



Project no. 600663

PRELIDA

Preserving Linked Data
ICT-2011.4.3: Digital Preservation

D2.6 Report on the consolidation and dissemination workshop

Start Date of Project: 01 January 2013

Duration: 24 Months

UNIVERSITY OF INNSBRUCK

Version: final

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

Document Information

Deliverable number: D2.6
Deliverable title: Report on the consolidation and dissemination workshop
Due date of deliverable: 10|2014
Actual date of deliverable: 11|2014
Author(s): José María García, Sotiris Batsakis, Carlo Meghini
Participant(s): UIBK, HUD, CNR
Workpackage: WP2
Workpackage title: Organizational Support
Workpackage leader: CNR ISTI
Est. person months: 3
Dissemination Level: PU (Public)
Version: final
Keywords:

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level
1	2014/10/24	draft	José María García (UIBK)	First draft
2	2014/10/31	final	Sotiris Batsakis (HUD) & Carlo Meghini (CNR)	Revisions on scientific outcome section.

Executive Summary

The present document provides details on the final workshop of PRELIDA, held in Riva del Garda, Italy, between the 17th and 18th of October, 2014. The workshop was open to interested parties in addition to the PRELIDA Working Group members, and it was the third and last 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 discussed, by illustrating the themes that have been discussed and that will be part of the final outcomes of PRELIDA.



Table of Contents

Document Information	1
Executive Summary	1
1. Introduction.....	3
2. List of Participants	4
3. Programme of the Workshop.....	5
4. Scientific Outcome.....	7
Related initiatives	7
Ingesting Linked Data datasets and managing relevant changes	7
Additional use cases	8
Organizational and economic issues	8
5. On-line Resources	9
6. Conclusions.....	10
Acknowledgements	11
References	11

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. Afterwards, in April 2014 the Midterm Workshop took place in Catania, where the main outcomes of PRELIDA during the first year were presented to WG members in order to foster discussions about the main issues detected and introduced in the state of the art [1] and gap analysis [2] deliverable.

This report presents the results of the third and final workshop of PRELIDA, named “dissemination and consolidation”, which was held in Riva del Garda, Italy, between the 17th and 18th of October, 2014. The aim of this workshop was to present the initial version of the PRELIDA roadmap [3] to a wider community (i.e. not only WG members), and obtain useful feedback from the discussions and participation of the attendants of the workshop.

The rest of the document is structured as follows. In Section 2 we list the participants of the workshop. Section 3 presents the programme of the workshop. Then, Section 4 elaborates on the scientific outcome of the presentations and discussions. Finally, Section 5 enumerates the on-line resources available about the workshop and Section 6 concludes the report.

2. List of Participants

The workshop was attended by a number of invited experts in Digital Preservation and Linked Data, including Working Group members of PRELIDA and people representing the PRELIDA beneficiaries, for a total of 29 participants. The list of attendants is included in the following:

Grigoris	Antoniou	University of Huddersfield
Ashkan	Ashkpour	International Institute of Social History, DANS
Sotirios	Batsakis	University of Huddersfield
Martin	Brümmer	University of Leipzig
Peter	Burnhill	EDINA, University of Edinburgh
Vittore	Casarosa	ISTI-CNR
Mariano	Consens	University of Toronto
Giorgos	Flouris	FORTH-ICS
José María	García	University of Innsbruck
David	Giaretta	Alliance for Permanent Access
Dan	Gilleen	Artefactual Systems
Sarah	Giuliani	University of Innsbruck
Paul	Groth	University of Amsterdam
Maria	Guercio	Sapienza University of Rome
Yunhyong	Kim	University of Glasgow
Maurizio	Lunghi	Fondazione Rinascimento Digitale
Hugo	Manguinhas	Europeana Foundation
Carlo	Meghini	ISTI-CNR
Albert	Meroño Peñuela	DANS
Dumitru	Roman	SINTEF, University of Oslo
Elena	Simperl	University of Southampton
Steffen	Staab	Universität Koblenz-Landau
Peter	Stanchev	IMI - BAS, Kettering University
Yannis	Stavrakas	ATHENA R.C. / IMIS
Śławek	Staworko	University of Edinburgh, University of Lille
Lars G.	Svensson	Deutsche Nationalbibliothek
Hideaki	Takeda	National Institute of Informatics
Costantino	Thanos	ISTI-CNR
Mark	Williams	Dartmouth College

3. Programme of the Workshop

The workshop lasted two full days, in co-location with a major conference in the Linked Data area as the 13th International Semantic Web Conference (ISWC)¹ in order to attract participants. Essentially, the programme was structured in two main parts:

- During the first part, which comprised the first day, the main PRELIDA outcomes to date were presented to the audience, focusing on the PRELIDA State of the Art (D3.1) and the draft roadmap (D4.2). Furthermore, different related projects were invited to present their achievements and relation to PRELIDA.
- The second day was devoted to foster discussions among all participants about the main topics that will be part of the final roadmap to be delivered by PRELIDA. There were presentations and discussions about technical issues, additional use cases, and finally a round table regarding organizational and economic issues of Linked Data preservation.

In the closing part, the PRELIDA Coordinator introduced the final outcomes of the project and asked for further collaboration for the successful development of the final state of the art and roadmaps to be produced in PRELIDA.

The detailed programme of the workshop is given below.

Oct. 17th, 2014

09:30 – 10:30 *Welcome, Overview, Aims (Carlo Meghini)*

10:30 – 11:00 Coffee break

11:00 – 13:00 *PRELIDA Outcomes*

- State of the Art Analysis deliverable (David Giarretta)
- Draft Roadmap (Grigoris Antoniou and Sotiris Batsakis)

13:00 – 14:00 Lunch break

14:00 – 16:00 *Setting the Context: Related Projects (1)*

- DIACHRON (Giorgos Flouris and Yannis Stavarakas)
- The CEDAR project from a technical perspective (Albert Meroño Peñuela)

16:00 – 16:30 Coffee break

16:30 – 18:30 *Setting the Context: Related Projects (2)*

- Reference Rot: Threat & Remedy (Peter Burnhill)
- The Media Ecology Project (Mark Williams)
- Privacy-Aware Preservation (Mariano Consens)

20:00 Social Dinner

¹ http://iswc2014.semanticweb.org/co-located_events



Oct. 18th, 2014

09:00 – 10:30 Ingesting and managing changes (Chair: Sotiris Batsakis)

- Experiments with evolving RDF (Slawek Stavorko and Peter Buneman)
- Discussion

10:30 – 11:00 Coffee break

11:00 – 13:00 Use case: Beyond DBpedia and Europeana (Chair: Grigoris Antoniou)

- The CEDAR project use case (Ashkan Ashkpour)
- Discussion

13:00 – 14:00 Lunch break

14:00 – 15:30 Round table on organizational and economic issues (Chair: David Giarretta)

- Organizational issues: stakeholders, scope of preservation systems, responsibilities
- Economic issues: preservation economic models, linked data preservation economic viability, Intellectual property and licences, liability and obligations

15:30 – 16:00 Conclusion (Carlo Meghini)

4. Scientific Outcome

The Opening Workshop was comprised of 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].

During the Midterm Workshop, those deliverables were presented and the discussions were more focused in addressing the concrete issues in archiving Linked Data by following the guidelines for conformance set by OAI. Three main themes were discussed, namely ingestion of Linked Data dataset, managing changes, and accession to an archived Linked Data dataset. The outcomes of that workshop were used as input for the Draft Roadmap [3] that was presented in the Dissemination and Consolidation Workshop.

In the Dissemination and Consolidation Workshop, apart from describing the latest deliverables of the project already mentioned, the discussions were focused on four different blocks:

- Presentation of related initiatives, posing additional challenges and offering alternative solutions related to the preservation of Linked Data.
- Technical issues to consider for preserving Linked Data, such as ingestion or change management.
- Additional use cases worth considering within the project.
- Other organizational and economic issues that needs to be tackled.

Further details on these blocks are presented in the rest of this Section.

Related initiatives

In this block, which occupied most of the first day of the workshop, several projects and initiatives were presented in order to set the context of PRELIDA and present alternatives and synergies with the work being carried out. DIACHRON, an FP7 integrated project, was presented extensively during the first session of this block, including technical details about their approach to preserve evolving data. The CEDAR project technicalities were also introduced during the session. This project focuses on the preservation of linked socio-historical data obtained from historical census data in the Netherlands. Challenges such as data evolution, archival and preservation measures that are present in these two projects were discussed during the session, providing a valuable input for PRELIDA activities.

In the second session of this block, other initiatives were showcased, including recent projects in related areas such as Hiberlink, focusing on reference rot in Linked Data datasets; the Media Ecology Project, which offers a Linked Data infrastructure to support research on multimedia archives; and a recent initiative from the University of Toronto related to privacy issues in Linked Data datasets and how preservation plans can be augmented by considering these issues. These related projects and initiatives put the focus on technical and organizational challenges that were also extensively discussed during the rest of the workshop, and will be taken into account for the remaining activities being carried out in PRELIDA.

Ingesting Linked Data datasets and managing relevant changes

Ingesting datasets and managing changes were the main technical issues identified for LOD preservation. Ingestion poses the challenge of identifying the designated community and the systems boundary, which is also a complex issue. Although Linked Data have not the volume of Big Data still the idea of ingesting the complete closure of data and definitions was met with scepticism and specific scalability problems were mentioned with respect to DBpedia. So complete closure for vocabularies

and a more limited coverage of external datasets or only internal data is a more reasonable approach. Annotating using vocabularies such as PROV, VoiD is also a good practise not only for LOD preservation but for preservation of other types of data as well. Lack of accurate preservation cost estimation methods and business models complicates the task of selecting what to preserve and for how long.

Dealing with changes is complex since changes in formats, technology, internal data and external datasets and definitions but also in other resources (PDF, images) representing linked data and designated community background knowledge must be taken into account. The most general case (for every type of date) is a form of Web preservation and cannot be fully addressed. For a more specific case where only RDF/OWL data is preserved, detecting changes (using crawling or notifications) is feasible and changes can be merged with existing previous versions. Related approaches were proposed/presented at the workshop, (e.g., labelling validity of RDF triples with intervals instead of timestamp for each instance). Diachron project also examines the detection of concept evolution, which is highly relevant to PRELIDA objectives. Compliance with W3C serialization standards was again identified as an important issue that all archives must do, in conjunction with proper annotation of data with respect to semantics (OWL version/profile) and compatible tools (Reasoners/SPARQL query engines). Overall preservation and managing changes will be more efficient if an ecosystem of archives is developed, dealing with a subset of Linked Data and referencing other archives for external links instead of replicating them. This reflects better the fact that LOD are distributed and will help to achieving a more reliable and more economic efficient solution in the long term.

Additional use cases

The objective of this block was to discuss other potential use cases that could lead to additional challenges for preservation of Linked Data as opposed to those already identified. In relation to the first block, on related initiatives, in this session the CEDAR project was presented from a more generic perspective, concluding that the outcomes of PRELIDA will be of use for that project and other Linked Data initiatives that publishes datasets so that they can define a proper preservation plan.

Organizational and economic issues

Finally, the last block of the workshop comprised a round table and further discussion with all workshop participants about organizational and economic issues related with Linked Data preservation. First, José María García, Albert Meroño y Mariella Guercio shortly presented the main organizational and economic issues that need to be taken into account for preserving Linked Data. These presentations kick-started a fruitful discussion with the rest of attendants, where other considerations were added to the table. Thus, stakeholders and responsible people were identified as key issues to consider when designing a preservation plan of a Linked Data dataset. Funding needs to be secured to ensure the long-term sustainability of these datasets.

Moreover, trust issues are also deemed as important for the community, where a federated approach for repositories could help in solving some of the discussed issues. However, this approach also goes deeper into the scope challenge, i.e. which granularity level has to be account for in the preservation plan, regarding functionality and associated datasets to be preserved. Consequently, service agreements between archives and LOD providers need to be put into place.

Finally, attendees also discussed the convenience of defining preservation plans not only for open data, but also for closed or proprietary data. In turn, Linked Data is not the only paradigm used to publish data on the Web, so similar measures should be taken in order to ensure long-term preservation of other kinds of open data.



5. On-line Resources

On-line resources related to the final PRELIDA Workshop are at the moment publicly hosted on the PRELIDA web-site², including the scientific programme of the workshop, giving the main objectives of the workshop and a list of potentially interesting topics.

The slides used by the presenters have been collected and will be made publicly available in the PRELIDA Slideshare account and disseminated using different dissemination channels, including links from the PRELIDA web-site.

² www.prelida.eu



6. Conclusions

During this final PRELIDA workshop, different views about how to tackle various challenges for preservation of Linked Data were presented to a wide number of participants from the two target communities of PRELIDA. Not only PRELIDA outcomes, but also related projects and initiatives were showcased, showing different approaches to address those challenges.

The workshop comprised several discussions about technical, organizational and economic issues that need to be taken into account when preserving Linked Data. All the insights and feedback obtained during the active participation of every attendant will be used as an important input to produce the final results of PRELIDA, namely the consolidated state of the art and the final roadmap deliverables.



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] PRELIDA Deliverable D4.2 First version of roadmap. Available from the PRELIDA web site: prelida.eu