Project no. 600663

PRELIDA

Preserving Linked Data
ICT-2011.4.3: Digital Preservation

D2.2 Final report on the opening 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:	D2.2
Deliverable title:	Final report on the opening workshop
Due date of deliverable:	06|2013
Actual date of deliverable:	07|2013
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:	1
Keywords:	

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level
1	July 22	draft	Carlo Meghini	Initial draft

Abstract

The present document provides details on the first workshop of PRELIDA, held in Tirrenia (Pisa) from the 25th to the 27th of June, 2013. The workshop was reserved to the PRELIDA Working Group members, and it was the first 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 and a brief abstract of each talk. 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

Document Information 1
Abstract 1
Executive Summary 3
1 Introduction 4
2 List of Participants 5
3 Programme of the Workshop 6
4 Abstracts of Presentations 8
5 Scientific outcome 13
Introduction 13
Accessibility and Re-use 14
Access models for LD 14
Storage models for LD 15
Dataset packaging 15
Quality assessment of LD 15
Coping with change 16
Persistent URIs 16
Annotating RDF triples 17
Dataset Versioning 17
Dataset Disappearance 17
Sustainability via data marketplaces 18
Panelist presentations 18
Discussion 20
Conclusions 20
6 On-line Resources 22
7 Conclusions 23
8 References 24

Executive Summary

A 1-2 page long executive summary providing an overview of deliverable contents including objectives, results/findings and evaluation of the work.

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 have been invited to a face-to-face workshop in order to present their views on the preservation of Linked Data and engage in discussions amongst themselves and with the beneficiaries of PRELIDA. This workshop is the opening workshop of PRELIDA.

More specifically, the focus of the opening workshop has been to present the state of the art in Digital Preservation solutions and Linked Data technologies and usage; and to initiate discussion between the two communities regarding the challenges of preserving Linked Data and possible ways of addressing them.

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 a short abstract for each presentation by the Working Group members
- Section 5 gives an account of the scientific outcome of the presentations and of the ensuing discussions
- Section 6 indicates where the on-line resources about the workshop can be found
- Section 7 concludes.

2 List of Participants

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

Robert Sharpe (TESSELLA)
Richard Cyganiak (DERI, NUI Galway)
Jamie Shiers (CERN)
Mariella Guercio (Sapienza University - Rome)
Soren Auer (Leipzig University)
Lars Svensson (DNB)
Jan Brase (DataCite)
Brian Matthews (STFC)
Vassilis Christophides (FORTH-ICS)
Elena Simperl (University of Southampton)
Antoine Isaac (Europeana)
Fabrizio Gagliardi (Microsoft Research and ACM)
Phil Archer (W3C)
José Borbinha (INESC)
Peter Buneman (University of Edimburgh)
Orit Edelstein (IBM Haifa)

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

David Giaretta (APA)
Krystina Giaretta (APA)
Grigoris Antoniou (HUD)
José Garcia (UIBK)
Carlo Meghini (ISTI CNR)

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

Francesca Borri
Sara Manca
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 the been structured in three main Sections:

- An opening session, which has been devoted to present PRELIDA to the Working Group members and to introduce the main themes: (1) Linked Data, (2) Digital Preservation and (3) the Preservation of Linked Data. The presentations in this session have been given by the PRELIDA beneficiaries (themes (1) and (2)) and by a guest speaker, Vassilis Christophides, who has a role in both scientific communities and is therefore a natural candidate to introduce the preservation of linked data.
- A section devoted to presentations by the Working Group members. This section has been divided into three main sessions, which mirror the three main themes introduced above.
- A panel section, where the participants have discussed a hot topic in the area of PRELIDA, namely data marketplaces and digital preservation.

In the closing session, 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.

Tuesday, June 25

Opening session (Chairman: Carlo Meghini)

14:30	15:00	Carlo Meghini	Welcome and Introduction to PRELIDA
15:00	16:00	David Giaretta	Digital Preservation
16:00	17:00	José Garcia	Linked Data
17:00	17:30	Coffee break	
17:30	18:30	Vassilis Christophides	Making Open/Linked Data Diachronic
18:30		Carlo Meghini	Closing of the 1st day

Wednesday, June 26

Session: *Linked Data* (Chairman: Grigoris Antoniou)

8:30	9:00	Antoine Isaac	Europeana: Linked Data at work in Cultural Heritage
9:00	9:30	Elena Simperl	Human computation in the Linked Data management lifecycle
9:30	10:00	Jamie Shears	Data Preservation at the Exabyte Scale: Challenges and Opportunities
10:00	10:30	Soren Auer	Linked-Data Life-cycle and diachronic referencing, evolution and archiving of Linked Data
10:30	11:00	Discussion	

11:00	11:30	Coffee break	

Session: *Preservation & Linked Data* (Chairman: Vassilis Christophides)

11:30	12:00	Peter Bunemann	Provenance, Annotation, Archiving and Citation - why can't we make it work for linked data?
12:00	12:30	Brian Matthews	Publication of facility investigations
12:30	13:00	Robert Sharpe	ENSURE Linked Data Registry
13:00	13:30	Discussion	
13:30	15:00	Lunch	

Session: *Preservation* (Chairman: David Giaretta)

15:00	15:30	Phil Archer	The 10 DOs and DON'Ts for persistent URIs
15:30	16:00	Mariella Guercio	Controlled vocabularies, metadata standards and linked data for digital preservation: the case of Sapienza Digital Library
16:00	16:30	Jan Brase	DataCite - persistent links to scientific data
16:30	17:00	Orit Edelstein	ENSURE: Enabling knowledge sustainability, sustainability and recovery for economic value
17:00	17:30	Discussion	
17:30		Carlo Meghini	Closing of the 2nd day
20:00		Social Dinner	

Thursday, June 27

Panel: *Data Marketplaces and digital preservation*

9:30	11:30	Fabrizio Gagliardi, Vassilis Christophides, David Giaretta, Robert Sharpe, Elena Simperl	
11:30	12:00	Coffee break	
12:00	12:30	Carlo Meghini	Next Steps
12:30		Carlo Meghini	Closing of the Workshop

4 Abstracts of Presentations

Welcome and Introduction to PRELIDA *Carlo Meghini*

Carlo Meghini illustrated the structure of PRELIDA [1], its objectives and the instruments to achieve them, with an emphasis on the role of the Working Group. He then went on to introduce the present workshop, motivating its programme and explaining the main expectations form the PRELIDA beneficiaries. He concluded with an outlook of the next 18 months, stressing the contribution that the Working Group members were expected to give to the success of PRELIDA.

Digital Preservation *David Giaretta*

David Giaretta gave an introduction to the fundamental concepts of digital preservation. He started from the main points of the 2030 vision given in the report "Riding the wave" [2], created by the High level Expert Group on Scientific Data, in which David had the role of rapporteur. He then went on to illustrate the threats that make preservation necessary, and how these threats are affected by the present context, which is characterized by the exponential growth of the volume of data that hold value to our society. Next, David introduced a major tool to address preservation, namely the OAIS reference model [3], providing an overview of the functional and information models of OAIS. Special emphasis was given to trust, which was illustrated through a number of questions that an OAIS must be able to properly address. Finally, David concluded by elaborating on the issues posed by the preservation of Linked Open Data.

Linked Data *José Garcia*

Jose Garcia's presentation was about data. He opened his presentation with Big Data. He offered some figures on the amount of data currently produced or consumed by the devices that pervade our life, pointing out the difficulties on performing any form of deduction at this level of scale. He then argued that data are best looked at from a streaming perspective, shifting the inference problem to "logical reasoning in real time on multiple, heterogeneous, gigantic and inevitably noisy data streams in order to support the decision process". He then focused on Open Data, giving definitions and basic principles underlying the movement for open data, and showing some success stories on the usefulness of making open data available to applications in machine-readable formats. He finally moved to Linked Open Data, stating the four principles and the five levels of quality classification of Linked Open Data. He offered a historical perspective on the Linked Data Cloud, starting from the small cloud in 2007, featuring a dozen datasets, to the almost three hundreds datasets in 2011. He concluded his presentation talking about the data economy and showing successful applications of Open data.

Making Open/Linked Data Diachronic *Vassilis Christophides*

After introducing the data economy, Vassilis presented the Diachron view on preserving linked data. Diachron[1] is a 3-year IP project in Preservation (full title "Managing the Evolution and Preservation of the Data Web") started on April 1st, 2013. Diachron aims at injecting preservation into the complex lifecycle that data undergo in the new scenarios created by the web and the data economy. Specifically, Diachron aims at preserving (semi-)structured, interrelated, evolving data by keeping them constantly accessible and reusable from an open framework such as the Data Web. This objective calls for effective and efficient techniques to manage the lifecycle of web data involving the

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

many actors that have a role in the new data lifecycle, such as data producers, curators, brokers and consumers. In particular, Diachron embraces the view of *Pay-as-you-go* data preservation, in order to spread the costs of preservation among key players in a community of interest. As a result, we will have so-called diachronic data, that is data enhanced with temporal and provenance annotations. Achieving its goals and views will require Diachron to face a number of challenges, described and commented by Vassilis in the core part of his presentation, such as: data quality, data appraisal, provenance, annotation, evolution, curation, citation, long-term accessibility, archiving and longitudinal querying.

Europeana: Linked Data at work in Cultural Heritage *Antoine Isaac*

Antoine Isaac introduced Europeana, highlighting its mission, network and role in the Cultural Heritage landscape. He stressed that Europeana harvests, processes and makes available metadata about cultural heritage artifacts, and does so by following an open data policy; to this end, Europeana has persuaded its contributing institutions to adhere to the Creative Common zero license for the metadata, which allows any re-use of the metadata. He then went on to illustrate the linking that Europeana does, both internally in order to increase the level of connection amongst the metadata it collects from possibly different sources, and externally to authorities (such as GEMET, Geonames and DBpedia) providing shared identifiers for well-known entities. The presentation then turned into the preservation issues that Europeana is currently experiencing, due to the frequent changes that occur in the descriptions that it collects from sources. Several themes were introduced for further discussion, all centered around how to properly deal with the changes that occur in Linked Data. Some of the possible solutions mentioned were named graphs, RDF quadruples, and versioned URI; also the MEMENTO project was mentioned as a way of dealing with the diachronic nature of descriptions.

Human computation in the Linked Data management lifecycle *Elena Simperl*

Human computation is an approach towards problem solving in which a task that a machine or an algorithm cannot solve with the desired quality (cost, efficiency, accuracy), is outsourced to humans. In the area of semantic technologies, typical tasks that humans do better than machines are domain modeling, data source integration and semantic mark-up of digital artifacts. These tasks are knowledge intensive or require an amount of contextual information that is not easily encoded in a representation or in an algorithm. The talk reviews the dimensions of human computation and connect with the theme of the workshop by discussing the challenges arising in using human computation for linked data management: translation of high-level tasks into micro tasks that human can do; how to combine with automatic tools; how to embed into an existing application; how to re-use the data obtained in this kind of activity.

A perspective on the preservation of Linked Data *Richard Cyganiak*

Linked Data preservation is easier that the preservation of other types of data because RDF allows to provide knowledge about the terms used in a representation within the representation itself (for this reason RDF it is sometimes called as self-describing). As such, representation information and context (two types of knowledge recommended by OAIS for preservation) of a Linked Data dataset may be found in the data set itself, expressed in an explicit and machine-processable way. At the same time, Linked Data preservation is harder because it is tied to a particular technical infrastructure, namely the infrastructure implementing the web architecture. For instance, if the domain name is lost, a dataset can no longer be Linked Data (cf. TimBL's four principles), even though the data may still be useful. Similar to preserving the web for humans. And the problem is there even if the web infrastructure is

up and running, and all URIs are working, because there may be changes in some resource's description that may break the descriptions that refer to it and therefore rely on it.

Data Preservation at the Exabyte Scale: Challenges and Opportunities *Jamie Shears*

This talk summarizes the status, strategy and goals of Long-Term Data Preservation across High-Energy Physics experiments, institutes and sites worldwide. It outlines the plan for collaborative, sustainable solutions (for the coming decades at least) as well as metrics by which our progress can be measured. It then argues that all (usefully) preserved data is linked data in some form and discusses how inter-disciplinary collaboration is the best approach for turning a problem statement into a solution. Background: DPHEP website: http://www.dphep.org/, DPHEP Blueprint: http://arxiv.org/pdf/1205.4667.pdf.

Linked-Data Life-cycle and diachronic referencing, evolution and archiving of Linked Data *Soren Auer*

The LOD2 project is briefly presented, based on the Linked Open Data lifecycle. The project aims at researching and developing tools for supporting storage, authoring, interlinking and enrichment of Linked Open Data. LOD2 does not focus on preservation, which is more an objective of Diachron, where quality is also tackled. Several dimensions of quality are reviewed, and the relations between them and with other quality metrics such as 5-stars are discussed. Different type of statistics for a dataset are considered, some of which are easy to preserve, for instance those concerning the size of the data set. By preserving those statistics it is possible to reconstruct the history of a dataset from a dimensional point of view. The talk touches upon the lifting of data published data portals. The portals that perform such activity may be important for the preservation, as they offer an access point to a wealth of data that are aggregated in this way. Finally, several types of preservation are discussed, based on the different linked data to be preserved.

Provenance, Annotation, Archiving and Citation - why can't we make it work for linked data? *Peter Bunemann*

The presentation analyses the problems created by the preservation of databases, and reviews one possible solution, based on annotating the data elements with time. The solution is part of the XArch archive management system[2] that allows one to create, populate, and query archives of multiple database versions. XArch is based on a nested merge approach that efficiently stores multiple database versions in a compact archive. The system allows one to create new archives, to merge new versions of data into existing archives, and execute both snapshot and temporal queries using a declarative query language. XArch also allows users to explicitly state who believes a certain tuple, and the talk reviews how time and believe annotations propagate from the database structures to query results. RDF is then considered as a data model, and it is shown through examples that the passage from XArch to an RDF-based representation may encounter serious difficulties. Nested RDF is finally reviewed as a possible way of solving the problems posed by annotating RDF triples.

Publication of facility investigations *Brian Matthews*

The journal articles, e-prints, reports and other similar artefacts which can be considered "documents" are the well established means of research communication, with a clear identity as well as an aid for tracking the state and the trends of a research discourse. We propose that other emerging types of

[2] http://xarch.sourceforge.net/

intellectual entities, called Investigations in our research domain but having counterparts in other domains (e.g. experiments, observations, studies etc.), have some essential features of their lifecycle similar to the document-like entities and are natural candidates for being first-class members of a research discourse, and thus suitable subjects for publication and preservation. These investigations provide a record of the totality of a research process, in our case from the point of view of an operator of a large-scale scientific facility.

We consider the representation and construction of Investigation-like entities as linked compound research objects, considering the tools and methods required for their representation, and publication, and take a look at issues associated with their long-term preservation. We consider, the implementation of a domain-specific metadata models implemented in the facilities data catalogues, supported by notions of data publication including assigning persistent identifiers (e.g. DOIs), citation and preservation that now move towards the open repository model for data sharing empowered by Linked Data, metadata modeling and semantic technologies. These observations lead to better understanding of what artifacts (or structured collections of artifacts as "Research Objects") are likely to represent a research discourse in near future, and what challenges this presents.

ENSURE Linked Data Registry *Robert Sharpe*

This talk will describe how linked data could be used to help solve digital preservation problems (rather than being about preservation of linked data *per sé*). It will discuss current registry initiatives (in particular PRONOM) and what are the issues with using such registries (inflexibility of the data model and thus ability to deal with issues of local importance, slow ability to turn around data updates and lack of ability to synchronize information with other organizations). These issues can be addressed via a registry based off linked data. It will discuss the sort of information needed by the ENSURE (and other projects) within a registry and how the information space can be managed and extended. It will also touch on some of the technical issues in trying to maintain flexibility and yet still provide a user friendly system.

The 10 DOs and DON'Ts for persistent URIs *Phil Archer*

Designing and managing URIs for persistence requires careful planning and an explicit commitment to long term preservation that can credibly outlive the organisation making the commitment. As part of their work for the European Commission's ISA Programme, W3C and PwC carried out a study on persistent URIs, the highlights of which will be presented.

Controlled vocabularies, metadata standards and linked data for digital preservation: the case of Sapienza Digital Library *Mariella Guercio*

If preservation can be defined as interoperability over time independently from the technological platforms, the capacity of creating and keeping open and contextualized data able to be *communicated to future users* is crucial and is strictly connected with the exploitation of linked data potentialities. Linked Data are data published on the web, machine-readable and intelligible, expressed as a bit-stream made of characters and mark-up, typically are first developed within a disciplinary environment but they are potentially open to be shared outside the specific original domain as part of large and intermingled contexts. The massive development of information and digital resources and the low level of manual metadata available today and in the future can find solution for an effective management only thanks to technology based on Linked Data. But many crucial questions are still open and many requirements have to be fullfilled. The presentation is organized in two parts dedicated to some introductory general remarks and open questions on digital preservation and linked, data and to the analysis of the first steps of Sapienza Digital Library as a case study to discuss.

DataCite - persistent links to scientific data *Jan Brase*

The presentation reviews the main features of the Datacite system, which offers to its user a service for citing research products such as papers, datasets, collections and in general everything that may be used as a foundation for deriving a scientific result. Datacite uses DOI for persistent identification and at present has register 1.7 millions DOIs. It does not host any product, but only a metadata record about the product. The metadata schema has been developed by Datacite and is based on several standards. Datacite supports a querying system that allows users to view the most recently added products, and also provides several kinds of statistics. More at http://www.datacite.org/

ENSURE: Enabling knowledge sustainability, sustainability and recovery for economic value *Orit Edelstein*

The presentation reviews the architecture and underlying principles of the ENSURE preservation system, a new generation preservation system currently being researched and developed by the FP7 Integrated Project ENSURE[3] (Enabling kNowledge Sustainability Usability and Recovery for Economic value). The key aspect of the ENSURE architecture is the separation of its functionality into two layers: the configuration layer and the system run-time. The configuration layer takes as input requirements and parameters from the users and reasons on them in order to derive an optimal architecture. The reasoning is carried out by the Preservation Plan Optimizer, which evaluates cost vs. risk, economic value and quality of the possible solutions. The user is proposed a series of alternative architectures and can choose one of them for deployment. The deployment is performed by the run-time layer, which includes a lifecycle manager for the preserved digital assets, a content-aware long-term data protection module, a preservation-aware storage service and finally a preservation infrastructure. The preservation infrastructure copes with the evolution of the current deployed architecture in order to counter the effects of obsolescence. It relies on the Linked Data registry (presented elsewhere) in order to acquire information on obsolescent components and possible replacements.

[3] http://ensure-fp7-plone.fe.up.pt/site

5 Scientific outcome

This Section presents some initial considerations regarding the preservation of Linked Data. These considerations are outlined in the form of an article and result from a ri-elaboration of the considerations and ideas that have emerged during the presentations and the discussions at the workshop. The re-elaboration is done by the PRELIDA beneficiaries and will be subsequently submitted to the Workshop participants for feedback.

Introduction

The preservation of Linked Data is a special case of the preservation of formal knowledge. Although formal knowledge plays a major role in preservation (as exemplified by the OAIS Information Model), its preservation has received a moderate amount of attention so far. In what follows, we will address the problem of preserving Linked Data by first considering the more general problem of preserving formal knowledge, and then considering the specific features of Linked Data and discussing how these features affect preservation. This approach will place our study at a level of abstraction where the preservation problem can be analysed independently of the many technical details that are involved in the implementation levels and that may obscure, if considered too early, the core aspects.

Unlike natural language, formal knowledge has a well-defined meaning and as such it best lends itself to preservation, which requires the communication of the meaning of the preserved knowledge in time. In addition, any formal knowledge representation language has the expressivity required for representing the meaning of its expressions. Such meaning is normally given in terms of axioms, occasionally accompanied by semantic conditions. This property of formal knowledge, sometimes called as *self-description*, is relevant to preservation because it permits to use the language of the knowledge to be preserved to also represent the meaning of its terms. In OAIS terms, this amounts to say that Representation Information can be given in the same language as the content object.

However, preservation requires the expression of several kinds of knowledge about the preserved object, not only its meaning. The OAIS Information model goes in some details in illustrating what knowledge needs to be represented and preserved about a certain object; in particular, it prescribes provenance, context, reference and fixity as necessary to preservation, by including them in the Preservation Description Information. Now, the provenance of a piece of knowledge is meta-knowledge, in the domain of discourse where the knowledge to be preserved belongs, and so are all the other categories, for that matter. Meta-knowledge cannot be represented easily in formal knowledge languages, in fact none of the languages in use nowadays, from description logics to first-order logic, allows to do that. Here we therefore loose the ability of using one and the same language for preservation purposes, and we need some device to be able to represent and reason the required meta-knowledge. A commonly adopted device is reification, whereby a well-formed formula is regarded as an individual of the domain of discourse and as such denotable in the language by a constant symbol that can be used to express whatever knowledge needs to be expressed for preserving the original, object-level knowledge. But other devices, such as nested representations, may also be possible.

In sum, we can conclude that the preservation of formal knowledge is made easier by its having a precise meaning as well as the machinery to express it, but the ability of using one and the same language for preserving knowledge is defeated by the necessity of representing meta-knowledge.

Let us now consider the case for Linked Data (LD).

LD datasets can be seen as formal knowledge bases expressed in the RDF language. As such, they inherit the above introduced features regarding preservation. Indeed, the meaning of LD datasets is given in vocabularies (also known as ontologies), which are LD datasets themselves, whereas

metadata about LD are not LD themselves, as it will emerge during the discussion below. In addition, LD possess two basic characteristics that have a strong impact on their preservation, namely:

- LD are tightly coupled with the web architecture infrastructure. This creates a dependency of the representation on an operational environment, somehow defeating the declarative nature of knowledge. For instance, if a domain ceases to exist, a LD dataset referring that domain ceases being LD, even though the knowledge that it contains is still valid and possibly useful. To quote [6] "de-referenceable URIs depend on the provision of an online service, one that cannot be maintained without some agency funding the relevant server infrastructure. Such funding is itself ultimately dependent on a decision that the cost is less than the benefit, a balance that is very much subject to change in either direction over time".
- LD are distributed, since a LD dataset can use URIs whose meaning is defined in another dataset, stored at a different location. Distribution is a key factor of the success of LD, which has allowed the dramatic growth of LD datasets that we have witnessed in the last few years; yet, distribution does not bode particularly well for preservation, because it defeats self-description. Whenever the axioms (triples) capturing the meaning of the terms used in a LD dataset are part of a different dataset (and this is generally the rule), then the former dataset is not self-descriptive, and its interpretation and correct usage (such as inference) depend on the existence and accessibility of the latter; if the latter dataset is no longer available, then it will not be possible to correctly interpret the former dataset, however well curated it be. Inn terms of preservation, this implies that the preservation of a LD datasets depends on the preservation of all the datasets that it links to, up to an arbitrary level of nesting. Since linking to other datasets is a recommendation of LD, which may gain 5 stars to a dataset, it must be noted that LD and preservation, at present, are not pushing towards the same direction, in spite of the widely recognized necessity of preserving LD.

Another important aspect of LD that may have a negative impact on preservation is their relative immaturity. Although widely and increasingly popular, LD datasets are a rather emergent asset in the IT landscape, and as such we do not yet have the necessary understanding and tools for properly managing several aspects of their usage. This is a somewhat minor problem that time will alleviate, provided of course the right actions are taken by the community developing LD tools. In what follows, several recommendations will be made in this direction.

In the rest of this Section, we will further elaborate on these and other related features of LD, examining in some detail the current status of LD development in each of them, the impact that they have on preservation. The exposition is organized along the three major aspects of preservation: accessibility and re-use, coping with change and sustainability, which has been discussed in the context of a panel. Some conclusions are finally offered.

Accessibility and Re-use

Keeping LD accessible over the long-term is the basic goal of preservation. This section reviews the basic aspects of LD accessibility that still need to be solved, and as such have an impact on the preservation of LD.

Access models for LD

A LD dataset can be accessed in several different ways. LD principles recommend de-referenceable URIs, but many LD datasets have SPARQL endpoints, or RDF dumps. In addition, there is the notion of embedding RDF into web pages, using RDFa [5] or microdata[4]. This plurality is a consequence of

[4] http://www.whatwg.org/specs/web-apps/current-work/multipage/microdata.html

the immaturity of LD, and it is not necessarily a blessing, because it implies using more resources than actually needed for tooling LD management and development.

At the workshop, Richard Cyganiak has recommended to expose LD datasets as RDF dumps, because they can be accessed without crawling, which may turn out to require a long time or a high amount of computational resources. Dumps may have the best cost/benefits ration and archiving them is useful *per sé*.

Soren Auer pointed out that there may be different types of preservation as we should distinguish these three types: Dataset dumps, which are available as large files somewhere in the internet; the second one are Incremental updates: for example, for DBpedia and Linked Geo Data publish incremental updates, additions and deletions of triples basically. And there might be a different preservation strategy for the LD in the third area: RDFa, Microdata and websites which are enriched with that.

Storage models for LD

It has been observed by Vassilis Christophides that there is an issue with the different formats of RDF graphs. The same graph may be obtained in turtle, in N3 and in XML syntax. The question arises what should a LD repository do in the circumstance. Should it keep all different formats with their respective provenance metadata, or select a canonical format? This is a preliminary question. We may push further and ask ourselves whether we want to organize the storage level of the web of data in terms of the entities (resources) that the descriptions are about, or we want to stick to the file metaphor that is in place nowadays. Even the evolution problem is much easier to address if repositories were entity-bases as opposed to being file-based.

Dataset packaging

Another aspect of dataset usage, impacting on both accessibility and re-use, is how to package individual versions of a dataset in an explicit, machine-readable way. Some datasets (e.g. DBPedia) are packaged and versioned, but they are exceptions. The question is how can the various parts of a dataset and its surrounding information be packaged and held together in an explicit, machine-readable way? What metadata needs to be recorded about these packages to preserve context and make them findable? The potential benefit of addressing these questions is to enable tooling for setting up a local copy of a published/archived dataset including all its dependencies. But there is a potential benefit also for preservation purposes, since such a package would fit the OAIS notion of information package and be used at the various stages of the OAIS functional model.

The OKFN[5] data packages do an important contribution to this goal. The metaphor again is that of software packages: when one installs a software package, there is a utility that only needs to be told where to find the package, and then the rest is automatically done; the utility knows how the package is structured, and where to upload the various parts of the package in the current environment in order to make everything work properly. There should be something similar for data packages.

In SNIA there is a format called Self-contained Information Retention Format (SIRF)[6] that is defined for encapsulating and labeling data for preservation purposes. SIRF does not address LD but it might be worth looking at what they do because they try to achieve self-containment and self-description of data packages.

Quality assessment of LD

Assessing the quality of a LD dataset is paramount in order to establish the economic value of a dataset and use it to take important decisions on the dataset preservation. At appraisal time, it must be decided whether it is worth to preserve a certain dataset, either in absolute terms or relatively to

[5] http://www.dataprotocols.org/en/latest/data-packages.html

[6] http://snia.org/SIRF

another dataset. When obsolescence threats the accessibility of the dataset, it must be decided whether it makes sense to protect the data from the threat, and what is the best way to do it. All these decisions are based, among other things, on estimations of quality parameters of the dataset. These aspects came up in the presentation of the ENSURE architecture, which includes a module for assessing the economic value of data.

In his presentation, Soren Auren talked about the work that the DIACHRON project is doing on LD quality. The research challenge is to establish measures for assessing the quality, the authority, the reliability of data resources, and match this with use case requirements. There different aspects of LD are gathered in 4 large related groups that partially overlap: the contextual dimension, including the relevance, the completeness and the amount of data. Then we have the representational dimension, including aspects such as conciseness, consistency and understandability. We have intrinsic characteristics, for example accuracy, conciseness, consistency, interlinking. And then some dimensions which are related with accessibility. The next question is how can we actually measure each aspect. Research on these topics is still on-going.

Coping with change

Change lies at the core of the preservation problem. In the context of LD, change may impact on URIs, which may in time point to different resources or to nothing at all, and on datasets, which may in time be updated or be taken off-line. Each of these problems is discussed in a separate section below.

Persistent URIs

LD are bound to use HTTP URIs (for simplicity we will drop the reference to the schema from now on) to denote resources in statements. A persistent URI is a URI that can be de-referenced in the long term and as long as it can be de-referenced, it references the same resource. Persistent URIs are therefore paramount for the preservation of LD.

Concerning the long-term de-referenceability of URIs, Phil Archer has described the W3C URI Persistence Policy[7], which consists of an explicitly expressed will to make persistent certain resources owned by the W3C as well as the URIs that refer to them, and to secure hosts that pledge to take up such resources and the machinery to reference them should the W3C not be able to continue doing so (the W3C hosts are MIT, ERCIM and Keio University).

Concerning the persistence of reference of URIs, there is an important study commissioned by the European Commission [6] that has been conducted by Phil Archer and presented by him at the workshop. The study reviews a number of case studies on the definition of schemas for persistent URIs, and proposes a set of good practices.

An interesting practice that has been illustrated and that is very relevant to preservation, is the one adopted by W3C for the documents it produces and that are intended to be available over the long-term. For anyone of these documents, several URI are minted:

- an un-versioned URI that always references the latest version of the document, eventually the last one when the document is finalized (*e.g.*, http://www.w3.org/TR/prov-overview/)
- a versioned URI that references a specific version of the document, and that from time to time coincides with the latest version (*e.g.*, http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/, where the versioning information is given by the date)
- a number of URIs, each referencing a specific representation of a version of the document (*e.g.*, http://www.w3.org/TR/2013/NOTE-prov-overview-20130430.html). These must be

[7] http://www.w3.org/Consortium/Persistence.html

properly related one another, by using the appropriate means of the language in which they are expressed.

All these URIs are created for specific purposes and must be used in a way consistent with such purpose. The W3C is formally engaged to make them persistent, and their combination addresses all possible issues of reference and persistence that have been encountered so far.

Annotating RDF triples

It has been pointed out that one way of coping with the change in a LD dataset is to make the data diachronic, by annotating them with time. The point was raised in Vassilis Christophides' presentation and further expanded in Peter Bunemann's talk.

This approach seems to be viable as far as the annotation of whole vocabularies is concerned, since a vocabulary is typically considered as a resource, endowed with a URI, which can be used as a subject in statements expressing the time validity of the vocabulary, and as an object in statements expressing the belonging of properties to the vocabulary.

However, there is an obvious difficulty in annotating single triples with time, as well as with any other information that may be relevant for preservation such as provenance. The only option at the moment seems reification, introduced in the first RDF vocabulary, but not really popular in the LD community. Named graphs, currently under study in the semantic web community, may be a more acceptable way of addressing the problem, which remains at the moment open.

Dataset Versioning

The distributed nature of LD makes the problem harder than usual, because a change in a LD dataset affects not only that datasets, but also all datasets that link to it. For instance, the removal of an axiom from a vocabulary may block inferences that were previously licensed. Similarly, the addition of an axiom to a vocabulary may licence inferences that were not possible before. The same problem also shows up in the preservation of the web.

If there are use cases in which these situations are undesirable, an obvious solution is to **version** datasets, so that the datasets that link to it are not forcefully brought to link to the new one, but can keep linking the one before the change.

The LD community does not cope well with the versioning of datasets, at the moment. For instance, the data behind a SPARQL end-point can change over night because the publisher has updated them, and the users of the end-point are not made aware of this fact. This clearly poses a challenge for preservation, and it is also an opportunity for the LD community to come up with a better way of managing the versions of the data. The Memento[8] work by Herbert van de Sompel is very relevant.

The problem is particularly important for vocabularies, which by their nature are datasets that are going to be referenced most. LOV[9] already archives versions of 100s of vocabularies.

Dataset Disappearance

Although it does not happen too often, sometimes LD datasets go offline. If we look at the LD cloud, between any two updates there are a couple of datasets that have disappeared. Quite often the bits are persisted somewhere, for instance as a dump of the dataset, and they remain accessible; yet there is no standard way of accessing these data, which implies that *ad hoc* techniques have to be employed from time to time, defeating the benefits of having a standard.

[8] http://www.mementoweb.org/

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

A possible solution is to create the machinery (tools and infrastructures) required for creating and accessing repositories of off-line datasets, so that these datasets are acted upon as if they were still on the web.

Sustainability via data marketplaces

Sustainability is a key issue in preservation, because preservation aims at achieving long-term functionality, and ensuring long-term functionality requires a means to sustain in time the technical efforts. The theme of sustainability has been addressed by a panel on a theme that hints a direction for sustainability based on the current scenarios. The theme has been illustrated by the five panellists, who were Fabrizio Gagliardi, Vassilis Christophides, Elena Simperl, Robert Sharpe and David Giaretta. The following discussion has been moderated by Fabrizio Gagliardi who also draw the final conclusions.

Panelist presentations

The data produced by projects funded by the EU must be made openly accessible and must be preserved over the long term. **Fabrizio Gagliardi** in his opening presentation at the panel has been offering the perspective of scientific organizations dealing with very high volumes of very structured data. These organizations are now pushed to care for preservation, which means to endow the infrastructures that hold the data with services targeted to that end. This clearly opens the sustainability issue, because creating and setting up and maintaining a preservation service cost money. Fabrizio's vision is to use modern technology and the availability of virtual computing services on the web to create an environment where people can store the data in an affordable way - ideally free of charge - develop tools and made them available - potentially with a charge, because tools cost money to produce - making them available to the virtual infrastructure, and then in a way create some kind of open market for people developing data, producing the data, people developing the tools, and then people using them. The data and the services must be available under different costing schemes, from completely free to pay for use. If that is enough to sustain a kind of virtual market, the question for us is how to get there, what we can do to make the process happen possibly with the support of the funding agencies. That is for me a way to make the data preserved in a sustainable way: if there is a business behind that, if people can make money by doing that, then there is a chance that people will invest, will continue to maintain the data over a long term, and technology will adapt, we will develop new technology. If there's a business behind we can go regularly and in a few years important data – important because people are using, so they are paying to use them – will migrate to new technologies. And in fact the only thing we have seen so far in preserving data, working is when data are accessed, so basically when data every two-three years are migrated to new technologies, they are refreshed, they are recycled: that's the only way so far I've seen digital data being preserved. Anything else in a CD, on a shelf, magnetic tape even in a sophisticated recording system, has a limited amount of life.

Vassilis Christophides, in the second presentation, shared the vision of sustainability of preservation through data marketplaces put forward in the previous presentation but argued that we lack business models to make the vision happen. He noted that although preservation is placed at a precise point of the data lifecycle (not necessarily at the end point, though), the data have to be handled with preservation in mind during their whole lifecycle. In other words, to be able to preserve the data, preservation actions must be taken at every stage of the data lifecycle. This is the notion of diachronicity of data. Secondly, we need to frame LD preservation as a sustainable economic activity, involving of course not only economic, but also social and technical aspects. The notion that has to be studied is that of the preservation of a data ecosystem. The questions that we must ask ourselves are well-known: Who benefits from use of the preserved data? Who selects what data to preserve? Who owns the data? Who preserves the data? Who pays both for data and preservation services? We can answer these questions for the scientific data lifecycle, at least in the countries that invest for research

the business model is clear. The scientists create the data and get rewarded by being cited. The consumers of the data use them to create wealth in form of jobs, taxes and investments. Then the problem is how do we decide what to preserve. In the data marketplace scenario, everything that has a value will be preserved, because it will be continuously used. But unfortunately the current market is shaped for big players, those that have enough money to afford to pay the collection of data. In order to create a bigger market we have to provide tools allowing small and medium providers to enter the game. So we need pricing models that are more fine-grained, and here is a very interesting discussion that Peter initiated a long time ago about micro payments: how we can actually pay only for the part of the data that we are actually interested and not the entire collection. This is a technical issue that needs to be questioned.

David Giaretta started his presentation with the questions that he was asked by the EC unit that is funding research in preservation: Who pays? Why? What to be preserved? And what's the value? These are the questions that may be difficult to answer for any kind of data that is seats in between fundamental data (such as standards) and useless data (such as much on what is my computer at present). But whenever the decision is taken that something has a certain value and needs to be preserved, then there are a certain number of things that one should be doing and exactly what one should be doing is the knowledge brought in by research in preservation. So we know that emulation does not allow to do anything new with what is being preserved. Migration on the other hand is costly and not tremendously effective for scientific data. It all depends on the kind of data one is looking at. Linked Data let one decides quickly about their quality because links to the data can be used to estimate how valuable the data are. In addition, LD are born with the notion of sharing, sharing vocabularies, sharing descriptions from authorities by linking to them, and this is a big advantage for preservation, we do not need to re-create everything from scratch, we can just link to what already exists and thereby we re-use it. So in terms of sustainability LD may have some advantages over other types of data because they are born with re-use in mind.

Robert Sharpe presented the Digital Archive System that is developed and used by Tessella to preserve the data of their customers around the world, thereby offering one very useful snapshot of the preservation market at the moment. He also analysed some current trends in storage, including a cloud-based solution recently launched by the company for US customers.

Elena Simperl resumed the discussion on the lifecycle of scientific data, offering her experience in various projects and her viewpoint as scientist producing these datasets. She essentially agreed with the considerations made by Vassilis about establishing a marketplace for scientific data.

She then talked about the usefulness of LD for preservation, arguing that LD are the way to in terms of sharing descriptions and vocabularies, because there is nothing at the moment that matches LD for doing the same at web scale efficiently. However it has to be said that one should not expect that by just going to the linked open data cloud they will find exactly what you are looking for. There was some discussion on this point, because the current situation should not be taken as immutable. If preservation organizations worldwide decide that LD is the way to go, then they will add a lot of high quality data and vocabularies to the LD cloud, and this may drastically change the situation.

Finally, Elena elaborated on the preservation of LD. Her view is that before we start thinking about how to preserve, what to preserve, what are the business models for preservation, we should clarify as the LD community questions of what are the business models for publishing and maintaining LD. There have been several discussions over the last months about whether there is a commercial value in LD all together, and the LD community is not so sure anymore which is the LOD cloud, as we know it is still very useful as it used to be, or whether it requires a major updo since it has not bee updated in a long time. The LD community needs to take a step back and think about the business models, and this reflection should also include what should be preserved.

The debate then started with the typical question that is being asked about LD: why did you do it in the first place, how much it cost, what was the return on investment. Phil is preparing a report on that, coming up at the end of this year, and PRELIDA may want to have a look at it.

Discussion

The discussion then addressed how LD could change the data market. At the moment, what is being observed is that there are vertical sectors that are dominated by vey big players and feature a long tail of very small players that strive to survive. Can LD have the power of changing this situation, and re-distribute wealth to more players? Elena Simperl observed that after the big players (such as twitter) established a market of social media data, a number of smaller players came up as re-sellers of the data. This is for instance the case with Bloomberg, which after selling big data to big companies ahs now opened an app-store where small companies can submit applications and are thereby granted access to their data. This opens up a market for small companies that was not there before.

Finally, Grigoris Antoniou raised the question what happens if LD are not preserved, what would be the consequences. He observed that from some previous comments it seems that the LD itself believes it is time to re-think about the publishing model that they have been following so far, because it has led to a LOD cloud whose quality value is questioned by the very members of the community. Elena Simperl addressed the question by proposing to start a research amongst the publishers and the consumers of the LOD dataset, raising the issue of LD preservation to both of them, with obviously different questions. Producers would have to be made aware of the need of preservation, of its costs and implications, and it would be interesting to hear from them to what extent they would be willing to invest or participate in the preservation of their LD assets. Consumers would have to be made aware that the LD datasets that they use may disappear any day, and it would be interesting to hear from them how they would react to the event, what kind of loss they would experience and to what extent they would be willing to invest their resources to make this not happen. This could be done for the top datasets, using mailing lists and other tools that are already in place; the communities would be interested in being involving and probably willing to share their thoughts and experiences. David Giaretta proposes to use the Research Data Alliance and set up a LD Working Group in it, on sustaining data marketplaces.

Soren Auer pointed out that we should review what is needed for the preservation of LD that is not already offered as a service, perhaps not under the “preservation” label. For instance, dropbox and googledocs offers facilities for storing permanently (or almost permanently) possibly large quantity of files, and there are services offering persistent URIs or persistent identifiers. We should do an analysis of what needs to be added, if any, on top of these services, in order to achieve the preservation of LD. Carlo Meghini notices that this is a very good suggestion which is in fact already in the agenda of PRELIDA, named “gap analysis”.

Conclusions

Data marketplaces are an opportunity for funding the preservation of the data that have a clear market value. At the moment, every vertical marketplace is dominated by few big players; besides being problematic from the economical point of view, this situation may in the long-term have a negative impact on the coverage of the data being preserved. Public money has been employed to a certain extent to endow major scientific data infrastructures with preservation services, but the effort needs to be consolidated and expanded to the other communities.

LOD datasets play a minor role in the data marketplaces, at the moment, and this may be due to their being freely available but also to the unclear economical value of these data. It has been proposed that the LOD community takes care of evaluating the status of the most used datasets, to the end of evaluating their usage, quality and economical value. PRELIDA may help in this effort by making an

effort to contact major LOD producers and consumers in order to investigate their awareness of preservation issues and their willingness to invest in the preservation of LOD.

6 On-line Resources

On-line resources related to the first PRELIDA Workshop are for the moment hosted on the private area of the PRELIDA web-site at:

http://prelida.isti.cnr.it/opening/index.php?content=1

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, giving the list of presentations, each endowed with the audio-visual recordings of the presentation and with the slides used by the presenter.

These information will later be made available to the general public the present report, by publishing them on the public section of the web site.

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

7 Conclusions

As it can be appreciated from Section 5, the workshop was very useful to bring together experts of Preservation and of Linked Data, and to have them exposing and discussing their views on the main issues concerning the preservation of Linked Data. In particular,

- The introductory session played the role of a tutorial on Linked Data and Preservation setting the context for the future discussion; in this section, preservation experts could learn the main features of Linked Data, while Linked Data experts could learn the basic objectives of preservation and the main features of the OAIS model, the reference model for preservation.
- The introduction of the DIACHRON project by Vassilis Christophides was the first bridge between the two areas, showing a wider picture where preservation is just one component of the data lifecycle, and Linked Data is just one of the types of data addressed by the project. The presentation touched upon all the themes that came forward in later discussions, such as the temporal annotation of data and datasets.
- The session on Linked Data introduced several contexts of usage of Linked Data, while the presentation of Richard Cyganiak attacked the preservation problem in a direct way, highlighting the open issues that the Linked Data community faces on the route to preservation. Some important problems that Linked Data have to solve in order to be preserved were highlighted by Peter Bunemann.
- The session on Preservation offered a survey on several important aspects of the problem. The presentations on the ENSURE system, in particular, provided an overview of the functionality that must be covered by a comprehensive solution to the problem, and highlighted the role that Linked Data might play in that context. Phil Archer offered a thorough overview about how the W3C tries to achieve the long-term persistence of its resources, with a particular emphasis on identifiers. A somewhat alternative view was offered by Datacite, a proprietary solution to the citability of scientific work. Mariella Guercio highlighted the challenges that must be won in order to make full usage of Linked Data in memory institutions devoted to preservation, such as archives.
- Finally, the panel on data marketplaces offered a view on an emerging notion for the exchange of data, where Linked Data can play an important role and the sustainability of preservation may be attained based on a market-driven model.

Overall, the workshop was very helpful in defining the main topics that PRELIDA will have to address in order to achieve its goals. A good climate was established amongst the participants, who were willing to engage in discussions and did not hesitate to ask fundamental questions or question basic assumptions that were in their opinion unwarranted. This is very encouraging for the next PRELIDA workshop, whose programme will be defined on the basis of the scientific outcome of the first workshop, as outlined in Section 5.

8 References

[1] PRELIDA FP7 CSA n. 600663. Description of Work.

[2] High level Expert Group on Scientific Data. Riding the wave. October 2010 http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

[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] DIACHRON.

[5] Ben Adida, Mark Birbeck Shane McCarron, Ivan Herman. RDFa Core 1.1 - Second Edition. W3C Proposed Edited Recommendation 25 June 2013. http://www.w3.org/TR/rdfa-core/

[6] Phil Archer. Study on Persistent URIs. On-line version: http://philarcher.org/diary/2013/uripersistence/ PDF version: http://joinup.ec.europa.eu/sites/default/files/D7.1.3%20-%20Study%20on%20persistent%20URIs_0.pdf