Digital Access
to
Textual Cultural Heritage

DATeCH
2014

Conference Proceedings

Madrid, May 19–20, 2014

Published by ACM ISBN 978-1-4503-2588-2/14/05)

Contents

Foreword by the programme chairs

We are delighted to present the program of the first international conference on Digital Access to Textual Cultural Heritage (DATeCH 2014). The aim of establishing this conference is to bring together researchers in the complementary fields of Document Image Analysis and Recognition, Computational Linguistics and Digital Humanities as well as content holders and practitioners working on the creation, transformation and exploitation of historical documents in digital form. We strongly believe that there are very significant benefits in gathering such a multi-disciplinary group of experts, combining experiences and discussing ways forward for tackling the significant challenges and opportunities presented by historical documents.

For this inaugural edition, we were very pleased to receive 49 submissions from 23 countries from 6 continents (no submissions from Antarctica this time). We gathered 147 reviews with the invaluable help of 44 reviewers (including members of the Program and Organising committees as well as additional reviewers). Based on these reviews, 16 papers were accepted for oral presentation. While not a primary aim for this first edition of DATeCH, it is worth noting that the oral paper acceptance rate of 33% is in line with other premier conferences and workshops in related fields. We also accepted 16 papers to be presented as posters based both on reviewers' recommendations as well as on the fact that, due to their nature, some papers lend themselves better to one-to-one discussion.

The program of DATeCH2014 is structured into five oral sessions and a poster session within the overall programme of the Digitisation Days event, which also comprises a number of panel sessions and presentations from commercial organisations. The program is broadly divided into the main themes of Document Analysis and OCR, Linguistic Processing and Encoding, Post-Correction, Best Practices and Experiences, and Enrichment. We sincerely hope that you will find interesting papers presented and discussed during the conference and that the proceedings will also serve as reference, well after DATeCH2014, for the valuable ideas and the good work produced by our combined communities.

We would like to sincerely thank all the authors who submitted their work to DATeCH2014 and especially the program committee members and all additional reviewers for their outstanding and timely work. We are looking forward to meeting all of you at DATeCH2014 during the Digitisation Days event in Madrid.

With warmest wishes,

Apostolos Antonacopoulos
University of Salford

Klaus U. Schulz
Ludwig-Maximilians-
Universität München

Foreword by the organisation chairs

The Impact Centre of Competence in Digitisation (http://www.digitisation.eu) is a European initiative that disseminates the best technology for the digitisation of historical text. Currently, the Impact Centre gathers over 35 institutions, including the most important European National Libraries, distinguished research institutions and leading companies delivering services in the digitisation sector.

Succeed is a support action funded by the European Union that promotes the take up and validation of research results in mass digitisation with focus on the textual content. Succeed improves the availability of tools and resources, fosters the transfer of knowledge, the creation of research consortia and explores the role of emerging business models, funding opportunities and public-private partnerships to improve large-scale text digitisation techniques.

The Impact Centre and Succeed organise the Digitisation Days, a major event in text digitisation field. The Digitisation Days are conceived as a meeting point for librarians, researchers and companies, to present an up-to-date vision of the most recent advances in technology for the digitisation of text, to showcase successful experiences in their application, and to explore the challenges for the near future of digitisation. In this framework, the DATeCH international conference has been conceived as a meeting point, where the latest and future technologies for the digitisation of historical text will be presented and discussed.

The DATeCH conference and the Digitisation Days take place at the Biblioteca Nacional de España in Madrid, a location surrounded by the famous art museums and the neighboring historic center. Therefore, the event provides an excellent opportunity not only to learn about the challenges and the state of the art in digitisation but also to enjoy an environment plenty of culture and history.

Rafael C. Carrasco
Universidad de Alicante

Francis Ballesteros
Fundación Biblioteca
Virtual Miguel de Cervantes

DATeCH 2014 organisation

Programme chairs

Apostolos Antonacopoulos (University of Salford) and Klaus U. Schulz (Ludwig-Maximilians-Universität München)

Organisation chairs

Rafael C. Carrasco (Universidad de Alicante) and Francis Ballesteros (Biblioteca Virtual Miguel de Cervantes and Impact Centre of Competence)

Organisation committee

Isabel Martínez, Enrique Mollà, Ester Boldrini, Rossana Pinna Inmaculada Caturla, Silvia G. Ponzoda, Rafael González, Fernando Pérez, Marta Mengual.

With the cooperation of

succeed

impact
digitisation.eu
centre of competence

Programme Committee

- Aly Conteh, The British Library
- Basilis Gatos, Demokritos National Center for Scientific Research
- Bruce Robertson, Mount Allison University
- Christoph Ringlstetter, Ludwig-Maximilians Universität
- Christopher Blackwell, Furman University
- Claudine Moulin, Universität Trier
- David Doermann, University of Maryland
- Enrique Vidal, Universitat Politècnica de València
- Francois Bry, Ludwig-Maximilians Universität
- Gregory Crane, Universität Leipzig
- Günter Mühlberger, Universität Innsbruck
- Joan Andreu Sánchez, Universitat Politècnica de València
- Laura Mandell, Texas A&M University
- Lou Burnard, TEI Board
- Malte Rehbein, Universität Passau
- Marco Büchler, Göttingen Centre for Digital Humanities
- Martin Müller, Northwestern University
- Neel Smith, College of Holy Cross
- Rose Holley, National Archives of Australia
- Simone Marinai, Università degli Studi di Firenze
- Stefan Gradmann, Institut für Bibliotheks- und Informations-wissenschaft, Humboldt-Universität zu Berlin
- Stoyan Mihov, Bulgarian Academy of Sciences
- Thierry Paquet, Université de Rouen
- Tomaz Erjavec, Institut Jožef Stefan

Session 1

Document Analysis and OCR

Session 2

Linguistic processing and encoding

Session 3

Postcorrection

Session 4

Best practices and experiences

Session 5

Enrichment

Session 6

Posters