Introduction to the focused issue on the 17th International Conference on Theory and Practice of Digital Libraries (TPDL 2013)

Trond Aalberg · Christos Papatheodorou

Published online: 6 June 2014
© Springer-Verlag Berlin Heidelberg 2014

Over the human life-span, seventeen years are an amount of time which marks the long transition from infancy into maturity. Definitely, conference maturity is not to be measured with the same metrics, but looking back in the areas of establishment and growth of the European Conference on Research and Advanced Technology for Digital Libraries, started in 1997, it would be worthy to ask what is making this conference a reference point of a community? We will argue that one of the unique impacts of this conference on the digital library domain is in the quality of sharing it offers and in its particular role to inspire. While its first editions helped the establishment and the professional recognition of the digital library community, isn't it a great compliment to this community that this domain sprouted in a whole range of new areas, which grew strong and now meet at their specialized well-established events—digital curation and preservation, metadata or evaluation of information services to mention just a few.

In 2013, the conference took place in Valletta, Malta during September 22–26, 2013 and it had the challenging task to re-connect its community in a climate of economic austerity. Once again, the conference called for sharing and inspiration: the general theme “Sharing meaningful information” addressed the challenges of interoperability and long-term preservation in the environment of the web of data. The advent of the technologies enhances the exchange of information with rich semantics, facilitates metadata interlinking

with user contributed data and offers new services outlook-ing to the development of a web of data. The conference, aiming to address the challenges of this environment, invited submissions in a wide range of research topics, clustered in four broader areas: Foundation, Infrastructures, Content and Services. The formulation of the topics once again aimed to bring into the conference a range of interdisciplinary meth-ods, and to engage with it the academic and practitioner communities.

This focused issue of the International Journal of Digital Libraries brings together extended versions of 8 among the best papers from TPDL 2013. Authors of papers that received the best review scores from the program committee were invited to submit revised and extended versions of their works. A new peer review process took place for the extended papers before including a subset of them in this issue. The topics of the papers are diverse, but they are a representative collection of both traditional topics that have been of concern in the DL community as well as emerging topics where our knowledge will develop in the next years.

The paper “*Unsupervised Document Structure Analysis of Digital Scientific Articles*” by Stefan Klampfl, Michael Granitzer, Kris Jack and Roman Kern presents a processing pipeline that performs both physical and logical layout analysis of scientific articles in PDF format. The pipeline uses a number of unsupervised machine learning techniques and heuristics to extract the body text and the table of contents. The results show that the proposed solution outperformed a state-of-the-art system in terms of quality. Work in this area is important given the extensive use of the PDF format and a big variety of layouts in use and the need to access the actual body text and table of contents to provide for more advanced browsing and searching in digital scientific content. This is a revised and extended version of the paper that received the TPDL 2013 Best Paper award.

T. Aalberg
Norwegian University of Science and Technology,
Trondheim, Norway
e-mail: Trond.Aalberg@idi.ntnu.no

C. Papatheodorou (✉)
Ionian University, Corfu, Greece
e-mail: papatheodor@ionio.gr

The paper “*Who and What Links to the Internet Archive*” by Yasmin AlNoamany, Ahmed AlSum, Michele C. Weigle and Michael L. Nelson analyzes the use of the Internet Archive’s (IA) Wayback Machine, which is the largest and oldest public web archive. It presents a study that analyzes the archive’s web access logs aiming to discover what users are looking for, why they come to IA, where they come from, and how pages link to IA. This is a revised and extended version of the paper that received the TPD L 2013 Best Student Paper award.

The paper “*A System for High Quality Crowdsourced Indigenous Language Transcription*” by Ngoni Munyaradzi and Hussein Suleman presents a crowdsourcing method to transcribe manuscripts from a rare collection that contains artwork, notebooks and dictionaries of the indigenous people of Southern Africa. Non-expert volunteers are invited to use a tool to transcribe pages of handwritten text in now-extinct languages with a specialized notation system. The paper presents details of the online tool as well as results from experiments that were conducted to determine the quality and consistency of transcriptions. The results show that volunteers are able to produce reliable transcriptions of high quality and the system achieves much better accuracy than previous automatic methods based on machine learning.

The paper “*Metadata Management, Interoperability and Linked Data Publishing Support for Natural History Museums*” by Giannis Skevakis, Konstantinos Makris, Varvara Kalokyri, Polyxeni Arapi and Stavros Christodoulakis presents the architecture, deployment and evaluation of the infrastructure developed in the Natural Europe project that allows the curators to publish, semantically describe, manage and disseminate cultural heritage objects. The main motivation for this work is to facilitate sharing and exploitation of the rich knowledge about Earth’s biodiversity and natural history that is maintained by natural history museums. Moreover, the article discusses the methodology for transition to the semantic web and the publishing of natural history museums’ metadata as linked open data.

The paper “*Word Occurrence Based Extraction of Work Contributors from Statements of Responsibility*” by Nuno Freire addresses the identification of all contributors of an intellectual work, when they are recorded in bibliographic data but in unstructured form. The identification of work contributors mentioned in statements of responsibility in library records is a typical motivation for the application of information extraction techniques. This paper presents an approach developed for the specific application scenario of

the ARROW rights infrastructure being deployed in several European countries to assist in the determination of the copyright status of works that may not be under public domain. Evaluation shows that it performs reliably across languages and bibliographic datasets.

The paper “*Profiling Web Archive Coverage for Top-Level Domain and Content Language*” by Ahmed AlSum, Michele C. Weigle, Michael L. Nelson and Herbert Van de Sompel defines the concept of web archive profile that is a set of characteristics that discriminate a web archive from others. The profiles are exploited by the Memento Aggregator, which performs a federated search over web archives to select the most probable of them, so that a URI—archived at a specific datetime—could be retrieved efficiently.

The paper “*Evaluating Distance-Based Clustering for User (Browse and Click) Sessions in a Domain-Specific Collection*” by Jeremy Steinhauer, Lois M. L. Delcambre, Marianne Lykke, Marit Kristine Ådland presents a method for improving information retrieval in domain-specific collections by exploiting clustering algorithms over user sessions from a click log. The sessions are grouped together if they answer to the same questions. Then, new sessions are classified in real-time to improve the retrieval performance. The paper investigates the effectiveness of the machine learning distance measures used by various clustering algorithms and presents a user study to evaluate the quality of the clusterings produced by the usage of such measures.

The paper “*Sustainability of Digital Libraries: a Conceptual Model and a Research Framework*” by Gobinda Chowdhury aims to develop a conceptual model and a research framework for studying the economic, social and environmental sustainability of digital libraries.

Many people have contributed to this special issue. A special thanks to all reviewers, who contributed high-quality feedback to the authors as well as important feedback that guided the editors in the decisions. It is a great pleasure to acknowledge the help of Christoph Becker, Lillian Cassel, Donatella Castelli, Klaus-Peter Clas, Ingo Frommholz, Clyde Lee Giles, Marcos André Gonçalves, Jaap Kamps, Michael Khoo, Andreas Rauber, Reagan Moore, Heiko Schuldt, Michalis Sfakakis, Giannis Tsakonas, Maja Žumer.

We are also grateful to N. Adam, E. J. Neuhold and R. Furuta for encouraging us to produce this focused issue, and to I. Frommholz for his assistance in managing this issue which we hope represents an important contribution to the research and practice of digital libraries.