Project no. 600663

PRELIDA

Preserving Linked Data
ICT-2011.4.3: Digital Preservation

D2.5 Final report on the midterm workshop

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

Consiglio Nazionale delle Ricerche
Istituto di Scienza e Tecnologie dell'Informazione

Version [draft,1]

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

---

Document Information

Deliverable number:	D2.5
Deliverable title:	Final report on the midterm workshop
Due date of deliverable:	03|2014
Actual date of deliverable:	04|2014
Author(s):	Carlo Meghini
Participant(s):	CNR ISTI
Workpackage:	WP2
Workpackage title:	Organizational Support
Workpackage leader:	CNR ISTI
Est. person months:	3
Dissemination Level:	PU (Public)
Version:	3
Keywords:	

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level
1	6.5.14	draft	Carlo Meghini	Initial draft
2	8.5.14	draft	Josè Maria Garcia	Revision
3	9.5.14	final	Carlo Meghini	Delivered version

Abstract

The present document provides details on the second workshop of PRELIDA, held in Catania from the 2nd to the 4th of April, 2014. The workshop was reserved to the PRELIDA Working Group members, and it was the second event in which the members met amongst themselves and with the project in order to discuss the preservation of Linked Data. The programme of the workshop is provided, along with the list of participants. The scientific outcome of the workshop is finally presented, by illustrating the themes that have been discussed and that will form the research agenda of the next PRELIDA developments.

Table of Contents

1. Introduction

PRELIDA aims at building bridges across the Digital Preservation and Linked Data communities, with the view of:

(a) making the Linked Data community aware of the already existing outcomes of the Digital Preservation community; and

(b) working out challenges of preserving Linked Data that pose new research questions for the preservation community. These challenges are related to intrinsic features of Linked Data, including their structuring, interlinking, dynamicity and distribution.

In order to achieve these goals PRELIDA has set up a Working Group composed of leading researchers and representatives of key sectors within the Digital Preservation and Linked Data communities. The Working Group is presented in Deliverable D2.1.

The members of the Working Group (WG) have been invited to the opening PRELIDA Workshop, which took place in Pisa in June 2013. During the first workshop, the WG members presented their views on the preservation of Linked Data and engaged in discussions amongst themselves and with the beneficiaries of PRELIDA. This reports gives an account of the second workshop of PRELIDA, named "midterm", which was held in Catania from the 2nd to the 4th of April, 2014.

The focus of the midterm workshop has been to discuss two deliverables that had been produced by PRELIDA during its first year:

- the initial state of the art in the Preservation of Linked Data [1]
- the gap analysis on tools and methods for the Preservation of Linked Data [2]

The present report gives an overview of the workshop, and it is structured as follows:

- Section 2 gives the list of participants to the workshop
- Section 3 gives the programme of the workshop
- Section 4 gives an account of the scientific outcome of the presentations and of the ensuing discussions
- Section 5 indicates where the on-line resources about the workshop can be found.

2. List of Participants

The workshop was reserved to the Working Group members of PRELIDA. The following members have participated:

Dimitris Kontokostas (University of Leipzig)
Paul Groth (VU University Amsterdam)
Sébastien Peyrard (BNF)
René van Horik (DANS)
Andrea Scharnhorst (DANS)
Milena Dobreva (University of Malta)
Maurizio Lunghi (Fondazione Rinascimento Digitale)
Mariano Consens (University of Toronto)
Marat Charlaganov (VU University Amsterdam & DANS)
Antoine Isaac (Europeana)

In addition, the following people from the PRELIDA beneficiaries have participated:

David Giaretta (APA)
Krystina Giaretta (APA)
Grigoris Antoniou (HUD)
Sotiris Batsakis (HUD)
José M. García (UIBK)
Carlo Meghini (ISTI CNR)

Finally, the ISTI CNR persons that have been in charge of the organizational aspects are:

Francesca Borri
Anna Molino

3. Programme of the Workshop

The workshop has lasted two full days, but in order to allow the participants to stay only two nights, it has spanned three days. The programme has been structured in three main Sections:

- An opening session devoted to present the work done from the Opening Workshop to the present time to the Working Group. The main focus concerned the PRELIDA State of the Art (D3.1) and the report on the Gap Analysis (D4.1).
- Four breakout sessions devoted to the discussion on four main issues elicited both by WG members and PRELIDA beneficiaries.
- A final discussion session, in which proposals about next steps for the work to be carried on in the project have been made.

In the closing part, the PRELIDA Coordinator has illustrated to the Working Group members what it is expected from them for the successful prosecution of the PRELIDA activities.

The detailed programme of the workshop is given below.

Wednesday, April 2nd

Opening session: "Work done so far"

14:00	14:15	Carlo Meghini	Welcome and Opening of the workshop
14:15	15:00	David Giaretta, René van Horik, Andrea Scharnhorts	PRELIDA State of the Art
15:00	15:45	Sotiris Batsakis	PRELIDA Gap Analysis
15:45	16:15	Coffee break	

Plenary: Discussion and Planning

16:15 18:15 Possible issues for Day 2:

- Research challenges
- Use cases
- Technological impact

18:15 Closing of Day 1

Thursday, April 3rd

9:00	11:00	Discussion on Ingesting Linked Data
11:00	11:30	Coffee break
11:30	13:00	Discussion on Ingesting Linked Data (cont.)
13:00	14:30	Lunch break
14:30	16:00	Discussion on Managing Change

16:00	16:30	Coffee break	
16:30	18:30	Discussion on Managing Change (cont.)	
	18:30	Closing of Day 2	
20:00		Social Dinner	

Friday, April 4th

9:30	11:30	Wrap up of the previous discussione	
11:30	12:00	Coffee break	
12:00	12:45	Carlo Meghini	Proposals and Next Steps
12:45	13:00	Carlo Meghini	Closing of the Workshop

4. Scientific outcome

The Opening Workshop was organized around presentations from the members of the PRELIDA Working Group around the theme of preserving Linked Data, and ensuing discussions. These presentations highlighted a number of issues, which have been used to create the two most important deliverables of the first year of PRELIDA, namely the State of the Art [1] and the Gap Analysis [2].

One of the most important results of the Opening Workshop, has been to raise awareness of the preservation problem within the Linked Data community, while at the same time informing the members of this community about the main conceptual tool produced by the digital preservation community, namely the OAIS Reference model.

The Midterm Workshop has built on these results, by being more focussed in addressing the concrete issues in archiving Linked Data by following the guidelines for conformance set by OAIS. Accordingly, the discussion in the Midterm Workshop has been centred around three main themes:

- ingesting a Linked Data dataset
- managing the changes that can impact on a Linked Data dataset
- accessing an archived Linked Data dataset

An account of each theme is given in the rest of this Section. In what follows, by "Linked Data dataset" (LDD for short) we mean a 5-star dataset, that is one expressed in RDF with links to a significant number of other web resources, including datasets but also web pages, documents, and in general anything that can be identified by an HTTP IRI. By making this choice we place ourselves in the most general and technically challenging case.

Ingesting a Linked Data dataset

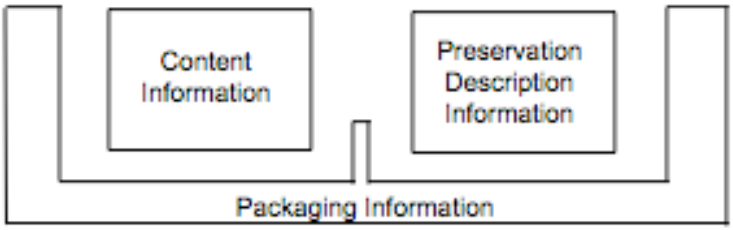
According to the Reference Model, "an OAIS is an archive, consisting of an organization, which may be part of a larger organization, of people and systems that has accepted the responsibility to preserve information and make it available for a designated community." In light of this, one of the goals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

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

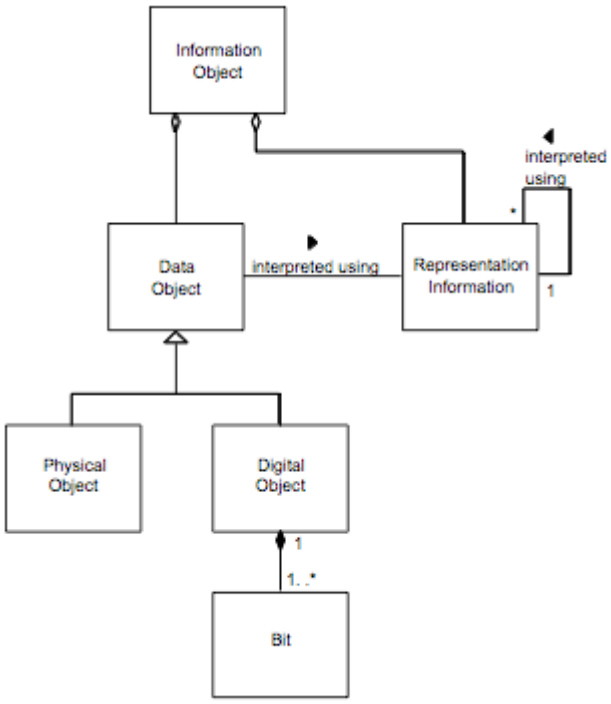
Indeed, linked data shifts away from the paradigm of self-containedness often assumed by archives. By their nature Linked Data refer to other resources outside the graph they belong, so that when archiving one graph, one has to decide what to do with the links going out of the graph. The question then naturally arises, what to do with the links. This question is a special case of a more general question concerning how to make a Submission Information Package (SIP) out of a LDD. A LDD is not an OAIS, it is, at best, just the content part of an OAIS. What do we need to add to a LDD such that it can be accepted by an OAIS as a SIP?

Making an SIP out of an LDD

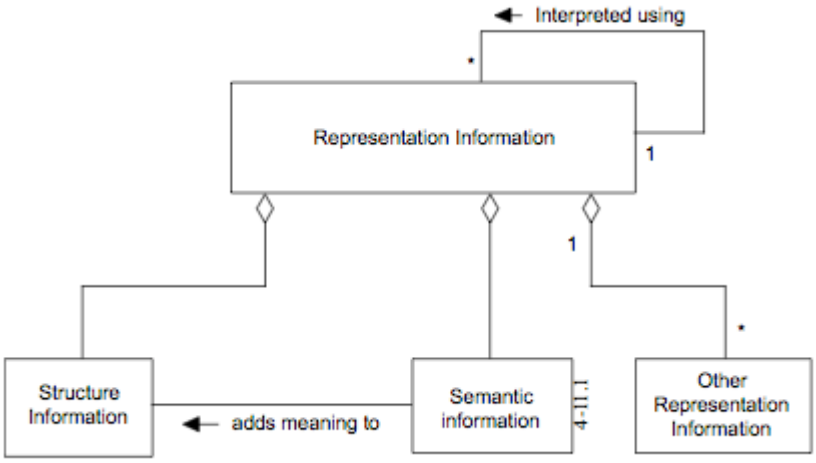
In the OAIS Reference Model, the unit of archiving and of exchange with the outside world is the Information Package (IP). An IP is structured as shown by the figure below (all the figures in this Section are taken from a publicly available version of the OAIS Reference Model [3]):



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



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



Structure Information is required to transform the bits of the Data Object into "something meaningful" to the designated community, while Semantic Information is required to support the members of the designated community in the interpretation of that "something meaningful".

Now, assuming that the data object to be preserved is an LDD, we can establish the following mappings between Representation Information and Linked Data:

- 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 (named graphs in RDF/XML require tricks, JSON-LD may not cover everything), and there is no evidence that some serializations are better than others. PRELIDA will try and establish a contact with the Data best practices W3C group, asking the group if they can make a recommendation on serialization for archive, or make all serializations kept fully compatible over time.
- Semantic Information, on the other hand, consists of two parts: the semantics of RDF (as given in http://www.w3.org/TR/rdf11-mt/) plus the semantics of the specific RDF vocabulary the graph is built on. The former is archived by the W3C in form of documents containing the various recommendations. The latter is given by the vocabularies (or ontologies, or terminologies) referred to by the Data Object.
- In addition, another important part of preserving the semantics of RDF datasets is preserving inference engines implementing *calculi* for computing implicit triples, both those coming from the semantics of RDF and those coming from the semantics of the user-defined vocabularies. One might simply preserve the documents where the *calculi* are described. However, re-implementing such *calculi* in order, *e.g.* to query an RDF graph, may not be the best option. A SPARQL engine may fit in the picture as "Other Representation Information".

Assuming W3C can successfully preserve their recommendations and the related papers, we are left with two problems:

- preserving the vocabularies used in an LDD, other than those defined by the W3C, and
- preserving RDF engines.

The latter problem has been no further discussed, as there is nothing specific to Linked Data about it. It amounts to preserving a service, and the repertoire of digital preservation solutions offers several options for doing it, including emulation.

The former problem has been discussed in an extensive way. Vocabularies exhibit a number of characteristics that make them somewhat special:

- they are small, compared to the size of LDD built on top of them; to wit, http://lov.okfn.org/dataset/lov/ stores many vocabularies, but the file gathering them all is only 8.4 megabytes, 64740 triples
- creating a vocabulary requires a large investment of human work, which is not necessarily the case with other types of LDD;
- mapping to vocabularies is a large investment of human work too - and creates a global understanding or at least a reference to something shared by others - this is the gist of the matter for LD
- vocabularies are linked from data (in statements). The strategies to retrieve vocabulary data will be different to what happens for XML documents and XML schemas.

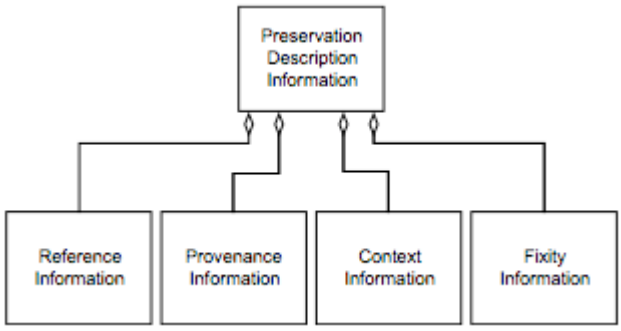
All this seems to indicate that ingesting vocabularies along with the content data is a viable option, and indeed the archivists in the discussion were interested in this solution, which must not be confused with the archiving of the web site where vocabularies are published. The positive effects of ingesting the vocabularies along with the data are: (1) the SIP is self-contained, and therefore its usability is maximized; (2) the content data in the SIP will always be aligned to the involved vocabularies. There are also negative effects, though: (1) any change to a vocabulary makes the SIP obsolete, even though the content part proper has not changed, which is rather paradoxical; this effect might be mitigated by the adoption of a specific change policy by vocabularies: never change the meaning of a term, but always create a new term to reflect the change. But there is no guarantee that the agents in charge of vocabularies will in fact adopt this policy. (2) Vocabularies may refer to other vocabularies, and so on, for an arbitrary level of recursion. Where to stop ingesting? OAIS may help to address this problem: in OAIS the recursion of Representation Information stops at the vocabularies that are part of knowledge base of the designated community. In the LD case, the knowledge base of the designated community can be at least partially expressed by a set of HTTP URIs of the known vocabularies. These URIs can be used to govern the ingestion of vocabularies.

The alternative solution is to ingest just the links to the vocabularies. The positive effect of this approach is that the SIP is somehow resilient to the changes that occur in the linked vocabularies, since the SIP can always point to the most recent version of a vocabulary, even one that did not exist when the SIP was created. But the SIP is no longer self-contained, and moreover the content data may no longer be aligned to the pointed vocabulary, if the latter has changed since the time when the former was created.

A mixed solution has also been discussed: an archive should ingest the domain-specific, more fragile (in terms of sustainability) vocabularies, while linking to the most common, more robust vocabularies, maintained by trustworthy institutions.

In any case, there has been general agreement that the archive needs to guarantee some form of completeness for versions of vocabularies, which is harder in LD than in XML Schema. It has been argued that there should be a VOID description that gives the versions of vocabularies used. It could be generated during ingestion or prior.

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:



In the case of LDD, reference is typically done via HTTP IRIs, while fixity information (defined in OAIS as "The information that documents the authentication mechanisms and provides authentication keys to ensure that the Content Information object has not been altered in an undocumented manner. An example is a Cyclical Redundancy Check (CRC) code for a file") pertains to a lower level of abstraction than the one discussed here.

According to the OAIS specification, Context Information is "the information that documents the relationships of the Content Information to its environment. This includes why the Content Information was created and how it relates to other Content Information objects". Now, RDF in general and Linked Data in particular, are designed to give a proper representation of context information (in the sense just given) by recommending IRIs for identifying both resources and their relationships, and by providing triples for establishing the connections between resources via relationships. By identifying the LDD to be preserved via an HTTP IRI, the full expressive power of RDF can be used to express Context Information. We note that the new developments of RDF, specifically RDF 1.1 N-Quads [8] provides the basic syntactic machinery for this purpose, while vocabularies such as VoID [4], in combination with DCAT [5], provide the classes and the properties for an RDF-based representation of the Context Information of a LDD.

The discussion at the Workshop has mostly focused on Provenance Information, defined in OAIS as "The information that documents the history of the Content Information. This information tells the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. Examples of Provenance Information are the principal investigator who recorded the data, and the information concerning its storage, handling, and migration". The PROV vocabulary [6] is recommended by the W3C for expressing Provenance Information. PROV is designed precisely to represent how the RDF was made, what is the history of this dataset before and after ingest. PROV is about documenting the provenance of an object, not about offering a metamodelling mechanism. In practice one sees in PROV links to a long list of named graphs [7]. Many providers are not very advanced in documenting the operations they have on the data. Even the more advanced ones still rely on scripts. The idea of PROV is to describe largely the operations and then point to the procedural world. The question is whether PROV is mature enough to be used for archival purposes. It is a part of the solution perhaps, but it does not solve the problem by itself. PROV contains information how the RDF graph is produced, not necessarily all documentations used in the production process. Here we see an opportunity of collaboration with W3C working group on Data on the Web Practices[1]. They are going to work about expressing the quality of datasets, based on DCAT and will be gathering requirements. They would be keen on feedback from the preservation community to define best practices. PRELIDA could suggest such

[1] http://www.w3.org/2013/dwbp/

requirements: for example what makes a good dataset from a preservation perspective, which properties you should use in a VoID/DCAT dataset description to give it archival quality.

Managing evolution

An information system (and therefore a Linked Data dataset) is a symbolic model of some slice of reality. As such it is a living object whose continuous evolution reflects the evolution of the modeled reality. On the other hand, an archive preserves something static: a snapshot, something a certain community agrees should not disappear. There must be a way of reconciling these two aspects, defining a set of tasks and assigning responsibilities for carrying out those tasks amongst the involved actors: the owner of the data, the OAIS and the designated community.

The evolution problem has two aspects:

- the *detection* of change, and
- the *management* of change.

We will discuss both issues in the following, starting with change management.

Change management

We started with an analysis of the changes that may have an impact on preservation. There are several types of such changes:

- Changes in the technology used by the archive to preserve the data. For 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. Also this type of changes is rather uncontroversial from the preservation point of view: 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. We must note that research in digital preservation provides us with solid methods and tools for dealing with this kind of problems, and the application of these methods and tools to LDDs does not pose any additional problem.
- Changes in the Content Data being preserved. For example, the DBPedia LDD is continuously updated with new triples. 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.
- Changes in the Representation Information or to the Preservation Description Information that are added to the Content Data for preservation purposes. For example, as a consequence of a migration to a new serialization format of the preserved LDD, new Representation Information is created that describes the new format. Or, a new role is created in the designated community to handle the responsibility for the preserved LDD, and this new role is recorded in the Context Information of the LDD. This case is similar to the previous one, in that a new Submission Information Package is created which must be ingested and properly related to the previous one. But there is the additional issue that the update to the data may propagate to the ontologies defining the terms used in the metadata. For instance, a new format must be added to the ontology of file formats, or a new role must be added to the ontology of the organization roles. Changes in the ontologies are discussed next.
- Changes in the vocabularies used in the LDD or in the additional information stored with it. For instance, 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 archive is updated by the authority maintaining it: a new term for planet is introduced and properly axiomatized, whereas the old term is deprecated. Or, as a consequence of political evolution, East Germany

and West Germany no longer exist because they are (re)united into Germany; in order to reflect this new situation, the gazetteer that was in use in the archive is updated by the authority maintaining it: a new term for Germany is introduced, whereas the old terms for East and West Germany are deprecated. We first note that if the referenced data that have changed, have been ingested in the archive for self-containedness (an alternative that has been previously considered), then the problem reduces to managing changes in the Content Data, discussed above. But in this case there is one additional issue to be dealt with: the change in the referenced vocabularies may impact other parts of the preserved LDD. For instance, the concept of planet may be used in the Representation Information of the LDD, or the term "East Germany" may be used in the Context Information of the LDD. Since both terms are deprecated, they will soon be obsolete; as a consequence, the Designated Community will no longer understand the Representation Information and the Context Information containing them. This is an open problem in digital preservation. Below we discuss how it can be attacked in the preservation of Linked Data.

- Changes in web resources other than those discussed here. For these, web archiving solutions have been indicated.
- Changes in the knowledge base of the designated community. For example, the term "planet" has acquired a new meaning as described above, but in this case there is no formal ontology defining it in a formal way; the term is only defined in the textbooks of the designated community and directly used, *e.g.* in some Representation Information. This case is similar to the previous one with the difference that there is no ontology to be updated: this fact simplifies one aspect, but leaves the same propagation problem as in the previous case. Additionally, the detection problem becomes somehow harder: the change in the knowledge base may go unnoticed for some time, since there is no digital representation of it.

The change management problem that remains open concerns therefore the propagation of ontology change to the archived descriptions (Representation Information or Preservation Description Information) that contain it. The way of tackling this problem strictly depends on the requirements of the designated community. In particular:

- If the designated community requires accessing and using the preserved data on the basis of the new term, then the occurrences of the old term have to be replaced, and this implies a re-rewriting of parts of the LDD. Techniques for doing so have been researched in the context of RDF. 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 [10] or COntoDiff/CODEX [11][12]. A more general approach to concept evolution can be found in [9]. The modified data have to be re-ingested, and it is the responsibility of the archive to maintain the proper connection between the previous and the updated data.

  Of course there are lighter ways of coping with these changes. For instance, an archive may just add to the Representation Information or to the Preservation Description Information a reference to a book explaining the difference between the current and the previous notion of planet, or to an historic atlas showing a map of Europe at the time when East and West Germany existed. Or, it may just indicate that there has been a change in the context (vocabulary) that may matter for the designated community.
- If the designated community requires accessing and using the preserved data on the basis of both the old and the new term, then mappings have to be created and used in the access function of the archive. This problem reduces to mapping the new vocabulary (*i.e.*, the

language including the new term(s)) to the old one, and for doing this a number of techniques developed in the last decade in the context of data integration on the web, can be employed.

Change Detection

The issue of change detection has been just touched upon during the workshop.

Detecting all possible changes described in the previous Section is a very expensive and time-consuming process that can be approached only on a community-based way, and the larger community the bigger are the chances of success.

From an archival point of view, a valid option is that the owner of the data also provides tools to detect changes, which archive could run to create documentation about the context of interpretation (if the vocabulary changes). Diff tools for ontologies are an example.

The SCIDIP-ES project[2], for instance, explores a (notification) service for the Alliance of Permanent Access, with a business plan: people can subscribe to areas of interest. People knowing that a vocabulary changes could triggers alerts to be changed. This approach could well complement the harvesting that LOV[3] does. LOV harvest 3 to 400 vocabularies on a daily basis, make a diff and keep the vocabulary if it has changed. They version the vocabularies. BNF considers capturing a snapshot of this website to keep a secured backup copy for those 3 to 400 vocabularies.

On file formats, Preservica[4] offers a Linked Data registry as a resource that can be used to be notified about the obsolescence of file formats.

Accessing a Linked Data dataset

Technically, Linked Data require the web infrastructure in order to be accessed. Clearly, preserving the access service would require preserving the web infrastructure (HTTP clients and servers) and this is something not doable by a single archive. An archive is not responsible for keeping a content negotiation service alive. It is responsible of keeping the data, and the context info for putting the service live again. However in some case, an archive's business model may trigger it to revive a defunct de-referencing service (on its namespace) and maintain it. In a reasonable scenario, an organization passes an LDD in an SIP to an archive and gets it back as a DIP that the organization deploys in the appropriate way to make the data accessible according to the Linked Data paradigm.

An interesting question is how it is possible to cope with the disappearance of Linked Data providers. In the favorable case, some archive can take back the domain name to keep redirects, or the archive can be put in some other place online (*e.g.* Internet Archive). However, there also some un-favorable cases, such as that by Kasabi (from Talis) who died but the owner kept their namespace, so the data has to be reconstructed because the URIs in the data do not resolve to anything useful anymore. In order to avoid these situations, there should be best practices for minting URIs with preservation concerns in mind.

CLOCKSS[5] have a model of triggered content when the archived content (e-publications) disappears on the web (publisher dies) make this data available again. DANS[6] has a number of cases on maintaining services, which need someone to take responsibility for *e.g.* an online analysis tool for social science research data: http://nesstar.dans.knaw.nl/webview/

---

[2] SCIDIP-ES.eu

[3] http://lov.okfn.org/

[4] http://preservica.com/

[5] http://www.clockss.org/clockss/Home

6 http://www.dans.knaw.nl/en

Conclusions

In the workshop, the problem of archiving a Linked Data dataset has been discussed, as well as the issues involved in coping with changes and with accessing the dataset. Making a LDD preservable still has some open issues. Several of these issues are being addressed by interest groups within the W3C and in other organizations involved with the preservation of digital information. These have been indicated in due course.

At the end of the discussion, the Working Group developed a list of desirable actions and features for a preservable LDD. This list is just a starting point that is used to identify all the features a LDD archive should consider. In practice, recommendations would probably not have all this.

1. Selection and appraisal of data: identify the boundaries of the LDD that has to be preserved, perhaps using a Concise Bounded Description as defined in [13].
2. Gather every RDF datasets (using quads to identify RDF graphs) that are relevant for the LDD to be preserved. Default strategy is complete closure. Both for vocabularies (ontologies) and instances. There are vocabularies that describe the provenance of crawl/imports/ingests of linked data (*e.g.,* see http://ldif.wbsg.de/#provenance).
3. Whenever a LDD is collected into a SIP, the owner of the LDD should be alerted that any change in that LDD is relevant for the collector, who is made part of a list of subscribers that have to be notified of any change. For notifications when a dataset changes: ResourceSync[7] can be used (it is used for instance for dbPedia synchronization, a lightweight alternative for LDD (using VOID) is offered by dady[8].
4. Submit data in a standard serialization (such as N-quads). Consider conversion between formats.
5. Include VoID/DCAT/PROV description in Representation Information. Also Resource Shape-like [14] data validation instructions. And the corresponding ontologies (DCAT and VoiD ontologies, *etc.*).
6. Specification documents should be also preserved, (*i.e.,*RDFS, OWL, serialization specs).
7. Time-stamps for the crawls of the collected datasets. Perhaps several ones: the snapshot time, the date of last modification, *etc* (the snapshot time is most important).
8. Reasoners and SPARQL engines (triple store) are also to be preserved for accessing purposes.
9. Submit every representation (HTML+RDFa, JSON) served in content negotiation. It has to be negotiated between producer and archive, in the light of what is wanted in the Dissemination Information Package. For the HTML part we could rely on existing web archiving (*e.g.* Korea national library has done work on this).

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

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

[8] https://code.google.com/p/dady/wiki/Demos

Acknowledgements

The PRELIDA beneficiaries are grateful to the participants to the workshop who engaged in the discussions that took place at the workshop, freely contributing their knowledge and experience to create a common understanding of the problems involved in preserving linked data.

References

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

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

[3] Consultative Committee for Space Data Systems. REFERENCE MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS). Magenta Book CCSDS 650.0-M-2. June 2012. Available at http://public.ccsds.org/publications/archive/650x0m2.pdf

[4] Keith Alexander, Richard Cyganiak, Michael Hausenblas, Jun Zhao. Describing Linked Datasets with the VoID Vocabulary. W3C Interest Group Note 03 March 2011 http://www.w3.org/TR/void/

[5] Fadi Maali, John Erickson. Data Catalog Vocabulary (DCAT). W3C Recommendation 16 January 2014. http://www.w3.org/TR/vocab-dcat/

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

[7] Fabien Gandon, Olivier Corby. Name That Graph, or the need to provide a model and syntax extension to specify the provenance of RDF graphs. http://www.w3.org/2009/12/rdf-ws/papers/ws06/

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

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

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

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

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

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

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

5. On-line Resources

On-line resources related to the second PRELIDA Workshop are for the moment hosted on the private area of the PRELIDA web-site (www.prelida.eu).

The resources include:

- the scientific programme of the workshop, giving the main objectives of the workshop and a list of potentially interesting topics;
- the agenda of the workshop, endowed with the audio-visual recordings of the presentations with the slides used by the presenters.

This information will be made available to the general public the present report, by publishing them on the public section of the web site and on social media. In addition, the slides used in the presentations are available on Slideshare.

A revised version of the present report will also be published, the revision concerning the collection of feedback by the Working Group members and the consequent modifications