TU WIEN

DocEng 2016

ACM Symposium on Document Engineering

September 13-16

VIENNA

AUSTRIA

Welcome

It is both an honor and a pleasure to hold the 16th ACM Symposium on Document Engineering, DocEng 2016, at the TU Wien, Austria, organized by the Computer Vision Lab (CVL). DocEng is the leading international ACM symposium for researchers, practitioners, developers, and users to explore cutting-edge ideas and to exchange techniques, tools, and experiences in the domain of document engineering. It aims at bringing together researchers in the fields of computer vision, multimedia technologies, image processing, image analysis, information and systems analysis, electronic publishing, business process analysis, and business informatics. The symposium is intended as convention of renowned experts in all areas of document engineering of both academia and industry to present and discuss recent progress and advances in the fields of: document models and structures, document representation and standards, distributed documents, collaborative documents and the sharing economy, document internationalization, multilingual representations, document authoring tools and systems, document presentation (typography, formatting, layout), automatically generated documents, content customization, variable printing, documents for mobile devices, web document processing and interaction, document repositories, massive collections of documents, digital libraries and archives, secure document workflows, collaborative authoring and editing, culture-dependent layouts, and many more.

Our call for papers attracted submissions from 27 countries (Australia, Austria, Brazil, Canada, China, France, Germany, Greece, India, Indonesia, Iran, Italy, Japan, Korea, Macao, Malta, Netherlands, New Zealand, Romania, Russian Federation, Slovakia, Spain, Switzerland, Tunisia, United Kingdom, United States, Vietnam). All papers were carefully reviewed by a minimum of three Program Committee members, upon which decisions for acceptance were based on correctness, presentation, technical depth, scientific significance and originality. The Program Committee accepted 11 of 35 reviewed full paper submissions (31%) and 12 of 36 reviewed short paper submissions (33%) for oral presentation, for a combined acceptance rate of 32%. A further 10 short paper submissions were accepted for poster presentation.

Robert Sablatnig
Symposium Chair

Tamir Hassan
Program Chair

Overview

Wednesday, September 14 — EI 8

08:45 Registration
09:30 Session 1: Welcome and Introduction
09:45 Session 2: Keynote I - Günter Mühlberger
10:45 Coffee Break
11:15 Session 3: Layouts and Publishing
12:30 Session 4: BoF - How It Works
12:45 Lunch Break
14:15 Session 5: XML & Data Modelling
15:30 Coffee Break
16:00 Session 6: ProDoc: Doctoral Consortium
17:00 Session 7: Text Analysis I: Similarity
19:00 Welcome Reception, TUtheSky

Thursday, September 15 — EI 8

09:30 Session 8: Keynote II - Peter Biľak
10:30 Coffee Break
11:00 Session 9: Text Analysis II: Classification
12:15 Session 10: BoF: The Results
12:45 Lunch Break
14:15 Session 11: Workshop Session Recap
14:45 Session 12: Poster Lightning Talks
15:05 Coffee & Poster Session
16:00 Session 13: Text Analysis III: Summarization
16:45 Session 14: SIGWEB Presentation
19:30 Conference Banquet, Rathaus (City Hall)

Friday, September 16 — EI 8

09:30 Session 15: Applications & Security
10:45 Coffee Break
11:15 Session 16: Visual Document Analysis
12:45 Session 17: Closing Notes
13:00 Lunch Break

Keynotes

Günter Mühlberger
Wednesday, September 14, 09:45

Research Infrastructures (RIs) are one of the key concepts in Horizon 2020, the European Commission's Research programme. A budget of EUR 2.7bn is available for projects under the RI programme. The talk will describe some of the main characteristics of RIs and introduce the H2020 Recognition and Enrichment of Archival Documents (READ) project which is dedicated to setting up a highly specialized service platform and making available some of the state-of-the-art technology in pattern recognition and document engineering, namely Handwritten Text Recognition, Automatic Writer Identification, and Keyword Spotting. Archives and libraries, as well as humanities scholars and the general public will be enabled to use the service platform which will improve access to cultural heritage, advance research in humanities and encourage a broad audience to investigate their personal family history.

Peter Biľak
Thursday, September 15, 09:30

In this talk, Peter Biľak will examine the ways that current publishing practices are rooted in the 19th century, and how in order to move forward, we may have to go back to the roots and reconnect with readers. He will also talk about his recent project, *Works That Work* magazine, which set out to rethink publishing paradigms, starting with its financing, distribution and production. Works That Work aims to discuss design outside of the traditional design discourse, and Peter Biľak will argue for widening the understanding of the design discipline.

Workshops & Tutorials

Tuesday, September 13

09:30 Session 1

Tutorial:	Table Modelling, Extraction and Processing	von Neumann
Tutorial:	Document Engineering Issues in Malware Analysis	Gödel
Workshop:	DChanges: Modeling, Detection, Storage and Visualization	Zemanek

11:00 Coffee Break

11:30 Session 2

Tutorial:	Table Modelling, Extraction and Processing	von Neumann
Tutorial:	Document Engineering Issues in Malware Analysis	Gödel
Workshop:	DChanges: Modeling, Detection, Storage and Visualization	Zemanek

13:00 Lunch

14:00 Session 3

Workshop:	Future Publishing Formats	Gödel
Workshop:	DChanges: Modeling, Detection, Storage and Visualization	Zemanek

15:30 Coffee Break

16:00 Session 4

Workshop:	Future Publishing Formats	Gödel
Workshop:	DChanges: Modeling, Detection, Storage and Visualization	Zemanek

Workshops

DChanges: Modeling, Detection, Storage and Visualization
Giole Barabucci, Uwe M. Borghoff, Angelo Di Iorio, Sonja Schimmler, Ethan Munson

Workshop on Future Publishing Formats
Michael Piotrowski

Tutorials

Table Modelling, Extraction and Processing
Max Göbel, Tamir Hassan, Ermelinda Oro, Roya Rastan.

Document Engineering Issues in Malware Analysis
Robert Brandon, Charles Nicholas

Wednesday *September 14*

09:30-09:45 Session 1: Welcome and Introduction

09:45-10:45 Session 2: Keynote I

Chair: *Robert Sablatnig*

09:45 *Günter Mühlberger*
Research Infrastructures, or How Document Engineering, Cultural Heritage, and Digital Humanities Can Go Together

10:45 Coffee Break

11:15-12:30 Session 3: Layouts and Publishing

Chair: *Steven R. Bagley*

11:15 *Frank Mittelbach*
A general framework for globally optimized pagination

11:45 *David Schölgens, Sven Müller, Christine Bauer, Roman Tilly and Detlef Schoder*
Aesthetic Measures for Document Layouts: Operationalization and Analysis in the Context of Marketing Brochures

12:15 *Lei Liu, Rares Vernica, Tamir Hassan, Niranjan Damera Venkata, Yang Lei, Jian Fan, Jerry Liu, Steve J. Simske and Shanchan Wu*
METIS: A Multi-faceted Hybrid Book Learning Platform

12:30-12:45 Session 4: BoF - How It Works

Chair: *Charles Nicholas*

12:45-14:15 Lunch Break (incl. BoF)

14:15-15:30 Session 5: XML & Data Modelling

Chair: *Ethan V Munson*

14:15 *Nikolaos Lagos and Jean-Yves Vion-Dury*
Digital Preservation Based on Contextualized Dependencies

14:45 *Vincent Barrellon, Pierre-Edouard Portier, Sylvie Calabretto and Olivier Ferret*
Schema-aware extended Annotation Graphs

15:15 *Marcio Moreno, Rafael Brandão and Renato Cerqueira*
NCM 3.1: A Conceptual Model for Hyperknowledge Document Engineering

15:30 Coffee Break

16:00-17:00 Session 6: ProDoc: Doctoral Consortium

Chair: *Cerstin Mahlow*

16:00 *Tobias Gradl*
A Language Theoretical Framework For The Integration Of Arts And Humanities Research Data

16:30 *Alan Guedes*
Towards supporting multimodal and multiuser interactions in multimedia languages

17:00-18:15 Session 7: Text Analysis I: Similarity

Chair: *Peter King*

17:00 *Nava Ehsan, Frank Tompa and Azadeh Shakery*
Using a Dictionary and n-gram Alignment to Improve Fine-grained Cross-Language Plagiarism Detection

17:30 *Xiangru Wang, Seyednaser Nourashrafeddin and Evangelos Milios*
Relaxing Orthogonality Assumption in Conceptual Text Document Similarity

18:00 *Ian Knight and David Brailsford*
Enhancing the Searchability of Page-Image PDF Documents Using an Aligned Hidden Layer from a Truth Text

19:00 Welcome Reception, TUtheSky

Thursday *September 15*

09:30-10:30 Session 8: Keynote II

Chair: *Tamir Hassan*

09:30 *Peter Biľak*
Design Is Not What You Think It Is

10:30 Coffee Break

11:00-12:15 Session 9: Text Analysis II: Classification

Chair: *Michael Piotrowski*

11:00 *Salvatore Trani, Diego Ceccarelli, Claudio Lucchese, Salvatore Orlando and Raffaele Perego*
SEL: a Unified Algorithm for Entity Linking and Saliency Detection

11:30 *Jan Oevermann and Wolfgang Ziegler*
Automated Intrinsic Text Classification for Component Content Management Applications in Technical Communication

11:45 *Mario Kubek and Herwig Unger*
Centroid Terms as Text Representatives

12:00 *Phanucheep Chotnithi and Atsuhiro Takasu*
Frequent Multi-Byte Character Subtring Extraction using a Succinct Data Structure

12:15-12:45 Session 10: BoF - The Results

Chair: *Charles Nicholas*

12:45-14:15 Lunch Break

14:15-14:45 Session 11: Workshop Session Recap

Chair: *Sonja Schimmler*

14:45-15:05 Session 12: Poster Lightning Talks

14:45 *Luciano Cabral, Manoel Neto, Artur Borges, Rafael Lins, Rinaldo Lima, Rafael Ferreira, Steven Simske and Marcelo Riss*
Multilingual News Article Summarization in Mobile Devices – Demo

14:47 *Daan Leijen*
Rendering Mathematic Formulas for the Web in Madoko

14:49 *Roya Rastan, Hye-Young Paik and John Shepherd*
A PDF Wrapper for Table Processing

14:51 *Alexey Shigarov, Andrey Mikhailov and Andrey Altaev*
Configurable Table Structure Recognition in Untagged PDF documents

14:53 *Tobias Gradl and Andreas Henrich*
Extending data models by declaratively specifying contextual knowledge

14:55 *Alessandro Calefati, Ignazio Gallo, Alessandro Zamberletti and Lucia Noce*
Using Convolutional Neural Networks for Content Extraction from Online Flyers

14:57 *Tobias Swoboda, Matthias Hemmje, Mihai Dascalu and Stefan Trausan-Matu*
Combining Taxonomies using Word2Vec

14:59 *Junki Tanijiri, Manabu Ohta, Atsuhiro Takasu and Jun Adachi*
Important Word Organization for Support of Browsing Scholarly Papers Using Author Keywords

15:01 *Baoli Li*
Selecting Features with Class Based and Importance Weighted Document Frequency in Text Classification

15:03 *Giorgos Sfikas, Georgios Louloudis, Nikolaos Stamatopoulos and Basilis Gatos*
Bayesian mixture models on connected components for Newspaper article segmentation

15:05-16:00 Coffee & Poster Session

16:00-16:45 Session 13: Text Analysis III: Summarization

Chair: *Dick Bulterman*

16:00 Rodolfo Ferreira, Rafael Ferreira, Rafael Lins, Hilário Oliveira, Marcelo Riss and Steven Simske
Appling Link Target Identification and Content Extraction to improve Web News Summarization

16:15 *Jamilson Batista Antunes, Rafael Dueire Lins, Rinaldo Lima, Steven J. Simske and Marcelo Riss*
Towards Cohesive Extractive Summarization through Anaphoric Expression Resolution

16:30 *Hilário Oliveira, Rinaldo Lima, Rafael Lins, Fred Freitas, Marcelo Riss and Steven Simske*
Assessing Concept Weighting in Integer Linear Programming based Single-document Summarization

16:45-17:30 Session 14: SIGWEB Presentation

Chair: *Dick Bulterman*

19:30 Conference Banquet, Rathaus (City Hall)

Friday *September 16*

09:30-10:45 Session 15: Applications & Security

Chair: *David F. Brailsford*

09:30 *Rodrigo Laiola Guimaraes, Priscilla Avegliano and Lucas Villa Real*
A Lightweight and Efficient Mechanism for Fixing the Synchronization of Misaligned Subtitle Documents

10:00 *Laurent Denoue, Scott Carter and Matthew Cooper*
DocuGram: Turning Screen Recordings into Documents

10:15 *Eya Ben Charrada and Stefan Mussato*
An Exploratory Study on Managing and Searching for Documents in Software Engineering Environments

10:30 *Margaret Sturgill and Steven Simske*
Mass Serialization Method for Document Encryption Policy Enforcement

10:45 Coffee Break

11:15-12:45 Session 16: Visual Document Analysis

Chair: *Steven Simske*

11:15 *Prerana Jana, Anubhab Majumdar, Sekhar Mandal and Bhabatosh Chanda*
Generation of Searchable PDF of the Chemical Equations segmented from Document Images

11:45 *Emilio Granell and Carlos David Martinez Hinarejos*
A Multimodal Crowdsourcing Framework for Transcribing Historical Handwritten Documents

12:15 *Lucia Noce, Ignazio Gallo, Alessandro Zamberletti and Alessandro Calefati*
Embedded Textual Content for Document Image Classification with Convolutional Neural Networks

12:45-13:00 Session 17: Closing Notes

13:00-14:30 Lunch Break

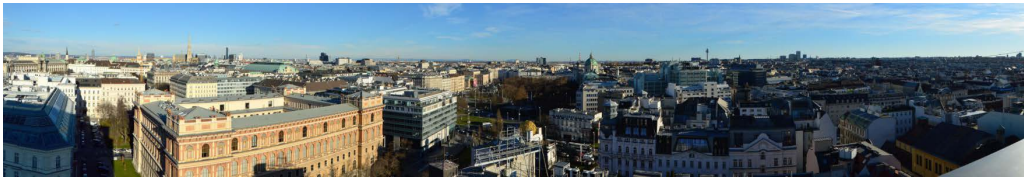
Events

Welcome Attendance
Wednesday, September 14 19:00

The welcome attendance will take place at TUtheSky, which is located in the 11th floor at the Campus Getreidemarkt near the city center. Enjoy the great view while eating some finger food and drinking Austrian wine and beer. Also take the possibility to have nice talks with all other participants of DocEng 2016.

Location
TUtheSky

Getreidemarkt 9
1060 Vienna



Conference Dinner
Thursday, September 15 19:30

The conference banquet will be held at the city hall. The city hall is one of the most splendid amongst the numerous monumental buildings in Vienna. Designed by Friedrich Schmidt (1825 - 1891), it was erected between 1872 and 1883. Don't miss the announcement of best paper and best student paper winners and join your DocEng colleagues for a dinner of wining, dining, and shining examples of research quality.

Location
City Hall

Rathausplatz 1
1010 Wien

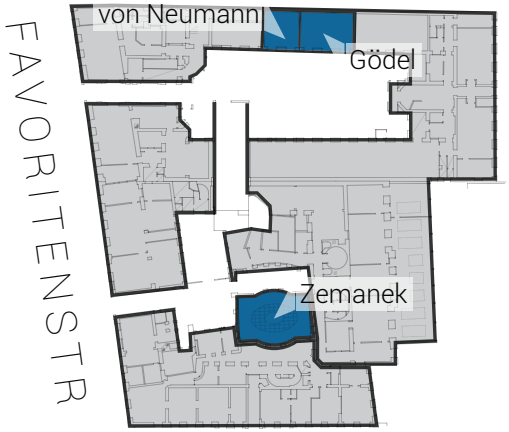
Venue - TU Wien

Legend

- U4 Underground
- Lokalbahn
- Bus
- Tramway (Bim)
- TU Wien
- TU Wien - DocEng16

Workshops

Favoritenstr. 9-11
1040 Vienna



Conference Hall

Gußhausstr. 27-29
1040 Vienna

Organization

Symposium Chair	Robert Sablatnig *TU Wien, Austria*
Program Chair	Tamir Hassan *HP Labs, Austria*
Workshop and Tutorials Chair	Sonja Schimmler *University of the Federal Armed Forces, Germany*
Doctoral Consortium Chair	Cerstin Mahlow *Institut für Deutsche Sprache, Germany*
BOF Chair	Charles Nicholas *University of Maryland, USA*
Local Chair	Florian Kleber *TU Wien, Austria*
	Stefan Fiel *TU Wien, Austria*
Publicity Chair	Markus Diem *TU Wien, Austria*
	Fabian Hollaus *TU Wien, Austria*
Steering Committee Chair	Steven Simske *HP Labs, USA*
Steering Committee	David Brailsford *University of Nottingham, UK*
	Dick Bulterman *CWI, Netherlands*
	Matthew Hardy *Adobe, USA*
	Peter King *University of Manitoba, Canada*
	Kim Marriot *Monash University, Australia*
	Ethan Munson *University of Wisconsin-Milwaukee, USA*
	Charles Nicholas *University of Maryland, USA*
	Maria da Graca C. Pimentel *Universidade de Sao Paulo, Brazil*
	Cécile Roisin *Université Pierre Mendes and INRIA, France*
	Jean-Yves Vion-Dury *Xerox Research Centre Europe, France*
	Anthony Wiley *OpenText, USA*

Program Committee

Apostolos Antonacopoulos *University of Salford, UK*
Vlad Atanasiu *University of Fribourg, Switzerland*
Steven R. Bagley *University of Nottingham, UK*
Helen Balinsky *HP Labs, UK*
Jean-Luc Bloechle *Sugarcube IT, Switzerland*
Uwe M. Borghoff *Universität der Bundeswehr München, Germany*
David F. Brailsford *University of Nottingham, UK*
Anne Brüggemann-Klein *Technische Universität München, Germany*
Pablo Cesar *CWI, Netherlands*
Paolo Ciccarese *Harvard Medical School/Mass. General Hospital, USA*
Michael L. Collard *The University of Akron, USA*
Niranjan Damera-Venkata *HP Labs, USA*
Markus Diem *TU Wien, Austria*
Angelo Di Iorio *University of Bologna, Italy*
Stefano Ferilli *University of Bari, Italy*
Stefan Fiel *TU Wien, Austria*
Pierre Geneves *CNRS, France*
Gersende Georg *French National Authority for Health, France*
C. Lee Giles *Pennsylvania State University, USA*
Matthew Hardy *Adobe, USA*
Tamir Hassan *HP Labs, Austria*
Fabian Hollaus *TU Wien, Austria*
Andrew Hunter *HP Labs, UK*
Nathan Hurst *Shutterstock, USA*
Rolf Ingold *University of Fribourg, Switzerland*
Peter R King *University of Manitoba, Canada*
Florian Kleber *TU Wien, Austria*
Alberto Laender *Universidade Federal de Minas Gerais, Brazil*
Monica Landoni *USI, Switzerland*
Baoli Li *Henan University of Technology, China*
Lei Liu *HP Labs, USA*
Marcus Liwicki *University of Fribourg, Switzerland*
John Lumley *jwL Research, UK*
Cerstin Mahlow *Institut für Deutsche Sprache, Germany*
Simone Marinai *University of Florence, Italy*
Kim Marriott *Monash University, Australia*
Evangelos Milios *Dalhousie University, Canada*
Mirella M. Moro *UFMG, Brazil*
Ethan V. Munson *University of Wisconsin-Milwaukee, USA*
Charles Nicholas *University of Maryland, USA*
Ermelinda Oro *ICAR-CNR, Italy*
Giorgio Orsi *University of Birmingham, UK*
Maria da Graca Pimentel *Universidade de Sao Paulo, Brazil*
Michael Piotrowski *Leibniz Institute of European History, Germany*
Stefan Pletschacher *University of Salford, UK*
Cécile Roisin *Université Grenoble Alpes, France*
Sebastian Roennau *Ravensburger AG, Germany*
Robert Sablatnig *TU Wien, Austria*
Sonja Schimmler *Universität der Bundeswehr München, Germany*
Patrick Schmitz *UC Berkeley, USA*
Badarinath Shantharam *HP India*
Ryan Shaw *University of North Carolina at Chapel Hill, USA*
Steven Simske *HP Labs, USA*
Fouad Slimane *EDSI Tech, Switzerland*
Margaret Sturgill *HP Labs, USA*
Cheng Thao *University of Wisconsin-Whitewater, USA*
Frank Tompa *University of Waterloo, Canada*
Fabio Vitali *University of Bologna, Italy*
Jean-Yves Vion-Dury *Xerox Research Centre Europe, France*
Christine Vanoirbeek *EPFL, Switzerland*
Erik Wilde *Siemens, USA*
Anthony Wiley *OpenText, USA*
Raymond Wong *UNSW, Australia*

Participating Countries