Report on the 2nd Linked Archives International Workshop (LinkedArchives 2022) at TPDL 2022

Carla Teixeira Lopes
University of Porto/INESC TEC
Portugal
ctl@fe.up.pt

Cristina Ribeiro
University of Porto/INESC TEC
Portugal
mcr@fe.up.pt

Franco Niccolucci
PIN - Servizi Didattici e Scientifici per l'Università di Firenze scrl
Italy
franco.niccolucci@gmail.com

María Poveda-Villalón
Universidad Politécnica de Madrid
Spain
mpoveda@fi.upm.es

Nuno Freire
ROSSIO Infrastructure, NOVA FCSH
Portugal
nunofreire@fcsh.unl.pt

Abstract

The 2nd edition of the International Workshop on Archives and Linked Data ran in conjunction with the 26th International Conference on Theory and Practice of Digital Libraries (TPDL 2022). TPDL 2022 was an in-person event in Padua with an online-only registration for non-speakers. Archives, the guardians of large volumes of historical and current information, are showing a growing interest in linked data and semantic web technologies. These technologies can be used to link archives' information with data from other cultural heritage institutions and more informal sources. In a world of linked data, users can access richer interfaces where the archival records are available in their context, with explicit metadata. The workshop gathered about 20 researchers and specialists engaged in initiatives crossing linked data technologies, archives, and cultural heritage in general, discussing advances and challenges in this area. The workshop was a successful event, with active participation in the discussions that followed the presentations. The participants found the workshop interesting (40%) or extremely interesting (60%), and everyone who answered a post-workshop survey would be interested in another edition of the workshop.

Date: 20 September, 2022.

Website: https://linkedarchives.inesctec.pt/.

1 Introduction

Cultural Heritage deals with treasures that are expected to survive generations. Many digital initiatives have explored segments of this global asset, with much more to uncover. They tend to be oriented by the organizations that have traditionally curated these valuable objects: libraries, museums, and archives.

The Linked Archives International Workshop started from the perspective of archives, the guardians of immense volumes of historical and current information driven by the need to keep a record of our past processes, achievements, and results. The growing interest in archival records and the availability of technologies that can take large volumes of data and process them leads archives into the world of linked data. In this vision, the archives' information is joined with data from other cultural heritage institutions and more informal sources. Later, users can explore archives in rich interfaces where the data are available in their context, with explicit metadata.

The two editions of the workshop were sponsored by a Portuguese national project, EPISA—Entity and Property Inference for Semantic Archives[1], a collaboration between two university groups and Torre do Tombo, the Portuguese National Archives. The project is experimenting with models and prototypes for archival information systems based on linked data and ready for the semantic web. As the elements of the project team made contact with organizations and groups with similar interests, they perceived the need to discuss models, technologies, and infrastructures. TPDL 2021 presented the opportunity to do this in a European-centred community. The 1st edition of the workshop was a success, with over 20 submissions. Nearly 90 people attended, and most of the contributions are published as CEUR proceedings [Lopes et al., 2022].

TPDL 2022 welcomed the 2nd edition of the International Workshop on Archives and Linked Data.

2 Organization

The workshop aimed to gather researchers and specialists engaged in initiatives that cross Archives and the Semantic Web and those planning similar efforts in other cultural heritage organizations. We adopted an interdisciplinary point of view, to stimulate the dialogue between the technically-oriented communities, researchers from the digital humanities, and specialists from cultural heritage institutions.

The organizers invited 33 scholars and specialists to join the members of the Organising Committee in the Program Committee. Three members of the Program Committee reviewed each paper, and each member reviewed, at most, one paper.

The number of submissions (8) was lower than in the previous edition. Two reasons may have contributed to this reduction, the move from an online event to an in-person one (initially, remote participation was not being considered by TPDL 2022) and the existence of a simultaneous event having the same target audience, the conference of the International Council on Archives. Six of the eight papers were accepted, two as short and four as full papers.

The workshop was a half-day event that started with a keynote by Kerstin Arnold from the Archives Portal Europe Foundation entitled "No Archive is an Island – A Tale of Exploring a Brave

[1] http://episa.inesctec.pt/

New World". We organized the six papers in three blocks according to their main focus: *Artificial Intelligence and Archives*, *Infrastructures for Archives and Linked Data*, and *Models for Linked Archives*. All the papers were assigned a 15-minutes presentation followed by a question-answering and discussion period.

After the workshop, we published the slides of the presentations on the workshop's webpage. Contributions are published in the TPDL 2022 proceedings for the Workshops and Doctoral Consortium [Candela and Silvello, 2022].

The call for papers announced that extended versions of the best papers would be selected to be published in the ACM Journal on Computing and Cultural Heritage (JOCCH). After a second round of reviews by the Organising Committee, the authors of one paper were invited to submit an extended version to the journal.

3 Participation

Four countries are associated with the accepted submissions, as seen in Figure 1, and European countries predominate. Note that submissions with contributions from more than one country are counted for each involved country. There was one submission with authors from both the United Kingdom and Italy.

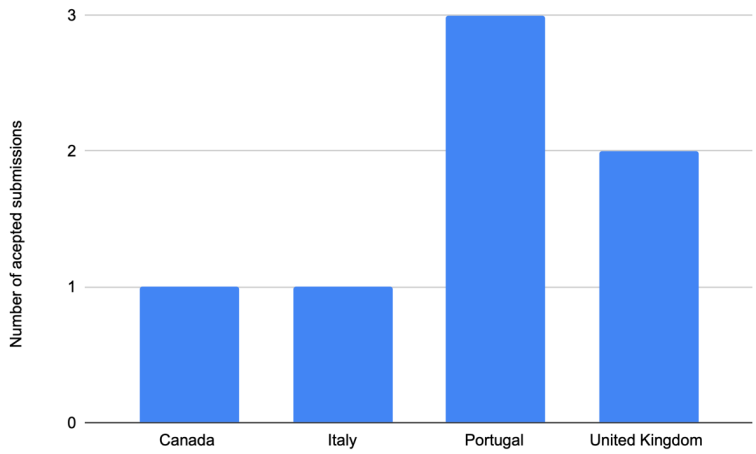

Figure 1: Number of accepted submissions per country.

An analysis of the institutions associated with the accepted submissions shows that, from the six institutions, 5 (83%) were Higher Education Institutions. One of the institutions (22%) was a national archive. The United Kingdom National Archives and the University of Minho were already present in the previous edition of the workshop. The institutions associated with accepted submissions are presented in Table 1.

Table 2 identifies two research projects associated with the accepted submissions. Both projects are also associated with works presented in the 1st edition of the workshop, although by different authors.

Two authors have presented works in both editions of the workshop.

Table 1: Institutions associated with accepted submissions.

Name	Country	Type
The National Archives	United Kingdom	Archives
University College of London	United Kingdom	Higher Education
University of Alberta	Canada	Higher Education
University of Minho	Portugal	Higher Education
University of Pisa	Italy	Higher Education
University of Porto	Portugal	Higher Education

Table 2: Projects associated with the submissions.

Acronym/Name	Title/Short Description	Countries
EPISA	Entity and Property Inference for Semantic Archives	Portugal
Project Omega	Towards a single pan-archival linked data catalogue	United Kingdom

As there was no registration for this specific workshop, we conducted a post-workshop survey to estimate the number of participants and their level of satisfaction with the workshop. We had approximately 15 people physically present and about seven attending online. 40% of the in-person and 40% of the remote participants have not attended other sessions in TPDL 2022, which shows a specific interest in the topics of the workshop.

In the post-workshop survey, participants evaluated the workshop as interesting (40%) or extremely interesting (60%), and everyone showed interest in another edition of the workshop. One of the participants said it would be good to know in advance who presents remotely, and another said it would be beneficial to have the papers, or at least their abstracts, available before the workshop.

4 Keynote

Kerstin Arnold, the Chief Operating Officer of the Archives Portal Europe[2], started the workshop with her presentation entitled "No Archive is an Island—A Tale of Exploring a Brave New World" [Arnold, 2022]. Contrary to the original plan, the presentation was delivered remotely due to Kerstin's inability to travel.

Kerstin discussed the importance of linking archives and presented the development and current status of Archives Portal Europe. In this context, Kerstin talked about the role of standardization on archives inter-connection, discussed several types of standardization, and described the role of aggregation and joint access points. To wrap up, Kerstin reflected on the appearance of the new Records in Context standard and other initiatives such as the Common European Data Space for Cultural Heritage.

The recording of the keynote presentation is available on the workshop's webpage.

[2] http://www.archivesportaleurope.net

5 Contributions

Contributions were organized into three blocks of presentations according to the nature of their content.

5.1 Artificial Intelligence and Archives

This session had two presentations on the application of artificial intelligence for information extraction in archival data: one on archival finding aids and another on digitized archival materials.

The research by Luis Filipe Cunha and José Carlos Ramalho approaches the problem of Named Entity Recognition (NER) in the Portuguese language [Cunha and Ramalho, 2022]. The authors started from a pre-trained Bidirectional Encoding Representation from Transformers (BERT) model for the Portuguese vocabulary and adapted it for NER. The resulting model was compared with an earlier model by the authors, and it improved F1-score. The authors also present a tool for annotating named entities in textual corpora, in which the NER model supports the process of annotation by the end user.

An ongoing project addressing the description of archival descriptions assisted by artificial intelligence was presented by Mariana Dias and Carla Teixeira Lopes [Dias and Lopes, 2022]. The system aims to assist the archivist by applying Optical Character Recognition (OCR) on the archival materials, followed by NER to identify the main concepts in the text. The extracted concepts are then used to populate an ontology. The project has already trained a NER model using a Bidirectional Long Short-Term Memory Network with a Conditional Random Field Layer (BiLSTM-CRF) Neural Network model and pre-trained contextual string embeddings. Ongoing work is populating the ontology based on programmed mapping and instantiation rules for the recognized concepts. The project uses the ArchOnto ontology from the EPISA project, but the authors indicate that their approach can be generalized to other linked data models.

5.2 Infrastructures for Archives and Linked Data

This session has two presentations about infrastructural frameworks: one for traditional archives and another for web archives.

The first presentation addressed the problem of content drift detection from the perspective of web archives. Content drift occurs when a website's content changes and moves away from its original topic. Brenda Reyes Ayala, Qiufeng Du, and Juyi Han presented a method for detecting content drift on the live web-based on title comparison [Ayala et al., 2022]. The authors use cosine similarity on the titles of live websites and their corresponding archived versions. They evaluated their method in three web archive collections manually checked for content drift, and their approach achieved an F-measure of 90.7.

The ongoing development of an infrastructure for integrating linked data in the archival management process was presented by Sérgio Nunes, Tiago Silva, Cláudia Martins, and Rita Peixoto [Nunes et al., 2022]. The authors presented the technological architecture of the infrastructure, which is based on RDF data, and an overview of its workflows. The use cases for archival professionals addressed by the infrastructure are searching, browsing, and archival management and

these include abstractions for working with linked data. Although the infrastructure is experimental, its functionality is currently under evaluation by archival professionals.

5.3 Models for Linked Archives

This session had two presentations on semantic models in archives: one proposing an extension of the Records in Context Ontology (RiC-O) and another evaluating a semantic model with the support of archival professionals.

Daria Mikhaylova and Daniele Metilli were motivated by the information needs of architects regarding architectural archives and proposed an extension of the RiC-O that is specifically designed for architectural archival records [Mikhaylova and Metilli, 2022]. The authors presented the design of their ontology, focusing on how it models the architectural project, its different phases, and the specific types of records that are found in an architectural archive. The authors also presented and discussed the application of the ontology to the personal archive of the Italian architect and engineer Dino Tamburini (1924–2011). The authors plan to publish the archive online and make it available through the National Archival System of Italy and Europeana.

Alex Green and Faith Lawrence reported on the latest progress on The Pan-Archival Catalogue at the UK National Archives [Green and Lawrence, 2022]. This new catalog is based on linked data and will bring together descriptions of both physical and digital records from a variety of sources. The presentation focused on the results of an evaluation with end users regarding the data model for the catalog. The authors concluded that although a linked data catalog offers many technical advantages, archival professionals must also adopt new practices.

6 Concluding Remarks

The Linked Archives International Workshop was motivated by the need to examine current technologies and data models for managing archives and describing their assets. The organizers maintained the broad perspective adopted in the first edition of the workshop, considering all aspects of linked data in cultural heritage. The call for contributions to the workshop was disseminated in cultural heritage communities and information technology circles dealing with semantic web solutions. The workshop tried to attract scholars, cultural heritage organizations, and the general public addressing the use of linked data. It is clear from current research that specialized data from archives and museums can gain new meaning in the context of the global information infrastructure. Also, new forms of expression in the arts and non-traditional archival assets require a fresh view of their representation and access.

Within the time limits of the workshop, participants discussed several topics familiar to those taking a linked data approach to cultural heritage. Cultural heritage has traditionally been organized in sectors with specialized practices, namely libraries, archives, and museums. The distinction between assets in these sectors is currently blurred. Moreover, creators have more fluid profiles and so do patrons and users in general. The semantic models open the spectrum of concepts common to all these communities: a location can be associated to an event, but also to the place of birth of a person, the address of an organisation or the place where an artifact originated. There is considerable experience with semantic models, namely in the museum community,

where the CIDOC-CRM model, a detailed event-based data model represented as an ontology, originated. The use of CIDOC-CRM has now been extended to other domains.

The main limitation for those working on linked data for cultural heritage is the availability of solid technologies. Heritage institutions require long-term solutions and graph databases, and their applications must manage the dependency on technologies that may not be mature enough besides evolving rapidly. The examples presented in the workshop have highlighted these challenges, namely when dealing with projects that involve vast collections, as is typical of the archives.

Although the workshop was aimed at a specialized audience, the diversity of participants made clear that linked data concerns people that range from the administration of heritage organizations to the creators of innovative technical solutions. The discussions led us to conclude that cultural heritage stakeholders are taking a fresh view on managing their assets as part of a more comprehensive information infrastructure. On the other hand, users in diverse areas are curious about the new applications of assets held by libraries, archives, and museums.

Overall, the workshop allowed people with similar concerns and a great diversity of experiences to meet. Most topics highlighted in the sessions will continue under discussion. We hope new opportunities will emerge to encourage a shared understanding of the problems ahead and the demonstration of current results.

Acknowledgments

This workshop was proposed within the scope of the Portuguese EPISA project - DSAIPA/ DS/ 0023/2018, financed by National Funds through FCT—Foundation for Science and Technology I.P..

References

Kerstin Arnold. Keynote: No Archive is an Island – A Tale of Exploring a Brave New World. In Leonardo Candela and Gianmaria Silvello, editors, *Workshops and Doctoral Consortium of the 26th International Conference on Theory and Practice of Digital Libraries 2022*, page 57, 2022. URL https://ceur-ws.org/Vol-3246/07_keynote.pdf.

Brenda Reyes Ayala, Qiufeng Du, and Juyi Han. Detecting Content Drift on the Web Using Web Archives and Textual Similarity. In Leonardo Candela and Gianmaria Silvello, editors, *Workshops and Doctoral Consortium of the 26th International Conference on Theory and Practice of Digital Libraries 2022*, pages 77–85, 2022. URL https://ceur-ws.org/Vol-3246/10_Paper3.pdf.

Leonardo Candela and Gianmaria Silvello, editors. *Proceedings of the 26th International Conference on Theory and Practice of Digital Libraries - Workshops and Doctoral Consortium*, number 3246 in CEUR Workshop Proceedings, Aachen, 2022. URL http://ceur-ws.org/Vol-3246/.

Luís Filipe Costa Cunha and José Carlos Ramalho. Fine-Tuning BERT Models to Extract Named Entities from Archival Finding Aids. In Leonardo Candela and Gianmaria Silvello, editors,

Workshops and Doctoral Consortium of the 26th International Conference on Theory and Practice of Digital Libraries 2022, pages 58–69, 2022. URL https://ceur-ws.org/Vol-3246/08_Paper1.pdf.

Mariana Dias and Carla Teixeira Lopes. Mining Typewritten Digital Representations to Support Archival Description. In Leonardo Candela and Gianmaria Silvello, editors, *Workshops and Doctoral Consortium of the 26th International Conference on Theory and Practice of Digital Libraries 2022*, pages 70–76, 2022. URL https://ceur-ws.org/Vol-3246/09_Paper2.pdf.

Alex Green and Faith Lawrence. The Shock of the New: Testing the Pan-Archival Linked Data Catalogue with Users. In Leonardo Candela and Gianmaria Silvello, editors, *Workshops and Doctoral Consortium of the 26th International Conference on Theory and Practice of Digital Libraries 2022*, pages 108–115, 2022. URL https://ceur-ws.org/Vol-3246/13_Paper6.pdf.

Carla Teixeira Lopes, Cristina Ribeiro, Franco Niccolucci, Irene Rodrigues, and Nuno Freire. Report on the 1st Linked Archives International Workshop (LinkedArchives 2021) at TPDL 2021. *SIGIR Forum*, 55(2), 2022. ISSN 0163-5840. doi: 10.1145/3527546.3527562. URL https://doi.org/10.1145/3527546.3527562.

Daria Mikhaylova and Daniele Metilli. An Extension of RiC-O for Architectural Archives. In Leonardo Candela and Gianmaria Silvello, editors, *Workshops and Doctoral Consortium of the 26th International Conference on Theory and Practice of Digital Libraries 2022*, pages 98–107, 2022. URL https://ceur-ws.org/Vol-3246/12_Paper5.pdf.

Sérgio Nunes, Tiago Silva, Cláudia Martins, and Rita Peixoto. EPISA Platform: A Technical Infrastructure to Support Linked Data in Archival Management. In Leonardo Candela and Gianmaria Silvello, editors, *Workshops and Doctoral Consortium of the 26th International Conference on Theory and Practice of Digital Libraries 2022*, pages 86–97, 2022. URL https://ceur-ws.org/Vol-3246/11_Paper4.pdf.