ICSOFT 2006

First International Conference on Software and Data Technologies

Proceedings

SETÚBAL, PORTUGAL· SEPTEMBER 11 - 14, 2006

Volume 2

ORGANIZED BY

INSTICC

SPONSORED BY

ENTERPRISE IRELAND

INSTITUTO POLITÉCNICO DE SETÚBAL

IN COOPERATION WITH

OMG
OBJECT MANAGEMENT GROUP

INSTICC PRESS

Proceedings of ICSOFT
First International Conference on Software and Data Technologies
ISBN: 972-8865-69-4
ISBN (13 digits): 978-972-8865-69-6
http://www.icsoft.org

ICSOFT 2006

Proceedings of the
First International Conference on
Software and Data Technologies

Volume 2

Setúbal, Portugal

September 11 – 14, 2006

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

Sponsored by
Enterprise Ireland
Polytechnic Institute of Setúbal

In Cooperation with
Object Management Group (OMG)

Hosted by
School of Business of the Polytechnic Institute of Setubal

Edited by Joaquim Filipe, Boris Shishkov and Markus Helfert

Printed in Portugal

ISBN: 972-8865-69-4

ISBN (13 digits): 978-972-8865-69-6

Depósito Legal: 246927/06

http://www.icsoft.org

secretariat@icsoft.org

BRIEF CONTENTS

KEYNOTE LECTURES

Leszek A. Maciaszek

Macquarie University

Australia

Juan Carlos Augusto

University of Ulster at Jordanstown

U.K.

Tom Gilb

Norway

Dimitris Karagiannis

University of Vienna

Austria

Brian Henderson-Sellers

University of Technology

Australia

Marten J. van Sinderen

University of Twente

The Netherlands

TUTORIAL

Tom Gilb

Norway

ORGANIZING AND STEERING COMMITTEES

Conference Chair

Joaquim Filipe, INSTICC/Polytechnic Institute of Setúbal, Portugal

Program Co-chairs

Markus Helfert, Dublin City University, Ireland

Boris Shishkov, University of Twente, The Netherlands

Proceedings Production

Paulo Brito, INSTICC, Portugal

Marina Carvalho, INSTICC, Portugal

Helder Coelhas, INSTICC, Portugal

Bruno Encarnação, INSTICC, Portugal

Vitor Pedrosa, INSTICC, Portugal

Webdesign and Graphics Production

Marina Carvalho, INSTICC, Portugal

Secretariat and Webmaster

Mónica Saramago, INSTICC, Portugal

PROGRAM COMMITTEE

Hamideh Afsarmanesh, University of Amsterdam, The Netherlands

Jacky Akoka, CNAM & INT, France

Tsanka Angelova, Uniccord Ltd, Bulgaria

Keijiro Araki, Kyushu University, Japan

Lora Aroyo, Eindhoven University of Technology, The Netherlands

Colin Atkinson, University of Mannheim, Germany

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

Elisa Baniassad, Chinese University of Hong Kong, China

Mortaza S. Bargh, Telematica Instituut, The Netherlands

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

Noureddine Belkhatir, LSR-IMAG University of Grenoble, France

Fevzi Belli, University Paderborn, Germany

Alexandre Bergel, Trinity College, Ireland

Mohamed Bettaz, Philadelphia University, Jordan

Robert Biddle, Carleton University, Canada

Maarten Boasson, University of Amsterdam, The Netherlands

Wladimir Bodrow, University of Applied Sciences Berlin, Germany

Marcello Bonsangue, LIACS - Leiden University, The Netherlands

Jonathan Bowen, London South Bank University, U.K.

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

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

Barrett Bryant, University of Alabama at Birmingham, U.S.A.

Cinzia Cappiello, Politecnico di Milano, Italy

Antonio Cerone, UNU-IIST, China

W. K. Chan, Hong Kong University of Science and Technology, China

Kung Chen, National Chengchi University, Taiwan

Samuel Chong, Accenture, U.K.

Chih-Ping Chu, National Cheng Kung University, Taiwan

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

Rolland Colette, University Paris1 Pantheon Sorbonne, France

Alfredo Cuzzocrea, University of Calabria, Italy

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

David Deharbe, UFRN/DIMAp, Brazil

Serge Demeyer, University of Antwerp, Belgium

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

Nikolay Diakov, CWI, The Netherlands

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

Jin Song Dong, National University of Singapore, Singapore

Brian Donnellan, National University of Ireland, Ireland

Jürgen Ebert, University Koblenz, Germany

Paul Ezhilchelvan, University of Newcastle, U.K.

Behrouz Far, University of Calgary, Canada

Bernd Fischer, University of Southampton, U.K.

Gerald Gannod, Arizona State University, U.S.A.

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

Dragan Gasevic, Simon Fraser University, Canada

Nikolaos Georgantas, INRIA Rocquencourt, France

Paola Giannini, Università del Piemonte Orientale, Italy

Paul Gibson, National University of Ireland, Maynooth, Ireland

Wolfgang Grieskamp, Microsoft Research, U.S.A.

Daniela Grigori, University of Versailles, France

Klaus Grimm, Daimlerchrysler Ag, Germany

Rajiv Gupta, University of Arizona, U.S.A.

Tibor Gyimothy, University of Szeged, Hungary

Naohiro Hayashibara, Tokyo Denki University False, Japan

Jang Eui Hong, Chungbuk National University, Korea

Shinichi Honiden, National Institute of Informatics, Japan

PROGRAM COMMITTEE (CONT.)

Ilian Ilkov, IBM Nederland B.V, The Netherlands

Ivan Ivanov, SUNY Empire State College, U.S.A.

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

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

Khaled Khan, University of Western Sydney, Australia

Roger King, University of Colorado, U.S.A.

Christoph Kirsch, University of Salzburg, Austria

Paul Klint, Centrum voor Wiskunde en Informatica (CWI) en University of Amsterdam, The Netherlands

Alexander Knapp, Ludwig-Maximilians-Universität München, Germany

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

Michael Kölling, University of Kent, U.K.

Dimitri Konstantas, University of Geneva, Switzerland

Jens Krinke, FernUniversität in Hagen, Germany

Tei-Wei Kuo, National Taiwan University, Taiwan

Rainer Koschke, University of Bremen, Germany

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

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

Kuan-Ching Li, Providence University, Taiwan

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

Shaoying Liu, Hosei University, Japan

Zhiming Liu, UNU-IIST, China

Andrea De Lucia, Università di Salerno, Italy

Christof Lutteroth, University of Auckland, New Zealand

Broy Manfred, Institut für Informatik, TU München, Germany

Tiziana Margaria, University of Göttingen, Germany

Johannes Mayer, Ulm University, Germany

Fergal McCaffery, University of Limerick, Ireland

Hamid Mcheick, University of Quebec at Chicoutimi, Canada

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

Dimitris Mitrakos, Aristotle University of Thessaloniki, Greece

Roland Mittermeir, Universitaet Klagenfurt, Austria

Birger Møller-Pedersen, University of Oslo, Norway

Mattia Monga, Università degli Studi di Milano, Italy

Aldo De Moor, Vrije Universiteit Brussel, Belgium

Peter Müller, ETH Zurich, Switzerland

Paolo Nesi, University of Florence, Italy

Elisabetta Di Nitto, Politecnico di Milano, Italy

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

Rory O'Connor, Dublin City University, Ireland

Claus Pahl, Dublin City University, Ireland

Witold Pedrycz, University of Alberta, Canada

Massimiliano Di Penta, RCOST - University of Sannio, Italy

Steef Peters, Vrije Universiteit Amsterdam, The Netherlands

Mario Piattini, University of Castilla-La Mancha, Spain

Arnd Poetzsch-Heffter, University of Kaiserslautern, Germany

Andreas Polze, Hasso-Plattner-Institute, Univ. Potsdam, Germany

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

Jolita Ralyte, University of Geneva, Switzerland

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

Anders P. Ravn, Aalborg University, Denmark

Marek Reformat, University of Alberta, Canada

Arend Rensink, University of Twente, The Netherlands

Stefano Russo, Federico II University of Naples, Italy

Shazia Sadiq, The University of Queensland, Australia

Kristian Sandahl, Linkoeping Universtiy, Sweden

Bradley Schmerl, Carnegie Mellon University, U.S.A.

Andy Schürr, Darmstadt University of Technology, Germany

Isabel Seruca, Universidade Portucalense, Portugal

COMMITTEE (CONT.)

Marten van Sinderen, University of Twente,
The Netherlands

Joao Sousa, Carnegie Mellon University, U.S.A.

George Spanoudakis, City University, U.K.

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

Larry Stapleton, ISOL Research Centre, Ireland

Stoicho Stoichev, Technical University-Sofia, Bulgaria

Kevin Sullivan, University of Virginia, U.S.A.

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

Ramayah Thurasamy, Universiti Sains Malaysia, Malaysia

Yasar Tonta, Hacettepe University, Turkey

Yves Le Traon, France Télécom R&D, France

Enrico Vicario, University of Florence, Italy

Bing Wang, University of Hull, U.K.

Kun-Lung Wu, IBM Watson Research, U.S.A.

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

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

Hongji Yang, De Montfort University, U.K.

Yunwen Ye, University of Colorado, U.S.A.

Yun Yang, Swinburne University, Australia

Gianluigi Zavattaro, University of Bologna, Italy

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

Alessandro Aldini, Università degli Studi di Urbino, Italy

Pete Andras, University of Newcastle, U.K.

Xiaoshan Li, UNU-IIST, China

Shih-Hsi Liu, University of Alabama at Birmingham, U.S.A.

Michele Pinna, University of Bologna, Italy

Riccardo Solmi, University of Bologna, Italy

Hongli Yang, UNU-IIST, China

Chengcui Zhang, University of Alabama at Birmingham, U.S.A.

Liang Zhao, UNU-IIST, China

Wei Zhao, University of Alabama at Birmingham, U.S.A.

SELECTED PAPERS BOOK

A number of selected papers presented at ICSOFT 2006 will be published by Springer, in a book entitled Software and Data Technologies. 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 2006 program committee members.

FOREWORD

This volume contains the proceedings of the first International Conference on Software and Data Technologies (ICSOFT 2006), organized by the Institute for Systems and Technologies of Information, Communication and Control (*INSTICC*) in cooperation with the Object Management Group (*OMG*), sponsored by Enterprise Ireland and the Polytechnic Institute of Setúbal and hosted by the School of Business of the Polytechnic Institute of Setubal.

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 2006 received 187 paper submissions from more than 39 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 23 papers were selected to be published and presented as full papers, i.e. completed work (8 pages in proceedings / 30' oral presentations), 44 additional papers, describing work-in-progress, were accepted as short paper for 20' oral presentation, leading to a total of 67 oral paper presentations. There were also 26 papers selected for poster presentation. The full-paper acceptance ratio was thus 12%, and the total oral paper acceptance ratio was 35%.

In its program ICSOFT includes a panel to discuss the future of software development, by six distinguished world-class researchers; furthermore, the program is enriched by one tutorial and six keynote lectures. These high points in the conference program, involving top researchers worldwide, experts in different knowledge areas, have definitely contributed to reinforce the overall quality of the conference.

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 Mónica Saramago for her 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 conference banquet was organized for the afternoon and evening of September 13 (Wednesday) in order to promote this kind of social networking.

We wish you all an exciting conference and an unforgettable stay in the lovely city of Setúbal. We hope to meet you again next year for the 2nd ICSOFT, in Barcelona (Spain), details of which will be shortly made available at http://www.icsoft.org.

Joaquim Filipe
INSTICC/Polytechnic Institute of Setúbal, Portugal
(Conference Chair)

CONTENTS

INVITED SPEAKERS

KEYNOTE LECTURES

INFORMATION SYSTEMS AND DATA MANAGEMENT

FULL PAPERS

SHORT PAPERS

POSTERS

KNOWLEDGE ENGINEERING

FULL PAPERS

SHORT PAPERS

POSTERS

SPECIAL SESSION ON METAMODELLING – UTILIZATION IN SOFTWARE ENGINEERING

A BAYESIAN NETWORK TO STRUCTURE A DATA QUALITY MODEL FOR WEB PORTALS

Angélica Caro[1], Coral Calero[2], Houari Sahraoui[2,3], Ghazwa Malak[3], Mario Piattini[2]
[1]*Universidad del Bio Bio, Departamento de Auditoria e Informática, Chillán, Chile*
mcaro@ubiobio.cl
[2]*Alarcos Group – Computer Science Dept., Universidad de Castilla-La Mancha*
Paseo de la Universidad, 4 – 13071 Ciudad Real (Spain)
email: {Coral.Calero, Mario.Piattini}@uclm.es
[3]*Dept. d'Informatique et de Recherche Opérationnelle, Université de Montréal*
CP 6128 succ. Centre Ville, Montréal QC H3C 3J7 Canada
{sahraouh, rifighaz}@iro.umontreal.ca

Keywords: Data Quality, Information Quality, Web Portals, Bayesian Network, Data Quality Model.

Abstract: The technological advances and the use of the internet have favoured the appearance of a great diversity of web applications, among them Web portals. Through them, organizations develop their businesses in a highly competitive environment. One decisive factor for this competitiveness is the assurance of its data quality. In previous works, a data quality model for Web portals has been developed. The model is represented as a matrix that links the user expectations of data web quality to the portal functionalities. Into this matrix a set of 34 attributes where classified. However, the quality attributes on this model have not an operational structure, necessary to be used actual assessment. In this paper we present how we have structured these attributes by means of a probabilistic approach, using Bayesian Networks. The final objective is to use the Bayesian network obtained for evaluating the quality of a data portal (or a subset of its characteristics).

1 INTRODUCTION

During the past decade, an increasing number of organizations have established Web portals to complement, substitute or widen existing services to their clients. In general, portals provide users with access to different data sources (providers) (Mahdavi et al., 2004), as well as to on-line information and information-related services (Yang, 2004). Moreover, they create a working environment where users can easily navigate in order to find the information they need to perform their operational or strategic tasks and make decisions (Collins, 2001). The users of Web portals need to ensure that this information is appropriate for the use they make of it.

In the literature, the concept of Information or Data Quality (DQ) is often defined as "fitness for use", i.e., the ability of a data collection to meet user requirements (Strong et al., 1997; Cappiello et al., 2004). Recently, due to the particular nature of Web applications the research community started studying the subject of data quality on the Web (Gertz et al., 2004).

However, there are no works on data quality that address the particular context of Web portals, in spite of the fact that some work highlights the data quality as one of the relevant factors in the quality of a portal (Moraga et al., 2004; Yang, 2004). Likewise, except for few work in the data quality area, like (Wang and Strong, 1996; Burgess et al., 2004; Cappiello et al., 2004), most of the works not targeted the quality from the data consumers perspective (Burgess et al., 2004).

In a previous work, we have developed a Portal Data Quality Model (PDQM), focused on the data consumer's perspective (Caro et al. 2006). This model are composed of 34 DQ attributes.

The definition of a model does not mean that it can be operational, i.e., it can be used to assess the quality of web portals in our case. To reach this goal, we need to define a structure that allows from the one hand, to evaluate each attributes using measures and, from the other hand, to combine attribute evaluations to access the portal quality.

Considering the uncertainty inherent to the quality perception, we propose to use a probabilistic approach (Bayesian network) to structure, refine and represent our model.

This rest of this paper is organized as follows. Section 2 presents a brief summary of PDQM. The description of Bayesian networks (BN) and their use to structure our model is presented in Section 3. Section 4 shows the process used for representing a new version of PDQM as a Bayesian network. Finally, Section 5 summarizes and concludes the paper.

2 PDQM

PDQM is a data quality model for Web portals focused in three key elements:

Data consumer perspective. Represented by DQ expectations of data consumer on Internet, stated in (Redman, 2000). These expectations are organized into six categories: Privacy, Content, Quality of values, Presentation, Improvement, and Commitment.

Web DQ attributes. We have identified DQ attributes which have been proposed for different domains in the context on the Web. The idea was to take advantage of work already carried out in the Web context and apply it to Web portals.

Web portal functionalities. Web portals present basic functionalities to data consumer deploying their tasks. Under our perspective, the consumer judges the data by using the application functionalities. So, we used the web portal functions proposes in (Collins, 2001) considering them as basics in our model. These functions are: Data Points and Integration, Taxonomy, Search Capabilities, Help Features, Content Management, Process and Action, Collaboration and Communication, Personalization, Presentation, Administration, and Security.

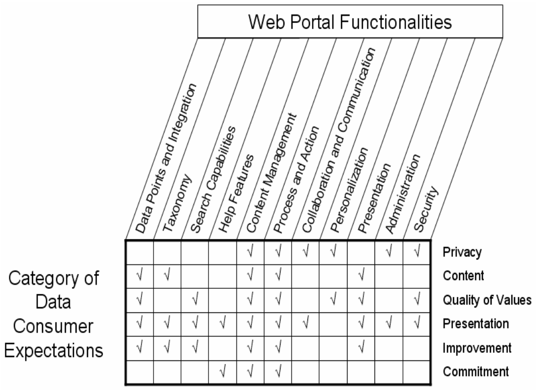
We produced the PDQM model through a three-phase process. In the next subsections we explain each of these phases with their results.

2.1 Web DQ Attributes

The first phase consisted in gathering Web DQ attributes from the literature. For this we have made a systematic review of the relevant literature. Then, we selected works proposed for different domains in the Web context (Web sites (Katerattanakul and Siau, 1999; Eppler et al., 2003; Moustakis et al., 2004), integration of data (Naumann and Rolker, 2000; Bouzeghoub and Peralta, 2004), e-commerce (Katerattanakul and Siau, 2001), Web information portals (Yang et al., 2004), cooperative e-services (Fugini et al., 2002), decision making (Graefe, 2003), organizational networks (Melkas, 2004) and DQ on the Web (Gertz et al., 2004)). The idea was to take advantage of the work already carried out in the Web context and apply it to Web portals. As result and after summarizing the collected initial set of attributes, we obtained 41 DQ attributes (see the top of Table 1).

2.2 Definition of a Classification Matrix for Web DQ Attributes

In the second phase, we have built a matrix for the classification of the DQ attributes obtained in previous phase. This matrix relates two basic aspects considered in our model: the data consumer perspective by means their DQ expectations on Internet (Redman, 2000) and the basic functionalities in a Web portal. On this matrix we carried out an analysis of what expectations were applicable in each different functionality of a Web portal, represented in Figure 1 with a "√" mark.



Web Portal Functionalities

Category of Data Consumer Expectations

	Data Points and Integration	Taxonomy	Search Capabilities	Help Features	Content Management	Process and Action	Collaboration and Communication	Personalization	Presentation	Administration	Security
Privacy					√	√	√	√		√	√
Content	√	√			√	√			√		
Quality of Values	√		√		√	√		√	√		√
Presentation	√	√	√	√	√	√	√		√	√	√
Improvement	√	√	√		√	√			√		
Commitment				√	√	√					

Figure 1: Matrix to classify the Web DQ attributes.

2.3 Classification of Web DQ Attributes in the Matrix

In the third phase, we used the obtained matrix to classify the Web DQ attribute identified in phase 1. Then for each relationship between functionality and expectation, we assigned the DQ attributes that could be used by the data consumer to evaluate the DQ in a portal. We did it by studying the appropriateness of each attribute (based on its definition), in relation to the objective of each portal

Table 1: Data quality attributes assigned for functionality.

Functionalities	Accessibility	Accuracy	Amount of data	Applicability	Attractiveness	Availability	Believability	Completeness	Concise Representation	Consistent Representation	Cost effectiveness	Customer support	Currency	Documentation	Duplicates	Ease of operation	Expiration	Flexibility	Granularity	Interactive	Internal consistency	Interpretability	Latency	Maintainable	Novelty	Objectivity	Ontology	Organization	Price	Relevancy	Reliability	Reputation	Response time	Security	Specialization	Source's information	Timeliness	Traceability	Understand ability	Validity	Value-added	Total of Attributes
Data Points and Integration	✓	✓	✓				✓		✓	✓		✓	✓				✓								✓					✓	✓							✓	✓	✓		15
Taxonomy	✓		✓				✓		✓			✓				✓														✓	✓							✓	✓	✓		11
Search Capabilities	✓		✓				✓	✓	✓			✓	✓			✓									✓					✓	✓							✓	✓			13
Help Features	✓		✓						✓			✓				✓															✓								✓	✓		8
Content Management	✓		✓	✓		✓	✓	✓	✓			✓	✓	✓	✓	✓						✓			✓	✓		✓		✓	✓	✓		✓	✓	✓		✓	✓	✓	✓	24
Process and Action	✓	✓	✓	✓		✓	✓	✓	✓			✓				✓	✓					✓				✓				✓	✓	✓		✓	✓			✓	✓	✓		21
Collaboration and Communication						✓						✓										✓								✓				✓					✓			6
Personalization		✓						✓					✓				✓																					✓	✓	✓		7
Presentation			✓		✓		✓		✓			✓	✓			✓	✓	✓				✓								✓	✓				✓			✓	✓	✓		15
Administration			✓						✓	✓						✓																		✓					✓			6
Security	✓	✓	✓					✓	✓							✓						✓												✓					✓	✓		10
Number of References	7	4	9	2	1	3	6	5	9	1	0	8	5	1	1	8	4	1	0	0	0	5	0	0	3	2	0	1	0	7	7	2	0	5	3	1	0	7	11	8	1	

functionality and the user DQ expectation. On Table 1, we have summarized the attributes assigned for functionality.

On the other hand some attributes, have not assigned, because in our analysis they are result not be important or visible for the data consumer. And therefore, the first version of PDQM has 34 DQ attributes.

3 BAYESIAN NETWORKS

A BN is a directed acyclic graph, whose nodes are the uncertain variables and edges are the causal or influential links between variables. A conditional probability functions model the uncertain relationship between each node and its parents (Neil et al., 2000). In our context, BNs offer an interesting framework with which it is possible to: **(1)** Represent the interrelations between attributes in an intuitive and explicit way by connecting influencing factors to influenced ones. Such a representation facilitates the comprehension of the model, its validation, its evolution and its exploitation, **(2)** Circumvent the problems of subjectivity uncertainty, **(3)** Actually use the obtained network to predict/estimate the quality of a portal, and **(4)** Isolate responsible factors in the case of low quality.

Another interesting property of the Bayesian approach is the fact that it considers the probability as being a dynamic entity that can be updated as more data arrive (self learning mechanism). New data may naturally improve the degree of belief in certain propositions (Baldi et al., 2003). Consequently, a BN model is particularly adapted to the changing domain of web portals.

Building a BN for a particular quality model can be done in two stages: (1) build the graph structure and (2) define the node probability tables for each node of the graph. In this paper we focus on the first stage (see next section). To this end, we use the approach proposed by (Malak et al., 2006) for building BN for web quality models.

4 STRUCTURING PDQM USING A BAYESIAN NETWORK

In its current state PDQM is defined as a set of DQ attributes without a structure that allows it to be used as an evaluation framework for web portals. To structure PDQM (in the form of a BN), we have decided to use the draft of the standard ISO/IEC 25012 (ISO-25012, 2006). Our choice is basically motiveted by two facts. First, ISO/IEC 25012 defines DQ requirements and describes DQ characteristics for any computer system application (i.e.: e-government, e-business, e-commerce). Second, the attributes of the standard are already structured in a hierarchy which can be used for our model. Thus, PDQM can be seen as a specialization of this standard.

In ISO/IEC 25012, there are three different ways of viewing DQ: Internal DQ, External DQ and DQ in Use. It categorises internal and external DQ attributes into six characteristics (functionality, reliability, usability, efficiency, maintainability and portability), which are further subdivided into subcharacteristics. DQ in use is categorized into 4 characteristics effectiveness, productivity, safety and satisfaction), which are refined into sub-characteristics.

The ISO/IEC 25012-guided generation of a BN for PDQM was performed following a three-phase process. In the first 2 phases, we matched the DQ attributes of PDQM with the sub-characteristics of ISO/IEC 25012. These 2 phases produced a hierarchy-like model. In the third phase, we studied and integrated the influence relationships that can exist between the attributes of PDQM. The final result was a graph-like model. The 3 phases are explained together with their results in the next 3 subsections.

4.1 Names Matching

In the first phase, we started to build the BN structure by identifying which attributes in PDQM are also included in the ISO/IEC 25012 as sub-characteristics. For doing this, we have matched the attributes in PDQM with the sub-characteristics in the standard by means of the coincidence between their names. For instance, the *Accuracy* attribute (which is part of PDQM) is present in the standard as a sub-characteristic of the *Functionality* characteristic. Then, in PDQM structure, we have considered, at the Internal/External DQ category, the Functionality characteristic and inside this, the Accuracy sub-characteristic.

This BN contains 4 levels: the model (PDQM), the quality views (I/E_DQ and DQ_in_use), the characteristics that have sub-characteristics present in PDQM, and the attributes of PDQM that match sub-characteristics in the standard (see Figure 3).

4.2 Definition Matching

In the first phase, we used a direct name mat-ching to structure PDQM attributes. The aim of this second phase was to complete the structure obtained. This was done by finding correspondences between the not-yet-assigned attributes and the characteristics in the standard. To this end, we used definition matching between PDQM and ISO/IEC 25012. To illustrate this phase, let's take the following example. For instance, the attribute *Flexibility* was associated to the characteristic Portability and we add this last to PDQM. This obeys to the following analysis. ISO/IEC defines Portability as: "*The capability of data to be transferred from one technological environment to another*". In PDQM Flexibility is defined as: "*The extent to which data are expandable, adaptable, and easily applied to others needs*". So, in our opinion, flexibility is necessary for the Portability of data. The summary of the matching performed in the two first phases is presented in Table 2 (three first columns).

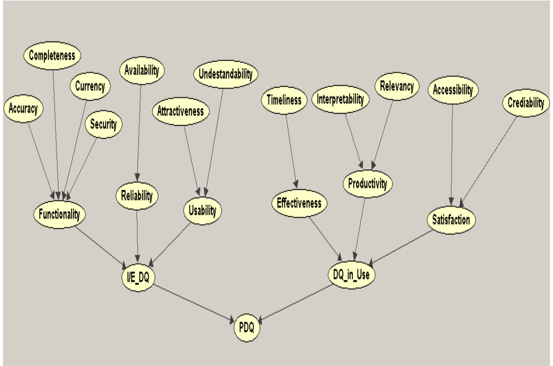


Figure 3: BN obtained in the phase 1.

Table 2: Attributes for the next level in the PDQM.

Rest of PDQM's DQ Attributes / Basic Structure of PDQM (second phase)			Amount of data	Concise Representation	Consistent Representation	Documentation	Expiration	Interactive	Objectivity	Reliability	Organization	Reputation	Specialization	Source's information
Internal/ External Data Quality	Functionality	Accuracy												
		Completeness												
		Currency					✔							
		Security												
	Reliability	Availability												
	Usability	Attractiveness									✔			
		Understandability	✔	✔	✔	✔					✔			
		Ease of operation	✔			✔		✔						
	Efficiency	Duplicates												
		Response time												
	Portability	Flexibility												
Data Quality In Use	Effectiveness	Timeliness												
		Applicability		✔	✔				✔				✔	
		Validity			✔					✔		✔	✔	
	Productivity	Interpretability		✔		✔					✔		✔	
		Relevancy					✔						✔	
	Safety	Customer support												
		Traceability												✔
	Satisfaction	Accesibility						✔						
		Credibility					✔		✔	✔		✔		✔
		Novelty					✔							
		Value-added												

4.3 Causal Relationships Establishing

The third phase was dedicated to the search for causal relationships between the attributes of PDQM. We used the definitions of the attributes to establish these relationships.

For instance, the PDQM attribute Concise Representation is defined as "the extent to which data are compactly represented without elements superfluous or not related". Then, based on this definition, we established a causal relation with the ISO subcharacteristics *Understandability*, *Applicability* and *Interpretability*. Indeed, if the data are compactly represented and without unnecessary elements, they can be more easily understood, applied and interpreted (Katerattanakul and Siau, 2001). Table 2 shows the attributes that were incorporated in this phase. And the Figure 4 shows the final BN generated.

4.4 Probability Definition

To be operational, the BN obtained needs to be supplemented with the probabilities. As stated by Malak et al. in (Malak et al., 2006), there are two types of probabilities that must be defined: input-node probabilities and intermediate-node probabilities.

Intermediate-node probabilities are obtained through tables that define conditional probabilities of the different values that can be taken by quality characteristic of the node knowing the values of the characteristics of the parent nodes. These tables are defined using expert judgment and refined by the self-earning mechanism as the new portals are evaluated.

Characteristics that are represented by input-nodes are those that can be directly measured from the web portals. Input-node probabilities are produced by a transformation of numerical-value measures into probabilities. Consider the node *Response_Time* in our model. The actual time can be classified into three categories: short, medium, and long. Using a fuzzy logic-based clustering algorithm, we can derive a probabilistic classifier that calculate respectively the probabilities that a response time value of a particular web belongs to each of the categories (Malak et al., 2006).

5 CONCLUSIONS

In this paper, we have proposed an operational model for web portal quality assessment. This model is defined as a Bayesian network that was build using the non-structured PDQM model. PDQM is a DQ model containing 34 attributes that were selected specifically for web portals. The BN model was obtained following a three-phase process guided by the ISO/IEC 25012 standard.

The choice of a BN-like model is motivated by the fact that many issues in quality assessment are circumvented: threshold value definition, metric combination, and uncertainty.

We are currently working on the definition of the parameter of the network, i.e., probabilities, using a hybrid approach that combines expert judgment with learning mechanisms.

One of the advantages of our model is its flexibility. Indeed, the model a global framework that can be adapted for both the goal and the context of the evaluation. From the goal perspective, the user can choose the sub-network that evaluates the characteristics he is interested in. From the context point of view, the parameters (probabilities) can be changed to consider the specific context of the evaluated portal. This operation can be done using available historical data from the organization.

To evaluate our model, we are currently designing an experimental study. This study will concern a large number of portals and will involve a set of portal-user subjects. The goal of the study is to compare the subjective judgments of the subjects with the evaluation results produced by our model.

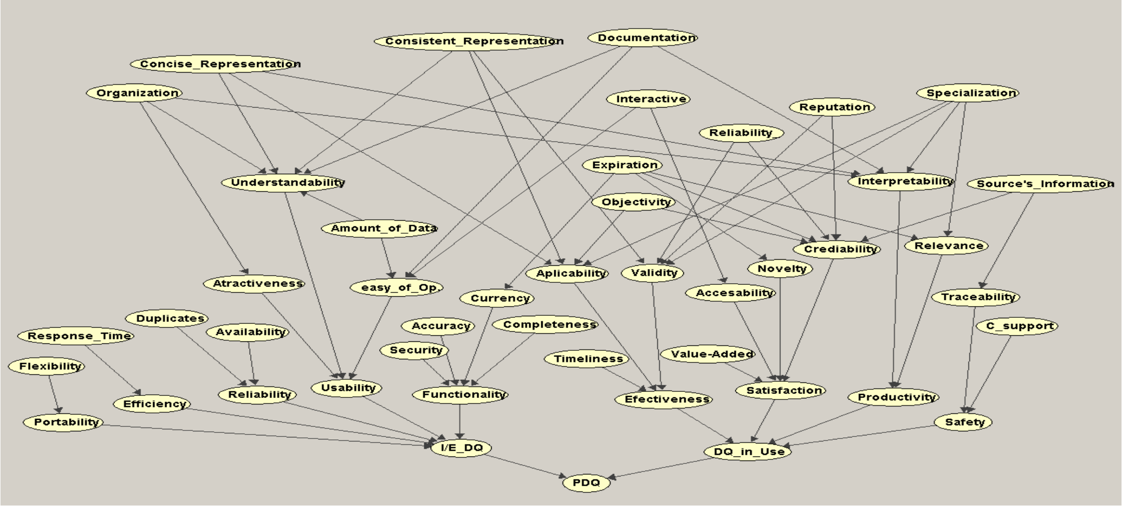


Figure 5: PDQM represented in a BN.

ACKNOLEDGEMENTS

This research is part of the following projects: CALIPO (TIC2003-07804-C05-03) supported by the Dirección General de Investigación of the Ministerio de Ciencia y Tecnología (Spain) and DIMENSIONS (PBC-05-012-1) supported by FEDER and by the "Consejería de Educación y Ciencia, Junta de Comunidades de Castilla-La Mancha" (Spain). This work was performed during the stay of Houari Sahraoui in the University of Castilla-La Mancha under the "Programa Nacional De Ayudas Para La Movilidad de Profesores en Régimen de año sabático", from Spanish Ministerio de Educación y Ciencia, REF: 2004-0161.

REFERENCES

Baldi, P. et al. (2003). Modeling the Internet and the Web; Probabilistic Methods and Algorithms. Wiley

Bouzeghoub, M. and V. Peralta (2004). A Framework for Analysis of data Freshness. International Workshop on Information Quality in Information Systems, (IQIS2004), Paris, France, ACM.

Burgess, M., et al. (2004). Quality Measures and The Information Consumer. Proceeding of the Ninth International Conference on Information Quality.

Cappiello, C., et al. (2004). Data quality assessment from the user´s perspective. International Workshop on Information Quality in Information Systems, (IQIS2004), Paris, Francia, ACM.

Caro , A., et al. (2006). Defining a quality model for portal data. International Conference on Web Engineering, ICWE-2006, Palo Alto, California, USA.

Collins, H. (2001). Corporate Portal Definition and Features, AMACOM.

Eppler, M., et al. (2003). Quality Criteria of Content-Driven Websites and Their Influence on Customer Satisfaction and Loyalty: An Empirical Test of an Information Quality Framework. Proceeding of the Eighth International Conference on Information Quality.

Fugini, M., et al. (2002). Data Quality in Cooperative Web Information Systems.Personal Communication. citeseer.ist.psu.edu/fugini02data.html.

Gertz, M., et al. (2004). "Report on the Dagstuhl Seminar "Data Quality on the Web"." SIGMOD Record vol. 33, Nº 1: 127-132.

Graefe, G. (2003). Incredible Information on the Internet: Biased Information Provision and a Lack of Credibility as a Cause of Insufficient Information Quality. Proceeding of the Eighth International Conference on Information Quality.

ISO-25012 (2006). " ISO/IEC 25012: Software Engineering - Software Quality Requirements and Evaluation (SQuaRE) - Data Quality Model (Draft)."

Katerattanakul, P. and K. Siau (1999). Measuring Information Quality of Web Sites: Development of an Instrument. Proceeding of the 20th International Conference on Information System.

Katerattanakul, P. and K. Siau (2001). Information quality in internet commerce desing. Information and Database Quality. M. Piattini, C. Calero and M. Genero, Kluwer Academic Publishers.

Mahdavi, M., et al. (2004). A Collaborative Approach for Caching Dynamic Data in Portal Applications. Proceedings of the 5th conference on Australian database.

Malak, G, Sahraoui, H, Badri, L, Badri, M. (2006). A Proposal of a Probabilistic Framework for Web-Based Applications Quality, Proceedings of the 10th ECOOP Workshop on Quantitative Approaches in Object-Oriented Software Engineering, (QAOOSE06).

Melkas, H. (2004). Analyzing Information Quality in Virtual service Networks with Qualitative Interview Data. Proceeding of the Ninth International Conference on Information Quality.

Moraga, M. Á., et al. (2004). Comparing different quality models for portals. To appear on Online Information Review, 2006.

Moustakis, V., et al. (2004). Website Quality Assesment Criteria. Proceeding of the Ninth International Conference on Information Quality.

Naumann, F. and C. Rolker (2000). Assesment Methods for Information Quality Criteria. Proceeding of the Fifth International Conference on Information Quality.

Neil, M., Fenton, N.E., Nielsen, L., (2000). Building large-scale Bayesian Networks. The Knowledge Engineering Review, 15(3). 257-284

Pressman, R. (2001). Software Engineering: a Practitioner's Approach. 5/e, McGraw-Hill.

Redman, T. (2000). Data Quality: The field guide. Boston, Digital Press.

Strong, D., et al. (1997). "Data Quality in Context." Communications of the ACM Vol. 40, Nº 5: 103 -110.

Wang, R. and D. Strong (1996). "Beyond accuracy: What data quality means to data consumers." Journal of Management Information Systems; Armonk; Spring 1996 12(4): 5-33.

Yang, Z., et al. (2004). "Development and validation of an instrument to measure user perceived service quality of information presenting Web portals." Information and Management. Elsevier Science 42: 575-589.