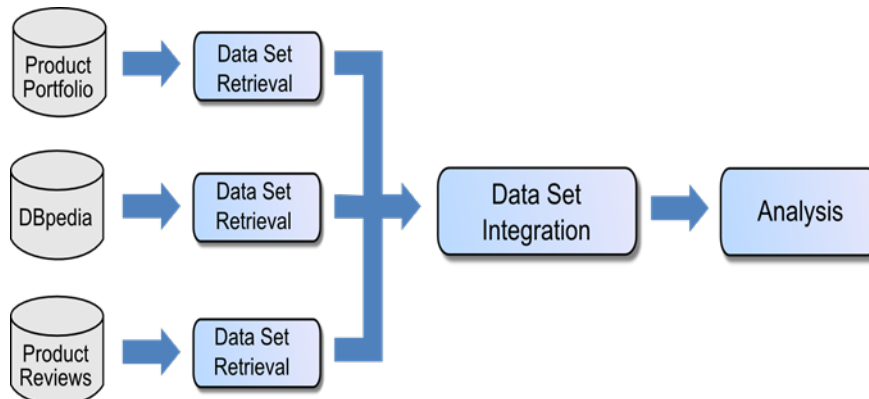
For this use case tools for combining and visualizing various socioeconomic datasets are developed. Challenges that must be addressed for this use case are (a) Update of datasets with new versions or insertion of new ones, (b) Estimation of frequency of change, (c) Detection of various types of changes on the new version of the datasets which can be: schema changes (e.g., new dimension on a multidimensional dataset), data updates, availability issues, dataset linkage evolution and data source quality assessment.

#### *Enterprise Structured Data Use Case*

---

<sup>26</sup> <http://www.diachron-fp7.eu/>

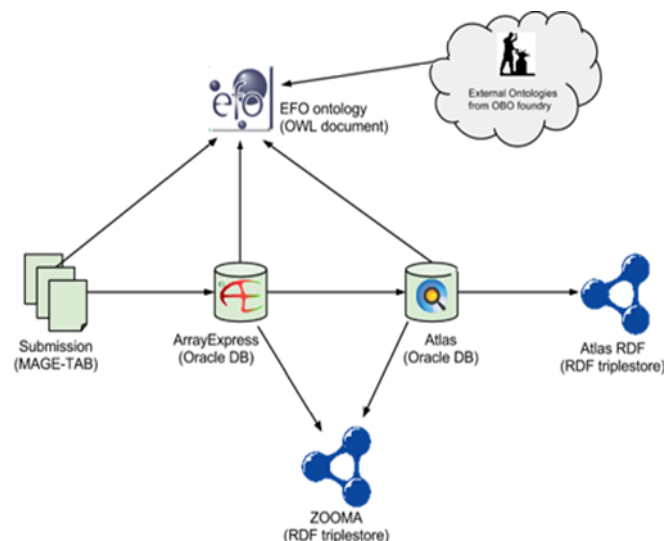
The objective in this use case is Enterprise Data integration with LOD and Web data. Datasets are (a) The Daimler Car Model Portfolio (b) Related Entities from DBpedia and (c) Product reviews from the Web or content from the enterprise Web and social channels. Use cases include dealing with (i) Evolving portfolio with Additions/Deletions and Change of (RDF) properties and (ii) Evolving content from the Web.



DIACHRON Enterprise data use case (Source: DIACHRON Project presentation at PRELIDA workshop)

### Scientific Linked Data Use Case

In this use case involving the EMBL EFO ontology<sup>27</sup>, ontology evolution and dataset annotation are the related challenges. Evolution in this case is a complex issue since analysis of the EFO ontology indicates that evolution is caused by (a) monthly releases (b) external dependencies that need to be monitored for changes and (c) repairing needs. Annotation dataset issues include detection of similar changes for auto suggestion to curators, and longitudinal queries (i.e., across versions) for visualization of dataset evolution.



DIACHRON scientific data use case (source DIACHRON presentation, PRELIDA Workshop)

<sup>27</sup> <http://www.ebi.ac.uk/efo/>

In these three use cases data can be either numerical-statistical data, which are mostly present in the Open Data Scenario or ontological-categorical data, in the other two scenarios. Data models are Multi-dimensional (online analytical processing-OLAP-Cubes) model, Entity-Relationship model and Ontologies. Also various data formats are used such as (a) Tabular data in CSV, XLS files or other tabular formats, (b) Hierarchical structured formats like XML and JSON, (c) Semi structured formats like HTML, (d) Domain specific formats like the MAGE-TAB format used for biological data submitted to EBI-EMBL and (e) RDF datasets. Although not all data is LD, evolution is an issue in all use cases, similar to the DBpedia use case. Changes can be:

- Changes in the data values. This kind of changes appear in datasets from all three pilot scenarios.
- Structural changes. These changes are dependent on the data model of each dataset.
  - Multi dimension model:
    - Addition/deletion/modification of a dimension
    - Addition/deletion/modification of a dimension values
    - Merge/split of dimensions
  - Ontological datasets present the following changes:
    - Changes in class hierarchy
    - Addition/deletion/modification in class/property
    - Addition/deletion/modification in class instance
- Changes in the metadata information of a dataset.
- Changes in the links between datasets.

This in turn leads to the identification of related challenges for DIACHRON project which are:

- Need for Preservation (Archiving)
  - Rapid evolution of data - Data change at different granularity levels
  - The Data Web changes without any notification to consumer applications
  - Reason for preservation: Cross reference, Provenance, Accountability
- “Understanding” Evolution (Evolution)
  - Structured data keep increasing on the web - Change Detection is feasible
  - Preservation can be achieved through evolution tracking
  - Changes become queryable – Longitudinal queries are feasible
  - Application unawareness
- Temporal and provenance annotation to enrich datasets (Annotation)
- Diachronic citations to ensure proper cross reference (Annotation)
- Data contain errors and dataset retrieval not always easy (Acquisition)
  - Assess the quality of the data and their sources through various metrics
  - Provide repairing and cleaning services

### 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 scheme to Linked Data yields the following:

- *LD is 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. 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.
- *LD is 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.
- *LD is 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.
- *LD is 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:

- The persistence the preserved objects depends 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, a major issue in preservation, is further complicated by the fact that LOD is distributed and typically not centrally controlled.
- LOD is 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 the various definitions and natural language terms used can be ambiguous (and formal semantics may not solve this problem if definitions are not accurate).

Additional issues are:

- 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-referencing 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 has to be assessed.
- Preservation requires the expression and recording of several kinds of metadata about the preserved object. For preserving Linked Data this means that metadata needs to 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<sup>28</sup>) 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? Notice that, openness of data (in the ‘legal’ sense) is not intrinsically related to privacy (in the usual sense of ‘personal privacy’), i.e., you can have privacy problems for closed data as well.
- *Storage*: Highest quality is storage in “Trusted Digital Repository”. But which other models can be used?
- *Metadata and Definitions*: Representation Information is required to enable the designated community to understand the meaning of LOD objects. Are LOD objects “self-descriptive”? That depends on the definition of the Designated Community (DC). If the DC is defined as *not* understanding the associated ontologies then the object that needs that Representation Information may be loosely termed “not self-descriptive”, and there is an additional preservation risk.

## 4 Technical challenges

This section deals with several issues related to digital preservation of Linked Data identified using the use cases of the previous section. In what follows, by “Linked Data dataset” (LDD for short) we mean a 5 star dataset<sup>29</sup>, 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

---

<sup>28</sup> <http://www.w3.org/TR/swbp-n-aryRelations/>

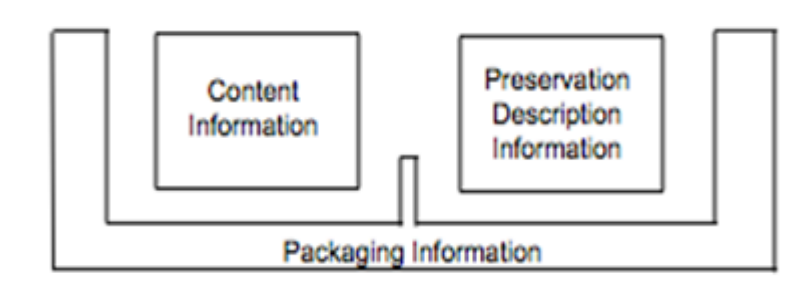
<sup>29</sup> <https://webfoundation.org/2011/11/5-star-open-data-initiatives/>

identified by an HTTP IRI . Class and property 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 OAIS model compliance

According to the OAIS Reference Model<sup>30</sup>, “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.

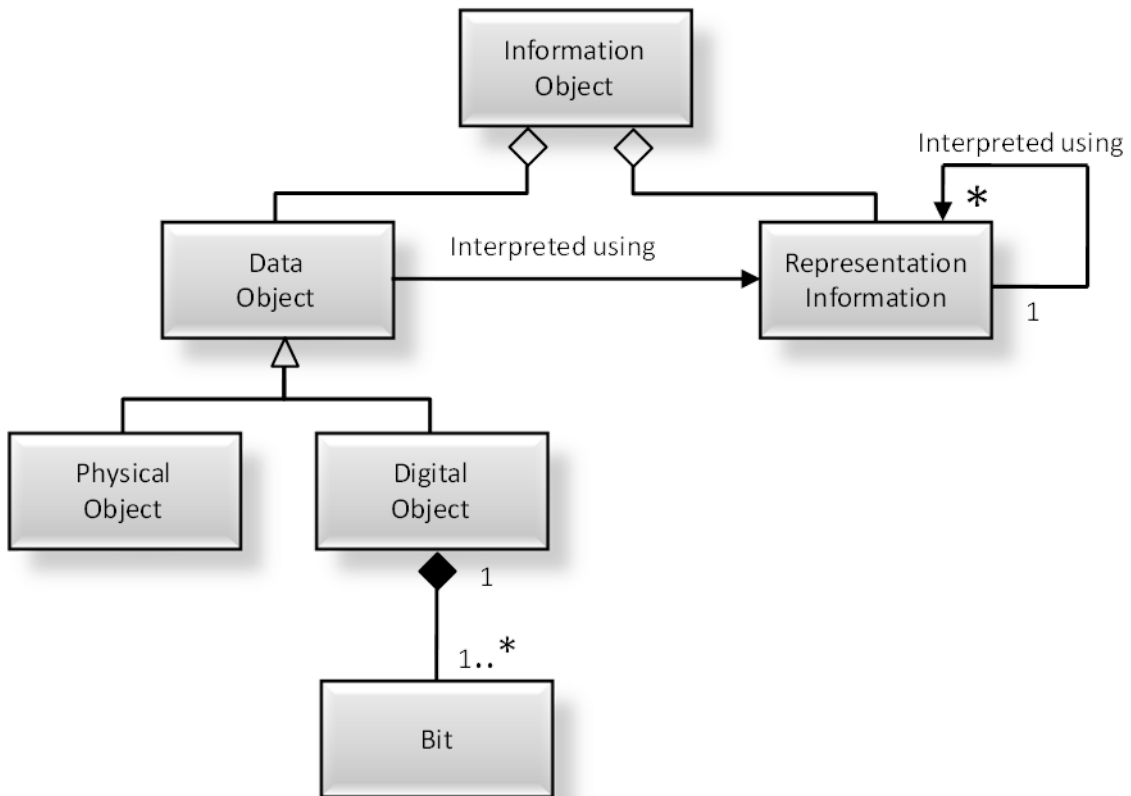


OAIS Information package

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

---

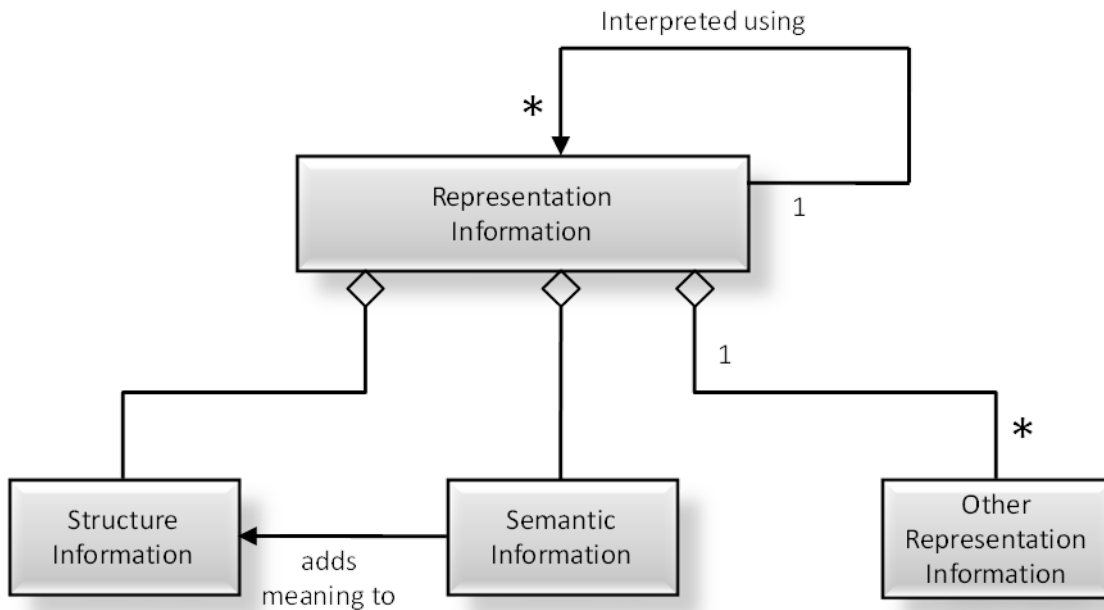
<sup>30</sup> [http://dpconline.org/docs/lavoie\\_OAIS.pdf](http://dpconline.org/docs/lavoie_OAIS.pdf)



OAIS Content Information Object

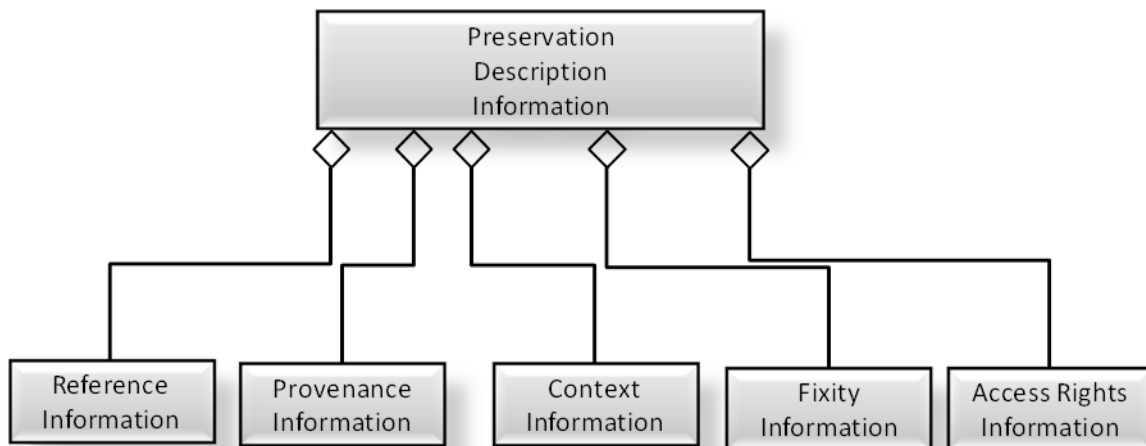
For Content Information the Data Object is the data to be preserved and Representation Information is information “needed to make the Content Data Object understandable to the Designated Community”. Representation Information is composed of various parts:





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 is defined by the archive – it is not an attempt to foretell the future. An archive partly relies on current requests of its funders and communities, partly it relies on the gut feeling of archivists, and there will be always an arbitrary element in archiving. Different archives which are preserving the same digitally encoded information may well have defined their Designated Community for that information differently.

This may be even more sensitive in the case of Linked Data, which is typically created with an idea of sharing in mind, and as such tends 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 make statements about. If an archive defines its Designated Community too narrowly, then producers of LD may choose not to deposit their data with that archive. For this reason an archive which wishes to preserve LD in a way which is likely to satisfy its stakeholders should be careful about the way in which it defines the Designated Community.

By its nature, Linked Data refers to other external resources, 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 is essentially the issue of clearly defining the data to be preserved, the preservation aims, Designated Community and by implication the Representation Information needed. In terms of OAIS the aim would be to create an Archival Information Package (AIP) which (logically) contains everything needed for the preservation of the object of interest. This then in turn helps to define what the Producer (in OAIS terminology) must provide to the OAIS. Formally this is described as being transferred to the OAIS in one or more Information Packages called Submission Information Packages (SIPs).

## 4.2 Ingesting a LD dataset

This section deals with the problem of how to construct an Archival Information Package (AIP) that includes an LD dataset as its Content Data.

### 4.2.1 Self-containedness

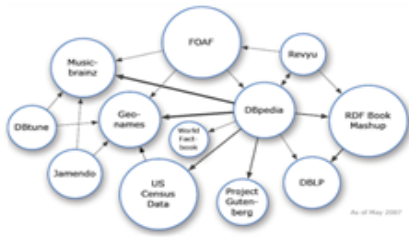
Ideally, an archive would like to ingest SIPs to create a self-contained AIP, thus avoiding any dependence on resources that are out of the archive's control. In the case of an LD dataset, this means to obtain, for instance by crawling the web, all the resources that the LD dataset links to (via IRIs).

- 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<sup>31</sup>. However, it is doubtful whether this strategy can be applied to all the related RDF resources, because it exposes the archive to the risk of archiving a large portion of the LOD cloud<sup>32</sup>. 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 example for the DBpedia use case archiving scalability issues must be taken into account because of the exponential growth of the LD cloud (illustrated in the following figure) which is of particular importance for datasets such as DBpedia that are both big and highly interconnected.

---

<sup>31</sup> For example <http://lov.okfn.org/dataset/lov/> stores many vocabularies, but the file gathering them all is only 8.4 megabytes, 64740 triples.

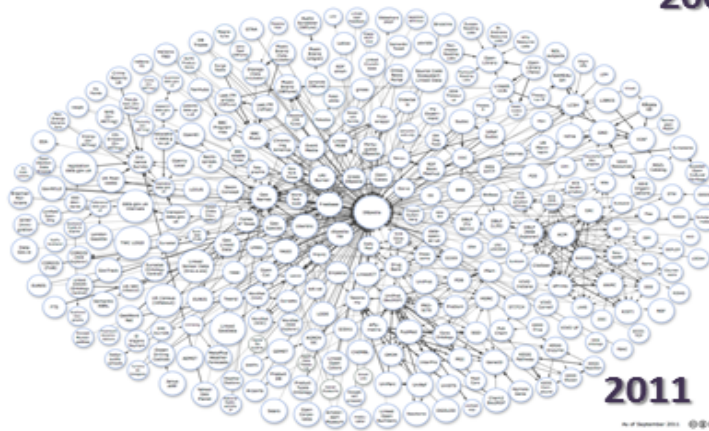
<sup>32</sup> State of the LOD Cloud in 2014 (Retrieved in September 2014):  
<http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/>



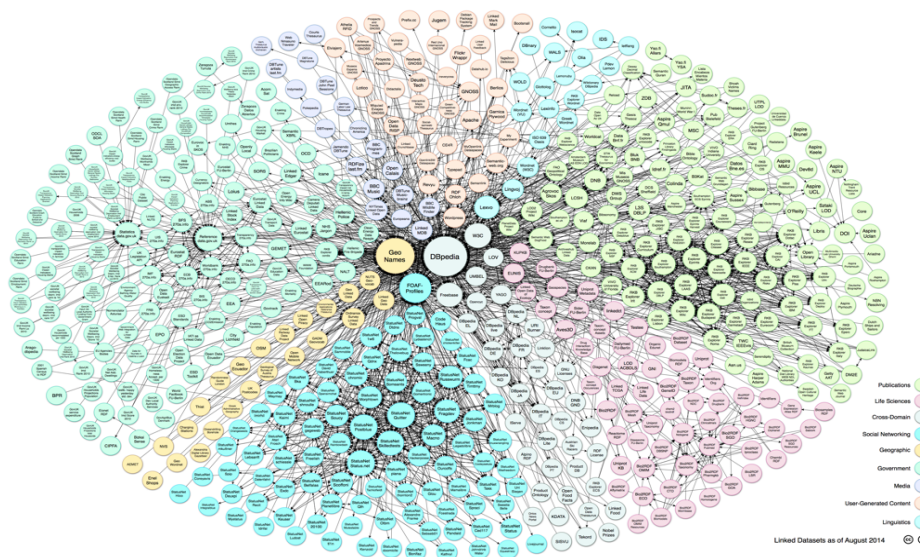
2007



2009



2011



2014 (As of: 2014-08-30)

LD cloud growth (source: Diachron presentation at Prelida Workshop & Richard Cyganiak and Anja Jentzsch)

- For all other kinds of resources, there exist the same boundary problem, and in addition there can be problem with non-open data or digital rights management (DRM), when resources are not freely accessible (More specifically, DRM is the digital mechanisms that control-lock-access to data. Closed data may still be technically available on the web, even with a non-open license. Again, then, the boundaries of the AIPs 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) or to the HTTP-based Memento framework<sup>33</sup> for time-based access to non-RDF resources.

#### 4.2.2 Serialization

Once the content data of the AIP is defined, a serialization format has to be chosen for the LD dataset being ingested as well as the related RDF resources. It has been suggested that RDF 1.1 n-quads<sup>34</sup> [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 AIP 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<sup>35</sup> and JSON-LD<sup>36</sup> may not cover everything), and there is no evidence that some serializations are better than others. The Data Best Practices W3C group<sup>37</sup> is best placed to make a recommendation on whether archives should focus on a particular serialization, or make all serializations kept fully compatible over time.

#### 4.2.3 LD Dataset Description

The AIP needs to have a description of the content data, and for this VoID<sup>38</sup>, DCAT<sup>39</sup> and PROV<sup>40</sup> have been proposed as suitable vocabularies for describing (respectively):

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

---

<sup>33</sup> H. Van de Sompel, M. Nelson, R. Sanderson. HTTP Framework for Time-Based Access to Resource States – Memento. RFC 7089. <http://tools.ietf.org/html/rfc7089>

<sup>34</sup> <http://www.w3.org/TR/n-quads/>

<sup>35</sup> <https://dvcs.w3.org/hg/rdf/raw-file/default/rdf-dataset/index.html>

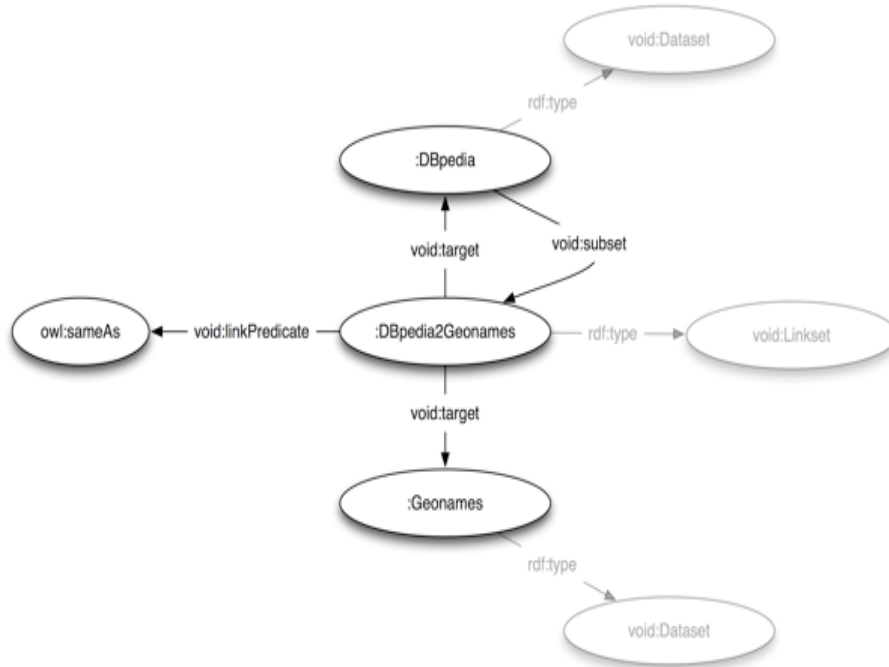
<sup>36</sup> <http://www.w3.org/TR/json-ld/>

<sup>37</sup> [http://www.w3.org/2013/dwbp/wiki/Main\\_Page](http://www.w3.org/2013/dwbp/wiki/Main_Page)

<sup>38</sup> <http://www.w3.org/TR/void/>

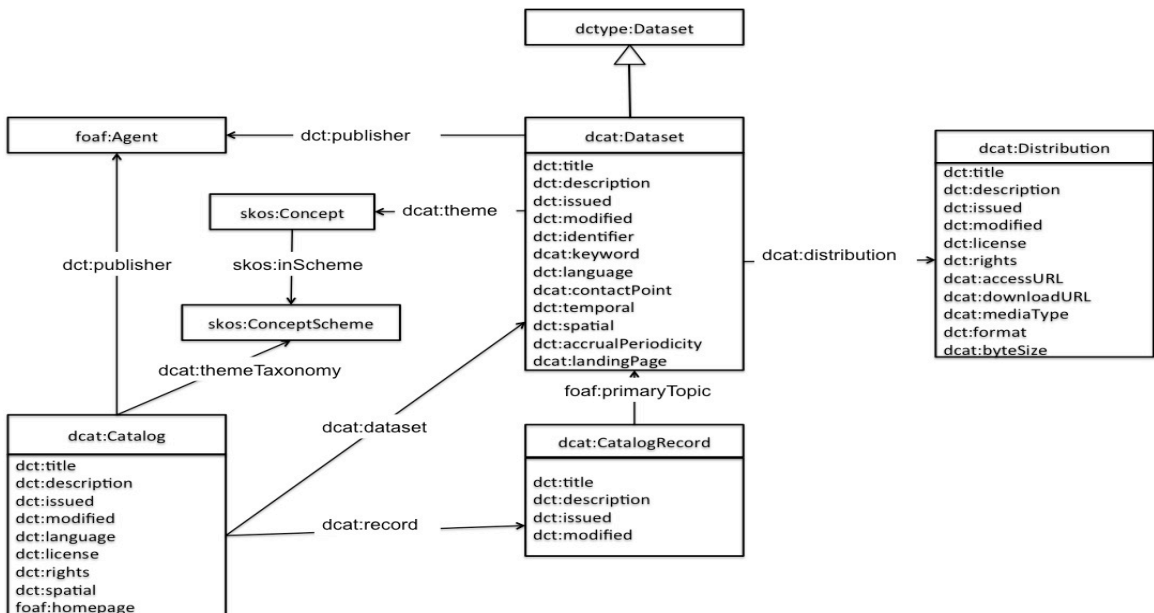
<sup>39</sup> <http://www.w3.org/TR/vocab-dcat/>

<sup>40</sup> <http://www.w3.org/TR/prov-o/>



Linked Dataset description example using VoID (source:W3C<sup>41</sup>)

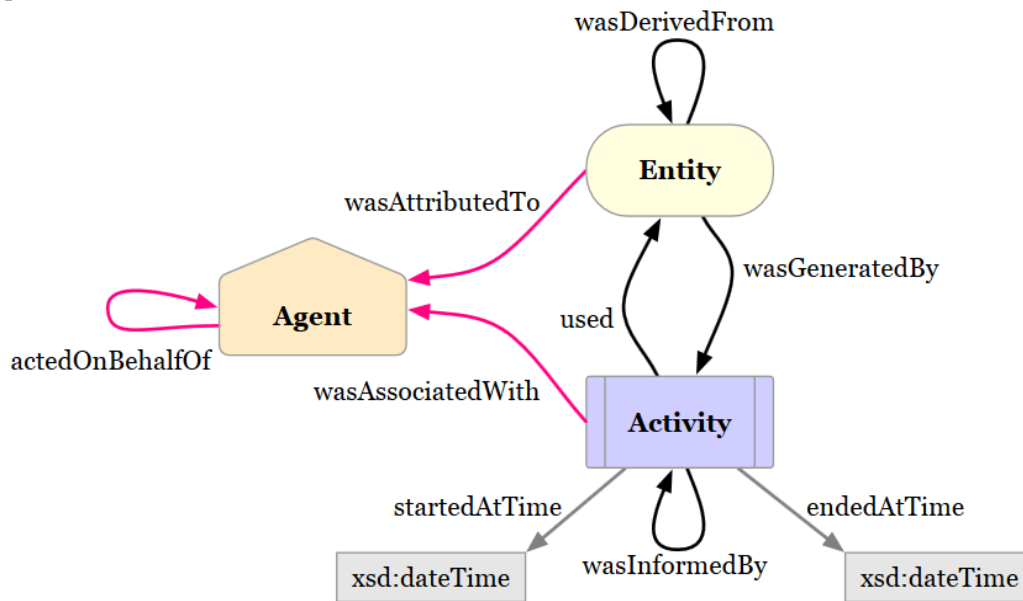
- the LD dataset in data catalogs



Data catalog vocabulary example using DCAT (source:W3C)

<sup>41</sup> <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 SIPs which are used to construct the AIPs. It belongs to Representation Information (VOID and DCAT) and to Preservation Description Information-PDI (PROV). The PROV vocabulary<sup>42</sup> is recommended by the W3C for expressing Provenance Information. PROV is designed precisely to represent how the RDF was made and what the history of this dataset is before and after ingest. PROV is about documenting the provenance of an object, not about offering a metamodeling mechanism. The Representation Information of, for example, the PROV encoding can be found in the W3C documents.

#### 4.2.4 Reasoner preservation

Semantic Information consists of two parts: the semantics of RDF<sup>43</sup> 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), typically in OWL format, referred to by the Data Object. Based on these reasoning can be applied to the LD.

Part of the preservation strategy must consider the preservation of related application software. Depending on the definition of the Designated Community, preservation of the software itself may not be a requirement. If it is required as part of the Representation Information, the specific software (especially if it is widely used), may be preserved elsewhere and pointed to instead of being held locally, and information about the specific reasoners used along with the preserved data and their

<sup>42</sup> Provenance Working Group. The PROV Namespace. W3C Document 19 May 2013. <http://www.w3.org/ns/prov>

<sup>43</sup> <http://www.w3.org/TR/rdf11-nt/>

version must be part of the archived information. Related software can be OWL/RDF reasoners<sup>44</sup> or SPARQL endpoints. This is also very important for OWL ontologies, since reasoners are equally important to the ontologies themselves. Since OWL semantics are the base for reasoning, semantics of OWL and metadata about the specific version used, must be preserved as well (Actually one could argue this is the main thing that has to be preserved: once one knows the OWL specification, he can re-create a reasoner). Specifically, the description of the OWL version used for description of classes and properties and reasoning are important here<sup>45</sup>.

### 4.3 LD dataset 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. In PRELIDA Deliverable D3.2, Consolidated State of the Art, these issues are discussed using an abstract and concise approach. In this document we look more extensively at more specific examples of the general issues in order to develop the roadmap by identifying specific issues.

#### 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*

---

<sup>44</sup> <http://www.w3.org/2001/sw/wiki/OWL/Implementations>

<sup>45</sup> [http://www.w3.org/standards/techs/owl#w3c\\_all](http://www.w3.org/standards/techs/owl#w3c_all)

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.

### **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 additional 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 choice of transformation is determined by balancing costs, risks, the transformational information properties identified and the preservation aims, as discussed in Deliverable D3.2. 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 Archival Information Package is created and properly related to the one of which its Content Information is a new version. Note that this would not be considered as what OAIS terms an “AIP Edition” nor an “AIP Version”.

### *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 which is the object of preservation do not change directly, but the change to the Representation Information, e.g. vocabularies, may have an influence on their semantics, making some statements obsolete or false. A clearer way of thinking about this is that what is important is the Content Information, i.e. the data plus the versions of the vocabularies etc., at the time of creation of the AIP in the archive. If the vocabularies change, then a completely new AIP must be created, which is related to the first.

### *Example*

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 to 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. Depending on the preservation aims the archive could have in place a mechanism to monitor the vocabularies of the preserved LD and, whenever a change occurs to one such vocabulary, the archive could either import or point to the appropriate vocabularies. Alternatively, the Producer (e.g., the data holder) may be alerted to the changes (perhaps by the archive), have occurred. The action in that case rests with the data holder who is in the best position to decide whether or how to change the data and if/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 (typically OWL ontologies) 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. As discussed in the PRELIDA mid-project workshop<sup>46</sup>, a service such as the Orchestration service developed by SCIDIP-ES<sup>47</sup> may be of use. This is a mechanism to help sharing information about changes. Those who make changes to a vocabulary inform the Orchestration service that a change has occurred and the Orchestration Service informs those who have subscribed to the appropriate topic.

## **4.3.5 Changes to web resources other than RDF/OWL.**

### *Description*

In the preserved LDD there may be URIs referring to web resources, which 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

---

<sup>46</sup> <http://www.prelida.eu/events/prelida-midterm-workshop>

<sup>47</sup> <http://int-platform.digitalpreserve.info/dashboard/orchestration-service/>

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 may be feasible, as discussed also in PRELIDA Deliverable D3.2.

#### *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 has not been expressed in a formal way in a single structure. The knowledge base may be thought of as what would be expected to be in the mind of a member of the Designated Community, which comes from many distributed sources, for example 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 case presented in Section 4.3.4, 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, although practical, human-based, partial solutions may be adequate in most cases. Related approaches are presented in Section 4.5.

#### *Required technology/standards*

Crawling strategies are more complex here since resources to be crawled are more varied (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. On the other hand a partial solution such as the SCIDIP-ES Orchestration service noted above may be adequate in some cases.

## **4.4 Dealing with changes**

As discussed above the change management problem that poses LD-specific challenges is the propagation of ontology changes to the archived descriptions (Representation Information or Preservation Description Information) that contain the ontology. Such changes may be looked at as requiring the creation of new AIPs. The question becomes how the changes are collected.

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 for the propagation of ontology changes to the archived descriptions that contain the ontology, the way of tackling this problem depends on the Preservation Aims of the archive (as described in Deliverable D3.2), which probably take into account the requirements of the designated community. Looking in particular from the point of view of the likely requirements of the designated community:

- If the designated community requires accessing and using the preserved data on the basis of the new classes and properties this involves no new activities – it means just using the new state of the web. If the practical implementation involves replacing the occurrences of the old classes and properties, then this implies a re-writing 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 [8,9,19,20]. A more general approach to concept evolution can be found in [6]. 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 classes and properties, 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 new classes and properties) 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. A detailed discussion is contained in Section 4.5.

### *Status*

Ways to detect changes have been discussed above. Technical solutions have been proposed, but standard policies and recommendations for 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.

## **4.5 Dataset Evolution and Preservation**

Dataset evolution is an important aspect for LD preservation since LD is dynamic and distributed. When an evolving dataset (such as a typical LOD) is preserved different versions of the dataset must be preserved as well. Archiving different versions of dataset only is not enough since data versions are related. More specifically these versions represent the evolution of similar concepts and this evolution is of great importance for archivists and archive users. For example, archivists may wish to submit SPARQL queries ranging over different versions of an RDF dataset. Therefore projects related to evolution of data are very important for preservation of dynamic data as well.

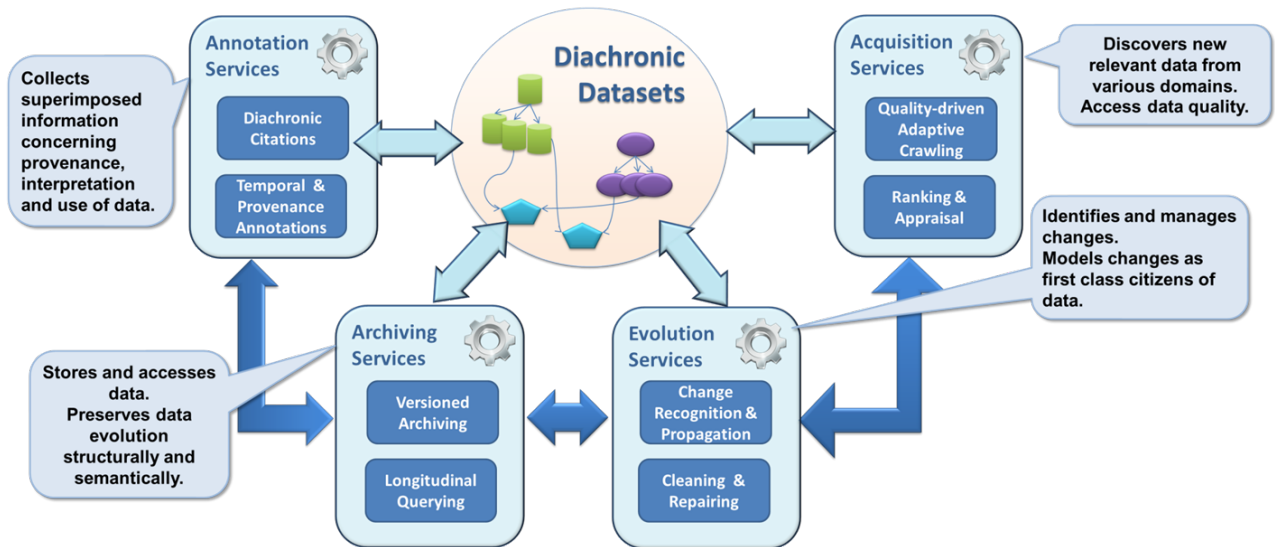
### **4.5.1 Challenges for data evolution.**

Challenges for data evolution are related to changes to data, since properties and relations of represented objects may change. It is often found that a large part of a dynamic dataset will change when two successive version are compared. Detection of changes and optimal storage are the basic challenges for data evolution.

But data evolution has additional challenges related to definitions, semantics and data schemas as well. Changes in schemas such as addition/removal of classes and properties (e.g., changes in OWL ontologies used for defining vocabularies) must be also detected in conjunction to changes to data. Even if definitions are not modified directly, changes to related definitions may cause indirect changes as well, for example modification to the meaning of classes and properties that are persisting across different versions. The EU DIACHRON project currently underway aims at addressing these challenges, and in the following we outline the approach it takes to do so. We should note that there exist alternative approaches to storage and scalability issues of evolving RDF data archiving, which are discussed in section 8 of this document on the research agenda.

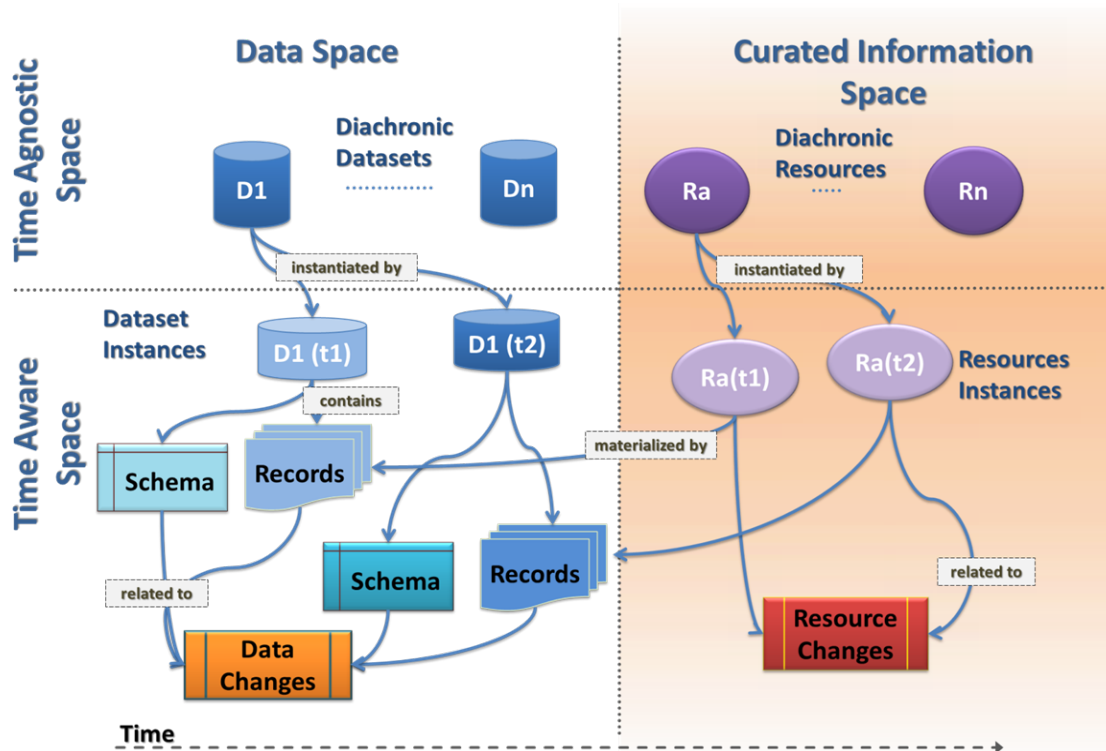
### **4.5.2 The DIACHRON approach.**

The EU DIACHRON project is highly relevant to evolution of datasets and although DIACHRON is not restricted to LD, the approach proposed for this project can be applied to LD as well. More specifically, evolution of dataset must be checked before archiving in order to detect changes in both data and schema and properly annotate the archived data. This is clearly illustrated in the DIACHRON processes, in which annotating and archiving services are integrated with dataset acquisition and evolution services that detect changes and also identify aspects of the datasets requiring repair.



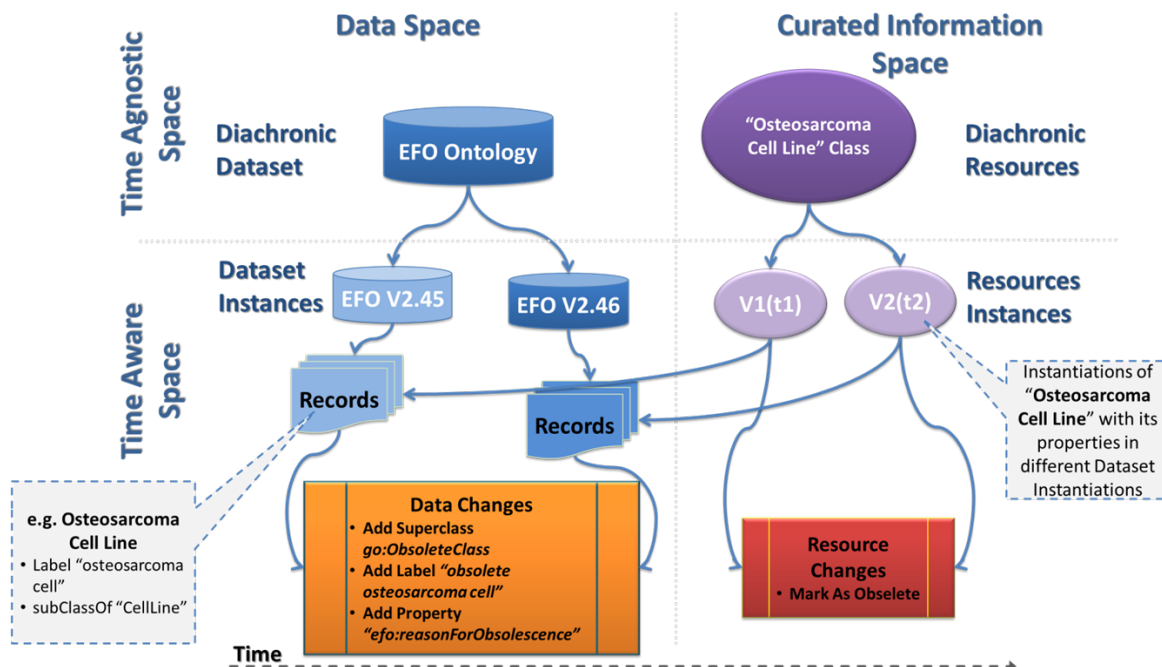
DIACHRON processes (source: DIACHRON presentation, PRELIDA workshop)

At the core, there is a Unified DIACHRONIC model for incorporating various data models and their evolution (see figure below). It is structured along the “time” and “information” dimensions. The time dimension distinguishes between time-aware and time-agnostic objects. Time aware objects incorporate evolution (changes) and temporal information, whereas time-agnostic objects represent unchangeable (diachronic) objects. In the Information Space, we have the Data Space where we capture basic changes on instances and schemas, and the Curated Information Space that represent a possibly more abstract view on these changes.



DIACHRON Data model (source: DIACHRON presentation, PRELIDA workshop)

An example where evolution and archiving services are applied to LD is the EFO ontology (see Section 3). EFO, the Experimental Factor Ontology, is a systematic description of many experimental variables and parts of biological ontologies, such as anatomy, disease and chemical compounds. A basic dataset of EMBL-EBI is also used in DIACHRON. The EFO Ontology becomes a Diachronic Dataset inside DIACHRON, and EFO Ontology versions are Datasets. Data changes are actual simple changes that occur between versions; e.g., classes or/and properties may be added or deleted. In the example below a super class has been added to a class, a new label and new property. Diachronic Resources are parts of the Dataset that users might be interested in monitoring through time at their “own level”, possibly more abstract. Resource changes might be simple changes of the same type as the data changes, but they might also be complex changes that occur as a combination of simple changes, e.g. the changes shown on the left comprise a complex change named “Class obsolete”. Detecting such changes and properly annotating archived versions is thus an important part of a preservation mechanism.



EFO ontology archiving example (source: DIACHRON presentation, PRELIDA workshop)

## 5 Organizational and Financial aspects of Linked Data preservation

As illustrated by the gap analysis and the use cases studies, dynamic and distributed environments (like LOD) are always complex to preserve for reasons inextricably connected with both the LOD nature and the preservation goals. The main challenges could find solutions if an adequate and accurate organizational infrastructure will be in place as early as possible. The questions to solve are in many cases conflicting and still open.

The LOD imply changes which are the most challenging issue for digital preservation: they have to be tracked, documented and maintained for future assessment and the links are essential LOD components, but according to the preservation rules and standards the main/significant links cannot be

preserved as simple references to external resources, but must be part of the ingestion process or, at least, well documented and assessed with reference to their impact.

The definition of what is significant implies a coordination among stakeholders and agreements with institutions of memory for ensuring continuity of access over time and sufficient documentation for presuming authenticity, while LOD are web based and not strictly related to the institutional control. Documentation, procedures, policies are recognized as crucial tools for preservation to be created and preserved in the creation phase, but the awareness for documenting persistency is not, at the moment, the focus of a lot of concrete actions in the LOD community.

The basic requirements imply the capacity of managing datasets and organizational changes by (a) early identification of representation information to collect, ingest, archive and in case transform according to the designated community involved (OAIS): this is not a technical question and it would be useful to add completeness and clarity to the definition of the Designated Community notion for LD cases, (b) early definition of boundaries to ensure a sustainable approach for ingestion into the repositories (related to an adequate description of the designated community): which links, which quantity of Preservation Description Information (PDI), which vocabularies (OAIS), (c) accurate choice of the Trusted Digital Repository (TDR) with reference to the governance, policies systems, certification processes (ISO 16363) and an eventual federated network and (d) clear identification of the profiles involved in the preservation processes and the crucial responsibilities (OAIS and ISO 16363).

In particular, according to ISO 16363 for certification processes, the competences for organizational infrastructure (3.3) imply the documentation of the LOD repository/provider reliability and include the ability to:

- evaluate the process by which a designated community is defined
- determine whether system documentation is adequate for all aspect of the TDR
- determine whether preservation plans are adequate and match the preservation policies
- determine if preservation policies are accurately captured in system workflows
- determine if workflows are adequately documented
- recognize whether an adequate level of detail has been recorded about system changes
- evaluate the organization's commitment to transparency and accountability

The chain of responsibility should be based on a governance system able to testify the commitment and the transparency of the LOD system and/or its provider (ISO 16363, requirements 3.1): they must be well defined, clearly ruled and well documented and include:

- who takes the main responsibilities,
- who defines the policies
- on which basis they are approved and disseminated,
- how to support the long-term persistency when the original LOD sets are not anymore curated (as when LOD providers disappear), etc.

Best practices must be identified (for more scenarios) and the repository should, at least, make evidence it is aware of these risks.

To ensure the continuity of the preservation services, the preservation strategy should define the type of repository for long-term like archives/repositories held by institutions of memory (against in-house solutions); in case of private archives/repositories policies must be in place and must be able to testify the dataset providers' awareness for this critical aspect. Policies and procedural workflows have a key role but must be further detailed in case of LOD (see the recommendations of APARSEN for policies as defined at the sections 3.1 and 3.3. of ISO 16363):

- general preservation policy
- policies for vocabularies, related changes and standards of reference

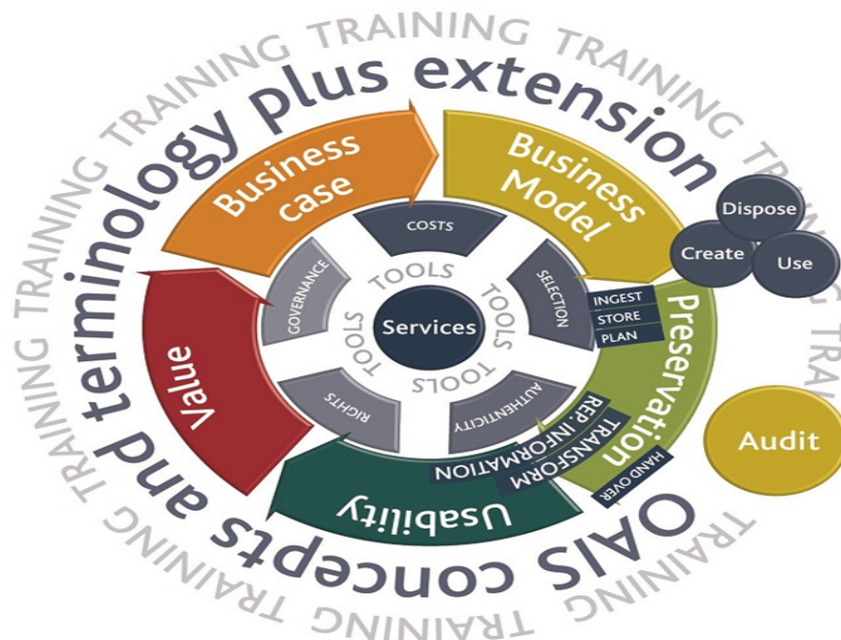


- policies for privacy
- policies for managing links and networks of SIPs/AIPs to ingest/manage
- policies for change management
- policies for appraisal and retention
- manuals which describe the fixity mechanisms
- updating system for all the policies included in the preservation plan

The level of granularity and functions to be preserved must be supported by a strategic preservation plan whose specific strategies are developed according to the datasets nature and function (to maintain the correct degree of data intelligibility). In addition, the economic sustainability has to be based on a cost model which takes into account the specific role of the stakeholders/providers and the custodians and risk assessment definition. Summarizing, organizational aspects of LD preservation are complex and many questions are still open:

- Who is going to pay for preserving LOD (also within an organization)?
- Why institutional repositories and institutions of memory should be interested? On which basis? Which role can be designed?
- Can a network of federated repositories be accepted and supported?
- Which level of service agreements is required between Institutions of Memory and LOD providers?

An important aspect to consider is the provision of the resource required for the preservation activities. An integrated view of preservation has been created by the APARSEN project illustrated by the following diagram:



APARSEN integrated view of digital preservation<sup>48</sup>

<sup>48</sup> A clickable version is available at:  
<http://www.alliancepermanentaccess.org/index.php/community/common-vision/>

A brief explanation is as follows: Preservation is judged on the basis of *Usability* – in terms of OAIS this is by the Designated Community but could be by other user communities. Usability creates *Value* – it is important to realize that adding Representation Information beyond that needed by the Designated Community, can add value by enabling more people to use that digital objects. *Business cases* can be constructed on the basis of the Value and *Business models* implement one or more of the Business Cases in order to generate resources, part of which could be used to support the preservation activities. Linked Data is of particular interest in that adding value through facilitating the combination with other Linked Data is a very natural process.

## 6 Using Linked Data for Digital Preservation

The main focus of this document is to bridge the gap between the Linked Data and the digital preservation communities, with the aim of ensuring the effective preservation of Linked Data. An additional benefit of the collaboration between the two communities is that the Digital Preservation community can use Linked Data for preserving different (i.e., non-Linked) Data. A general example is the use of a Registry/Repository of Representation Information<sup>49</sup> – the data object has a link to the appropriate piece of Representation Information, which in turn links to its own Representation Information.

A separate, specific example where LD is used for archiving historical data (Dutch Census Dara Archive-CEDAR [16]) is presented in section 6.1. Efficient representation of metadata for archived data, catalog and provenance information can be achieved using vocabularies such as VOID, DCAT and PROV as in the case of Linked Data preservation. Since these vocabularies can be used for describing a dataset that is not necessarily a Linked Dataset, the recommendations presented in section 4.2 for Linked Data are valid for other types of data as well. In addition to these recommendations, Linked Data vocabularies can be used for representing information related to privacy of data. This is a very important aspect of preservation, and section 6.2 shows one way Linked Data vocabularies can be used to ensure privacy in the context of preservation. The entire section 6 should be seen as indicative of the broad potential applicability of Linked Data technologies in the context of digital preservation solutions.

### 6.1 The CEDAR use case

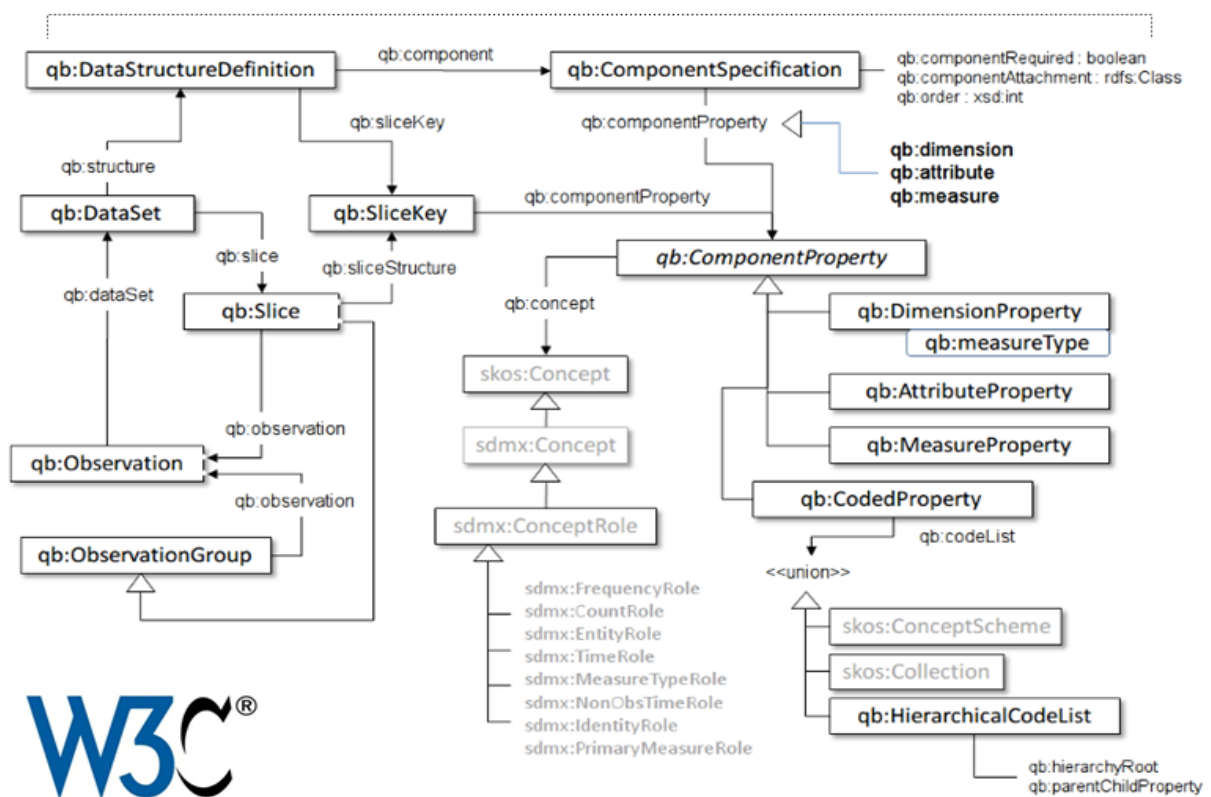
The CEDAR archive<sup>50</sup> contains Dutch population, occupation and housing census data from 1795 to 1971. This data is successively converted from analog to digital form, then to a tabular format (Excel) and finally to RDF. Conversion to RDF format is used because RDF is good for open data publishing on the Web, data in RDF is machine readable, thus supporting easy processing, visualizations, dynamic schemas and easy linking. This use case exemplifies the usefulness of LD for digital preservation, but also issues related to evolution and change presented in section 4. More specifically, since data covers three centuries, both concept definitions, use of language and dataset contents evolve over time. Concept drift is also an issue to be addressed, and is integrated in concept definitions as in case of the DIACHRON use case.

---

<sup>49</sup> See for example: <http://int-platform.digitalpreserve.info/dashboard/registry/>

<sup>50</sup> <http://www.cedar-project.nl/>

Another useful characteristic of using an RDF dataset for linking and integrating data corresponding to different time points is that longitudinal SPARQL queries (queries spanning across time) can be applied directly to the dataset. Such queries are very useful for researchers for understanding evolution of data and thus identifying trends and patterns in them. RDF representation is achieved by converting Excel tabular data to RDF using the W3C RDF Data Cube vocabulary<sup>51</sup> (QB). QB defines cubes as a set of observations that consist of dimensions, measures and attributes. Dimensions (qb:DimensionProperty) such as time period and area “identify the observation”. Measure (qb:MeasureProperty) such as population life expectancy are the observed phenomenon and Attributes (qb:AttributeProperty) group the unit of a measure (e.g., years) and additional metadata such as status (e.g., ‘estimated’).



W3C Data Cube vocabulary (QB)

RDF representation also combines additional vocabularies such as PROV for provenance information, demonstrating additional advantages of using LD vocabularies for digital preservation.

<sup>51</sup> <http://www.w3.org/TR/vocab-data-cube/#cubes-model>

```
cedar:BRT_1889_02_T1-S0-K17-h a qb:Observation ;
  cedar:population "12"^^xml:decimal ;
  maritalstatus:maritalStatus
    maritalstatus:single ;
  cedarterms:occupationPosition cedarterms:job-D ;
  sdmx-dimension:sex sdmx-code:sex-F ;
  cedarterms:occupation hisco:88030 ;
  sdmx-dimension:refArea gg:11150 ;
  prov:wasDerivedFrom
    cedar:BRT_1889_08_T1-S0-K17 ;
  prov:wasGeneratedBy
    cedar:BRT_1889_08_T1-S0-K17-activity .
```

CEDAR data example (source: CEDAR presentation, PRELIDA workshop)

## 6.2 Privacy Aware Preservation and Linked Data

The increasing volume and importance of data causes an increasing need for privacy frameworks that allow individuals to express their privacy preferences and service providers to interpret, enforce, and be held accountable for respecting individual's privacy concerns. Such examples of compliance directives are the HIPAA Privacy Rule<sup>52</sup> for medical information the Gramm - Leach - Bliley Act<sup>53</sup> for financial information and the EU Directive 95/46/EC<sup>54</sup> for personal data protection.

OAIS is closely related to privacy preservation since the PDI (Preservation Description Information) includes Access Rights Information. This information consists of:

- Access restrictions pertaining to the Content Information; including the legal framework, licensing terms, and access control
- access and distribution conditions stated in the Submission Agreement, related to both preservation (by the OAIS) and final usage (by the Consumer) and
- the specifications for the application of rights enforcement measures.

The importance of privacy issues for OAIS is illustrated by the fact that changes in OAIS from 2002 (CCSDS 650.0 - B - 1) to 2012 (CCSDS 650.0 - M - 2) contain the addition of Access Rights Information to PDI and the Removal of Annex A (existing archive examples) where out of the five examples removed, two deal with privacy issues, thus privacy is an important and evolving issue.

An example for the importance of privacy issues in digital archives is the Life Sciences Data Archive (LSDA). LSDA is responsible for collecting, cataloging, storing and making accessible the data of NASA-funded Life Sciences space flight investigations. The LSDA has strict security measures for data from human subjects which require sensitivity and secure handling due to the Human Data Privacy Act. Only mean pooled human data is made available to the public and in this case privacy is

---

<sup>52</sup> <http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/>

<sup>53</sup> <http://www.business.ftc.gov/privacy-and-security/gramm-leach-bliley-act>

<sup>54</sup> [http://europa.eu/legislation\\_summaries/information\\_society/data\\_protection/l14012\\_en.htm](http://europa.eu/legislation_summaries/information_society/data_protection/l14012_en.htm)

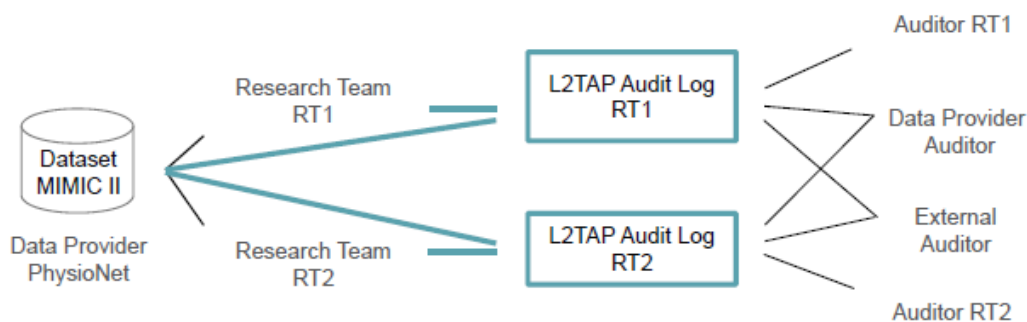
more important than usability. Another example of privacy sensitive archive is the US National Collaborative Perinatal Project (NCPP). NCPP was a multi-institutional, multiyear study of pregnant women; information on the children born from those pregnancies was collected to provide baseline information useful for later determining the causes of neurological diseases. The NIH NINDS project spent more than \$200 million over two decades to collect NCPP and it is unlikely that a study of this duration and magnitude will be repeated. Then NCPP transferred to the National Archives and Records Administration (NARA) after NARA and NIH resolved the privacy and access concerns. In this case NARA provides added value by addressing privacy concerns. More specifically, NARA provides NCPP as received from NINDS, and in addition it has created Public Use Files for the two data files containing personal identifiers in conformance with the Freedom of Information Act. Also NARA enforces restrictions on access to records whose release might result in unwarranted invasion of personal privacy.

Privacy protection is a challenge for social computing and data driven science, especially if size of data nowadays is taken into account. Consider big data biomedical research with massive datasets of human genome, biological imaging, and clinical information collected and aggregated from individual health records. Data subjects' privacy in clinical research is addressed by multiple legislations and regulations (e.g., U.S. Department of Health and Human Services or HHS). Linked Data can be useful for achieving this and a use case representing the applicability of L2TAP and SCIP [14,15] will be presented in the following.

### 6.2.1 Privacy awareness in OAIS using ontologies

Encoding privacy information for archived data (e.g., medical data) can be achieved using RDFS ontologies, as shown by the ontologies L2TAP and SCIP [14,15]. By applying RDFS/OWL ontologies, privacy log events can be published as Linked Data and since all privacy-related events are encoded in RDF, log integration via secure web access to all event descriptions can be achieved. Privacy policies can be also represented in RDF and these representations, in conjunction with dataset descriptions, allow for applying SPARQL queries for log construction, derivation of obligations and auditing of compliance checking. The SCIP ontology is designed to capture Contextual Integrity requirements, where contexts are defined as the participants and their roles, data attributes and purposes in addition to norms and information transmission principles.

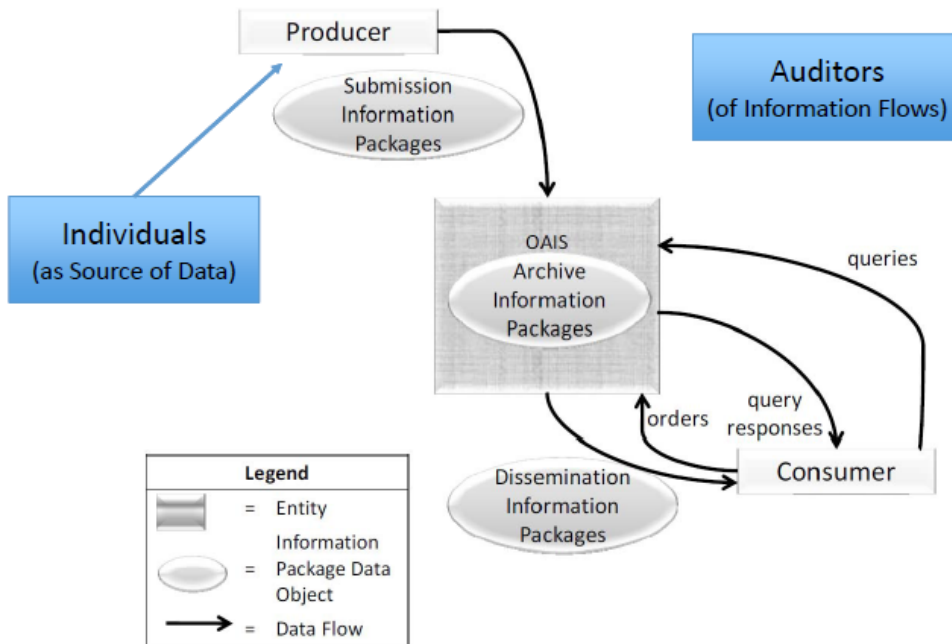
A scenario that RDFS ontologies L2TAP and SCIP are used for privacy is the following:



Medical research study scenario (source: Mariano Consens, PRELIDA workshop)

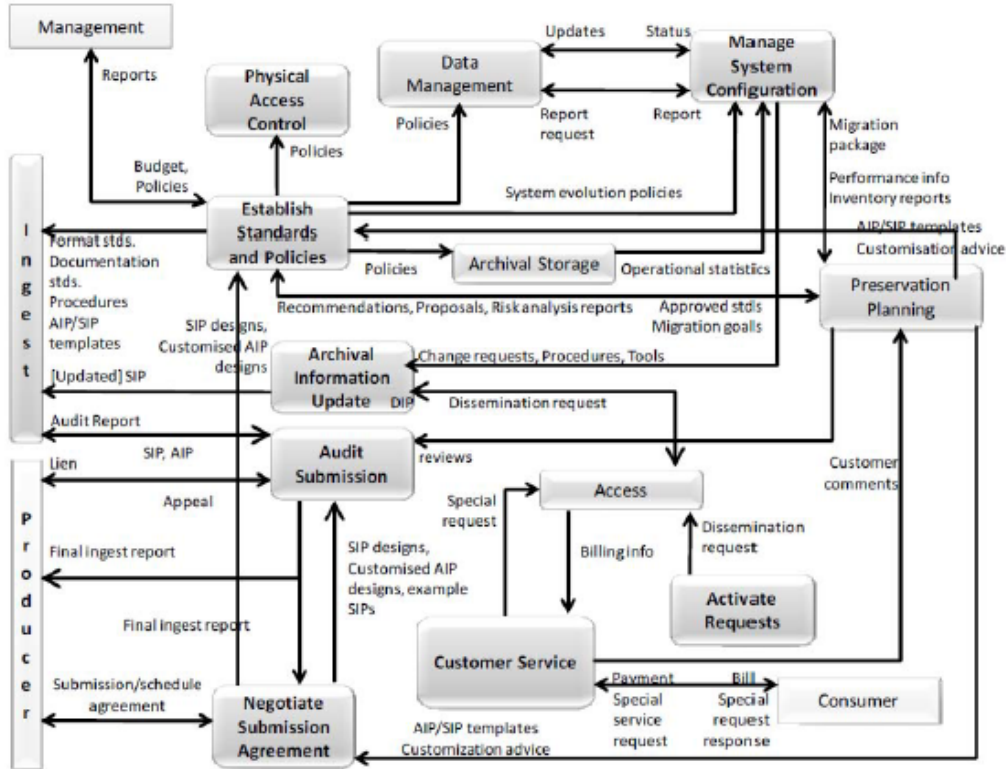
Research teams are interested in analyzing primary reasons for intensive care unit (ICU) hospitalization, and examine effectiveness of medication across patient demographics. MIMIC II is a public clinical database provided by PhysioNet of data on anonymized patient ICU admissions. Researchers must comply with MIMIC II data use agreement, as well as HHS, Hospital/University,

and other regulations. L2TAP ontology provides a set of classes and properties that can be used to represent and publish a log of privacy events as Linked Data. L2TAP related events in the log are Log Initialization and information about who is the logger, what time model is used and participant Registration using objects DataSubject, DataRequestor, DataSender, ObligationPerformer, ObligationWitness, PrivacyLogger, PrivacyExpert.

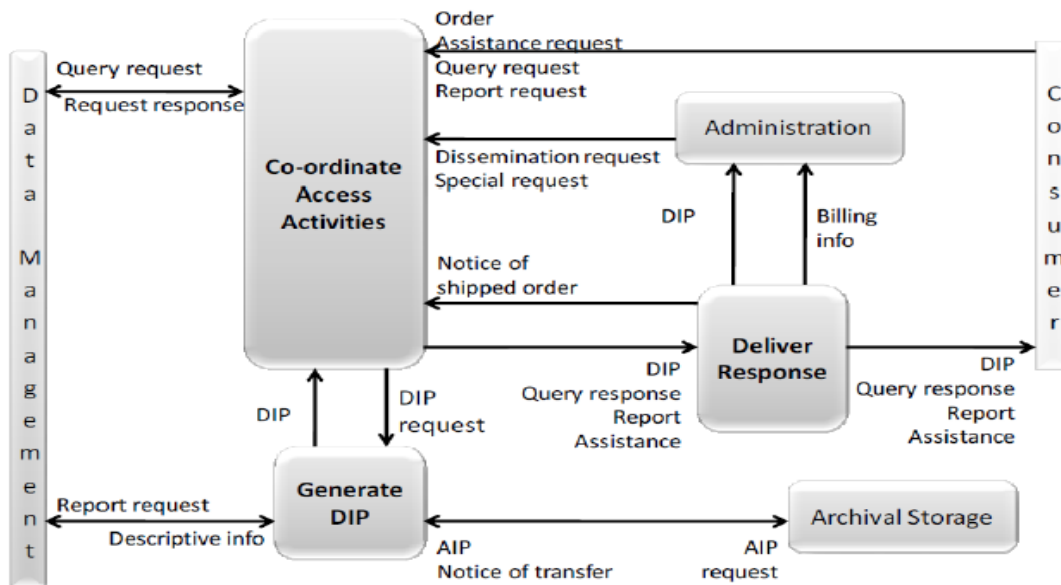


OAIS participants scenario using L2TAP ontology (source Mariano Consens, PRELIDA workshop)

The SCIP ontology is designed to capture in RDF Contextual Integrity requirements such as contexts (Participants and their Roles, Data Attributes, Purposes), norms and Information Transmission Principles. The first type of data that can be encoded in SCIP is privacy preferences, as specified by data providers.

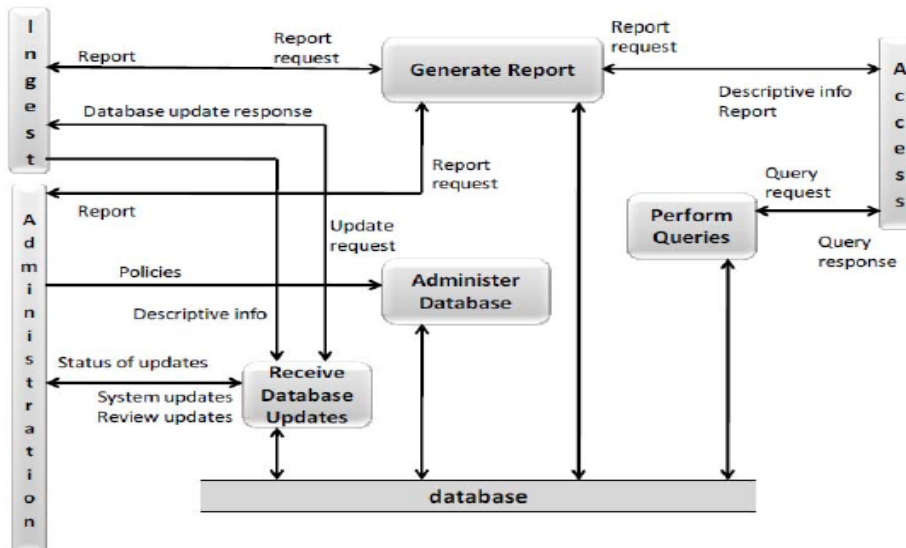


OAIS and privacy preferences using LD (source: Mariano Consens, PRELIDA workshop)  
 OAIS Access requests and responses from researchers are the second type of SCIP data:



Privacy related requests and responses in OAIS (source: Mariano Consens, PRELIDA workshop)

Obligation acceptance, obligation performance and access activities by the researchers are the last kind of data that can be expressed in RDF using the SCIP ontology.



Privacy activities in OAIS (source: Mariano Consens, PRELIDA Workshop)

Using RDF ontologies has thus the potential to address many of the privacy concerns raised in OAIS. Besides providing a standard language for expressing definitions and messages, LD-based privacy data has the additional benefit that SPARQL queries can be used for several tasks. The first one is log construction. It requires descriptions of the participants, policies, and access requests. Data providers are motivated to express the policies that govern data usage, and research institutions are motivated to facilitate researcher accountability since access requests can be logically derived from dataset description using SPARQL Construct queries. Crucially, LD-based solutions can accommodate multiple log scenarios raised by distribution, replication, and the existence of third party custodians.

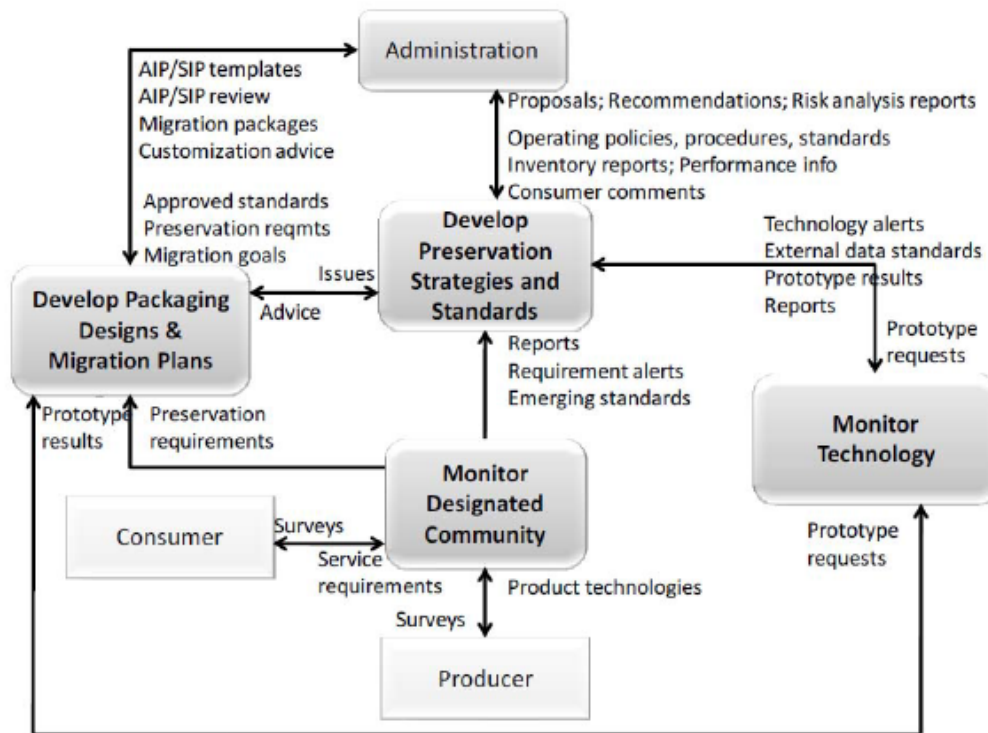
Second, SPARQL queries can be used for obligation derivation. More specifically, given an access request, we need a mechanism to log all applicable obligations. The process consists of the following three tasks: (a) find matches between an access request and privacy preferences, (b) generate the set of obligations and (c) construct the logical expression that describes how the individual satisfaction of each obligation contributes to the overall compliance of the originally matched access request.

Finally compliance checking can be achieved using SPARQL. Representative compliance queries are:

- a) Which access requests are not compliant at time  $t$ ?
- b) Which access requests have been discharged?
- c) What obligations are pending?

The procedure for compliance checking is as follows: (a) determine the individual satisfaction of each obligation (ASK query), (b) evaluate how the individual satisfaction of each obligation contributes to the overall compliance of an access request (multiple ASK queries) and (c) determine the access request compliance (SELECT query).





Monitoring changes in OAIS preservation

PRELIDA's list of desirable features for a Preservable (Link Data) Dataset archive also includes Privacy Policies so it is recommended to explore the re-use of Linked Data vocabularies such as SCIP for this task.

Note that privacy preferences are dynamic since privacy policies and the social context of norms change through time. Thus dealing with changes as in section 4 will also be an important issue for privacy awareness.

## 7 Assessment & 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 meant to identify the features an LDD archive should consider and is based on the gap analysis of Section 3 and the technical challenges of section 4. Based on the analysis of sections 3 and 4, these recommendations are selected for being both valuable for preservation and doable with existing methodologies and tools, or at least achievable with relatively little effort. Issues requiring significant research are discussed in the next section.

- Selection and appraisal of data: identify the boundaries of the LDD that has to be preserved, e.g., using a Concise Bounded Description strategy [10]. Although the system boundary must be clearly defined before archiving (Section 4.2), the optimal selection of this boundary with respect to scalability is a challenging task that will be part of the future research agenda.
- Gather every RDF datasets (using quads to identify RDF graphs) that are relevant for the LDD to be preserved. The ideal strategy is complete closure. Both for vocabularies (ontologies) and instances. There are vocabularies that describe the provenance of crawl/imports/ingests of

Linked Data<sup>55</sup>. For vocabularies the complete closure is the suggested strategy because of their importance and relatively small size (see Section 4.2.1). For other instances scalability issues must be taken into account.

- Submit data in a standard serialization (such as N-quads), also consider conversion between formats. Submit every representation (HTML+RDFa, JSON) served in content negotiation (see Section 4.2.2). 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).
- Add time-stamps for the crawls of the collected datasets (see Section 4.3), including among others date of last modification, and most importantly, data ‘snapshot’ time.
- Whenever an LDD is collected into an AIP, 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 have to be notified of any change (see Section 4.3). For notifications when a dataset changes: ResourceSync<sup>56</sup> can be used (it is used for instance for DBPedia synchronization). A lightweight alternative for LDD (using VoID) is offered by dady<sup>57</sup>.
- Include VoID/DCAT/PROV description in Representation Information (see Section 4.2.3). Also include data validation instructions as in Resource Shape[11]. And the corresponding ontologies (DCAT and VoID ontologies, etc.). This recommendation also applies to preservation of non-Linked data since LD is useful for such tasks (Section 6). Specification documents should be also preserved, (i.e., RDFS, OWL, serialization specs). Reasoners and SPARQL engines (triple store) are also to be preserved for accessing purposes as proposed in section 4.2.4.

Overall, besides the recommendations above a formal set of recommendations by organizations such as W3C may be the outcome of PRELIDA project. This 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. There is not yet an answer to the decision related to the scalability issue. This may lead to different approaches related to Web archiving and also research on optimal storage schemas. Both of these are the topics covered in the research agenda.

Finally, the distributed nature of Linked Data suggests that a distributed approach may be more appropriate than a traditional one for preservation of LDD (see Section 4.2.1). In such an approach, an OAIS can be spread over several archives, each storing a part of the Content Data. It needs to be explored further, whether this distributed structure would be more suitable to archive a LDD, where references to external entities can be managed as references to other OAIS managing those entities. This may be however a longer-term goal.

## 8 Research Agenda

Use cases and gap analysis illustrated the shortcomings to current approaches for LD preservation mainly because LD is dynamic and distributed. The use cases clearly illustrated both the need for research specific to LD preservation while at the same time showing the more widely applicable,

---

<sup>55</sup> See <http://ldif.wbsg.de/#provenance>

<sup>56</sup> <http://www.niso.org/workrooms/resourcesync/> (see also <http://www.openarchives.org/rs/toc> and <http://www.openarchives.org/pmh/>)

<sup>57</sup> <https://code.google.com/p/dady/wiki/Demo>

mutually beneficial advantages to be derived by maintaining close links between digital preservation and DP. A number of specific research topics are described in the following sub-sections. Of fundamental importance in most of these is the definition of the boundary of what is being preserved, which is particularly difficult given the natural distribution of LD; this is discussed in detail in the next section.

## 8.1 Defining the boundaries of LD archives

As discussed in PRELIDA Deliverable D3.2 “State of the Art”, an important issue for the research agenda is the optimal identification of the system boundary with respect to performance and scalability. Formulas estimating the appropriate boundary, based on reasoning completeness, size, growth rate and degree of interconnectedness of a dataset is important for LD preservation.

Here also, web archiving research may be useful: in the Hiberlink project and the Internet Robustness project<sup>58</sup>, various ideas are explored that closely relate to the problem of interconnectedness of Linked Data. Hiberlink focuses on web resources that are referenced in scholarly publications and Internet Robustness focuses on web-based legal literature and the blogosphere. The projects share the notion of a core collection that someone cares about (cf. a Linked Data set) and resources that are linked from it (cf. resources interconnected with the Linked Data set). They also share the notion of pro-actively archiving the linked resources at crucial moments in the lifecycle of the core collection. Both projects are exploring a variety of ways to achieve this. Harvard's amberlink approach is to cache linked resources along with the core collection<sup>59</sup>. In Hiberlink, the emphasis is into pushing linked resources into web archives. These perspectives can easily be transposed to the Linked Data world. In addition to the above the Apache Marmotta effort, an open implementation of the W3C Linked Data Platform specification, supports versioning of linked data and access to versions via the Memento protocol<sup>60</sup>.

Research topics therefore include:

- Identifying the limitations of the projects mentioned above in terms of defining the boundaries of LD objects to be preserved.
- Supplementing those projects' results in order to provide theoretically solid and practically implementable tools to help users define distributed LD objects suitable for preservation.

## 8.2 Change detection

Change detection and proper annotation of versions (e.g., the research objective of the DIACHRON project) are issues of great importance for LD preservation). For example changes in vocabularies are commonly seen in the LD world. Since the DIACHRON project will carry on beyond the end of PRELIDA, the outcome and problems identified may contribute to that project's research agenda.

Deliverable D3.2 State of the Art pointed out that general tools exist<sup>61</sup> to help in sharing information about changes. Moreover there are two major considerations with respect to change.

---

<sup>58</sup> <http://cyber.law.harvard.edu/research/internetrobustness>

<sup>59</sup> <http://amberlink.org/>

<sup>60</sup> <http://marmotta.apache.org/platform/versioning-module.html>

<sup>61</sup> SCIDIP-ES project tools and services available from:  
<http://int-platform.digitalpreserve.info/>

- 1) Changes which could affect how the LD will be understood, which would adversely affect preservation. For example if vocabularies change then the original meanings could be lost. Therefore the old vocabularies should be kept as Representation Information.
- 2) The changes (including those noted in (1)) may be of interest as entirely new objects of preservation and could lead the Producer (in OAIS terms) to submit further information for the archive.

In preservation terms (1) is always relevant but (2) may also be relevant, depending on the archive's Preservation Aims, for example with wish to preserve evolution of the concepts.

Therefore techniques for archiving, for example, vocabularies, is a recommendation that can be directly adopted, in conjunction with data directly linked to the preserved dataset. The specific topics for the research agenda include:

- Methods to adapt existing or develop new services to capture information about changes
- Techniques to track evolution of schema and vocabularies etc., including information about version, fixity and responsibility
- As discussed in D3.2 State of the Art, additional techniques the enable the use of cached linked information e.g. local caches of schema.

There are additional issues than must be examined as part of a research agenda and these are the relation of LD archiving with Web archiving which will be presented in section 8.1 and the optimal storage schema for the preserved data and the corresponding scalability issues (section 8.2).

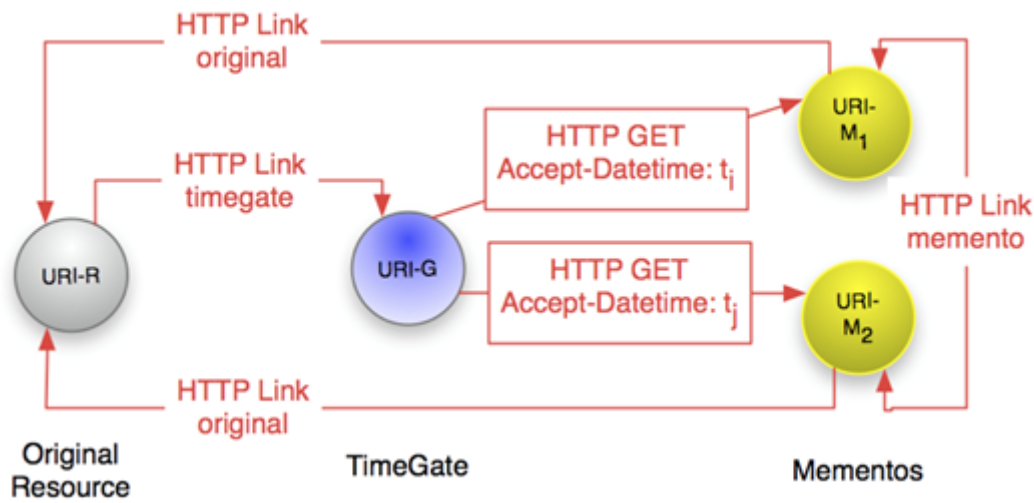
### 8.3 Web archiving and LD preservation

Scalability issues and frequent changes is an issue that Web archiving community is already dealing with. Like linked data, the “traditional” web is dynamic and link rot is a very common issue. Tools such as Memento are very useful for Web archiving. The Memento<sup>62</sup> "Time Travel for the Web" protocol is an interoperable approach to access web archives (IETF RFC 7089) adopted by several major public archives worldwide, including the Internet Archive. Memento [3] is also used to the Hiberlink project<sup>63</sup> for preservation of scholarly LD data. Memento keeps time stamped versions of URI contents in order to access them upon request even of the data is not available on the live web (or they have been modified).

---

<sup>62</sup> <http://www.mementoweb.org/>

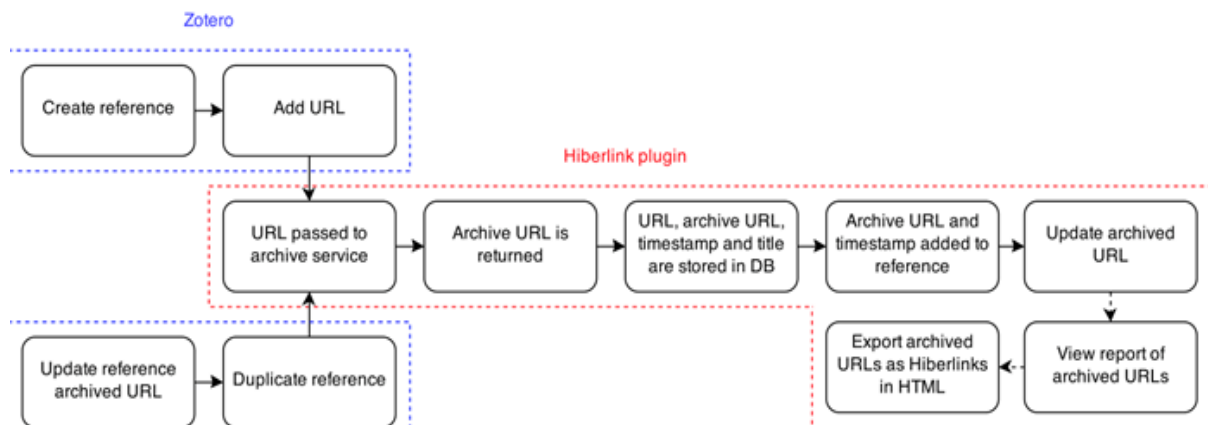
<sup>63</sup> <http://hiberlink.org/>



Web archiving using Memento

### 8.3.1 Web archiving using Memento

Reference rot is a major concern since even important documents such as legal decisions are typically not available even after a relatively short period of time [17]. Based on Memento, a plug-in for the Zotero tool<sup>64</sup> has been developed for preserving Web resources as part of the Hiberlink project.



Hiberlink plug-in for Web resources preservation (source: Hiberlink presentation, PRELIDA workshop)

Using Hiberlink or similar approaches e.g., the distributed approach adopted from CLOCKSS / LOCKSS<sup>65</sup> is an alternative to simple centralized repositories.

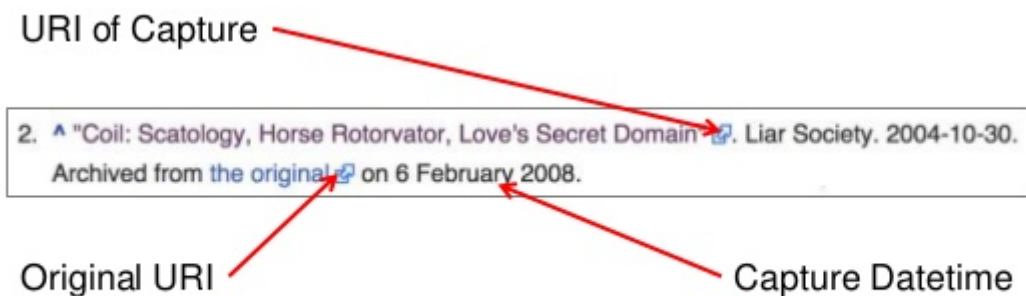
The most prominent research topic here is therefore:

<sup>64</sup> <https://www.zotero.org/>

<sup>65</sup> <http://www.clockss.org/clockss/Home>

- Whether and how a combination of these decentralized Web archiving approaches with OAIS can address preservation when the digital object of interest is the evolution of underlying digital objects – this is related to several topics identified above.

The Memento protocol, which is specified as RFC 7089, defines interoperability for access to resource versions based on a resource's generic URI and a desired datetime. Memento is fully aligned with the web architecture principles.



Capture and URI of links using Memento (source: Herbert Van de Sompel)

Memento's paradigm is distributed as the web itself, and hence can work in a hybrid environment of centralized and decentralized archives. Over the past five years, the Memento protocol has been adopted by several major publicly accessible web archives. Currently, there is a focus on getting it adopted for versioning systems such as wikis, software control system, evolving technical specification, etc. In addition to that its applicability to and relevance to Linked Data has been demonstrated in [18]. Also for over three years now, there has been a publicly accessible, Memento compliant DBpedia archive<sup>66</sup>. This archive is, as per the Memento protocol, integrated with DBpedia itself, in the sense that DBpedia URIs provide a "timegate" link to the archive (it suffices to look at DBpedia response headers to see them). This Memento-compliant interface for DBpedia achieves preservation of DBPEDIA RDF data (see DBpedia use case of section 3).

---

<sup>66</sup> Available at: <http://mementoweb.org/depot/native/dbpedia/>



#### Missing links representation using Memento (source: Herbert Van de Sompel)

The above characteristics make the Memento protocol a strong candidate to be considered for interoperability for time-based access to Linked Data archives. The problem of interconnectedness of Linked Data with other Linked Data also exist for regular web resources because they link to or embed remote resources. A Memento client can navigate across web archives and versioning systems to collect all temporally appropriate linked resources. Since this is the way the protocol works it can be directly used to navigate across Linked Data archives, and this kind of functionality Is not currently provided by OAIS compliant archives. A research topic is therefore:

- To what extent can Memento-based approaches can be considered as a component of a distributed version of OAIS compliant archives.

### 8.3.2 Web archiving to preserve results for linked data queries

Memento can be used to preserve LD but not applications such as the SPARQL access to archived Linked Data. On the other hand, there might be alternatives to SPARQL end-point preservation that combine approaches to RDF data archiving with alternative to SPARQL end-points.

Researchers at Ghent University have proposed Linked Data Fragments, a way to slice up Linked Data sets in a multitude of documents along the subject/predicate/object axes<sup>67</sup>. The idea is that in order to answer a certain query, a client obtains various Linked Data Fragments from one or more sources and further processes them locally. This approach has advantages from a preservation perspective: Linked Data Fragments are just documents at a URI and can be preserved way more easily than a SPARQL end point. Moreover, they can be made accessible under the Memento protocol to access prior versions. The integration of Linked Data Fragments is currently a topic of research.

## 8.4 The need for refreshing OAIS in a web environment?

OAIS is a preservation standard. In previous sections (Use cases, Technical Issues) it is has been demonstrated that Linked Data archiving has a lot of similarities with web archiving and less with current typical examples of national archives and libraries to which OAIS is often applied.

<sup>67</sup> <http://linkeddatafragments.org/>

However, as discussed in the OAIS-Futures forum<sup>68</sup> there is a core abstract standard, OAIS itself, plus, as identified in OAIS (OAIS section 1.5 *Road map for development of related standards*), associated standards. A fundamental question is whether there are concepts missing from the core document and/or whether there are additional standards needed.

For example OAIS itself does not specify whether a digital object is a single file or a distributed set of bits, nor whether Archival Storage consists of a monolithic store or a distributed set of stores. Distributed archives have been mentioned as a new type of archive. What new concepts do they introduce versus what new technology do they use? Do we need to define a new type of federation or are they just another special storage system?

Web resources, including Linked Data typically live on the web, (online use of LD, see Section 2.2) and thus it is expected that archives of such materials to be equally interconnected as the live web. Some of the presented use cases actually confirm this desire. Memento is about achieving this in an interoperable manner: Memento move web archives (including Linked Data archives) away from being destinations; they become infrastructure.

We can therefore identify a number of research topics:

- Investigating whether there are fundamentally new concepts required in the core OAIS standard or are there new associated standards concerning LD which need to be developed
- Identifying which pieces of metadata are missing from LD which are required for preservation
- Supplementing the previous topics, investigating the extent to which Memento could be integrated with OAIS or associated new standards

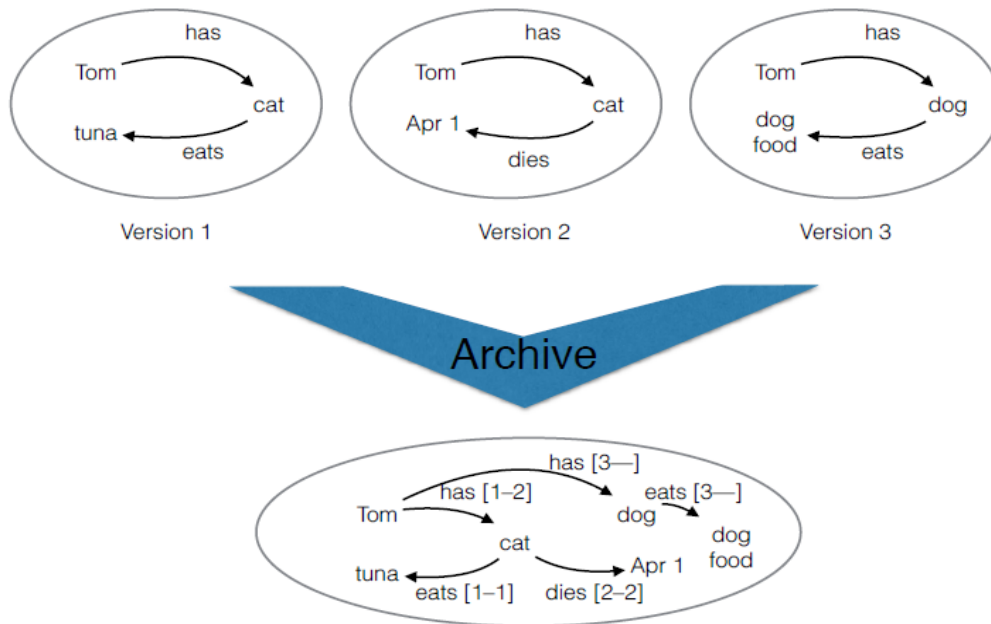
## 8.5 LD archiving and storage

The need for ingesting all contextual resources for a LD to archive is demonstrated by the presented use cases involving dynamic interconnected and large datasets such as DBpedia. On the other hand, scalability issues involving high volume data must be addressed as well. The main problem is that LD can be big and if many versions of the evolving RDF are archived severe scalability issues will arise. Projects such as DIACHRON address the issue of evolution and change detection (changes in data, schema and concept drift) but the issue of efficient storage and preservation of evolving RDF data is not being fully addressed. Storing evolving RDF data can be achieved by storing all versions or store the original databases and log the changes. These two approaches can be combined using a hybrid approach e.g. store the initial and every 10th version and also store log changes for the intermediate versions. Another approach is to use annotations, specifically never delete data but annotate its validity with time intervals, allowing one triple to 'serve across several versions'. This approach will be presented in the following. The first step is to annotate validity of triples using intervals instead of storing all versions.

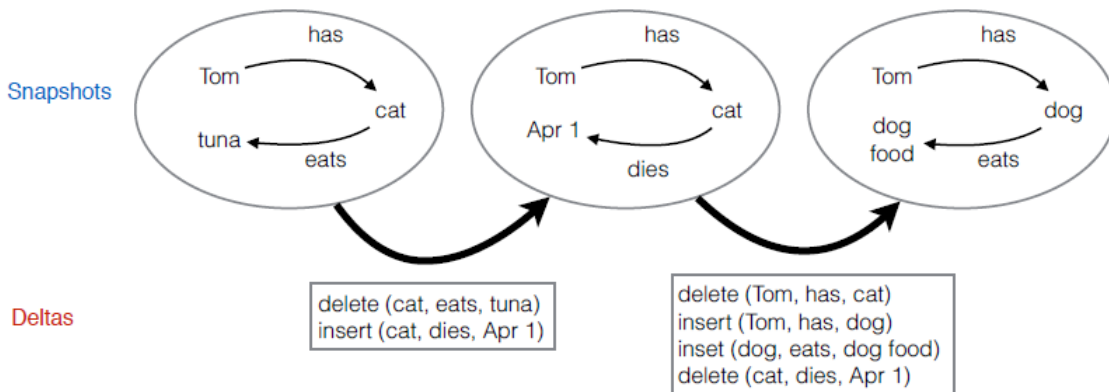
---

<sup>68</sup> Forum on OAIS Futures <http://www.iso16363.org/forums/forum/oais-futures/>



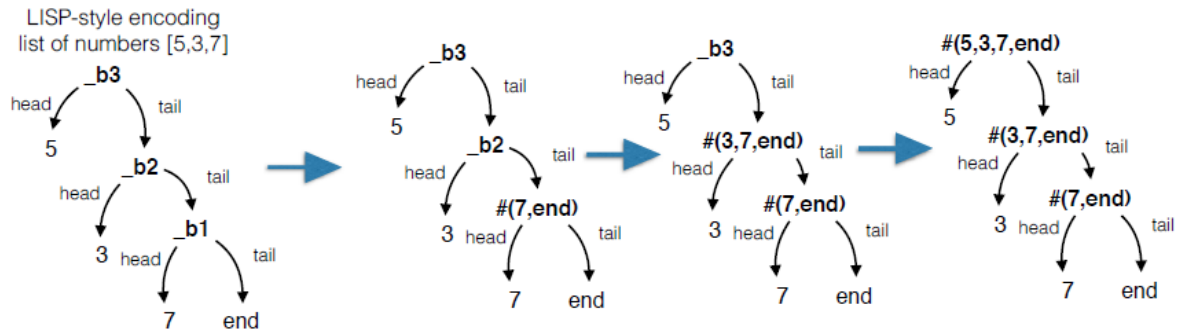


RDF annotation (validity intervals) (Source: S. Staworko, presentation at PRELIDA Workshop)  
 The input to archive (that can be stored using annotation) can be either ontology snapshots or deltas.



RDF snapshots and deltas (Source: S. Staworko, presentation at PRELIDA Workshop)

Annotating data is a relatively simple task if deltas are known since deleting a triple closes its interval and adding a triple opens a new interval, but it gets complicated when only snapshots are given. Then deltas must be computed by the corresponding snapshots and the main challenge is to identify objects that are the same across versions of the database. This is the Entity resolution problem which is a well-studied database problem in various settings (from duplicate elimination to record matching) and also addressed in related projects such as DIACHRON. In case of RDF blank nodes introduce complexities since two different nodes may correspond to the same entity. Blank nodes are used extensively e.g. for representing events, beliefs and data structures such as lists. A solution is to assign to blank nodes identifiers based on the properties and identities of linked entities and then apply entity resolution and interval based annotation.



Assigning identifiers to blank nodes (Source: S. Staworko, presentation at PRELIDA Workshop)

Annotated RDF improves space efficiency for storing evolving data but has not yet been applied to ontology evolution. Definitions of classes and properties may change as data does. In this case annotation must be applied to ontologies and vocabularies as well, and not only to instances; and data must be mapped to the corresponding definitions efficiently.

Querying and reasoning capabilities must be integrated to such a representation of evolving data and schema. Such capabilities are part of DIACHRON, but DIACHRON is based on dataset snapshots, and this is not the optimal solution with respect to storage requirements. Research topics are therefore:

- How can combining evolution management and space efficiency, using annotations can lead to optimal storage solutions.
- How can this in turn can be used to identify the system boundary for data ingestion, based on expected performance and storage requirements using the combined. But it is an open question whether this approach will be efficient for archiving large, dynamic and interconnected datasets such as DBpedia.
- Parallel to such efforts, distributed approaches based on novel Web archiving solutions can also be applied if scalability issues for such datasets cannot be addressed using a centralized approach.
- How distributed and large scale storage solutions can be audited and certified<sup>69</sup>

## 9 Summary and conclusions

This report examines and proposes solutions to issues related to the long term preservation of Linked 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 web of Linked Data.

The main approach in the digital preservation community is to document fixed digital objects and store them in a Trusted Digital 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 discuss about preservation without confusion. The research activities in the digital preservation

<sup>69</sup> see ISO 16363 <http://www.iso16363.org>

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 [2].

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.

Section 8 collects together a number of topics which should form the basis of the research agenda for the preservation of Linked Data.

Recommendations and the proposed research agenda can in turn 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 roadmap 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, D. Advanced digital preservation. pp 31-39: Springer, Berlin, 2011.

[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 Press, 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] Carothers, G. 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] Bykau, S., Mylopoulos, J., Rizzolo, F. and Velegrakis, Y. On Modeling and Querying Concept Evolution. Journal on Data Semantics, (1), pp. 31-55, 2012.

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

- [9] Hartung, M., A. Groß, and Rahm. E. Conto –diff: Generation of complex evolution mappings for life science ontologies. *Journal of Biomedical Informatics*, vol. 46, pages 15–32, 2013.
- [10] Stickler, P. CBD-Concise Bounded Description. W3C Member Submission 3 June 2005. <http://www.w3.org/Submission/CBD/>
- [11] Ryman, A. 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](http://prelida.eu).
- [13] PRELIDA Deliverable D3.1. State of the art. Available from the PRELIDA web site: [prelida.eu](http://prelida.eu)
- [14] Samavi, R. and Consens, M.P. Publishing L2TAP Logs to Facilitate Transparency and Accountability. In *Linked Data on the Web (LDOW2014)*, WWW Workshops, 2014.
- [15] Samavi, R. and Consens, M.P. L2TAP+SCIP: An audit - based privacy framework leveraging Linked Data. In *8th International Conference on Collaborative Computing (CollaborateCom2012)*, 2012.
- [16] Albert, M., Guéret, C., Ashkpour, A. and Scharnhorst, A. Publishing, Harmonizing and Consuming Census Data: the CEDAR Project. In *Workshop proceedings*. 2013.
- [17] Zittrain, J. , Z., Albert, K., and Lessig, L. Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations. *Harv. L. Rev. F.* 127 (2014): 176-176.
- [18] de Sompel, H.V., Sanderson, R., Nelson, M., Balakireva, L., Shankar, H., and Ainsworth, S. An http-based versioning mechanism for linked data. *CoRR* (2010).
- [19] Papavasileiou, V., Flouris, G., Fundulaki, I., Kotzinos, D., Christophides, V.: High-level change detection in RDF(S) KBs. *ACM Trans. Database Syst.* 38(1): 1 (2013)
- [20] Zeginis, D., Tzitzikas, Y., Christophides, V.: On Computing Deltas of RDF/S Knowledge Bases. *TWEB* 5(3): 14 (2011)