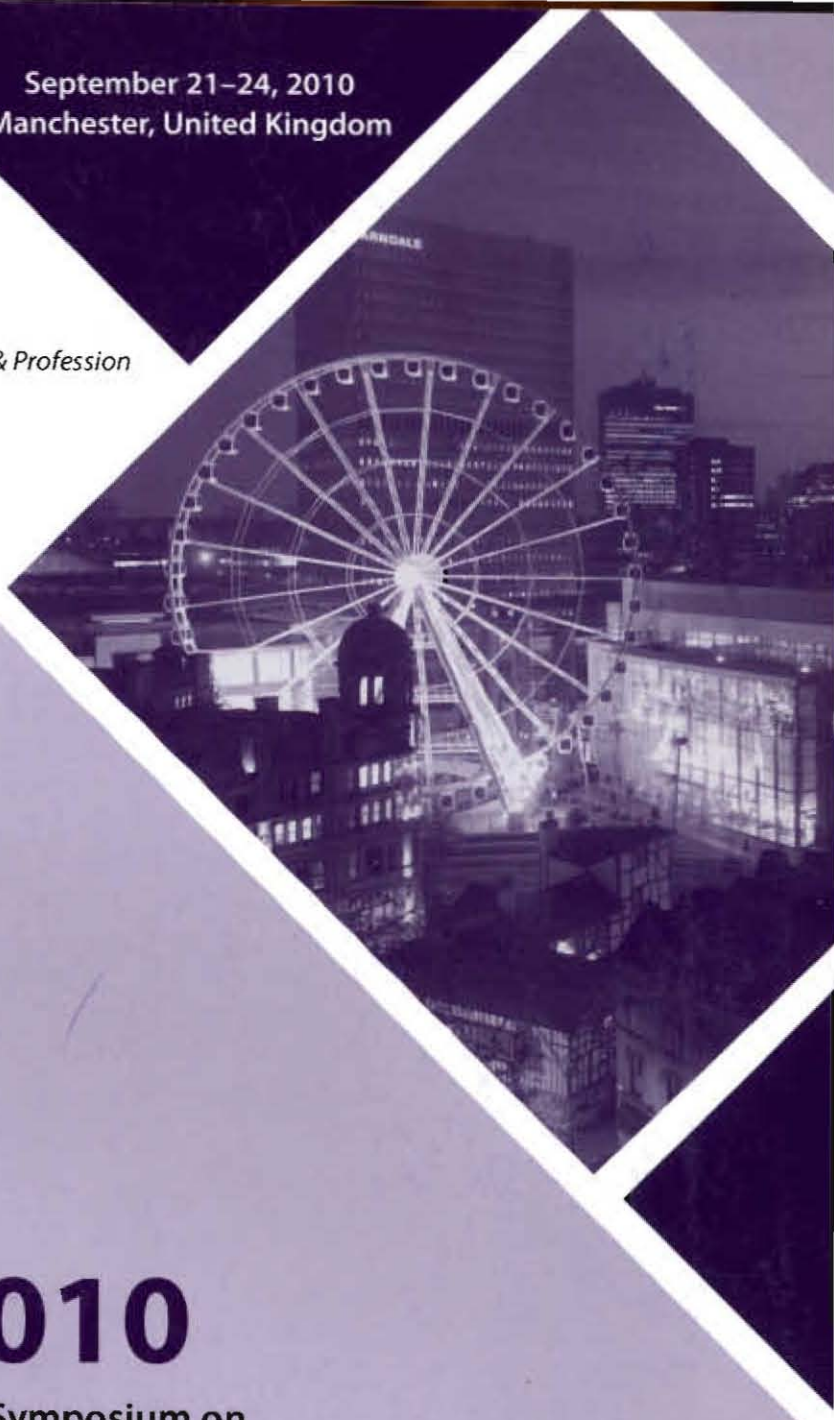


September 21–24, 2010
Manchester, United Kingdom



Association for
Computing Machinery

Advancing Computing as a Science & Profession



DocEng2010

Proceedings of the 2010 ACM Symposium on
Document Engineering

Sponsored by:

ACM SIGWEB

In cooperation with:

ACM SIGDOC

Supported by:

Hewlett-Packard & Xerox

Table of Contents

DocEng2010 Symposium Organization	ix
--	----

Keynote Address

Session Chair: Apostolos Antonacopoulos (*University of Salford*)

- **Exploring the World's Knowledge in the Digital Age** 1
Aly Kaloa Conteh (*British Library*)

Session 1: Systems

Session Chair: Ethan Munson (*University of Wisconsin-Milwaukee*)

- **Document Engineering for a Digital Library: PDF Recompression Using JBIG2 and Other Optimizations of PDF Documents** 3
Petr Sojka, Radim Hatlapatka (*Masaryk University*)
- **Multilingual Composite Document Management Framework for the Internet: An FRBR Approach** 13
Jean-Marc Lecarpentier, Cyril Bazin, Hervé Le Crosnier (*GREYC-CNRS UMR 6072*)

Session 2: Authoring

Session Chair: Peter King (*University of Manitoba*)

- **A Social Approach to Authoring Media Annotations** 17
Roberto Fagá Jr., Vivian Genaro Motti (*Universidade de São Paulo*),
Renan G. Cattelan (*Universidade Federal de Uberlândia*),
Cesar A. C. Teixeira (*Universidade Federal de São Carlos*), Maria da Graça C. Pimentel (*Universidade de São Paulo*)
- **Creating and Sharing Personalized Time-Based Annotations of Videos on the Web** 27
Rodrigo Laiola Guimarães, Pablo Cesar, Dick C. A. Bulterman (*Centrum Wiskunde & Informatica*)
- **"This Conversation Will Be Recorded": Automatically Generating Interactive Documents from Captured Media** 37
Didier A. Vega-Oliveros, Diogo Santana Martins, Maria da Graça C. Pimentel (*Universidade de São Paulo*)

Session 3: Tools

Session Chair: David Brailsford (*University of Nottingham*)

- **Document Imaging Security and Forensics Ecosystem Considerations** 41
Steven J. Simske, Margaret Sturgill, Guy Adams (*Hewlett-Packard Laboratories*), Paul Everest (*Hewlett-Packard Co.*)
- **XUIB: XML to User Interface Binding** 51
Lendle Tseng (*National Taiwan University*), Yue-Sun Kuo (*Tatung University*),
Hsiu-Hui Lee, Chuen-Liang Chen (*National Taiwan University*)
- **From Templates to Schemas: Bridging the Gap Between Free Editing and Safe Data Processing** 61
Vincent Quint, Cécile Roisin (*INRIA, Grenoble*), Stéphane Sire, Christine Vanoirbeek (*EPFL*)
- **Lessons from the Dragon: Compiling PDF to Machine Code** 65
Steven R. Bagley (*University of Nottingham*)

Session 4: E-books

Session Chair: Helen Balinsky (*Hewlett Packard Laboratories*)

- **Transquotation in EBooks** 69
Steven A. Battle (*Gloze Limited*), Matthew Bemius (*Rochester Institute of Technology*)
- **Table of Contents Recognition for Converting PDF Documents in E-Book Formats** 73
Simone Marinai, Emanuele Marino, Giovanni Soda (*Università di Firenze*)

Session 5: Editing

Session Chair: Dick Bulterman (*CWI, The Netherlands*)

- **Using Versioned Tree Data Structure, Change Detection and Node Identity for Three-Way XML Merging** 77
Cheng Thao, Ethan V. Munson (*University of Wisconsin-Milwaukee*)
- **A Model for Editing Operations on Active Temporal Multimedia Documents** 87
Jack Jansen, Pablo Cesar, Dick C. A. Bulterman (*Centrum Wiskunde & Informatica*)
- **Semantics-Based Change Impact Analysis for Heterogeneous Collections of Documents** 97
Serge Autexier (*German Research Center for Artificial Intelligence*), Norman Müller (*Jacobs University*)
- **Linking Data and Presentations: From Mapping to Active Transformations** 107
Olivier Beaudoux (*ESEO-GRI*), Arnaud Blouin (*INRIA*)
- **Blocked Recursive Image Composition with Exclusion Zones** 111
Hui Chao, Daniel R. Treter, Xuemei Zhang, C. Brian Atkins (*Hewlett-Packard Laboratories*)

Session 6: Document Systems

Session Chair: Simone Marinai (*University of Florence*)

- **Differential Access for Publicly-Posted Composite Documents with Multiple Workflow Participants** 115
Helen Y. Balinsky, Steven J. Simske (*Hewlett-Packard Laboratories*)
- **Assessing the Readability of Clinical Documents in a Document Engineering Environment** 125
Mark Truran (*Teesside University*), Gersende Georg (*Haute Autorite de Sante*), Marc Cavazza (*Teesside University*), Dong Zhou (*Trinity College Dublin*)
- **Optimized Reprocessing of Documents Using Stored Processor State** 135
James A. Ollis, David F. Brailsford, Steven R. Bagley (*University of Nottingham*)
- **APEX: Automated Policy Enforcement eXchange** 139
Steven J. Simske, Helen Balinsky (*Hewlett-Packard Laboratories*)

Session 7: Analysis

Session Chair: John Lumley (*University of Nottingham*)

- **Unsupervised Font Reconstruction Based on Token Co-Occurrence** 143
Michael P. Cutter, Joost van Beusekom (*Technical University of Kaiserslautern*), Faisal Shafait (*German Research Center for Artificial Intelligence (DFKI)*), Thomas Michael Breuel (*Technical University of Kaiserslautern*)
- **Document Structure Meets Page Layout: Loopy Random Fields for Web News Content Extraction** 151
Alex Spengler, Patrick Gallinari (*Université Pierre et Marie Curie*)
- **Comparison of Global and Cascading Recognition Systems Applied to Multi-Font Arabic Text** 161
Fouad Slimane (*University of Fribourg & University of Sfax*), Slim Kanoun, Adel M. Alimi (*REGIM, University of Sfax, National School of Engineers (ENIS)*), Jean Hennebert (*University of Fribourg & Business Information System Institute*), Rolf Ingold (*University of Fribourg*)
- **Automatic Selection of Print-Worthy Content for Enhanced Web Page Printing Experience** 165
Suk Hwan Lim (*Hewlett-Packard Laboratories, Palo Alto*), Liwei Zheng, Jianming Jin (*Hewlett-Packard Laboratories, China*), Huiman Hou (*China HP Co. Ltd.*), Jian Fan, Jerry Liu (*Hewlett-Packard Laboratories, Palo Alto*)

Session 8: Creation/Printing

Session Chair: Steven Simske (*Hewlett Packard Labs, USA*)

- **A New Model for Automated Table Layout** 169
Mihai Bilauca, Patrick Healy (*University of Limerick*)

- **PDF Profiling for B&W Versus Color Pages Cost Estimation for Efficient On-Demand Book Printing** 177
Fabio Giannetti, Gary Dispoto (*HP Laboratories*),
Rafael D. Lins, Gabriel de França Pereira e Silva (*Universidade Federal de Pernambuco*),
Alexis Cabeda (*Hewlett-Packard Brasil*)
- **Next Generation Typeface Representations: Revisiting Parametric Fonts** 181
Tamir Hassan (*Technische Universität Wien*), Changyuan Hu (*20-20 Technologies Inc.*),
Roger D. Hersch (*Ecole Polytechnique Fédérale de Lausanne*)

Document Engineering I: Posters

Session Chair: Michael Gormish (*Ricoh Innovations*)

- **DSMW: A Distributed Infrastructure for the Cooperative Edition of Semantic Wiki Documents** 185
Hala Skaf-Molli, G  r  me Canals, Pascal Molli (*LORIA/INRIA Nancy-Grand Est, University of Nancy*)
- **Open World Classification of Printed Invoices** 187
Enrico Sorio, Alberto Bartoli, Giorgio Davanzo, Eric Medvet (*University of Trieste*)
- **Diffing, Patching and Merging XML Documents: Toward a Generic Calculus of Editing Deltas** 191
Jean-Yves Vion-Dury (*Xerox Research Centre Europe*)
- **Contextual Advertising for Web Article Printing** 195
Shengwen Yang, Jianming Jin (*Hewlett-Packard Company, China*),
Joshi Parag, Sam Liu (*Hewlett-Packard Company, Palo Alto*)
- **Table Layout Performance of Document Authoring Tools** 199
Mihai Bilauca, Patrick Healy (*University of Limerick*)
- **Document Product Lines: Variability-Driven Document Generation** 203
M   Carmen Penad  s, Jos   H. Can  s (*Technical University of Valencia*),
Marcos R.S. Borges (*Federal University of Rio de Janeiro*), Manuel Llavador (*Technical University of Valencia*)
- **Degraded Dot Matrix Character Recognition Using CSM-Based Feature Extraction** 207
Abderrahmane Namane (*University S  ad Dahlab of Blida*),
El Houssine Soubari, Patrick Meyrueis (*University of Strasbourg*)
- **Picture Detection in Document Page Images** 211
Patrick Chiu, Francine Chen, Laurent Denoue (*FX Palo Alto Laboratory*)
- **Down to the Bone: Simplifying Skeletons** 215
Jannis Stoppe, Bj  rn Gottfried (*University of Bremen*)
- **Interactive Layout Analysis and Transcription Systems for Historic Handwritten Documents** 219
Oriol Ramos-Terrades, Alejandro H. Toselli, Nicolas Serrano, Ver  nica Romero,
Enrique Vidal, Alfons Juan (*Universidad Polit  cnica de Valencia*)
- **Document Conversion for Cultural Heritage Texts: FrameMaker to HTML Revisited** 223
Michael Piotrowski (*Law Sources Foundation of the Swiss Lawyers Society*)
- **Glyph Extraction from Historic Document Images** 227
Lothar Meyer-Lerbs, Arne Schuldt, Bj  rn Gottfried (*University of Bremen*)
- **Style and Branding Elements Extraction From Businessweb Sites** 231
Limei Jiao (*Hewlett-Packard Laboratories, Beijing*),
Suk Hwan Lim, Nina Bhatti (*Hewlett-Packard Laboratories, Palo Alto*),
Yuhong Xiong (*Innovation Works*), Jerry Liu (*Hewlett-Packard Laboratories, Palo Alto*)

Document Engineering II: Posters

Session Chair: Rolf Ingold (*University of Fribourg*)

- **FormCracker: Interactive Web-Based Form Filling** 235
Laurent Denoue, John Adcock, Scott Carter, Patrick Chui, Francine Chen (*FX Palo Alto Laboratory, Inc.*)
- **Semantics-Enriched Document Exchange** 239
Jingzhi Guo, Ming Sang Hou (*University of Macau*)

• Document and Item-Based Modeling: A Hybrid Method for a Socio-Semantic Web	243
Jean-Pierre Cahier, Xiaoyue Ma (<i>Université de Technologie de Troyes</i>), L'Hédi Zaher (<i>Cogniva Europe</i>)	
• Structure-Aware Topic Clustering in Social Media	247
Julien Dubuc, Sabine Bergler (<i>Concordia University</i>)	
• Pre-Evaluation of Invariant Layout in Functional Variable-Data Documents	251
John Lumley (<i>University of Nottingham</i>)	
• Towards a Common Evaluation Strategy for Table Structure Recognition Algorithms	255
Tamir Hassan (<i>Technische Universität Wien</i>)	
• Using Feature Models for Creating Families of Documents	259
Sven Karol, Martin Heinzerling, Florian Heidenreich, Uwe Aßmann (<i>Dresden University of Technology</i>)	
• Two New Aesthetic Measures for Item Alignment	263
Aline D. Riva, Alexandre K. Seki, João B. S. de Oliveira, Isabel H. Mansour, Ricardo Farias Piccoli (<i>PUCRS</i>)	
• Term Frequency Dynamics in Collaborative Articles	267
Sérgio Nunes (<i>University of Porto</i>), Cristina Ribeiro, Gabriel David (<i>University of Porto & INESC-Porto</i>)	
• A File-Type Sensitive, Auto-Versioning File System	271
Arthur Müller, Sebastian Rönna, Uwe M. Borghoff (<i>Universität der Bundeswehr München</i>)	
• Medieval Manuscript Layout Model	275
Micheal Baechler, Rolf Ingold (<i>University of Fribourg</i>)	
• Using Model Driven Engineering Technologies for Building Authoring Applications	279
Olivier Beaudoux (<i>ESEO-GRI</i>), Arnaud Blouin (<i>INRIA</i>), Jean-Marc Jézéquel (<i>INRIA/IRISA</i>)	
• On Helmholtz's Principle for Documents Processing	283
Alexander A. Balinsky (<i>Cardiff School of Mathematics</i>), Helen Y. Balinsky, Steven J. Simske (<i>Hewlett-Packard, Laboratories</i>)	
Author Index	287