Lecture Notes in Computer Science 2590
Edited by G. Goos, J. Hartmanis, and J. van Leeuwen

Springer
Berlin
Heidelberg
New York
Barcelona
Hong Kong
London
Milan
Paris
Tokyo

Stéphane Bressan Akmal B. Chaudhri
Mong Li Lee Jeffrey Xu Yu
Zoé Lacroix (Eds.)

Efficiency and Effectiveness of XML Tools and Techniques and Data Integration over the Web

VLDB 2002 Workshop EEXTT
and CAiSE 2002 Workshop DIWeb
Revised Papers

Springer

Volume Editors

Stéphane Bressan
Mong Li Lee
School of Computing, NUS, Department of Computer Science
3 Science Drive 2, Singapore 117543, Republic of Singapore
E-mail:{steph/leeml}@comp.nus.edu.sg

Akmal B. Chaudhri
IBM developerWorks
6 New Square Bedfont Lakes, Feltham, Middlesex TW14 8HA, UK
E-mail: akmal.b.chaudhri@uk.ibm.com

Jeffrey Xu Yu
Chinese University of Hong Kong
Dept. of Systems Engineering and Engineering Management
Shatin, New Territories, Hong Kong
E-mail: yu@se.cuhk.edu.hk

Zoé Lacroix
Arizona State University, Mechanical Aerospace Engineering
P.O. Box 876106, Tempe, AZ 85287-6106, USA
E-mail: zoe.lacroix@asu.edu

Cataloging-in-Publication Data applied for

A catalog record for this book is available from the Library of Congress

Bibliographic information published by Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie;
detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>.

CR Subject Classification (1998): H.4, H.2, H.3

ISSN 0302-9743
ISBN 3-540-00736-9 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag Berlin Heidelberg New York
a member of BertelsmannSpringer Science+Business Media GmbH

http://www.springer.de

Typesetting: Camera-ready by author, data conversion by DA-TeX Gerd Blumenstein
Printed on acid-free paper SPIN 10872475 06/3142 5 4 3 2 1 0

Preface

This volume comprises papers from the 1st VLDB Workshop on Efficiency and Effectiveness of XML Tools and Techniques (EEXTT 2002) and selected papers from the 2nd Workshop on Data Integration over the Web (DIWeb 2002).

Efficiency and Effectiveness of XML Tools and Techniques (EEXTT)

As XML emerges as the standard for data representation and interexchange on the World-Wide Web, a variety of XML Management Systems (XMLMS) differing widely in terms of expressive power and performance are becoming available. The majority of these systems are legacy systems, that is, relational database systems which have been extended to load, query, and publish data in XML format. A few are native XMLMS and capture almost all the characteristics of XML data representation. A large number of new techniques are being devised for the management of XML data. The EEXTT workshop focused on the evaluation of the efficiency, effectiveness, and performance of XML management systems, tools, and techniques.

This first workshop was organized in conjunction with the 28th International Conference on Very Large Data Bases, held in Hong Kong, China. The workshop Call for Papers (CFP) solicited contributions covering:

- Storage of XML data
- Compression of XML data
- Security for XML data
- Generation of XML data from legacy applications
- Indexing and retrieval of XML data
- XML query languages
- Query processing over XML data
- Interchange and integration of XML data
- Benchmarks for the management of XML data

After peer review, nine papers were selected for presentation. These were grouped into three main areas: languages, modeling and integration, and storage. A special invited session on benchmarking XML was also organized. This included papers presented by representatives from major research groups around the world working on performance benchmarks for XML tools and applications.

For this LNCS volume, all papers were revised based upon reviewer feedback and questions and issues raised during the workshop.

Workshop Program Committee:

Zohra Bellahsène	LIRMM (France)
Elisa Bertino	University of Milan (Italy)
Timo Böhme	University of Leipzig (Germany)
Philippe Bonnet	University of Copenhagen (Denmark)
Stéphane Bressan	National University of Singapore (Singapore)
Akmal B. Chaudhri	IBM developerWorks (USA)
Gillian Dobbie	University of Auckland (New Zealand)
Wynne Hsu	National University of Singapore (Singapore)
Tok Wee Hyong	Singapore Telecommunications (Singapore)
Zoé Lacroix	Arizona State University (USA)
Mong Li Lee	National University of Singapore (Singapore)
Ioana Manolescu	INRIA (France)
Ullas Nambiar	, Arizona State University (USA)
Michael Rys	Microsoft Corporation (USA)
Zahir Tari	RMIT University (Australia)
Jeffrey Xu Yu	Chinese University of Hong Kong (Hong Kong)

Data Integration over the Web (DIWeb)

Many approaches have been developed in the past to address data integration over distributed heterogeneous databases. However, the exploitation of Web data raises multiple new challenging issues. Mediation or multidatabase systems integrate many data sources generally using the notion of views to make them accessible through an integrated schema by a unique system. Most of these approaches are designed to integrate databases not Web data. But is the Web a database? Are the data sources available on the Web databases? Unlike data stored in databases, Web data is semistructured and its structure is both explicit and implicit. Access to Web data is performed through rather limited access, such as cgi forms, browsing, and extraction, as well as other applications such as search engines. The limitation and the variety of accesses to Web data further distinguish the Web from a database. In addition, access to Web data is costly, uncertain and unreliable when compared to database systems.

The 1st International Workshop on Data Integration over the Web, organized by Zohra Bellahsène, was held in conjunction with the International Conference on Advanced Information Systems Engineering (CAiSE) in Interlaken, Switzerland, in June 2001. In May 2002, DIWeb was again held in conjunction with CAiSE and took place in Toronto, Canada. The second DIWeb focused on XML, complex applications, and the Semantic Web. An invited talk by Alberto Mendelzon (University of Toronto), presented Tox, an XML management system. Several contributions covered various uses of XML, such as an XML data exchange language, a case tool to design XML views, and a caching technique to optimize XPath execution. Three papers addressed issues related to the integration of the Web for specific applications: scientific data, printed and digital data, and geographical data. Issues related to the Semantic Web were addressed in four

papers. Presented approaches included exploiting Web services, source capabilities, and a variety of explicit and implicit knowledge for Web sites and schema mapping. The workshop concluded with a discussion animated by Tamer Özsu (University of Waterloo), Joe Wigglesworth (Centre for Advanced Studies, IBM Toronto Lab), Louiqa Raschid (University of Maryland), and Anthony Tomasic (XML Media) during a panel entitled "Using Standards for Data Integration over the Web."

DIWeb 2002 lasted an entire day including an invited talk, ten presentations of scientific papers (accepted after peer-review), an invited paper, and a final panel. Five papers were selected to be included in this issue of Lecture Notes in Computer Science.

Workshop Program Committee:

Bernd Amann	INRIA (France)
Omar Boucelma	Université de Provence (France)
Stéphane Bressan	National University of Singapore (Singapore)
Subbarao Kambhampati	Arizona State University (USA)
Brigitte Kerhervé	Université du Québec (Canada)
Mong Li Lee	National University of Singapore (Singapore)
Bertram Ludäscher	San Diego Supercomputer Center (USA)
Alberto Mendelzon	University of Toronto (Canada)
Vincent Oria	NJ Institute of Technology (USA)
Louiqa Raschid	University of Maryland (USA)
Mark Roantree	Dublin City University (Ireland)
Shalom Tsur	BEA Systems (USA)

Thanks also to the following additional reviewers: Ilkay Altintas, Chaitan Baru, Ajay Hemnani, Amarnath Gupta, Zaiqing Nie, Ullas Nambiar, Eli Rohn, and Sreelakshmi Vaddi.

The organizers of both workshops would like to thank all program committee members, as well as all external referees, for their excellent work in evaluating the submitted papers. We are also very grateful to the organizers of VLDB and CAiSE for their help and support. Finally, our sincere thanks to Mr. Hofmann at Springer-Verlag for his guidance and enthusiasm in putting this volume together.

London, December 2002

Stéphane Bressan, Akmal B. Chaudhri,
Mong Li Lee, Jeffrey Xu Yu
(Program Chair and Co-chairs EEXTT)
Zoé Lacroix
(Program Co-chair DIWeb)

Table of Contents

Efficiency and Effectiveness of XML Tools and Techniques (EEXTT)

XML Languages

XML Modeling and Integration

XML Storage

Benchmarking XML

Data Integration over the Web (DIWeb)