ICSOFT 2007

Second International Conference on
Software and Data Technologies

Proceedings

BARCELONA, SPAIN • July 22-25, 2007

Volume: ISDM / EHST / DC

INSTICC PRESS

ORGANIZED BY

INSTICC

SPONSORED BY

MEMBER Workflow Management Coalition WfMC

IN COOPERATION WITH

JCREST

ICSOFT 2007

Proceedings of the
Second International Conference on
Software and Data Technologies

Volume ISDM/WsEHST/DC

Barcelona, Spain

July 22 – 25, 2007

Organized by
INSTICC – Institute for Systems and Technologies of Information, Control and Communication

Co-Sponsored by
Workflow Management Coalition – Process Thought Leadership

In Cooperation with
IICREST – Interdisciplinary Institute for Collaboration and Research on Enterprise Systems and Technology

Edited by Joaquim Filipe, Markus Helfert and Boris Shishkov

Printed in Portugal

ISBN: 978-989-8111-07-4

Depósito Legal: 261299/07

http://www.icsoft.org

secretariat@icsoft.org

BRIEF CONTENTS

INVITED SPEAKERS

Jan Dietz

Delft University of Technology

The Netherlands

David Lorge Parnas

University of Limerick

Ireland

Kalle Lyytinnen

Case Western Reserve University

Canada

Stephen Mellor

Australia

Bart Nieuwenhuis

K4B Innovation / University of Twente

The Netherlands

Tony Shan

Bank of America

USA

Brian Fitzgerald

Lero - the Irish Software Engineering Research Centre

Ireland

SPECIAL SESSION CHAIRS

SPECIAL SESSION ON METAMODELLING – UTILIZATION IN SOFTWARE ENGINEERING (MUSE)

Cesar Gonzalez-Perez, Neco, Spain

Brian Henderson-Sellers, University of Technology, Australia

SPECIAL SESSION ON E-HEALTH SERVICES AND TECHNOLOGIES (EHST)

Dimitri Konstantas, University of Geneva, Switzerland

Boris Shishkov, University of Twente, The Netherlands

DOCTORAL CONSORTIUM

DOCTORAL CONSORTIUM CHAIR

Markus Helfert, Dublin City University, Ireland

ORGANIZING AND STEERING COMMITTEES

CONFERENCE CHAIR

Joaquim Filipe, Polytechnic Institute of Setúbal / INSTICC, Portugal

PROGRAM CO-CHAIRS

Markus Helfert, Dublin City University, Ireland

Boris Shishkov, University of Twente, The Netherlands

PROCEEDINGS PRODUCTION

Vera Coelho, INSTICC, Portugal

Andreia Costa, INSTICC, Portugal

Bruno Encarnação, INSTICC, Portugal

Luís Marques, INSTICC, Portugal

Vitor Pedrosa, INSTICC, Portugal

Vera Rosário, INSTICC, Portugal

CD-ROM PRODUCTION

Paulo Brito, INSTICC, Portugal

WEBDESIGNER

Marina Carvalho, INSTICC, Portugal

GRAPHICS PRODUCTION

Helder Coelhas, INSTICC, Portugal

SECRETARIAT AND WEBMASTER

Mónica Saramago, INSTICC, Portugal

PROGRAM COMMITTEE

Jemal Abawajy, Deakin University, Australia

Silvia Abrahão, Valencia University of Technology, Spain

Muhammad Abulaish, Jamia Millia Islamia (A Central University), India

Hamideh Afsarmanesh, University of Amsterdam, The Netherlands

Jacky Akoka, CNAM, France

Rafa Al Qutaish, École de Technologie Supérieure - University of Quebec, Canada

Markus Aleksy, University of Mannheim, Germany

Tsanka Petrova Angelova, Uniccord Ltd., Bulgaria

Keijiro Araki, Kyushu University, Japan

Alex Aravind, University of Northern British Columbia, Canada

Colin Atkinson, University of Mannheim, Germany

Juan Carlos Augusto, University of Ulster at Jordanstown, U.K.

Elisa Baniassad, Chinese University of Hong Kong, China

Luciano Baresi, Politecnico di Milano, Italy

Joseph Barjis, Georgia Southern University, U.S.A.

Bernhard Beckert, University of Koblenz, Germany

Noureddine Belkhatir, University of Grenoble, France

Fevzi Belli, University Paderborn, Germany

Alexandre Bergel, Hasso-Plattner Institut, Germany

Sue Black, University of Westminster, U.K.

Maarten Boasson, Universiteit van Amsterdam, The Netherlands

Wladimir Bodrow, University of Applied Sciences Berlin, Germany

Marcello Bonsangue, University of Leiden, The Netherlands

Pere Botella, Universitat Politecnica de Catalunya, Spain

Lisa Brownsword, Software Engineering Institute, U.S.A.

Gerardo Canfora, University of Sannio, Italy

Cinzia Cappiello, Politecnico di Milano, Italy

Antonio Cerone, United Nations University, China

W. K. Chan, City University of Hong Kong, Hong Kong

Shiping Chen, CSIRO ICT Centre, Australia

T. Y. Chen, Swinburne University of Technology, Australia

Kung Chen, National Chengchi University, Taiwan, Province Of China

Samuel Chong, Accenture, U.K.

Peter Clarke, Florida International University, U.S.A.

Rolland Colette, Université Paris 1 Panthéon Sorbonne, France

Rem Collier, University College Dublin, Ireland

Kendra Cooper, The University of Texas at Dallas, U.S.A.

Alfredo Cuzzocrea, University of Calabria, Italy

Bogdan Czejdo, Loyola University, U.S.A.

Mehdi Dastani, Utrecht University, The Netherlands

Sergio de Cesare, Brunel University, U.K.

Clever de Farias, University of São Paulo, Brazil

Rogerio de Lemos, University of Kent, U.K.

Andrea De Lucia, Università di Salerno, Italy

Serge Demeyer, Universiteit Antwerpen, Belgium

Steven Demurjian, University of Connecticut, U.S.A.

Elisabetta Di Nitto, Politecnico di Milano, Italy

Massimiliano Di Penta, University of Sannio, Italy

Nikolay Diakov, Fredhopper B.V., The Netherlands

Oscar Dieste, Universidad Politécnica de Madrid, Spain

Jan L. G. Dietz, Delft University of Technology, The Netherlands

Jin Song Dong, National University of Singapore, Singapore

Jing Dong, University of Texas at Dallas, U.S.A.

Brian Donnellan, National University of Ireland, Ireland

Juan C. Dueñas, Universidad Politécnica de Madrid, Spain

Jürgen Ebert, Universität Koblenz-Landau, Germany

PROGRAM COMMITTEE (CONT.)

Paul Ezhilchelvan, University of Newcastle, U.K.

Behrouz Far, University of Calgary, Canada

Massimo Felici, The University of Edinburgh, U.K.

Rudolf Ferenc, University of Szeged, Hungary

Juan Fernandez-Ramil, The Open University, U.K.

Bernd Fischer, University of Southampton, U.K.

Gerald Gannod, Miami University, U.S.A.

Jose M. Garrido, Kennesaw State University, U.S.A.

Dragan Gasevic, Athabasca University, Canada

Nikolaos Georgantas, INRIA, France

Paola Giannini, Università del Piemonte Orientale, Italy

John Paul Gibson, Institut National des Télécommunications, France

Holger Giese, University of Paderborn, Germany

Karl Goeschka, Vienna University of Technology, Austria

Swapna Gokhale, University of Connecticut, U.S.A.

Jose Ramon Gonzalez de Mendivil, Universidad Publica de Navarra, Spain

Jesus M. Gonzalez-Barahona, Universidad Rey Juan Carlos, Spain

Daniela Grigori, University of Versailles, France

Klaus Grimm, Daimlerchrysler AG, Germany

Yann-Gaël Guéhéneuc, University of Montreal, Canada

Tibor Gyimothy, University of Szeged, Hungary

Michael Hanus, University of Kiel, Germany

Naohiro Hayashibara, Tokyo Denki University, Japan

Reiko Heckel, University of Leicester, U.K.

Christian Heinlein, University of Ulm, Germany

Markus Helfert, Ireland

Rattikorn Hewett, Texas Tech University, U.S.A.

Jang-Eui Hong, Chungbuk National University, Republic of Korea

Shinichi Honiden, Graduate School of Information Science and Technology,University of Tokyo, Japan

Ilian Ilkov, IBM Nederland B.V., The Netherlands

Ivan Ivanov, State Univesity of New York, Empire State College, U.S.A.

Stephen Jarvis, University of Warwick, U.K.

Damir Kalpic, Faculty of Electrical Engineering and Computing, Croatia

Krishna Kavi, University of North Texas, U.S.A.

Taghi Khoshgoftaar, Florida Atlantic University, U.S.A.

Roger (Buzz) King, University of Colorado, U.S.A.

Paul Klint, Centrum voor Wiskunde en Informatica, The Netherlands

Alexander Knapp, Ludwig-Maximilians-Universität München, Germany

Mieczyslaw Kokar, Northeastern University, U.S.A.

Rainer Koschke, University of Bremen, Germany

Jens Krinke, FernUniversität in Hagen, Germany

Padmanabhan Krishnan, Bond University, Australia

Martin Kropp, University of Applied Sciences Northwestern Switzerland, Switzerland

Tei-Wei Kuo, National Taiwan University, Taiwan, Province pf China

Yvan Labiche, Carleton University, Canada

Michele Lanza, University of Lugano, Switzerland

Eitel Lauria, Marist College, U.S.A.

Insup Lee, University of Pennsylvania, U.S.A.

Jonathan Lee, National Central University, Taiwan, Province of China

Yu Lei, The University of Texas at Arlington, U.S.A.

Hareton Leung, Hong Kong Polytechnic University, Hong Kong

Kuan-Ching Li, Providence University, Taiwan, Province of China

Man Lin, St. Francis Xavier University, Canada

Panos Linos, Butler University, U.S.A.

Hua Liu, Xerox Corp., U.S.A.

Chengfei Liu, Swinburne University of Technology, Australia

David Lorenz, University of Virginia, U.S.A.

Christof Lutteroth, University of Auckland, New Zealand

PROGRAM COMMITTEE (CONT.)

Jianhua Ma, Hosei University, Japan

Broy Manfred, Technische Universität München, Germany

Tiziana Margaria, University Potsdam, Germany

Katsuhisa Maruyama, Ritsumeikan University, Japan

Johannes Mayer, Ulm University, Germany

Tommaso Mazza, University Magna Græcia of Catanzaro, Italy

Fergal McCaffery, University of Limerick, Ireland

Hamid Mcheick, University of Quebec at Chicoutimi, Canada

Massimo Mecella, SAPIENZA - Università di Roma, Italy

Karl Meinke, Royal Institute of Technology, Sweden

Simão Melo de Sousa, Universidade da Beira Interior (UBI), Portugal

Emilia Mendes, The University of Auckland, New Zealand

Manoel Mendonça, Salvador University, Brazil

Raffaela Mirandola, Politecnico di Milano, Italy

Hristo Mirkov, MorganStanley, U.S.A.

Prasenjit Mitra, Pennsylvania State University, U.S.A.

Dimitris Mitrakos, Aristotle University of Thessaloniki, Greece

Birger Møller-Pedersen, University of Oslo, Norway

Mattia Monga, Università degli Studi di Milano, Italy

Sandro Morasca, Università degli Studi dell'Insubria, Italy

Maurizio Morisio, Politecnico di Torino, Italy

Markus Müller-Olm, WWU Münster, Germany

Paolo Nesi, University of Florence, Italy

Alan O'Callaghan, De Montfort University, U.K.

Rory O'Connor, Dublin City University, Ireland

Pasi Ojala, Nokia, Finland

Claus Pahl, Dublin City University, Ireland

Witold Pedrycz, University of Alberta, Canada

Steve Peters, Vrije Universiteit Amsterdam, The Netherlands

Mario Piattini, University of Castilla-La Mancha, Spain

Martin Pinzger, University of Zurich, Switzerland

Lori Pollock, University of Delaware, U.S.A.

Andreas Polze, Hasso-Plattner-Institute for Software Engineering at University Potsdam, Germany

Peter Popov, City University, U.K.

Wenny Rahayu, La Trobe University Australia, Australia

Jolita Ralyte, University of Geneva, Switzerland

Anders P. Ravn, Aalborg University, Denmark

Marek Reformat, University of Alberta, Canada

Arend Rensink, University of Twente, The Netherlands

Werner Retschitzegger, Johannes Kepler University Linz, Austria

Gustavo Rossi, LIFIA, Argentina

Guenther Ruhe, University of Calgary, Canada

Stefano Russo, Università di Napoli Federico II, Italy

Mortaza S. Bargh, Telematica Instituut, The Netherlands

Shazia Sadiq, University of Queensland, Australia

Francesca Saglietti, University of Erlangen-Nuremberg, Germany

Bernhard Schätz, TU München, Germany

Douglas Schmidt, Vanderbilt University, U.S.A.

Andy Schürr, Darmstadt University of Technology, Germany

Isabel Seruca, Universidade Portucalense, Portugal

Samir Shah, Penn State University, U.S.A.

Boris Shishkov, University of Twente, The Netherlands

Harvey Siy, University of Nebraska at Omaha, U.S.A.

Jacob Slonim, Dalhousie University, Canada

George Spanoudakis, City University, U.K.

Peter Stanchev, Kettering University, U.S.A.

Nenad Stankovic, University of Aizu, Japan

Larry Stapleton, ISOL Research Centre, Ireland

PROGRAM COMMITTEE (CONT.)

Richard Starmans, Utrecht University, The Netherlands

Leon Sterling, University of Melbourne, Australia

Junichi Suzuki, University of Massachusetts, Boston, U.S.A.

Ramayah T., Universiti Sains Malaysia, Malaysia

Yarar Tonta, Hacettepe University, Turkey

Mark van den Brand, Technical University of Eindhoven, The Netherlands

Marten van Sinderen, University of Twente, The Netherlands

Enrico Vicario, University of Florence, Italy

Aurora Vizcaino, University of Castilla-La Mancha, Spain

Christoph von Praun, IBM Research, U.S.A.

Christiane Gresse von Wangenheim, UNIVALI, Brazil

Bing Wang, University of Hull, U.K.

Edgar Weippl, Secure Business Austria, Austria

Danny Weyns, Katholieke Universiteit Leuven, Belgium

Ing Widya, University of Twente, The Netherlands

Dietmar Wikarski, Fachhochschule Brandenburg - University of Applied Sciences, Germany

Hongwei Xi, Boston University, U.S.A.

Haiping Xu, University of Massachusetts Dartmouth, U.S.A.

Hongji Yang, De Montfort University, U.K.

Tuba Yavuz-Kahveci, University of Florida, U.S.A.

Rym Zalila Mili, University of Texas at Dallas, U.S.A.

Kang Zhang, University of Texas at Dallas, U.S.A.

Du Zhang, California State University, U.S.A.

Xiaokun Zhang, Athabasca University, Canada

Jianjun Zhao, Shanghai Jiao Tong University, China

Hong Zhu, Oxford Brookes University, U.K.

Andrea Zisman, City University, U.K.

AUXILIARY REVIEWERS

Jonatan Alava, Florida International University, U.S.A.

David Arney, University of Pennsylvania, U.S.A.

Louise Avila, University of Pennsylvania, U.S.A.

Djuradj Babich, Florida International University, U.S.A.

Tibor Bakota, University of Szeged, Hungary

Nurlida Basir, University of Southampton, U.K.

Massimo Canonico, Università del Piemonte Orientale, Italy

Glauco Carneiro, Salvador University (UNIFACS), Brazil

Su-Ying Chang, Department of Computer Science and Information Engineering, National Taiwan, Taiwan

Shih-Chun Chou, Department of Computer Science and Information Engineering, National Taiwan, Taiwan

Daniela Cruzes, State University of Campinas (UNICAMP), Brazil

Marco D'Ambros, University of Lugano, Switzerland

Florian Deissenböck, TU Muenchen, Germany

Daniele Theseider Duprè, Università del Piemonte Orientale, Italy

Lavinia Egidi, Università del Piemonte Orientale, Italy

Ekaterina Ermilove, University of Amsterdam, The Netherlands

Hua-Wei Fang, Department of Computer Science and Information Engineering, National Taiwan, Taiwan

Massimo Ficco, CINI Lab "C. Savy", Italy

Christina von Flach, Federal University of Bahia (UFBa), Brazil

Rita Francese, University of Salerno, Italy

Lajos Fulop, University of Szeged, Hungary

Lajos Jenő Fülöp, University of Szeged, Hungary

Udo Gleich, Daimler Chrysler AG, Germany

Leonardo Grassi, University of Florence, Italy

AUXILIARY REVIEWERS (CONT.)

Andreas Griesmayer, United Nations University, Macau SAR China

Ralph Guderlei, Ulm University, Germany

Michael Haupt, Software Architecture Group, Hasso Plattner Institute, Germany

Stefan Henkler, University of Paderborn, Germany

Martin Hirsch, University of Paderborn, Germany

Florian Hoelzl, TU Muenchen, Germany

Bernhard Hohlfeld, Daimler Chrysler AG, Germany

Endre Horváth, University of Szeged, Hungary

Ping-Yi Hsu, Department of Computer Science and Information Engineering, National Taiwan, Taiwan

Judit Jasz, University of Szeged, Hungary

Joop de Jong, Delft University of Technology, The Netherlands

Elrmar Juergens, TU Muenchen, Germany

Madhan Karky, The University of Queensland, Australia

Steven van Kervel, Delft University of Technology, The Netherlands

Tariq M. King, Florida International University, U.S.A.

Peter Lammich, Westfälischen Wilhelms-Univer-sität, Germany

Massimiliano de Leoni, University Roma, Italy

Martin Leucker, TU Muenchen, Germany

Yun-Hao Li, Department of Computer Science and Information Engineering, National Taiwan, Taiwan

Adrian Lienhard, Software Composition Group, University of Bern, Switzerland

Ruopeng Lu, The University of Queensland, Australia

Heng Lu, The University of Hong Kong, Hong Kong

Viviane Malheiros, University of São Paulo (USP), Brazil

Sergio Di Martino, University of Salerno, Italy

Michael Meisinger, TU Muenchen, Germany

Samar Mouchawrab, Carleton University, Canada

Simon S. Msanjila, University of Amsterdam, The Netherlands

Sudarsanan Nesmony, The University of Queensland, Australia

Joseph Okika, Aalborg University, Denmark

Rocco Oliveto, University of Salerno, Italy

Jennie Palmer, University of Newcastle, U.K.

Ignazio Passero, University of Salerno, Italy

Gustavo Perez, University of Southern California (USC), U.S.A.

Christian Pfaller, TU Muenchen, Germany

Roberto Pietrantuono, DIS - Federico II University of Naples, Italy

Dan Ratiu, TU Muenchen, Germany

Giancarlo Ruffo, Università di Torino, Italy

Ruggero Russo, University Roma, Italy

Laís Salvador, Salvador University (UNIFACS), Brazil

Valeriano Sandrucci, University of Florence, Italy

Giuseppe Scanniello, University of Basilicata, Italy

Siraj Shaikh, United Nations University, Macau SAR China

Marwa Shousha, Carleton University, Canada

Istvan Siket, University of Szeged, Hungary

Carine Souveyet, Universite Paris 1, France

Michael Sowka, Carleton University, Canada

Bas Steunebrink, Utrecht University, The Netherlands

Tatiana Tavares, Catholic University of Pelotas (UCPel), Brazil

Matthias Tichy, University of Paderborn, Germany

Carlo Torniai, University of Florence, Italy

Kun-Yi Tsai, Department of Computer Science and Information Engineering, National Taiwan, Taiwan

Laszlo Vidacs, University of Szeged, Hungary

Stefan Wagner, TU Muenchen, Germany

Doris Wild, TU Muenchen, Germany

Tao Yue, Carleton University, Canada

Zhenyu Zhang, The University of Hong Kong, Hong Kong

Xiaohui Zhao, Swinburne University of Technology, Australia

SELECTED PAPERS BOOK

A number of selected papers presented at ICSOFT 2007 will be published by Springer, in a book entitled Software and Data Technologies II. This selection will be done by the conference chair and program co-chairs, among the papers actually presented at the conference, based on a rigorous review by the ICSOFT 2007 program committee members.

CO-SPONSOR

FOREWORD

This volume contains the proceedings of the second *International Conference on Software and Data Technologies* (*ICSOFT 2007*), organized by the Institute for Systems and Technologies of Information, Control and Communication (*INSTICC*) in cooperation with the Interdisciplinary Institute for Collaboration and Research on Enterprise Systems and Technology (IICREST), and co-sponsored by the Workflow Management Coalition (WfMC).

The purpose of this conference is to bring together researchers, engineers and practitioners interested in information technology and software development. The conference tracks are "*Software Engineering*", "*Information Systems and Data Management*", "*Programming Languages*", "*Distributed and Parallel Systems*" and "*Knowledge Engineering*".

Software and data technologies are essential for developing any computer information system, encompassing a large number of research topics and applications: from programming issues to the more abstract theoretical aspects of software engineering; from databases and data-warehouses to management information systems and knowledge-base systems; Distributed systems, ubiquity, data quality and other related topics are included in the scope of ICSOFT.

ICSOFT 2007 received 292 paper submissions from more than 56 countries in all continents. To evaluate each submission, a double blind paper evaluation method was used: each paper was reviewed by at least two internationally known experts from ICSOFT Program Committee. Only 41 papers were selected to be published and presented as full papers, i.e. completed work (8 pages in proceedings / 30' oral presentations), 74 additional papers, describing work-in-progress, were accepted as short paper for 20' oral presentation, leading to a total of 115 oral paper presentations. There were also 76 papers selected for poster presentation. The full-paper acceptance ratio was thus 14%, and the total oral paper acceptance ratio was 39%.

In its program ICSOFT includes panels to discuss aspects of software development, with the participation of distinguished world-class researchers; furthermore, the program is enriched by several keynote lectures delivered by renowned experts in their areas of knowledge. These high points in the conference program definitely contribute to reinforce the overall quality of the ICSOFT conference, which aims at becoming one of the most prestigious yearly events in its area. This year, ICSOFT was held back-to-back with ENASE (Evaluation of Novel Approaches to Software Engineering) working conference, in a joint effort to offer the research community the best possible environment for discussing and debating innovative aspects of Software Engineering. This was quite a rewarding experience, thanks to ENASE program chairs Leszek Maciaszek and Cesar Gonzalez-Perez and all other ENASE participants.

The program for this conference required the dedicated effort of many people. Firstly, we must thank the authors, whose research and development efforts are recorded here. Secondly, we thank the members of the program committee and the additional reviewers for their diligence and expert reviewing. I would like to personally thank the Program Chairs, namely Boris Shishkov and Markus Helfert, for their important collaboration. The local organizers and the secretariat have worked hard to provide smooth logistics and a friendly environment, so we must thank them all and

especially Ms. Monica Saramago for their patience and diligence in answering many emails and solving all the problems. Last but not least, we thank the invited speakers for their invaluable contribution and for taking the time to synthesize and prepare their talks.

A successful conference involves more than paper presentations; it is also a meeting place, where ideas about new research projects and other ventures are discussed and debated. Therefore, a social event including a conference diner was organized for the evening of July 24 (Tuesday) in order to promote this kind of social networking.

We wish you all an exciting conference and an unforgettable stay in the cosmopolitan city of Barcelona. We hope to meet you again next year for the 3rd ICSOFT, to be held in the historic city of Porto (Portugal), details of which will be shortly made available at http://www.icsoft.org.

Joaquim Filipe

INSTICC/Polytechnic Institute of Setúbal, Portugal

(Conference Chair)

CONTENTS

SHORT PAPERS

POSTERS

SPECIAL SESSION ON E-HEALTH SERVICES AND TECHNOLOGIES (EHST)

DOCTORAL CONSORTIUM

DATA QUALITY IN XML DATABASES
A Methodology for Semi-structured Database Design Supporting Data Quality Issues

Eugenio Verbo, Ismael Caballero
Indra Sistemas
UCLM- Indra Research and Development Institute
Ronda de Toledo s/n – 13004 Ciudad Real, Spain
{emverbo, icaballerom}@indra.es

Eduardo Fernandez-Medina, Mario Piattini
ALARCOS Research Group
Information Systems and Technology Departament
UCLM- Indra Research and Development Institute
Paseo de la Universidad 4 s/n – 13071 Ciudad Real, Spain
{Eduardo.FdezMedina, Mario.Piattini}@uclm.es

Keywords: Data Quality, XML databases, design methodology, semi-structured data.

Abstract: As the use of XML as a technology for data exchange has widely spread, the need of a new technology to store semi-structured data in a more efficient way has been emphasized. Consequently, XML DBs have been created in order to store a great amount of XML documents. However, like in previous data models as the relational model, data quality has been frequently left aside. Since data plays a key role in organization efficiency management, its quality should be managed. With the intention of providing a base for data quality management, our proposal address the adaptation of a XML DB development methodology focused on data quality. To do that we have based on some key area processes of a Data Quality Maturity reference model for information management process definition.

1 INTRODUCTION

Since the World Wide Web Consortium (Bray et al., 1998) approved the first edition of the XML standard in 1998, its use has spread up to become the standard *de facto* for data exchange due to the flexibility and richness that XML provides to capture semantic aspects of the application domain.

However, XML can also be used as a data storage technology. A good example of that kind of use would be the OpenDocument standard (OASIS, 2006) for the specification of electronic documents.

Due to the great amount of data that are transmitted in XML format, it is reasonable to think that storing that data directly in DBMS adapted to the characteristics of XML will improve the performance of the retrieval and preparation of data for its trasmission.

On the other hand, since data is one of the main assets that organizations hold (Huang et al., 1999), databases schemas should guarantee the quality of the contained data because the proper working of the information system (IS) could depend, in more or lesser extent, on this feature.

A way to achieve that goal could be to enrich data items with metadata that would serve as a basis to assess data quality according to the selected quality criteria or quality dimension. Doing so, data value quality, i.e. how adequately data values represent real world objects or facts, could be improved.

Although data quality is often associated to data value quality, even completely correct and valid data could be faulty if they are supported by an invalid data model (Levitin and Redman, 1995). Hence the possible solution to those problems could go through integrating quality aspects into the database design process in order to get a resulting product that satisfies the system quality requirements. Thus, data model would be designed in a proper way so that the

number of defects propagated to the final information product would not increase.

Our work comprises those concepts since we propose data quality integration inside some processes of CALDEA, a reference model defined in (Caballero and Piattini, 2007), in order to create a XML DB design methodology with data quality support.

The remainder of the paper is structured as follows: in section 2 the CALDEA reference model is presented, in section 3 our proposal is shown, and, lastly, in section 4 some conclusions are summarized.

2 CALDEA

CALDEA is a Data and Information Quality Management Maturity Model that can be used as a reference for assessing and improving data quality through the *Information Management Process* concept, a specialization of the Software Process (Fuggeta, 2000) for the Information and Data Quality Management. This IMP is composed of two kinds of subprocesses: a) data product fabrication processes (MfP) and b) data quality management processes (MnP), centered in data and information quality.

CALDEA is the reference model for this framework. It is structured in Key Area Processes (KPA) in the same way as CMMI is. Each one of these KPAs can belong to one or both of the two kinds of subprocess previously defined.

Each KPA defines a set of activities which identifies a collection of elements. For each activity, these elements are: a) input and output products, b) techniques and tools, c) workers and d) execution time. The CALDEA KPAs used in the proposed methodology and its corresponding acronyms can be seen in Table 1.

The URM KPA enumerates activities aimed to compile, understand and document user requirements in order to drive the information management process towards the user point of view.

The goal of the DSTM KPA is the identification, definition and characterization of data sources and destinations for the generated information products.

In the AIMPM KPA, it is done the management of the databases and data warehouses of the information system.

In the DIQM KPA, the information and data quality management processes of the IMP are implemented.

Table 1: CALDEA Definition level KPAs.

Acronym	Meaning
URM	User Requirements Management
DSTM	Data Sources and Data Targets Management
AIMPM	Database or Data Warehouse Acquisition, Development or Maintenance Project
DIQM	Data and Information Quality Management in IMP Components

3 XML DATABASE DESIGN METHODOLOGY WITH DATA QUALITY SUPPORT

3.1 Methodology Steps

We have defined a semi-structured DB design methodology divided in the following ten steps:

1. Define user requirements.
2. Define quality user requirements.
3. Design semi-structured DB schema.
4. Identify data quality dimensions of the DB application domain.
5. According to the data quality dimensions identified in the previous step, extend database schema with quality data extension mechanism.
6. Identify data sources.
7. Adapt data input format to the database schema.
8. Define context dependant measures to evaluate input data quality.
9. Establish a threshold for input data quality.
10. Apply quality measures to input data and only store those whose measure results are above the threshold previously defined.

These ten steps are the result of adapting different activities of CALDEA KPAs. In the following sections, we summarize tasks and techniques to be used during the application of the proposed methodology.

3.2 User Requirements Management (URM)

This KPA covers points 1 and 2 of the proposed methodology.

It is composed of a set of activities aimed to collect user requirements specification. Apart from compile traditional user requirements, the definition of user requirements related to data quality is also an important task that should be done.

The importance of this KPA must be highlighted because it is the basis on which the remainder of the effort on data quality will be built in later phases.

3.3 Data Sources and Data Targets Management (DSTM)

This KPA covers points 6 and 7 of the proposed methodology.

If resulting XML DB must satisfy a quality threshold, sources from which data is retrieved must also satisfy that quality threshold. If data is processed before storing it in the XML DB, processes in charge of that task must deal with data quality either maintaining data quality already present in data or improving it through an analysis of data.

If stored data is a product of another information system, it would be recommendable that received data were in a format that allows the integration of the new data with data already stored in the DB and that keeps quality data representation as, for example, the one presented in (Verbo et al., 2007). Following this approach, data sources could be compared and only that ones according to the quality requirements would be used.

3.4 Database Acquisition, Development or Maintenance (DADM)

This KPA covers point 3 of the proposed methodology.

DB development is done during this phase. One of the main tasks is the XML DB conceptual modeling. In the traditional approach, this task is done using E/R and UML diagrams. However, this is not the best choice for XML DBs modeling because that kind of diagrams does not capture all XML semantics like certain sorts of associations or type creation. As a possible solution to this problem, in (Marcos et al., 2001) an UML extension for XML Schema (XSD) representation is proposed.

3.5 Data and Information Quality Management (DIQM)

This KPA covers points 4, 5, 8, 9 and 10 of the proposed methodology.

The goal of this KPA is to determine which information and data quality aspects are involved in the information management process components and which are important to the context being studied. Applying this concept to our discourse domain, it means that after having created the XML DB conceptual model it is necessary to integrate in it all the elements that will help to guarantee the quality for the given solution.

CALDEA define two activities for this process:

1. (DIQM.1) Identify information and data quality dimensions starting from users data quality requirements.
2. (DIQM.2) Identify measures for each information and data quality dimension.

3.5.1 DIQM.1 Data Quality Dimensions

During this phase we must identify most important data quality dimensions for the application domain. Many authors like (Redman, 1996), (English, 1999) and (Strong et al., 1997) have explained how to identify these data quality dimensions and even how to measure certain characteristic data quality aspects in specific application domains and environments.

A major problem is that many of these proposals for data quality dimension selection involve the authors to define a set of dimensions that are valid as a reference for a specific context. Further evaluation of these frameworks reveals too much frequently that they have been defined specifically for a particular domain, which implies that they are highly context dependent (Eppler, 2001).

A possibility could be to develop an universal reference model valid for any context, but as (Lee et al., 2006) affirm, this is highly unlikely due to the fact that information and data quality are tightly related to particular problems that organizations have with their own information and data.

Due to this reason, instead of proposing a concrete set of dimensions to be handled during this KPA, the goal of our proposal is to define a structure that allows to represent quality data in an uniform way and with higher semantic meaning. To reach this objective, we have based on the approach proposed in (Wang et al., 1995), where it is shown an extension of the relational model to represent quality data. Its main contribution consists in using a conceptual data model extended with data quality attributes that store data related to data quality dimensions to improve overall system data quality. In order to get the highest detail as possible and since the relational cell is the minimal storage unit in the relational model, it is necessary to tag data quality at cell level. Tag data quality means that quality data is associated to a cell value.

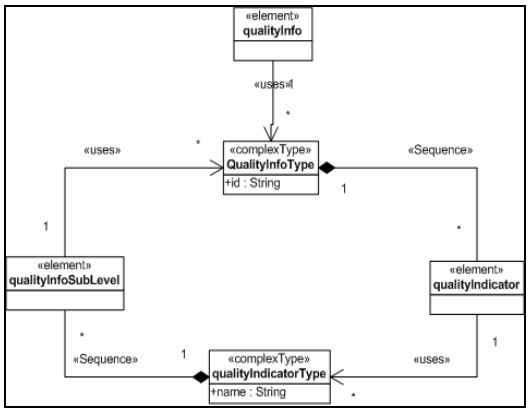

Figure 1: Data quality XSD extension diagram.

That proposal is based on the quality indicator concept formerly explained. A quality indicator gives objective information about certain data characteristics and about its transformation process.

We have developed an extension mechanism to represent quality data in XML DBs. To get that, we have created an XML Schema (see Figure 1 for a graphical representation and Figure 2 for the XSD source code) that includes the following elements:

- **QualityInfo**. This element acts as a grouping section of quality data for an XML DB component. It is optional since a component may not have associated quality attributes.
- **QualityIndicator**. This element contains metadata about quality data. On the one hand, it specifies the value assigned to that quality indicator and, on the other hand, it may contain "qualityInfo" elements, i.e., a quality indicator can have associated quality indicators. For example (see Figure 3), let us suppose a newspaper includes a set of news. Each piece of news has a source associated to it. This source can be a news agency. This would be the first level of quality indicators. In turn, a news agency may have a set of news sources that would correspond to the second level of quality indicators.

This structure allows to enrich the model obtained after the DB conceptual modeling to represent quality data on it. In Figure 3 "qualityInfo" elements are represented in bold font and "qualityIndicator" elements are in italics.

3.5.2 DIQM.2 Measures

Information and data quality dimensions definition represents an important step in the process of deciding which quality aspects and quality criteria are significant to the context of the problem to be resolved. According to the ISO 9126 standard (ISO, 1991), it represents an answer to the problem of identifying which data quality aspects must address a specific component. But this is not enough as there are other problems to solve like, for example, to know how good the studied component with respect to a concrete quality dimension is. To fulfill this answer, it is necessary to define measures, i.e., sets composed of a way of measure and a scale to obtain a value on that dimension (García et al., 2005).

```
<xs:schema xmlns:xs=
"http://www.w3.org/2001/XMLSchema">
  <xs:element
      name="qualityInfo"
      type="qualityInfoType"
      nillable="true"/>
  <xs:complexType
      name="qualityInfoType">
    <xs:sequence>
      <xs:element
        name="qualityIndicator"
        type="qualityIndicatorType"
        minOccurs="1"
        maxOccurs="unbounded"/>
    </xs:sequence>
    <xs:attribute
        name="id"
        type="xs:string"
        use="required"/>
  </xs:complexType>

  <xs:complexType
      name="qualityIndicatorType"
      mixed="true">
    <xs:sequence>
      <xs:element
        name="qualityInfo"
        type="qualityInfoType"
        minOccurs="0"
        maxOccurs="1"/>
    </xs:sequence>
    <xs:attribute
      name="name"
      type="xs:string"
      use="required"/>
  </xs:complexType>
</xs:schema>
```

Figure 2: Data quality XSD extension.

As stated in previous section, data quality dimension selection for a specific component may be highly context dependent. Consequently, quantitatively measure definition can also be highly context dependent. However, this paper tries to give a broad overview of XML DB development so the set of measures we have defined are generic since they can be applied to XML documents independently of the application context in order to optimize the schema design. Those measures try to

give a general understanding about the complexity of the XML documents stored in the XML DB.

```
<news>
  news_content
  <qualityInfo id="news_qi2">
    <qualityIndicator
        name="news_source_l1">
      Reuters
      <qualityInfo id="news_qi2">
        <qualityIndicator
            name="news_source_l2">
          John Smith
        </qualityIndicator>
      </qualityInfo>
    </qualityIndicator>
  </qualityInfo>
</news>
```

Figure 3: XML extended with DQ indicators.

To define measures we have followed the Goal-Question-Metric (GQM) methodology. The steps followed to get the resulting measures are shown in Table 2 and Table 3.

Table 2: Measure definition for the first goal.

Goal	Evaluate XML documents quality
Question	How does XML document complexity affects when manipulating it?
Metrics	NE(D), Number of Elements NA(D), Number of Attributes NEE(D), Number of Empty Elements NEA(D), Number of Empty Attributes NN(D), Number of Nodes NArc(D), Number of Arcs SC_{XML}(D), Structural Complexity

Table 3: Measure definition for the second goal

Goal	Evaluate quality of an XML document extended with quality data.
Question	How does quality data complexity of an XML document affects when manipulating it?
Metrics	QDV(D), Quality Data Volume DDQT(D), Depth of the DQXML Tree

Resulting measures can be divided into several groups according to their purpose:

1. ***Completeness measures.*** They give a notion about the degree of completeness for an XML document.
   - *Number of elements* (NE(D)). Defined as the number of elements in an XML document D.
   - *Number of attributes* (NA(D)). Defined as the number of attributes in an XML document D.
   - *Number of Empty Elements* (NEE(D)). Defined as the number of elements in an XML document D that has neither value of any type nor child elements.
   - *Number of Empty Attributes* (NEA(D)). Defined as the number of empty attributes in elements of an XML document D.
2. ***Complexity measures.*** They give a notion about the complexity of an XML document. The more complex an XML document is, the more difficult its processing will be.
   - *Number of Nodes* (NN(D)). Defined as the number of nodes needed to represent the XML document graph considering as a node any element, attribute or element value.
   - *Number of Arcs* (NArc(D)). Defined as the number of arcs needed to represent the XML document graph, considering an arc as any relation among parent and child elements, element attributes and element values.
   - *Structural Complexity* (SCXML(D)). Represents the structural complexity of an XML document contained in a XML DB. It is defined with the next formula:

$$SC_{XML} = NArc - NN + 1 \quad (1)$$

3. ***Associated quality data.*** They provide an estimation of the amount of quality data is associated to an XML document. As more associated quality data it has, XML document quality could be assessed more accurately.
   - *Quality Data Volume* (QDV(D)). Defined as the total number of "qualityIndicator" elements, in any nesting level, those elements of an XML document contains. As the result of this measure increases, more quality data is stored in the XML DB so a more precise data quality assessment could be done.
   - *Depth of the Data Quality Tree* (DDQT(D)). Defined as the maximum level of nested "qualityInfo" elements in an XML document. As data quality tree is deeper, more detailed data quality will be stored in the XML database.

4 CONCLUSIONS

Nowadays, information is one of the main assets that organizations hold. Data is the raw material where information is extracted from. It is logical to think that the more quality data achieves, the more quality could reach the resulting information improving accordingly organizational processes quality.

For many years the importance of data quality has been ignored when designing and developing databases in which organizations store their data. Our proposal tries to integrate data quality notions inside a DB development methodology in order to open a new research work that fulfill this blank.

On the other hand, new technologies related to XML have spread so widely due to the success of Service Oriented Architectures that XML have became the standard *de facto* to data exchange among agents. This situation has provoked that new approaches to semi-structured data storage optimization have arisen. Inside this field, XML DBs have been created with the goal of improving massive storage of XML documents.

Our research work is centered in developing new strategies for data quality treatment during XML DBs development phase. To reach this target, we have based on some Key Area Processes from the CALDEA framework to define a methodology that considers data quality as a basic aspect during the creation of a XML DB.

The explained approach treats aspects related to user quality requirements management, data source quality assessment, data quality management during the XML DB design phase and measure of different characteristics of data stored in a XML DB.

ACKNOWLEDGEMENTS

This research is part of the FAMOSO and ESFINGE projects supported by the Dirección General de Investigación of the Spanish Ministerio de Ciencia y Tecnología (Ministry of Science and Technology)(TIC2003-07804-C05-03).

REFERENCES

Bray, T., Paoli, J. & Sperberg-McQueen, C. M., 1998. *Extensible Markup Language (XML) 1.0. W3C Recommendation.*

Caballero, I. & Piattini, M., 2007. Assessment and Improvement of Data and Information Quality. IN AL-HAKIM, L. (Ed.) *Information Quality Management: Theory and Applications.* Hershey, PA, USA, Idea Group Publishing.

English, L., 1999. *Improving Data Warehouse and Business Information Quality: Methods for reducing costs and increasing Profits*, New York, NY, USA, Willey & Sons.

Eppler, M., 2001. A Generic Framework for Information Quality in Knowledge-Intensive Processes. In *Proceeding of the Sixth International Conference on Information Quality.*

Fuggeta, A., 2000. Software Process: A Road Map. . In FINKELSTEIN, A. (Ed.) In *Twenty-Second International Conference on Software Engineering (ICSE'2000).* Limerick, Ireland, ACM Press.

García, F., Bertoa, M. F., Calero, C., Vallecillo, A., Ruiz, F., Piattini, M. & Genero, M., 2005. Toward a consistent terminology for software measurement. *Information and Software Technology,* 48**,** 631-644.

Huang, K. T., Lee, Y. W. & Wang, R. Y., 1999. *Quality Information and Knowledge*, Upper Saddle River, NJ, USA, Prentice-Hall.

Lee, Y. W., Pipino, L. L., Funk, J. D. & Wang, R. Y., 2006. *Journey to Data Quality,* Cambridge, MA, USA, Massachussets Institute of Technology.

Levitin, A. & Redman, T., 1995. Quality Dimensions of a Conceptual View. *Information Processing and Management,* 31(1)**,** 81-88.

Marcos, E., Vela, B. & Cavero, J. M., 2001. Extending UML for Object-Relational Database Design. In *Fourth Int. Conference on the Unified Modeling Language, UML 2001*. Toronto (Canada), Springer-Verlag.

OASIS, 2006. *ISO/IEC 26300:2006 Information technology -- Open Document Format for Office Applications (OpenDocument) v1.0.* International Organization for Standardization.

Redman, T. C., 1996. *Data Quality for the Information Age,* Boston, MA, USA, Artech House Publishers.

Strong, D., Lee, Y. & Wang, R., 1997. Data Quality in Context. *Communications of the ACM,* Vol. 40, Nº 5**,** 103 -110.

Verbo, E., Caballero, I. & Piattini, M., 2007. *DQXSD: An XML Schema for Data Quality*. *Paper accepted for the 9th International Conference on Enterprise Information Systems (ICEIS).* Funchal, Madeira - Portugal.

Wang, R. Y., Reddy, M. P. & Kon, H. B., 1995. Toward quality data: An attribute-based approach. *Decision Support Systems.*