Preface

Welcome to the proceedings of the 9th Workshop of the Initiative for the Evaluation of XML Retrieval (INEX)!

Traditional IR focuses on pure text retrieval over “bags of words,” but the use of structure—such as document structure, semantic metadata, entities, or genre/topical structure—is of increasing importance on the Web and in professional search. INEX has been pioneering the use of structure for focused retrieval since 2002, by providing large test collections of structured documents, uniform evaluation measures, and a forum for organizations to compare their results.

2010 was an exciting year for INEX, in which a number of new tracks started. In total, nine research tracks were included, studying different aspects of focused information access:

**Ad Hoc Track:** investigated the effectiveness of XML-IR and Passage Retrieval for highly focused retrieval by restricting result length to “snippets” or discounting for reading effort, using Wikipedia as a corpus.

**Book Track:** investigated techniques to support users in reading, searching, and navigating full texts of digitized books, by constructing reading lists of books for a given topic, or by looking for book pages that refute or confirm a factual statement.

**Data-Centric Track:** investigated focused retrieval over a strongly structured collection of IMDb documents, containing information about various entities like movies, actors, directors.

**Interactive Track:** investigated the behavior of users when interacting with XML documents, working on large set of Amazon book data (formal book descriptions) augmented by LibraryThing data (user-generated data).

**Link-the-Wiki Track:** investigated link discovery in the Te Ara encyclopedia.

**Question Answering Track:** investigated real-world focused information needs formulated as natural language questions, using the collection structure to construct readable summaries of question context and lists of answers.

**Relevance Feedback Track:** investigated the utility of incremental passage level feedback by simulating a searcher’s interaction, with submissions in the form of a executable computer program.

**Web Service Discovery:** investigated techniques for discovery of Web services based on searching service descriptions provided in WSDL.

**XML-Mining Track:** investigated structured document mining, especially the classification and clustering of semi-structured documents.

The aim of the INEX 2010 workshop was to bring together researchers who participated in the INEX 2010 campaign. During the year participating organizations contributed to the building of a large-scale XML test collection by creating topics, performing retrieval runs, and providing relevance assessments.
The workshop concluded the results of this large-scale effort, summarized and addressed issues encountered, and devised a work plan for the future evaluation of XML retrieval systems. There proceedings report on the final results of INEX 2010. We received 42 submissions, already being a selection of work at INEX, and accepted a total of 37 papers based on peer-reviewing, yielding an 88% acceptance rate.

All INEX tracks start from having available suitable text collections. We gratefully acknowledge the data made available by: Amazon and LibraryThing (Interactive Track), New Zealand Ministry for Culture and Heritage (Te Ara, Link-the-Wiki Track), Microsoft Research (Book Track), the Internet Movie Database (Data Centric Track), and the Wikimedia Foundation (Adhoc, Relevance Feedback, and XML-Mining Track).

After many years at Schloss Dagstuhl, and a trip to Brisbane, Australia, in 2009, the INEX workshop returned to Europe and was held in The Netherlands. Thanks to the Amsterdam team for preserving the unique atmosphere of INEX—a setting where informal interaction and discussion occur naturally and frequently—in the unique location of Huize Bergen in Vught.

We thank the Dutch Association for Information Science (Werkgemeenschap Informatiewetenschap, WGI) for sponsoring the best student award, which was presented to Ning Gao (Peking University) for the paper entitled “Combining Strategy for XML Retrieval.”

Finally, INEX is run for, but especially by, the participants. It was a result of tracks and tasks suggested by participants, topics created by particants, systems built by participants, and relevance judgments provided by participants. So the main thank you goes to each of these individuals!

May 2011

Shlomo Geva
Jaap Kamps
Ralf Schenkel
Andrew Trotman
Steering Committee

Charles L. A. Clarke  
University of Waterloo, Canada
Norbert Fuhr  
University of Duisburg-Essen, Germany
Shlomo Geva  
Queensland University of Technology, Australia
Jaap Kamps  
University of Amsterdam, The Netherlands
Mounia Lalmas  
Yahoo Research Barcelona, Spain
Stephen Robertson  
Microsoft Research Cambridge, UK
Andrew Trotman  
University of Otago, New Zealand
Ellen Voorhees  
NIST, USA

Chairs

Shlomo Geva  
Queensland University of Technology, Australia
Jaap Kamps  
University of Amsterdam, The Netherlands
Ralf Schenkel  
Max-Planck-Institut für Informatik, Germany
Andrew Trotman  
University of Otago, New Zealand

Track Organizers

Ad Hoc

Paavo Arvola  
University of Tampere, Finland
Shlomo Geva  
Queensland University of Technology, Australia
Jaap Kamps  
University of Amsterdam, The Netherlands
Ralf Schenkel  
Max-Planck-Institut für Informatik, Germany
Andrew Trotman  
University of Otago, New Zealand

Book

Antoine Doucet  
University of Caen, France
Gabriella Kazai  
Microsoft Research Cambridge, UK
Marijn Koolen  
University of Amsterdam, The Netherlands
Monica Landoni  
University of Lugano, Switzerland

Data-Centric

Qiuyue Wang  
Renmin University, China
Andrew Trotman  
University of Otago, New Zealand
**Interactive (iTrack)**

Thomas Beckers  
University of Duisburg-Essen, Germany

Norbert Fuhr  
University of Duisburg-Essen, Germany

Ragnar Nordlie  
Oslo University College, Norway

Nils Pharo  
Oslo University College, Norway

**Link-the-Wiki**

Shlomo Geva  
Queensland University of Technology, Australia

Andrew Trotman  
University of Otago, New Zealand

**Question Answering (QA)**

Veronique Moriceau  
LIMSI-CNRS, University Paris-Sud 11, France

Eric SanJuan  
University of Avignon, France

Xavier Tannier  
LIMSI-CNRS, University Paris-Sud 11, France

**Relevance Feedback**

Timothy Chappell  
Queensland University of Technology, Australia

Shlomo Geva  
Queensland University of Technology, Australia

**Web Service Discovery**

James Thom  
RMIT University, Australia

Chen Wu  
Curtin University of Technology, Australia

**XML-Mining**

Chris De Vries  
Queensland University of Technology, Australia

Sangeetha Kutty  
Queensland University of Technology, Australia

Richi Nayak  
Queensland University of Technology, Australia

Andrea Tagarelli  
University of Calabria, Italy
## Table of Contents

### Ad Hoc Track

Overview of the INEX 2010 Ad Hoc Track ........................................ 1  
  *Paavo Arvola, Shlomo Geva, Jaap Kamps, Ralf Schenkel, Andrew Trotman, and Johanna Vainio*

The Potential Benefit of Focused Retrieval in Relevant-in-Context Task .................................................. 33  
  *Paavo Arvola and Johanna Vainio*

ENSM-SE and UJM at INEX 2010: Scoring with Proximity and Tag Weights .................................................. 44  
  *Michel Beigbeder, Mathias Géry, Christine Largeron, and Howard Seck*

LIP6 at INEX’10: OWPC for Ad Hoc track ........................................ 54  
  *David Buffoni, Nicolas Usunier, and Patrick Gallinari*

A Useful Method for Producing Competitive Ad Hoc Task Results .................................................. 63  
  *Carolyn J. Crouch, Donald B. Crouch, Sandeep Vadlamudi, Ramakrishna Cherukuri, and Abhijeet Mahule*

Relaxed Global Term Weights for XML Element Search ........................................ 71  
  *Atsushi Keyaki, Kenji Hatano, and Jun Miyazaki*

Searching the Wikipedia with Public Online Search Engines ........................................ 82  
  *Miro Lehtonen*

Extended Language Models for XML Element Retrieval ........................................ 89  
  *Rongmei Li and Theo van der Weide*

### Book Track

Overview of the INEX 2010 Book Track: Scaling Up the Evaluation Using Crowdsourcing ........................................ 98  
  *Gabriella Kazai, Marijn Koolen, Jaap Kamps, Antoine Doucet, and Monica Landoni*

LIA at INEX 2010 Book Track ........................................ 118  
  *Romain Deveaud, Florian Boudin, and Patrice Bellot*

The Book Structure Extraction Competition with the Resurgence Software for Part and Chapter Detection at Caen University ........................................ 128  
  *Emmanuel Giguet and Nadine Lucas*
Focus and Element Length for Book and Wikipedia Retrieval

Jaap Kamps and Marijn Koolen

Combining Page Scores for XML Book Retrieval

Ray R. Larson

OUC’s Participation in the 2010 INEX Book Track

Michael Preminger and Ragnar Nordlie

Data Centric Track

Overview of the INEX 2010 Data Centric Track

Andrew Trotman and Qiuyue Wang

DCU and ISI@INEX 2010: Adhoc and Data-Centric Tracks

Debasis Ganguly, Johannes Leveling, Gareth J.F. Jones, Sauparna Palchowdhury, Sukomal Pal, and Mandar Mitra

Automatically Generating Structured Queries in XML Keyword Search

Felipe da C. Hummel, Altigran S. da Silva, Mirella M. Moro, and Alberto H.F. Laender

UPF at INEX 2010: Towards Query-Type Based Focused Retrieval

Georgina Ramírez

BUAP: A First Approach to the Data-Centric Track of INEX 2010

Darnes Vilariño, David Pinto, Carlos Balderas, Mireya Tovar, and Saul León

Interactive Track

Overview of the INEX 2010 Interactive Track

Nils Pharo, Thomas Beckers, Ragnar Nordlie, and Norbert Fuhr

Using Eye-Tracking for the Evaluation of Interactive Information Retrieval

Thomas Beckers and Dennis Korbar

Link the Wiki Track

Overview of the INEX 2010 Link the Wiki Track

Andrew Trotman, David Alexander, and Shlomo Geva

University of Otago at INEX 2010

Xiang-Fei Jia, David Alexander, Vaughn Wood, and Andrew Trotman
Question Answering Track

Overview of the INEX 2010 Question Answering Track (QA@INEX) .......................................................... 269

Eric SanJuan, Patrice Bellot, Veronique Moriceau, and Xavier Tannier

The GIL Summarizers: Experiments in the Track QA@INEX’10 ................................................................. 282

Edmundo-Pavel Soriano-Morales, Alfonso Medina-Urrea, Gerardo Sierra Martínez, and Carlos-Francisco Méndez-Cruz

The Cortex Automatic Summarization System at the QA@INEX Track 2010 .............................................. 290

Juan-Manuel Torres-Moreno and Michel Gagnon

The REG Summarization System with Question Reformulation at QA@INEX Track 2010 ............................. 295

Jorge Vivaldi, Iria da Cunha, and Javier Ramírez

Relevance Feedback Track

Overview of the INEX 2010 Focused Relevance Feedback Track ................................................................. 303

Timothy Chappell and Shlomo Geva

Exploring Accumulative Query Expansion for Relevance Feedback ............................................................... 313

Debasis Ganguly, Johannes Leveling, and Gareth J.F. Jones

Combining Strategies for XML Retrieval ......................................... 319

Ning Gao, Zhi-Hong Deng, Jia-Jian Jiang, Sheng-Long Lv, and Hang Yu

Web Service Discovery Track

Overview of the INEX 2010 Web Service Discovery Track ................................................................. 332

James A. Thom and Chen Wu

Semantics-Based Web Service Discovery Using Information Retrieval Techniques ............................................. 336

Jun Hou, Jinglan Zhang, Richi Nayak, and Aishwarya Bose

The BUAP Participation at the Web Service Discovery Track of INEX 2010 .............................................. 347

María Josefa Somodevilla, Beatriz Beltrán, David Pinto, Darnes Vilariño, and José Cruz Aaron

XML Retrieval More Efficient Using Double Scoring Scheme ......................................................... 351

Tanakorn Wichaiwong and Chuleerat Jaruskulchai
# XML Mining Track

Overview of the INEX 2010 XML Mining Track: Clustering and Classification of XML Documents ........................................ 363  
*Christopher M. De Vries, Richi Nayak, Sangeetha Kutty, Shlomo Geva, and Andrea Tagarelli*

An Iterative Clustering Method for the XML-Mining Task of the INEX 2010 ........................................................... 377  
*Mireya Tovar, Adrián Cruz, Blanca Vázquez, David Pinto, Darnes Vilariño, and Azucena Montes*

PKU at INEX 2010 XML Mining Track ................................................. 383  
*Songlin Wang, Feng Liang, and Jianwu Yang*

**Author Index** ................................................................................. 397