

# Data Integration over the Web 2004

## Preface

Zohra Bellehène, LIRMM, Univ. Montpellier II  
Peter McBrien, Dept. Computing, Imperial College London

This is the third international workshop on *data integration over the web* (DIWeb), following on from those held in Interlaken (2001) and Toronto (2002), and all held in conjunction with the CAiSE conference in those years. The web sites for all DIWeb workshops may be reached via <http://www.doc.ic.ac.uk/~pjm/diweb/>.

The aim of this workshop is to provide a forum for articles that address the challenges in the design, maintenance and implementation of web based information systems that integrate data from heterogeneous data sources, and the opportunity for all involved to debate new issues and directions for research and development work in the future. The advent of the WWW has dramatically increased the need for efficient and flexible mechanisms to provide integrated views over multiple heterogeneous information sources. The field is facing new challenges raised by the penetration of the Internet to everyone's daily life, and the changes of the economical and financial environment where database systems are used.

Information on the WWW is created and maintained independently by different organizations and people. Thus WWW data sources containing related information may appear at different web sites in different formats, and with different access methods. This requires sophisticated data integration techniques to unify these data sources, and using new network architectures (e.g. P2P) for sharing data in a large scale context. Also, data warehousing of just in-house data may be no longer sufficient to support an organization's information needs. Therefore, warehousing of WWW data and its integration with traditional databases is an important problem.

The workshop attracted 20 papers, and the programme committee accepted five full papers and one short paper. The selected papers focused on the themes of XML, P2P, query processing, ontologies and declarative web site design. To set an overall context to the proceedings, Lenzerini gave an invited talk and paper on the principles of P2P integration, where he developed ideas from his 2002 PODS paper to tackle the challenges of P2P data integration. In the paper by Córcoles and González, an ontology is used to write *descriptions* of XML data sources, which are then used to control the rewriting of queries to execute on data sources. The paper focuses on dealing with the problems of handling *spatial* query operators such as *Area*, *ConvexHull*, *Intersect*, etc.

Karnstedt, Hose, and Sattler are also concerned with controlling the rewriting of queries, but in the context of *schema based* P2P systems with incomplete schema information. They study the use of *routing indexes* (which contain sets of *categories* which describe the schema or extent of schemas), and compare the relative efficiency of *data shipping* and *query shipping*.

Cooper and Davidson are concerned with providing a *component model* for web site design. Their tool allows a specification of entities and attributes to be used to build a relational, XML or hybrid data source, and they provide an algorithm to automate the choice of which model should be used.

In the paper by Zamboulis and Poulouvasilis, an extension to the *both-as-view* (BAV) based data integration system *AutoMed* is proposed to handle XML data. After describing a modelling of XML in AutoMed that does not require a DTD or XML Schema to be provided, the paper describes algorithms to restructure the schemas so that they may be integrated into a global schema, and then materialise that global schema.

Böhme and Rahm deal with the issue of storing XML data efficiently in a relational database, where the XML is supplied with only partial schema information, as is often found in *document centric* XML data. In particular they propose a node identifier numbering scheme that is compact, but reduces the impact of renumbering that other methods suffer from when XML data is *streamed* or inserted into a relational database holding XML data. Bianchini and De Antonellis describe a method for constructing a three layer *ontology* over XML data sources, and then using that ontology for data integration of those data sources.

As editors, we would like to thank the members of the programme committee and all the other referees who gave up their valuable time to review the papers, and helped in putting together an exciting programme. We would also like to thank the invited speaker Maurizio Lenzerini and the authors of papers. Finally, a special thank you to the members of local CAiSE 2004 organisation committee, without whom this workshop would not have been possible.