

# Advances in Database Technology – EDBT 2010

---

13th International Conference  
on Extending Database Technology  
Lausanne, Switzerland, March 22–26, 2010  
Proceedings

Editors:

Ioana Manolescu (INRIA, France)  
Stefano Spaccapietra (EPFL, Switzerland)  
Jens Teubner (ETH Zurich, Switzerland)  
Masaru Kitsuregawa (Tokyo University, Japan)  
Alain Léger (Orange – France Telecom R&D, France)  
Felix Naumann (Hasso Plattner Institute, Germany)  
Anastasia Ailamaki (EPFL, Switzerland)  
Fatma Özcan (IBM Almaden Research Center, USA)

Advances in Database Technology – EDBT 2010  
Proceedings of the 13th International Conference  
on Extending Database Technology  
Lausanne, Switzerland, March 22–26, 2010

Editors:

Ioana Manolescu  
Stefano Spaccapietra  
Jens Teubner  
Masaru Kitsuregawa  
Alain Léger  
Felix Naumann  
Anastasia Ailamaki  
Fatma Özcan

The Association for Computing Machinery  
2 Penn Plaza, Suite 701  
New York, NY, 10121-0701

ACM COPYRIGHT NOTICE. Copyright © 2010 by the Association for Computing Machinery, Inc. Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Publications Dept., ACM, Inc., fax +1 (212) 869-0481, or [permissions@acm.org](mailto:permissions@acm.org).

For other copying of articles that carry a code at the bottom of the first or last page, copying is permitted provided that the per-copy fee indicated in the code is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, +1-978-750-8400, +1-978-750-4470 (fax).

ACM ISBN: 978-1-60558-945-9

# Table of Contents

Foreword .....	ix
Program Committee Members .....	x–xi

## Invited Papers

Provenance for Database Transformations	
<i>Val Tannen</i> .....	1
Scalable Ontology-Based Information Systems	
<i>Ian Horrocks</i> .....	2

## Research Sessions

### XML and Semi-Structured Data

Feedback-driven Result Ranking and Query Refinement for Exploring Semi-structured Data Collections	
<i>Huiping Cao, Yan Qi, K. Selcuk Candan, and Maria Luisa Sapino</i> .....	3–14
Beyond Pages: Supporting Efficient, Scalable Entity Search with Dual-Inversion Index	
<i>Tao Cheng and Kevin Chang</i> .....	15–26
Processing XPath queries with forward and downward axes over XML Streams	
<i>Makoto Onizuka</i> .....	27–38

### OLAP and Decision Support

Correlation Aware Synchronization for Near Real Time Decision Support Systems	
<i>Ying Yan and Wen-Syan Li</i> .....	39–50
Turbo-Charging Hidden Database Samplers with Overflowing Queries and Skew Reduction	
<i>Arjun Dasgupta, Nan Zhang, and Gautam Das</i> .....	51–62
Region-based Online Promotion Analysis	
<i>Tianyi Wu, Yizhou Sun, Cuiping Li, and Jiawei Han</i> .....	63–74

### Distributed Databases

The Data Cyclotron Query Processing Scheme	
<i>Romulo Goncalves and Martin Kersten</i> .....	75–86
Gossiping Personalized Queries	
<i>Xiao Bai, Marin Bertier, Rachid Guerraoui, Anne-Marie Kermarrec, and Vincent Leroy</i> .....	87–98
Optimizing Joins in a Map-Reduce Environment	
<i>Foto Afrati and Jeffrey Ullman</i> .....	99–110

## Security and Privacy 1

Private Record Matching Using Differential Privacy <i>Ali Inan, Murat Kantarcioglu, Gabriel Ghinita, and Elisa Bertino</i> .....	111–122
The Hardness and Approximation Algorithms for L-Diversity <i>Xiaokui Xiao, Ke Yi, and Yufei Tao</i> .....	123–134
K-Symmetry Model for Identity Anonymization in Social Networks <i>Wentao Wu, Yanghua Xiao, Wei Wang, Zhenying He, and Zhihui Wang</i> .....	135–146

## XPath and XQuery Optimization

Let SQL Drive the XQuery Workhorse <i>Torsten Grust, Manuel Mayr, and Jan Rittinger</i> .....	147–158
Efficient Physical Operators for cost-based XPath Execution <i>Haris Georgiadis, Minas Charalambides, and Vasilis Vassalos</i> .....	159–170
Statistics-based Parallelization of XPath Queries in Shared Memory Systems <i>Rajesh Bordawekar, Lipyeow Lim, Anastasios Kementsietsidis, and Bryant Wei-Lun Kok</i> .....	171–182

## Query Processing and Optimization 1

Adaptive Join Processing in Pipelined Plans <i>Kwanchai Eurviriyankul, Norman W. Paton, Alvaro A. A. Fernandes, and Steven Lynden</i> .....	183–194
BSkyTree: Scalable Skyline Computation Using A Balanced Pivot Selection <i>Jongwuk Lee and Seung-won Hwang</i> .....	195–206
Stream Schema: Providing and Exploiting Static Metadata for Data Stream Processing <i>Peter M. Fischer, Kyumars Sheykh Esmaili, and Renée J. Miller</i> .....	207–218

## Optimization for Modern Hardware

Warm Cache Costing – A Feedback Optimization Technique for Buffer Pool Aware Costing <i>Ramanujam Halasipuram S and Edwin Seputis</i> .....	219–227
Position List Word Aligned Hybrid: Optimizing Space and Performance for Compressed Bitmaps <i>François Deliège and Torben Bach Pedersen</i> .....	228–239
Suffix Tree Construction Algorithms on Modern Hardware <i>Dimitris Tsirogiannis and Nick Koudas</i> .....	240–251

## Scientific Databases and OLAP

Reducing Metadata Complexity for Faster Table Summarization <i>K. Selcuk Candan, Mario Cataldi, and Maria Luisa Sapino</i> .....	252–263
Splash: Ad-Hoc Querying of Data and Statistical Models <i>Lujun Fang and Kristen LeFevre</i> .....	264–275
Anchoring Millions of Distinct Reads on the Human Genome within Seconds <i>Tien Huynh, Michail Vlachos, and Isidore Rigoutsos</i> .....	276–286

## Data Provenance

Techniques for Efficiently Querying Scientific Workflow Provenance Graphs <i>Manish Anand, Shawn Bowers, and Bertram Ludaescher</i> .....	287–298
Fine-grained and efficient lineage querying of collection-based workflow provenance <i>Paolo Missier, Norman W. Paton, and Khalid Belhajjame</i> .....	299–310
Lost Source Provenance <i>Jing Zhang and H.V. Jagadish</i> .....	311–322

## Probabilistic and Spatial Databases

Bridging the Gap Between Intensional and Extensional Query Evaluation in Probabilistic Databases <i>Abhay Jha, Dan Olteanu, and Dan Suciu</i> .....	323–334
Probabilistic Path Queries in Road Networks: Traffic Uncertainty Aware Path Selection <i>Ming Hua and Jian Pei</i> .....	335–346
Probabilistic Threshold k Nearest Neighbor Queries over Moving Objects in Symbolic Indoor Space <i>Bin Yang, Hua Lu, and Christian S. Jensen</i> .....	347–358

## Query Processing and Optimization 2

A Simple (yet Powerful) Algebra for Pervasive Environments <i>Yann Gripay, Frédérique Laforest, and Jean-Marc Petit</i> .....	359–370
Self-selecting, self-tuning, incrementally optimized indexes <i>Goetz Graefe and Harumi Kuno</i> .....	371–381
Minimizing Database Repros using Language Grammars <i>Nicolas Bruno</i> .....	382–393

## Spatial Databases

Efficient and Scalable Multi-Geography Route Planning <i>Vidhya Balasubramanian, Dmitri Kalashnikov, Sharad Mehrotra, and Nalini Venkatasubramanian</i> .....	394–405
Querying Trajectories Using Flexible Patterns <i>Marcos R Vieira, Petko Bakalov, and Vassilis J. Tsotras</i> .....	406–417
Querying Spatial Patterns <i>Vishwakarma Singh, Arnab Bhattacharya, and Ambuj K. Singh</i> .....	418–429

## Technologies for the Web

Indexing Relations on the Web <i>Sergio Mergen, Juliana Freire, and Carlos Heuser</i> .....	430–440
An Execution Environment for C-SPARQL Queries <i>Davide Francesco Barbieri, Daniele Braga, Stefano Ceri, and Michael Grossniklaus</i> .....	441–452
Rewrite Techniques for Performance Optimization of Schema Matching Processes <i>Eric Peukert, Henrike Berthold, and Erhard Rahm</i> .....	453–464

## Ranking and Nearest Neighbor

Fast Computation of SimRank for Static and Dynamic Information Networks <i>Cuiping Li, Jiawei Han, Guoming He, Xin Jin, Yizhou Sun, Yintao Yu, and Tianyi Wu</i> .....	465–476
Probabilistic Ranking over Relations <i>Lijun Chang, Jeffrey Xu Yu, Lu Qin, and Xuemin Lin</i> .....	477–488
Privacy Preserving Group Nearest Neighbor Queries <i>Tanzima Hashem, Lars Kulik, and Rui Zhang</i> .....	489–500

## Data Cleaning and Curation

Subsumption and Complementation as Data Fusion Operators <i>Jens Bleiholder, Sascha Szott, Melanie Herschel, Frank Kaufer, and Felix Naumann</i> .....	501–512
HARRA: Fast Iterative Hashed Record Linkage for Large-Scale Data Collections <i>Hung-sik Kim and Dongwon Lee</i> .....	513–524
Finding Misplaced Items in Retail by Clustering RFID Data <i>Leonardo Weiss Ferreira Chaves, Erik Buchmann, and Klemens Böhm</i> .....	525–536

## XML Keyword Search

Keyword Search for Data-Centric XML Collections with Long Text Fields <i>Arash Termehchy and Marianne Winslett</i> .....	537–548
Fast ELCA Computation for Keyword Queries on XML Data <i>Rui Zhou, Chengfei Liu, and Jianxin Li</i> .....	549–560
Suggestion of Promising Result Types for XML Keyword Search <i>Jianxin Li, Chengfei Liu, Rui Zhou, and Wei Wang</i> .....	561–572

## Personalization and Preferences

Feedback-Based Annotation, Selection and Refinement of Schema Mappings for Dataspaces <i>Khalid Belhajjame, Norman W. Paton, Suzanne Embury, Alvaro A. A. Fernandes, and Cornelia Hedeler</i> .....	573–584
PerK: Personalized Keyword Search in Relational Databases through Preferences <i>Kostas Stefanidis, Marina Drosou, and Evaggelia Pitoura</i> .....	585–596
Efficient Computation of Trade-Off Skylines <i>Christoph Lofi, Ulrich Güntzer, and Wolf-Tilo Balke</i> .....	597–608

## Security and Privacy 2

How to Authenticate Graphs Without Leaking <i>Ashish Kundu and Elisa Bertino</i> .....	609–620
Trustworthy Vacuuming and Litigation Holds in Long-term High-integrity Records Retention <i>Ragib Hasan and Marianne Winslett</i> .....	621–632
Algorithm-safe Privacy-Preserving Data Publishing <i>Xin Jin, Nan Zhang, and Gautam Das</i> .....	633–644

## Industrial Sessions

### Transactions and Distribution

BronzeGate: Real-time Transactional Data Obfuscation for GoldenGate <i>Shenoda Guirguis and Alok Pareek</i> .....	645–650
Logging Last Resource Optimization for Distributed Transactions in Oracle WebLogic Server <i>Tom Barnes, Adam Messinger, Paul Parkinson, Amit Ganesh, German Shegalov, Saraswathy Narayan, and Srinivas Kareenhalli</i> .....	651–656
DEDUCE: At the Intersection of MapReduce and Stream Processing <i>Vibhore Kumar, Henrique Andrade, Bugra Gedik, and Kun-Lung Wu</i> .....	657–662

### New Applications

An Experimental Study of Time-Constrained Aggregate Queries <i>Ying Hu, Wen-Chi Hou, Seema Sundara, and Jagannathan Srinivasan</i> .....	663–668
Xbase: Cloud-enabled Information Appliance for Healthcare <i>Wen-Syan Li, Jianfeng Yan, Ying Yan, and Jin Zhang</i> .....	669–674

### Data Warehousing and Analytics

Aggregation of asynchronous electric power consumption time series knowing the integral <i>Raja Chiky, Laurent Decreusefond, and Georges Hebrail</i> .....	675–680
A Plan for OLAP <i>Bernhard Jaecksch, Franz Faerber, and Wolfgang Lehner</i> .....	681–686
Augmenting OLAP Exploration with Dynamic Advanced Analytics <i>Benjamin Leonhardi, Bernhard Mitschang, Ruben Pulido de los Reyes, Christoph Sieb, and Michael Wurst</i> .....	687–692

## Demonstrations

### Demonstrations

Advanced Knowledge Discovery on Movement Data with the GeoPKDD system <i>Mirco Nanni, Roberto Trasarti, Chiara Renso, Fosca Giannotti, and Dino Pedreschi</i> .....	693–696
Timely YAGO: Harvesting, Querying, and Visualizing Temporal Knowledge from Wikipedia <i>Yafang Wang, mingjie zhu, Lizhen Qu, Marc Spaniol, and Gerhard Weikum</i> .....	697–700
PARINDA: An Interactive Physical Designer for PostgreSQL <i>Cristina Maier, Debabrata Dash, Ioannis Alagiannis, Anastasia Ailamaki, and Thomas Heinis</i> .....	701–704
BIAEditor - Matching Process and Operational Data for a Business Impact Analysis <i>Sylvia Radeschütz, Florian Niedermann, and Wolfgang Bischoff</i> .....	705–708
Pattern Detector: Fast Detection of Suspicious Stream Patterns for Immediate Reaction <i>Ira Assent, Hardy Kremer, Stephan Günemann, and Thomas Seidl</i> .....	709–712
BP-Ex – A uniform query engine for Business Process Execution traces <i>Eran Balan, Tova Milo, and Tal Sterenzy</i> .....	713–716

B-Fabric: The Swiss Army Knife for Life Sciences <i>Can Türker, Fuat Akal, Dieter Joho, Ralph Schlapbach, Christian Panse, Simon Barkow-Oesterreicher, and Hubert Rehrauer</i> .....	717–720
---	---------

## Tutorials

FPGAs: A New Point in the Database Design Space <i>Rene Mueller and Jens Teubner</i> .....	721–723
Querying the Deep Web <i>Andrea Calì and Davide Martinenghi</i> .....	724–727
Author Index .....	728–730



# Foreword

The papers in this volume were presented at the 13th International Conference on Extending Database Technology (EDBT 2010), held in Lausanne, Switzerland, March 22–26, 2010. Starting last year, EDBT is held jointly with the ICDT (International Conference on Database Theory) conference.

This year, the conference research track received a total of 307 submissions, out of which 54 were accepted. This amounts to an acceptance rate of about 18 %, confirming the reputation of the EDBT conference as being quite selective! The program committee consisted of 75 members, leading to 12–13 articles to be reviewed by each PC member. While such load is not per se inconsistent with previous editions of EDBT or other prestigious conferences such as, *e.g.*, SIGMOD or VLDB, some reviewers found this quite a heavy load, and significant care had to be taken to ensure appropriate external reviewers or load re-balancing. The submissions were thoroughly discussed, typically as soon as three reviews were available for the paper, even if this was before the review submission deadline; overall, 444 comments were exchanged through the electronic submission tool (EasyChair). The chair is particularly grateful to the PC members which have taken the time to carefully consider papers and debate each other’s arguments. Given the high number of interesting submissions, the PC has worked with the chair to select the best submissions, and did not consider issues such as the overall balance of various database research areas in the conference program. Thus, the proceedings you are holding now reflect the best and most interesting works submitted to our conference, and you can consider the way they split over various topics as a snapshot of the state of our research community as of September 2009.

The conference program is beautifully complemented by the other tracks. Ian Horrocks and Val Tannen have gracefully accepted our invitation to deliver keynote talks at the conference. Alain Léger and Masaru Kitsuregawa have put together an exciting industrial program. Fatma Özcan has chaired the demonstration track, while Felix Naumann has organized the selection of two interesting tutorials. Natassa Ailamaki has helped add to the conference a set of workshops on timely issues. The chairs all join together in acknowledging the careful work of their program committees and external reviewers, and thanking them for their effort.

Standing quite proud of our selection, we hope you will enjoy the program as much as we enjoyed putting it together, and are looking forward to an exciting conference in Lausanne!

Ioana Manolescu (Program Chair)  
Stefano Spaccapietra (General Chair)  
Jens Teubner (Proceedings Chair)  
Masaru Kitsuregawa (Industr. & Appl. Co-Chair)

Alain Léger (Industr. & Appl. Co-Chair)  
Felix Naumann (Tutorial Chair)  
Anastasia Ailamaki (Workshop Chair)  
Fatma Özcan (Demonstrations Chair)

# Program Committee Members

## Research

Amr El Abbadi	Torsten Grust	Jaroslav Pokorny
Gustavo Alonso	Vagelis Hristidis	Ravishankar Ramamurthy
Denilson Barbosa	Wynne Hsu	Philippe Rigaux
Klemens Boehm	Ihab Ilyas	Tore Risch
Angela Bonifati	Yaron Kanza	Domenico Sacca
Philippe Bonnet	Alfons Kemper	Mohand Said-Hacid
Luc Bouganim	Eamonn Keogh	Pierangela Samarati
Loreto Bravo	Martin Kersten	Thomas Seidl
Luca Cabibbo	Alexandros Labrinidis	Kyuseok Shim
Diego Calvanese	Philippe Lamarre	Altigran Da Silva
Alessandro Campi	Paul Larson	Divesh Srivastava
Malu Castellanos	Bertram Ludascher	Dan Suciu
Dario Colazzo	Stefan Manegold	Jens Teubner
Emiran Curtmola	Ioana Manolescu (chair)	Peter Triantafillou
AnHai Doan	Amélie Marian	Patrick Valduriez
Anne Doucet	Volker Markl	Maurice Van Keulen
Cedric DuMouza	Giansalvatore Mecca	Vassilis Vassalos
Piero Fraternali	Paolo Missier	Michalis Vazirgiannis
Christine Froidevaux	Mirella Moro	Jef Wijsen
Irini Fundulaki	Wolfgang Nejdl	Haruo Yokota
Ariel Fuxman	Raymond Ng	Masatoshi Yoshikawa
Helena Galhardas	Matthias Nicola	Jeffrey Yu Xu
Floris Geerts	Boris Novikov	Aoying Zhou
Aristides Gionis	Nicola Onose	
Lukasz Golab	Michalis Petropoulos	
Daniela Grigori	Neoklis Polyzotis	

## Industrial & Applications

Soren Auer	Alain Léger (chair)	Carsten Saathoff
Francois Bry	Wen-Syan Li	Tomas Vitvar
Orri Erling	Anne Monceaux	Alexander Wahler
Fabien Gandon	Miyuki Nakano	Haofen Wang
German Herrero	Massimo Paolucci	Kun-Lung Wu
Ruben Lara Hernandez	Dimitris Plexousakis	Yiannis Kompatsiaris
Masaru Kitsuregawa (chair)	Steffen Saab	

## Demonstrations

Laurent Amsaleg	Ohad Greenshpan	Stelios Paparizos
Christof Bornhoevd	Kristen LeFevre	Evaggelia Pitoura
Bjorn Tor Jonsson	Qiong Luo	Pinar Senkul
Hakan Hacigumus	Fatma Özcan (chair)	Alkis Simitsis

David Simmen  
Nesime Tatbul

Zografoula Vagenas  
Yannis Velegarakis  
Stratis Viglas

Cong Yu

# Provenance for Database Transformations

Val Tannen  
University of Pennsylvania, USA  
val@cis.upenn.edu

## ABSTRACT

Database transformations (queries, views, mappings) take apart, filter, and recombine source data in order to populate warehouses, materialize views, and provide inputs to analysis tools. As they do so, applications often need to track the relationship between parts and pieces of the sources and parts and pieces of the transformations' output. This relationship is what we call database provenance.

This talk presents an approach to database provenance that is based on two observations. First, provenance is a kind of annotation, and we can develop a general approach to annotation propagation that also covers other applications, for example to uncertainty and access control. In fact, provenance turns out to be the most general kind of such annotation, in a precise and practically useful sense. Second, the propagation of annotation through a broad class

of transformations relies on just two operations: one when annotations are jointly used and one when they are used alternatively. This leads to annotations forming a specific algebraic structure, a commutative semiring.

The semiring approach works for annotating tuples, field values and attributes in standard relations, in nested relations (complex values), and for annotating nodes in (un-ordered) XML. It works for transformations expressed in the positive fragment of relational algebra, nested relational calculus, unordered XQuery, as well as for Datalog, GLAV schema mappings, and tgdc constraints. Specific semirings correspond to earlier approaches to provenance, while others correspond to forms of uncertainty, trust, cost, and access control.

This is joint work with J.N. Foster, T.J. Green, Z. Ives, and G. Karvounarakis, done in part within the frameworks of the Orchestra and pPOD projects.

# Scalable Ontology-Based Information Systems

Ian Horrocks  
University of Oxford, UK  
ian.horrocks@comlab.ox.ac.uk

## ABSTRACT

Ontologies and ontology based systems are becoming increasingly important in meeting the demand for more powerful and flexible information systems. Requirements for such systems include the need to deal with incomplete and semi-structured information, to integrate information from heterogeneous sources, to employ richer and more flexible schemas, and for query answers to reflect both knowledge

and data. Provision of such enhanced capabilities must, however, be in addition to, and not instead of, the well-established features of existing database systems, in particular their robust scalability. Achieving this is, of course, extremely challenging. In this talk I will present some recent research efforts that tackle this problem, including investigations of tractable fragments, new algorithmic techniques, new optimisations and the exploitation of relational database technology.