Introduction to the focussed issue on Semantic Digital Archives

**Thomas Risse · Livia Predoiu · Annett Mitschick ·
Andreas Nürnberger · Seamus Ross**

Preservation and curation of digital materials is a significant contemporary cultural, economic and social issue, yet it is often neglected. For decades, the amount of content created digitally has grown dramatically and more recently exponentially as digital process have become core aspects of everyday business and life. Indeed, the complete life cycle of most information nowadays tends to remains digital. A selection of this content is expected to be of value for the future and can thus be considered being part of our cultural heritage and essential as evidence for accountability over time. As soon as these digital materials, whether they be publications, data sets, or records of transactions, become obsolete, but are still deemed to be of value in the future, they have to be transferred smoothly into appropriate archival information systems (AIS) where they can be kept accessible against a backdrop of changing technologies, information practices, and documentation standards.

T. Risse (✉)
L3S Research Center, Leibniz University Hannover,
Hannover, Germany
e-mail: risse@L3S.de

L. Predoiu
University of Oxford, Oxford, UK
e-mail: livia.predoiu@cs.ox.ac.uk

A. Mitschick
University of Dresden, Dresden, Germany
e-mail: annett.mitschick@tu-dresden.de

A. Nürnberger
Otto von Guericke University Magdeburg, Magdeburg, Germany
e-mail: andreas.nuernberger@ovgu.de

S. Ross
University of Toronto, Toronto, Canada
e-mail: seamus.ross@utoronto.ca

This focused issue is inspired from discussions during the Semantic Digital Archives (SDA) workshop series held in conjunction with the *International Conference on Theory and Practice of Digital Libraries*. Digital Preservation is a challenging research area that consists of many interdisciplinary sub-challenges. Two communities that traditionally have a strong interest in preserving information for the future are the library and the archiving community. Both have made valuable contributions to the management of huge amounts of knowledge and data. However, both are approaching this topic from different perspectives and while there has been much cross-fertilization between the two domains, there is a need for greater integration of research and collaboration in practice. At the same time opportunities enabled by the emergence of the Semantic Web has become another research area that provides promising technical solutions for knowledge representation and management. At the forefront of making the Semantic Web a mature and applicable reality is the linked data initiative, which already has started to be adopted within some segments of the library community. For instance, there is increasing understanding within these communities that semantic representations of contextual knowledge about cultural heritage objects enhances the organization and access of data and knowledge.

The eight papers collected in this issue look at different aspects of the implementation of semantic digital archives and archival information infrastructures, including examining interoperability and preservation, semantic long-term preservation models, ontologies and business processes, temporal evolutions, as well as a vision for knowledge-based Culturomics research. Five of the papers address issues regarding the preservation of digital objects and processes.

Libraries, archives and museums are increasingly implementing Linked Data technologies to enhance their cataloguing workflow. Catherine Ryan et al. in their paper "Linked

Data Authority Records for Irish Place Names" review current best practice in library cataloguing, examine how linked data is used to link collections and briefly describe the relationship between linked data, library data models and descriptive standards. The paper includes a description of the processes involved in creating a new linked data set including links to other linked data resources.

Preservation metadata is a mixture of provenance information, technical information about digital objects and rights of information. In "PREMIS OWL—A Semantic Long-Term Preservation Model" Sam Coppens and collaborators discuss a semantic representation of the PREMIS 2.2 data dictionary established by the Library of Congress (USA) which enables dissemination of the preservation metadata as Linked Open Data on the Web and, at the same time, support Semantic Web technologies in the preservation processes.

In "Assisting Digital Interoperability and Preservation through Advanced Dependency Reasoning" Yannis Tzitzikas et al. explore the problem of preserving digital material in an environment of ecosystem change. The authors present a rule-based approach for dependency management. They show that the human effort required for checking whether a task on a digital object is performable can be reduced when automatic reasoning mechanisms are deployed to support the process.

The preservation of processes in business and science requires the description of a multitude of information objects as well as their interconnections and relations. Rudolf Mayer and colleagues in "Using Ontologies to Capture the Semantics of a (Business) Process for Digital Preservation" present a formal model that enables the sematic description of these objects. They detail the overall architecture and demonstrate the usefulness of their approach for different domains.

An Open Linked Data architecture for maintaining historical archives is proposed in the paper "A linked open data architecture for the historical archives of the Getulio Vargas Foundation". Alexandre Rademaker and his colleagues discuss the benefits and present already achieved milestones. In addition, possibilities for extending the accessibility and usefulness of the data archive using Semantic Web technologies, natural language processing, etc., are shown.

Besides the preservation of digital objects also their access, exploration and analysis play an increasingly central role as is demonstrated in three other papers.

A general overview and vision for Culturomics research is given by Nina Tahmasebi and collaborators in "Visions and Open Challenges for a Knowledge-Based Culturomics". The aim of Culturomics research is to make sense of cultural and language phenomena over time through harnessing the new availability of massive amounts of data. The research vision they have developed comb ines statistical methods with knowledge-based approaches. The paper discusses possibilities and open challenges that arise due to the nature of the data, diversity of sources, changes in language overtime as well as temporal dynamics of information in general.

In "A Metadata Model and Mapping Approach for Facilitating Access to Heterogeneous Cultural Heritage Assets", Thomas Orgel and his colleagues address the issue that although there are massive amounts of rich cultural heritage content available the potential of its use for educational and scientific purposes remains largely untapped. The authors present a metadata model that enables the combination of federated search results from different cultural heritage data sources. They show that an easily configurable metadata mapping enables an on-the-fly execution by an automatic service. This allows to unfold the educational potential of these materials.

Advances in technology and culture expectations lead to changes in the language. It affects the user's possibility for finding and interpreting content created in the past. Holzmann and his collaborators in "Named Entity Evolution Recognition on the Blogosphere" address the special challenge of recognizing the evolution of names of named entities in the Blogosphere. They adapted their existing approach for the recognition of name entity evolutions (NEER) to work on noisy data such as are evident in the Blogosphere.

The papers in this special issue effectively demonstrate the wide range of use cases and challenges for semantics in digital archives and archival infrastructures. With the increasing quantities and diversity of types of information created, collected and archived every day and the wide variability in formats and applications further research is necessary to make data collections a valuable and viable asset for use today and in the future. The semantic nature of digital information is both an asset and a liability. The research drawn together in these papers provide evidence as to research results so far and an indications for possible future investigations.

Finally, we wish to thank the Editorial Board of the *International Journal of Digital Libraries* for the opportunity to improve the exchange between involved communities through this focused issue. We also wish to thank all the authors of this focused issue for their valuable contribution, together with the reviewers for the very constructive comments and feedback that, we believe, make this issue of very high interest for the readers of the *International Journal of Digital Libraries*.