Proceedings of the 1st International Workshop on **Semantic Digital Archives**

co-located with the 1st International Conference on Theory and Practice of Digital Libraries (TPDL 2011), formerly known as European Conference on Digital Libraries (ECDL) and held on the 29.09.2011 in Berlin, Germany

Editors:

Livia Predoiu, Otto-von-Guericke University of Magdeburg, Germany

Steffen Hennicke, Humboldt University of Berlin, Germany

Andreas Nürnberger, Otto-von-Guericke University of Magdeburg, Germany

Annett Mitschick, University of Dresden, Germany

Seamus Ross, University of Toronto, Canada

CEUR Workshop Proceedings http://ceur-ws.org ISSN 1613-0073

Vol-801
urn:nbn:de:0074-801-0

Preface

These proceedings are the result of an exciting workshop held in conjunction with the first international conference on Theory and Practice of Digital Libraries, TPDL 2011, formerly known as European Conference on Digital Libraries, ECDL. The name of the workshop, *Semantic Digital Archives – sustainable long-term curation perspectives of Cultural Heritage* (short: SDA 2011) already provides a first hint towards its general topics and goals: to promote and discuss sophisticated knowledge representation and knowledge management solutions specifically designed for improving Archival Information Systems (AIS) and Archival Information Infrastructures (AII).

Over the past couple of decades, digitally created content has come to permeate all aspects of our lives and the life cycle of these objects is increasingly exclusively digital. A portion of this content can be expected to have enduring value as it delivers insight into the contemporary trends and spirit of its time. Hence, it can be considered being part of our cultural and scientific heritage. This vast corpus needs to be appraised and items of enduring value selected, archived and kept accessible so that it can be made available in response to requests from information professionals, and the general public. Therefore, sustainable long-term curation perspectives for our digital cultural heritage are essential. Digital content poses many socio-cultural and technological challenges which create obstacles to long-term or indefinite preservation. Changing technologies and shifting user communities as well as the increasing complexity of digital content consisting of or being enriched with software and multimedia attachments are only a few examples. Dealing with these challenges was the central theme of the workshop.

The workshop aimed to involve and stimulate discussions between the digital archiving, the digital museums, the digital libraries and the semantic (web) technologies communities. Archives, museums and libraries share a natural bond as all three have a long history of experience with maintaining (storing and retrieving) a large amount of objects, data and information. Hence, there is a lot potential for cross-fertilization between these related fields. Furthermore, libraries already started to adopt semantic web technologies successfully as shown by various workshops and conferences on this topic that recently have emerged. Most remarkably, also a W3C incubator group on *library linked data* has been created. Hence, the workshop aimed at fostering discussions about experiences and best practices of employing semantic web technologies in the library domain yielding so called *semantic digital libraries* in order to inspire and boost the adoption of semantic web technologies in the area of digital archiving as well.

The area of semantic (web) technologies is a broad scientific discipline that focuses on providing promising technical solutions for knowledge representation and knowledge management. It provides knowledge representation languages and management technologies based on a solid artificial intelligence foundation and is supported by appropriate W3C recommendations and a large user community. At the

forefront of making the semantic web a mature and applicable reality is the linked data initiative. Using semantic (web) technologies in general and linked data in particular can be expected to mature the area of digital archives as well and technologically tighten the bond between digital libraries and digital archives. Furthermore, digital archives and their users have special requirements that can also inspire semantic (web) technologies research in general.

The workshop was well accepted by the community and was able to attract 23 submissions from which we selected 13 papers with the help of our program committee; giving an overall acceptance rate of 56%. The papers covered a broad range of relevant topics in the area of semantic digital archives, bringing together people from archives, museums, digital libraries and the semantic web as hoped and expected. A lot of different research projects are represented in these proceedings, e.g. the KEEP project (W. Bergmeyer), SHAMAN (J. Brunsman, K. Qian et al.), Semlib (C. Morbidoni et al.), ASPI (C. Cortese and G. Mantegari), Europeana (S. Hennicke et al.) and EUscreen (J. Oomen and V. Tzouvaras). Some of the papers that have been presented during the workshop are very data-oriented and focus on a specific kind of data to be preserved, maintained or kept accessible, like computer games (W. Bergmeyer), metadata on products in a company (J. Brunsmann), digital libraries in general (C. Morbidoni et al.), archival data in general (C. Cortese and Mantegari, S. Mazzini and F. Ricci, S. Hennicke et al.) and pictures of museum items (T. Wray and P. Eklund). Other papers focus on a general approach like the paper by A. Schröder et al. who present a novel and promising approach for semantic hierarchical storage management. Another example for a paper that focuses on a general approach is the paper by Kai Eckert who proposes a basic linguistic indexer for digital libraries.

The workshop started with an invited talk on *The KEEP emulation framework (W. Bergmeyer)* which is also contained as publication in this volume. In this publication, W. Bergmeyer presents the KEEP (Keeping Emulation Environments Portable) project which is a research project of the European 7th Framework Programe. During the workshop, a demo of the KEEP emulation software framework has been shown. This talk brought the general trend of emulation as a preservation strategy which is currently the method of choice when preserving software tools and multimedia systems into the discussions of the workshop. Afterwards, focusing on hardware as well, a *semantic extension of a hierarchical storage management system for small and medium-sized enterprises (A. Schröder et al.)* has been discussed. Since such a system saves costs, capacity and access time, it can be especially useful in large digital archiving frameworks and infrastructures in order to distribute, store and retrieve semantically coherent archival data.

In the submission about the *semantic exploration of archived product lifecycle metadata under schema and instance evolution (J. Brunsmann)*, J. Brunsmann brings a new view into the discussion since he considered a holistic approach for maintaining the life cycle of linked data describing obsolete product ideas within a company archive. Hence, he introduces an interesting application field for semantic digital archives.

The paper *Towards a semantic data library for the social sciences (T. Grotton et al.)* brings a very interesting preliminary approach for a linked library data infrastructure for statistical data in the social sciences into the discussion. Although this work does not consider digital archiving and is on a very preliminary state, it

provides an insight into a statistical semantic digital library infrastructure and hence stimulated the discussion on semantic digital libraries versus semantic digital archives. More information on semantic digital libraries is provided by the paper on *introducing the Semlib project: semantic web tools for digital libraries (C. Morbidoni et al.)* which describes an annotation system for digital libraries. The proposed system adds user interaction to digital libraries via annotation and provides semantic structure to such annotations as well.

The paper *LOHAI: Providing a baseline for KOS based automatic indexing (K. Eckert)* proposes a free, open source and easy to use indexer tool for KOSs. This tool can provide the fundament on which to build more ambitious tools; although it has been developed for digital libraries, it can be used in other contexts like digital archiving contexts as well.

The publication on *extending the digital archives of italian psychology with semantic data (C. Cortese and G. Mantegari)* discusses an approach for implementing a semantic digital archive using CIDOC CRM for ontology modeling. Similarly, the paper on *EAC-CPF Ontology and Linked Archival Data (S. Mazzini and F. Ricci)* presents a topic that is relevant for digital archiving. More specifically, the development of an ontology is described that corresponds semantically to the EAC-CPF schema which is an archival standard for modelling and describing individuals, families and corporate bodies that create, preserve, use and are responsible for and/or associated with records in a variety of ways. A related topic is discussed in the submission about the *conversion of EAD into EDM linked data (S. Hennicke et al.)* as it deals with integrating archival finding aids into the portal of the Europeanna project which is an ambitious european project aiming at integrating data and information of museum, archives and libraries in one semantic web enhanced portal.

With *Concepts and Collections: A case study using objects from the Brooklyn Museum (T. Wray and P. Eklund)*, an interesting approach for a browsing framework for digitised cultural collections based on Formal Concept Analysis has been presented. This framework has also been evaluated with a case study using real data of the Brooklyn museum which nicely demonstrates that appropriate NLP techniques can be used to extract formal contexts from textual resources.

By the paper *Publishing Europe's television heritage on the web (J. Oomen and V. Tzouvaras)*, the first results of the European project EUScreen that deals with aggregating television heritage from European television archives for the European digital library Europeana have been presented.

Another very interesting paper is *A security contextualization framework for digital long-term preservation (K. Qian et al.)* as it is concerned with semantic security policies for digital archives. The approach extends the OAIS standard with security related features. Hence, an often neglected but crucial aspect in digital archiving is considered when establishing policies and infrastructures for long-term preservation.

The submission on *DA-NRW: A distributed architecture for long-term preservation (M. Thaller et al.)* presents an ongoing project that aims at creating a digital archive or long-term repository for the German state of North-Rhine Westphalia. This system will have a kind of sandwich position having to ingest data of depositors being archives and act as a pre-aggregator for portals like the Deutsche

Digitale Bibliothek or Europeana. A similar topic is dealt with by the paper on *RDFa as a lightweight metadata interoperability layer between repository software and LOCKSS (F. Ostrowski)* as it considers the extension of the LOCKSS framework with RDF and ontologies using SPARQL endpoints.

We would like to thank all members of the program committee for supporting us in the reviewing process. Altogether, the diversity of the papers in these proceedings represent a multitude of interesting facets about the new, exciting and promising research field of semantic digital archives and semantic digital archiving infrastructures. Hence, these proceedings provide a good and conclusive overview of the current research in this area.

December, 2011

Livia Predoiu,
Steffen Hennicke,
Andreas Nürnberger
Annett Mitschick
Seamus Ross

Organization

Program Chairs

Livia Predoiu	Otto-von-Guericke University of Magdeburg, Germany
Steffen Hennicke	Humboldt University of Berlin, Germany
Andreas Nürnberger	Otto-von-Guericke University of Magdeburg, Germany
Annett Mitschick	University of Dresden, Germany
Seamus Ross	University of Toronto, Canada

Program Committee

Sören Auer	University of Leipzig, Germany
Kai Eckert	University Library of Mannheim, Germany
Armin Haller	CSIRO, Australia
Stijn Heymans	SemanticBits, USA
Pascal Hitzler	Wright State University, USA
Yannis E. Ioannidis	University of Athens, Greece
Christian Keitel	State Archive of Baden-Württemberg, Germany
Thomas Lukasiewicz	University of Oxford, UK
Knud Möller	DERI Galway, Ireland
Vit Novacek	DERI Galway, Ireland
Johan Oomen	Netherlands Institute for Sound & Vision, Netherlands
Jacco van Ossenbruggen	VU University Amsterdam, Netherlands
Daniel Pitti	University of Virginia, USA
Andreas Rauber	Vienna University of Technology, Austria
Thomas Risse	L3S Research Center Hannover, Germany
Sebastian Rudolph	Karlsruher Institut für Technologie, Germany
Francois Scharffe	University of Montpellier, France
Michael Seadle	Humboldt University of Berlin, Germany
Marc Spaniol	Max-Planck-Institut Saarbrücken, Germany

Additional Reviewers

Thomas Low	Otto-von-Guericke University of Magdeburg, Germany
Magnus Pfeffer	HdM Stuttgart, Germany

Table of Contents

Archiving Frameworks and Infrastructures