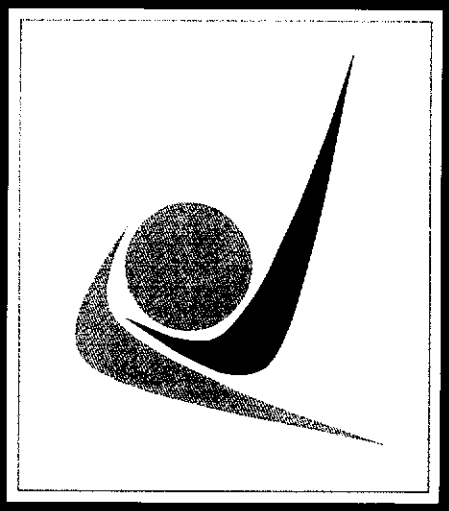

QSIC 2007

Proceedings of the Seventh International Conference on

Quality Software

Portland, Oregon, USA, October 11–12, 2007

Edited by Aditya Mathur, W. Eric Wong, and M. F. Lau

Sponsor:
- The University of Hong Kong, Hong Kong

Proceedings

Seventh International Conference on Quality Software

QSIC 2007

IEEE Computer Society Order Number P3035
ISBN 13: 978-0-7695-3035-2
ISBN 10: 0-7695-3035-4
ISSN 1550-6002

Additional copies may be ordered from:

IEEE Computer Society
Customer Service Center
10662 Los Vaqueros Circle
P.O. Box 3014
Los Alamitos, CA 90720-1314
Tel: + 1 800 272 6657
Fax: + 1 714 821 4641
http://computer.org/cspress
csbooks@computer.org

IEEE Service Center
445 Hoes Lane
P.O. Box 1331
Piscataway, NJ 08855-1331
Tel: + 1 732 981 0060
Fax: + 1 732 981 9667
http://shop.ieee.org/store/
customer-service@ieee.org

IEEE Computer Society
Asia/Pacific Office
Watanabe Bldg., 1-4-2
Minami-Aoyama
Minato-ku, Tokyo 107-0062
JAPAN
Tel: + 81 3 3408 3118
Fax: + 81 3 3408 3553
tokyo.ofc@computer.org

Individual paper REPRINTS may be ordered at: reprints@computer.org

Editorial production by Randall S. Bilof
Cover art production by Joe Daigle/Studio Productions
Printed in the United States of America by Applied Digital Imaging

IEEE Computer Society
Conference Publishing Services (CPS)
http://www.computer.org/cps

Contents

Proceedings of the Seventh International Conference on Quality Software

QSIC 2007

Position Papers by Distinguished Plenary Panel

Regular Papers

Session 1A: Software Testing 1

Session 1B: Distributed Systems

Session 2A: Software Security and Model Checking

Session 3A: Software Testing 3

Session 3B: Embedded Systems

Session 4A: Component-Based Systems

Session 4B: Software Quality

Short Papers

Session 1C: Short Papers—Software Quality and Reliability

Session 2B: Short Papers—Software Testing 2

Session 2C: Short Papers—Software Architecture and Components

Session 5: Short Papers—Systems Modeling, Model Construction and Checking

First International Workshop on Software Test Evaluation (STEV 2007)

Workshop Papers

Workshop Position Paper

Message from the Program Co-Chairs

QSIC 2007

Welcome to QSIC 2007—the Seventh International Conference on Quality Software! QSIC brings together researchers and practitioners working to improve the quality of software to present new results and exchange ideas.

We have prepared a strong and varied technical program spread over seven regular paper sessions, four short paper sessions, one distinguished plenary panel, and two keynote addresses. We are honored to have Professor Edward A. Lee, Chair of EECS and Robert S. Pepper Distinguished Professor, UC Berkeley, and Dr. Solom Heddaya, Head of Windows OS Reliability, Microsoft Corp., to be our distinguished keynote speakers. We received 92 submissions from 29 countries. Each submission was carefully evaluated by at least three reviewers. The reviews were used for selecting papers to be presented at the conference. Twenty-four regular papers have been accepted. These papers cover a broad spectrum of areas including software testing, verification, reliability, security, model checking, and quality assurance. Also included are papers on component-based software, distributed systems, and embedded systems. In addition, 25 submissions have been accepted as short papers. One associated workshop emphasizes "Software Test Evaluation." It includes six regular papers and one position paper.

We thank our distinguished keynote speakers, panelists, and all of the authors for sharing their ideas and results with us, and all of the members of the Program Committee and reviewers for their help in evaluating and selecting high-quality papers. Their participation has been crucial to the success of QSIC 2007. Special thanks to Professor T. H. Tse for his continuous support and encouragement, to Professor T. Y. Chen for his insightful suggestions, and to Mr. Lei Zhao for maintaining the conference website and assisting with registration.

On behalf of the Program Committee, we thank all participants for attending QSIC 2007 and hope that you enjoy the conference.

Aditya Mathur
Purdue University, USA

W. Eric Wong
University of Texas at Dallas, USA

Conference Committees

QSIC 2007

Steering Committee

Chair

T. H. Tse, *The University of Hong Kong, Hong Kong*

Members

T. Y. Chen, *Swinburne University of Technology, Australia*
Hans-Dieter Ehrich, *Technische Universitaet Braunschweig, Germany*
Huimin Lin, *Institute of Software, Chinese Academy of Sciences, China*
Peter C. Poole, *Emeritus Professor at the University of Melbourne, Australia*
C. V. Ramamoorthy, *University of California at Berkeley, USA*
Stephen S. Yau, *Arizona State University, USA*

General Chair

C. V. Ramamoorthy, *University of California at Berkeley, USA*

Program Committee

Co-Chairs

Aditya Mathur, *Purdue University, USA*
W. Eric Wong, *University of Texas at Dallas, USA*

Members

Doo-Hwan Bae, *Korean Advanced Institute of Science and Technology, Korea*
Xiaoying Bai, *Tsinghua University, China*
Fevzi Belli, *University of Paderborn, Germany*
Maarten Boasson, *University of Amsterdam, The Netherlands*
Kai-Yuan Cai, *Beijing University of Aeronautics and Astronautics, China*
Joao Cangussu, *University of Texas at Dallas, USA*
Alessandra Cavarra, *Oxford University, UK*
Keith C. C. Chan, *The Hong Kong Polytechnic University, Hong Kong*
Victor Chan, *Macao Polytechnic Institute, Macao*
W. K. Chan, *City University of Hong Kong, Hong Kong*
Jason Chen, *National Central University, Taiwan*
Jessica Chen, *University of Windsor, Canada*
S. C. Cheung, *Hong Kong University of Science and Technology, Hong Kong*
Byoungju Choi, *Ewha Womans University, Korea*
William C.-C. Chu, *TungHai University, Taiwan*

Takeshi Chusho, *Meiji University, Japan*
Kendra Cooper, *University of Texas at Dallas, USA*
Marcio Delamaro, *Centro Universitario Euripides de Marilia, Brazil*
Jin Song Dong, *National University of Singapore, Singapore*
Rachida Dssouli, *Concordia University, Canada*
Abdeslam En-Nouaary, *Concordia University, Canada*
Kokichi Futatsugi, *Japan Advanced Institute of Science and Technology, Japan*
Jerry Gao, *San Jose State University, USA*
Sudipto Ghosh, *Colorado State University, USA*
Arnaud Gotlieb, *IRISA-INRIA, France*
Wolfgang Grieskamp, *Microsoft Research, USA*
Aiman Hanna, *Concordia University, Canada*
Xudong He, *Florida International University, USA*
Michael Jiang, *Motorola Labs, USA*
Ho-Won Jung, *Korea University, Korea*
Victor Kuliamin, *Russain Academy of Sciences, Russia*
Fei-Ching Kuo, *Swinburne University of Technology, Australia*
Richard Lai, *La Trobe University, Australia*
Yu Lei, *University of Texas at Arlington, USA*
Xuandong Li, *Nanjing Univeristy, China*
Shaoying Liu, *Hosei University, Japan*
Yan Liu, *Motorola Labs, USA*
Jian Lu, *Nanjing University, China*
Jose Maldonado, *Universidade de Sao Paulo, Brazil*
Johannes Mayer, *Ulm University, Germany*
Hong Mei, *Peking University, China*
Atif Memon, *University of Maryland, USA*
Simanta Mitra, *Iowa State University, USA*
Takako Nakatani, *Wakayama University, Japan*
Tien Nguyen, *Iowa State University, USA*
Allen Nikora, *Jet Propulsion Laboratory, USA*
Jeff Offutt, *George Mason University, USA*
Hideto Ogasawara, *Corporate Research and Development Center, Toshiba Corporation, Japan*
Amit Paradkar, *IBM T.J. Watson Research Center, USA*
Andy Podgurski, *Case Western Reserve University, USA*
Isidro Ramos, *Universidad Politecnica de Valencia, Spain*
Motoshi Saeki, *Tokyo Institute of Technology, Japan*
Klaus-Dieter Schewe, *Massey University, New Zealand*
Wuwei Shen, *Western Michigan University, USA*
Paul Strooper, *The University of Queensland, Australia*
Kenji Taguchi, *National Institute of Informatics, Japan*
Barrie Thompson, *University of Sunderland, UK*
Jeff Tian, *Southern Methodist University, USA*
June Verner, *University of New South Wales, Australia*
Ji Wang, *Changsha Institute of Technology, China*
Qianxiang Wang, *Peking University, China*

Min Xie, *National University of Singapore, Singapore*
Dianxiang Xu, *North Dakota State University, USA*
Qiwen Xu, *University of Macau, Macau*
Hongji Yang, *De Montfort University, UK*
Y. T. Yu, *City University of Hong Kong, Hong Kong*
Jian Zhang, *Chinese Academy of Sciences, China*
Xiangyu Zhang, *Purdue University, USA*
Wenyun Zhao, *Fudan University, China*
Hong Zhu, *Oxford Brookes University, UK*
Mohammad Zulkernine, *Queen's University, Canada*

Organizing Committee

Co-Chairs

Warren Harrison, *Portland State University, USA*
Kal Toth, *Portland State University, USA*

Finance Chair

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

Publication Chair

M. F. Lau, *Swinburne University of Technology, Australia*

Registration Co-Chairs

M. F. Lau, *Swinburne University of Technology, Australia*
Lei Zhao, *University of Texas at Dallas, USA*

Web Master

Lei Zhao, *University of Texas at Dallas, USA*

Refinement of a tool to assess the data quality in Web portals

Angélica Caro[1], Coral Calero[2], Juan Enriquez de Salamanca[2] and Mario Piattini[2]

[1]*Department of Computer Science and Information Technologies, University of Bio Bio, Chillán, Chile*
mcaro@ubiobio.cl

[2]*Alarcos Research Group. Information Systems and Technologies Dep. UCLM-INDRA Research and Development Institute University of Castilla-La Mancha, Spain*
{Coral.Calero, Mario.Piattini}@uclm.es

Abstract

The Internet is now firmly established as an environment for the administration, exchange and publication of data. To support this, a great variety of Web applications have appeared, among these web portals. Numerous users worldwide make use of Web portals to obtain information for different purposes. These users, or data consumers, need to ensure that this information is suitable for the use to which they wish to put it. PDQM (Portal Data Quality Model) is a model for the assessment of portal data quality. It has been implemented in the PoDQA tool (Portal Data Quality Assessment tool), which can be accessed at http://podqa.webportalquality.com. In this paper we present the various refinements that it has been necessary to make in order to obtain a tool which is stable and able to make accurate and efficient calculations of the elements needed to assess the quality of the data of a web portal.

1. Introduction

A Web portal is a site that aggregates information from multiple sources on the Web and organizes this material in an easy user-friendly manner [7]. Over the past decade the number of organizations that provide portals has grown dramatically. These organizations provide portals that complement, substitute or extend existing services to their client base [8]. Numerous users worldwide make use of Web portals to obtain information for their work and to help with decision making. These users, or data consumers, need to ensure that the data obtained are appropriate for the use to which they need to be put. Likewise, the organizations that provide Web portals need to offer data that meet user requirements and help these users to achieve their goals. Therefore data quality represents a common interest between data consumers and portal provider.

Data Quality (DQ) is often defined as "fitness for use", i.e., the ability of a collection of data to meet user requirements [1, 6]. Moreover, the terms "data" and "information" are often used as synonyms. In this work we shall also treat them as being synonymous.

PDQM, created in our previous work [2], is a data quality model for Web portals centred on the point of view of data consumers and uses a probabilistic approach (based on Bayesian networks) for DQ evaluation. The PoDQA tool (http://podqa.webportalquality.com) is a tool which implements PDQM. The PoDQA tool was built using a 3-tiered architecture to separate the presentation, application(business), and storage components, using Visual Basic .NET technology. The main functions of PoDQA are to calculate the level of DQ for a given Web portal and to calculate the DQ ranking for a given Web portal domain. However, this has occurred after a process of refinement of the tool which will be the focus of this paper.

The paper is structured as follows. Section 2 briefly describes the PDQM, the DQ model that is supported by PoDQA. Section 3 explains the main characteristics of the PoDQA. The refinement of the tool is described in Section 4. Finally, Section 5 shows our conclusions.

2. PDQM: a DQ model for web portals

PDQM is a data quality model for Web portals which focuses upon the data consumer perspective. Its definition was carried out in 2 phases: (1) the identification of a set of DQ attributes through which to assess the DQ in Web portals (see Table 1) and (2) the organization of these attributes in an operational model. In order to do this, we used a probabilistic approach which implemented Bayesian networks (see Figure 1).

1550-6002/07 $25.00 © 2007 IEEE

Table 1
Data quality attributes of PDQM

Attractiveness	Documentation	Customer Support
Accessibility	Duplicates	Reliability
Accuracy	Ease of Operation	Reputation
Amount of Data	Expiration	Response Time
Applicability	Flexibility	Security
Availability	Interactivity	Specialization
Believability	Interpretability	Timeliness
Completeness	Novelty	Traceability
Concise Representation	Objectivity	Understandability
Consistent Representation	Organization	Validity
Currency	Relevancy	Value added

As we can see in Figure 1, we have organized the 33 DQ attributes into 4 DQ categories: Intrinsic DQ, Operational DQ, Contextual DQ and Representational DQ.

As an initial step in the implementation of the Bayesian network (BN) we have worked with the Representational DQ sub-network. For this sub-network we have defined the measures, indicators and the node probability tables necessary for the evaluation of the DQ. The probability tables must be defined according to the domain that we wish to assess. In our case, we are working within the context of University portals. In Figure 2 the BN generated for this sub-network, the quantifiable variables which measure the DQ attributes which are input nodes in the BN (nodes without parents and that are general for any domain) and some of the probability tables defined for each node in the sub-network (specific to the context of university portals) are shown.

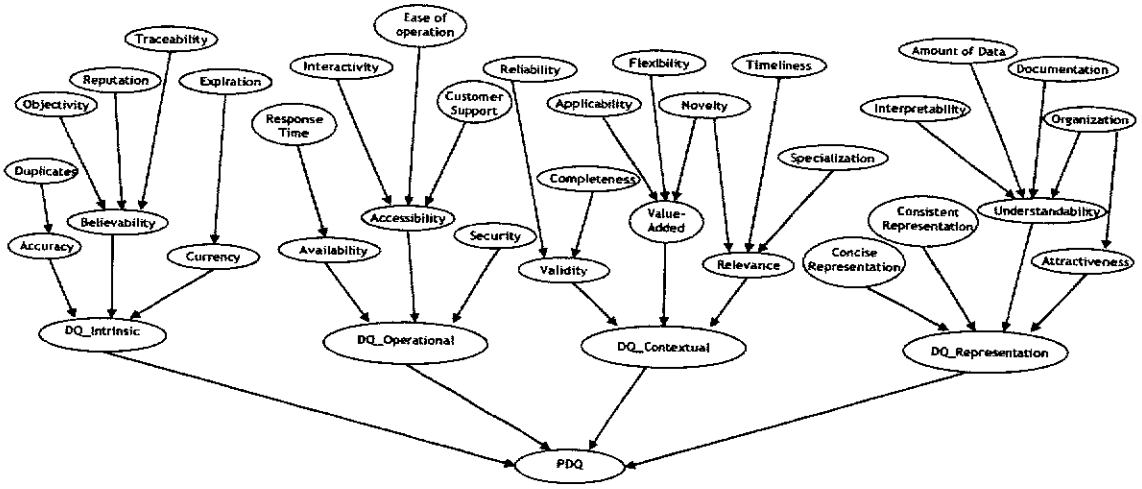

Figure 1. A Bayesian network representing the PDQM

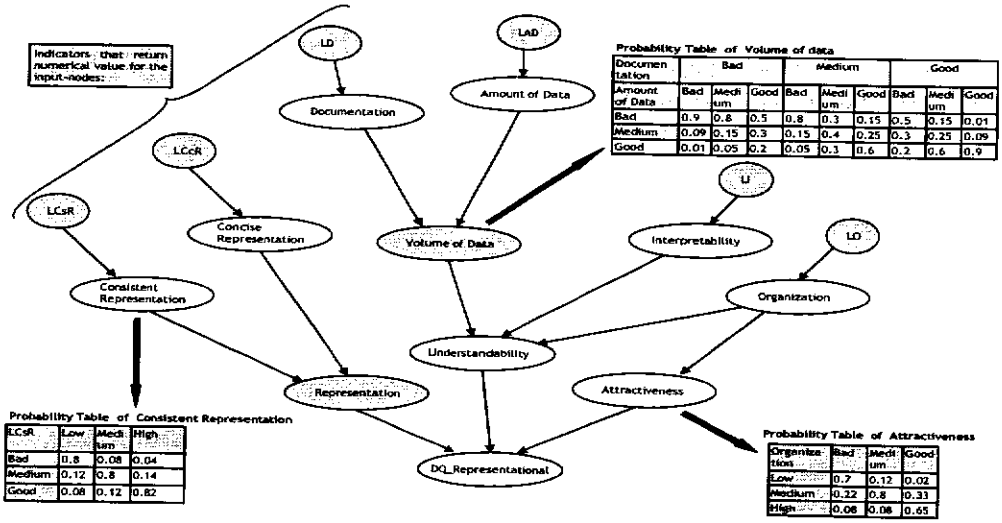

Figure 2. Example of the sub-network of Representational DQ

Each entry node of this sub-network is calculated by means of an indicator. Five of these in an objective way and one is calculated by using a questionnaire that must be answered by the user. In accordance with the definition of the PDQM, the indicators are defined as follows:

- *Level of Consistent Representation (LCsR).* The Consistent Representation attribute is defined as: *The extent to which data are always presented in the same format, are compatible with previous data and are consistent with other sources.* The measures selected for this attribute are centred on the consistency of the format and on compatibility with the pages. For this indicator we have defined measures based on the use of Style within the pages of the portal and on the correspondence between a source page and the destination pages.
- *Level of Concise Representation (LCcR).* The Concise Representation attribute is defined as: *The extent to which data are compactly represented without superfluous or non-related elements.* For this attribute we have considered measures associated with the amount and size of paragraphs and the use of tables to represent data in a compact form.
- *Level of Documentation (LD).* The Documentation attribute is defined as: *Quantity and utility of the documents with metadata.* The measures to evaluate this attribute are related to the basic documentation that a portal presents to data consumers. To calculate this indicator, we considered the simple documentation associated with the hyperlinks and images on the pages of the portal.
- *Level of Amount of Data (LAD).* The Amount of Data attribute is defined as: *The extent to which the quantity or volume of data delivered by the Web portal is appropriate.* We understand that from the data consumer's perspective the amount of data is concerned with the distribution of data throughout the pages in the portal. It is for this reason that, when measuring the amount of data, we have considered that data in text form (words), in hyperlink form (links) and in visual form (images).
- *Level of Interpretability (LI).* The Interpretability attribute is defined as: *The extent to which data are expressed in language and units appropriate for the consumer's capability.* For this attribute we have decided to use a check list for its measurement. Each item in the check list will be evaluated with a number from 1 to 10; these values need to be subsequently transformed into a value input for the BN.
- *Level of Organization (LO).* The Organization attribute is defined as: *The organization, visual settings or typographical features (colour, text, font, images, etc.) and the consistent combinations of these various components.* So, we used measures which verify the existence of a data group (tables, frames, etc.), the use of colours, titles and different fonts, etc, as a means to establish the level of organization of the data in the portal.

In general, the measures have been defined with the assistance of works by Nielsen, Ivory and Eppler [3-5]. As an example of the measures defined, Table 2 shows the measures for the *Level of Consistency* entry node.

Table 2. Measures for the LCsR indicator in the Representational DQ fragment.

Consistent Representation Attribute	
Base Measures	**Derived Measures**
• Pages Count (**PgC**) • Link Count (**LnC**) • Maximum of Pages with the Same Style (**MaSS**) • Link Text Correspondence (**LTC**)	• Source Destiny Correspondence Degree(**SDCD**): SDCD= LTC/LnC • Pages with the Same Style Degree (**PSSD**): PSSD= MaSS/PgC
Level of Consistent Representation (LCsR) Indicator	
LCsR = (PSSD * 0.5 + SDCD * 0.5)	

Once the sub-network was completely defined (with the indicators and the probability tables) we decided to implement it with a tool in such a way that any user could assess the DQ data in a given portal. This tool is named PoDQA and will be presented in the following section.

3. PoDQA: a portal DQ assessment tool

The objective of the PoDQA tool is to give to the user information about the DQ level in a given portal (at present this is only for the Representational DQ). This process cannot take place in real time because it is necessary to download and analyze all the pages of the portal, in order to calculate the defined indicators. In order to perform the measurements, the tool works exclusively with the public information in portals.

The tool considers different portal domains. In this way, DQ can be evaluated depending upon the domain to which the portal belongs. This is done by using the appropriate probability tables for each domain. Finally, the application not only gives information about DQ, but also suggests certain activities that could be applied in order to improve that web portal DQ.

The tool was built by using a 3-tiered architecture to separate the presentation, application, and storage components, using Visual Basic .NET technology. By means of the presentation tier the tool provides an interface for the user which allows him/her to develop an evaluation process. The data tier corresponds with the database in which the results of various evaluations are stored. Finally, the application tier is composed of two sub-applications. The first calculates the indicators for the portal under study, stores the results, generates the inputs for the second sub-application and notifies the user when the evaluation has been completed. The second sub-application loads and executes the appropriate BN (corresponding to the Web portal

domain) and sends the final results to the first sub-application to be stored.

3.1. Assessing the DQ_Representational

In order to obtain the score of the Representational DQ in a given Web portal the steps indicated in Figure 3 must be performed.

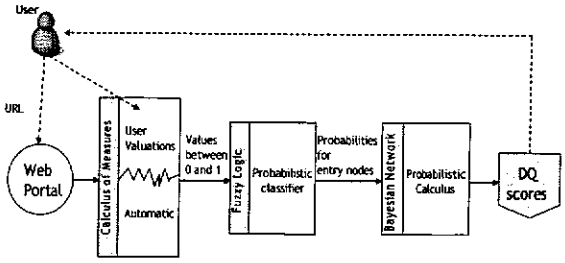

Figure 3. PoDQA Process to assess DQ

It is first necessary to calculate the measures associated with the indicators: LCsR, LCcR, LD, LAD, LI, LO (the objective measures are calculated automatically and the user's evaluations are requested). Each indicator measured will take a value of between 0 and 1. This value must be transformed into a set of probabilities for the corresponding categories. Then, by means of a probabilistic classifier (fuzzy logic-based clustering algorithm), the probabilities for each entry node in the BN are calculated. These probabilities are entered in the BN. From each piece of evidence, and by using the corresponding probability table, each node generates a result that is propagated, via a causal link, to the child nodes for the whole network to the level of the DQ.

As a result of the evaluation of the Representational DQ, the user will receive a description of the level of DQ and recommendations for the improvement of this level.

It is obvious that the correct calculation of the indicators of the entry nodes is fundamental in assuring that the DQ assessment functions correctly. Although these indicators have been defined in a complete way, when they were implemented it was necessary to perform certain refinement tasks. These tasks have assured both the correctness of the calculation and its efficiency when performed. The following section explains both the refinement tasks and the reasons behind the decisions which were made for each of them.

4. Refining PoDQA & the assessment process

As has previously been indicated, once the tool was finished certain refinement tasks were necessary. We formed three groups of tasks and as a result we have obtained a stable and efficient tool which assesses the Representational DQ level in a highly accurate manner.

The three groups of tasks are: refinement of the probability tables, treatment of the outlier indicators and determination of the depth of the measurements.

4.1. Refining the probability tables

The first refinement task consisted of the use of two different strategies through which to evaluate the Representational DQ in a given portal: one of these evaluated the DQ with a group of subjects and the other evaluated the DQ with PoDQA. The goal was to make a comparison between the results obtained from both in order to determine whether the evaluation made with PoDQA was similar to that of the DQ user's perception. That is, whether the model represented the data consumer's perspective.

For the first strategy we developed an experiment to obtain the judgments of a group of 79 subjects about Representational DQ in a university portal. In this experiment, the subjects were asked for their partial evaluations of each DQ attribute in the fragment and for their evaluation of the global DQ in the portal. For the second strategy we used PoDQA. In Table 3 the values obtained from both strategies are shown.

Table 3. Values obtained from experiment & PoDQA

Attribute Evaluated	Valuations					
	Low/Bad		Medium		High/Good	
	Subj.	Tool	Subj.	Tool	Subj.	Tool
Attractiveness	30%	34%	61%	44%	9%	22%
Organization	37%	26%	44%	66%	19%	8%
Amount of Data	18%	6%	49%	13%	33%	81%
Understandability	32%	52%	47%	23%	21%	25%
Interpretability	6%	43%	45%	49%	48%	7%
Documentation	16%	9%	49%	82%	34%	9%
Consistent Rep.	18%	81%	53%	13%	29%	6%
Concise Rep.	16%	6%	52%	13%	32%	81%
Portal	**17%**	**20%**	**68%**	**40%**	**16%**	**40%**

With regard to the final evaluation (Table 3, last row) it can be observed that while in the experiment the subjects evaluated the DQ at a Medium level (68%), the automatic evaluation of the DQ returned the same value at Medium and High levels (in both cases 40%). With regard to the partial values, that is, for each DQ attribute, the results are also very different.

In our opinion the reason for this was that the results given for the indicators were, in some cases, very extreme (see for example the values for LCcR and LCsR). Consequently, the nodes with most differences are the child nodes of the nodes that represent the indicators that take these extreme values. A preliminary interpretation of these results is that PDQM is more demanding than the subjects and needs to be adjusted.

By taking advantage of the flexibility of PDQM we consequently attempted to reduce these differences by ad-

justing the node probability tables and recalculating the level of DQ. The results obtained can be observed in Table 4.

Table 4. New Valuations obtained after node probability table adjustment

Attribute Evaluated	Valuations					
	Low/Bad		Medium		High/Good	
	Subj.	Tool	Subj.	Tool	Subj.	Tool
Attractiveness	30%	26%	61%	58%	9%	16%
Organization	37%	26%	44%	60%	19%	14%
Amount of Data	18%	8%	49%	17%	33%	75%
Understandability	32%	32%	47%	48%	21%	20%
Interpretability	6%	40%	45%	46%	48%	14%
Documentation	16%	11%	49%	73%	34%	15%
Consistent Rep.	18%	73%	53%	15%	29%	12%
Concise Rep.	16%	8%	52%	17%	32%	75%
Portal	**17%**	**18%**	**67%**	**58%**	**16%**	**24%**

As a result of this new configuration the general result of the automatic evaluation is closer to the subjects' evaluations. However, in spite of the fact that both evaluations gave their result as *Medium*, total coincidence between the values calculated does not exist (see last row in Table 4). Moreover, the partial values also have a better fit than in the first calculation, but do not totally coincide (see for example the differences between the Interpretability and Consistent Representation attributes for the valuations Low/Bad).

We again believe that the main reason for this is the extreme values of the indicators and thus decided to work on this aspect. We believed that the design of the portal evaluated may also influence this result. For example, in order to calculate the *Level of Amount of Data*, it is necessary to know the *distribution of words per page*. The measured portal presents values for this measure which can be considered as outliers (they take extreme values that do not follow a uniform distribution). Obviously, these values need to be removed from the calculation of the measure.

Another problem encountered during this phase was the lack of efficiency when performing the measurements (the results arrived up to three days after the measures were performed).

4.2. Deciding about the outliers of the measures

Our next task in the refinement of PoDQA consisted of the detection of the outlier indicators in order to remove them from the calculation. To do this, we decided to work by measuring several portals and then making decisions about the results obtained.

In Table 5, the five maximum values obtained for a set of measures that are used to calculate the indicators of a given portal are shown. As can be observed, in general, the last value does not follow the same distribution (in comparison to its ancestors) as the others.

Although these are the results for only one portal, we have developed the same study for 9 portals and the tendency was the same. We have also studied the minimum values for the ten portals and we have not found outliers so we have therefore decided to do not remove any value from the minimum.

Table 5. The five maximum values obtained for some measures that are used to calculate the indicators of a given portal

	MaWP	MaLP	MaIP	MaPP	MaTP
Portal 1	1º->1919	1º->116	1º->116	1º->97	1º->21
	2º->2080	2º->133	2º->121	2º->128	2º->23
	3º->2949	3º->198	3º->215	3º->168	3º->24
	4º->6049	4º->212	4º->305	4º->215	4º->48
	5º->50008	5º->332	5º->852	5º->823	5º->52

Thus, for the indicators we decided to eliminate the maximum value. If the maximum does not correspond to an outlier (see MaTP indicator in Table 5) its deletion will not greatly affect the final valuation of the indicator. If the value is an outlier (see the MaWP indicator in Table 5) then its deletion will allow a more accurate calculation of the indicator.

4.3. Deciding the depth of the calculus of measures

As has been previously mentioned, in the first version of PoDQA we found many problems of efficiency. PoDQA took, in some cases, three days to perform the calculation of the indicators. PoDQA even aborted its execution when it, for example, found dynamic pages.

Therefore, our objective in the third refinement task was to determine whether it was possible to reduce the number of links followed for the calculation. This means, whether it was possible to restrict the depth of links studied (although the study in width was carried out completely for each level) and determine whether the calculation was accurate enough.

To do this, we performed proofs in a portal that was completely measured without depth limits and after we carried out the same calculation, restricting the depth levels. Table 6 shows the results obtained for this portal, where the indicators were calculated by studying the complete portal, and only depth levels of 2, 3 and 4 were studied (by following the portal links).

If we observe these results, we can see that for levels three and four, by examining a lesser number of links (i.e. without examining the complete portal), we obtain practically the same results as when working with the complete portal depth. These values change in certain ways when working only to the second level. Once again, this experience has been repeated in the other nine portals and the tendency was the same.

Table 6. Results obtained for a portal and different levels

Links downloaded and analyzed	Level 2	Level 3	Level 4	Complete
	40	193	798	20622
Execution time	Quite instantaneous	Scarce	Scarce	10-12h
LAD	0.62 Medium	0.9 Good	0.89 Good	0.98965 Good
LCcR	0.82 Good	0.9 Good	0.9 Good	0.9919 Good
LCsR	0.1 Low	0.20 Low	0.12 Low	0.12577 Low
LD	0.5109 Medium	0.52 Medium	0.488 Medium	0.4877205 Medium
LO	0.38 Medium	0.623 Medium	0.489 Medium	0.43555 Medium

As there are no great variations in the calculation between levels 3 and 4, we decided to limit the calculation and did this to the fourth level of link depth. By using this constraint we have improved the time of calculus, obtaining up to four reports in 6 hours.

Another benefit obtained from limiting the number of links studied is that, in the case of web sites with dynamic pages, such as calendars, agendas, etc., the tool restricts the calculus and it is not necessary to work with the complete depth (for example, it is not necessary to measure calendar pages until the year 1 A.D.), and this also alleviates the time of calculations.

Thanks to the lightness of the calculation and to the possibility of managing the data in many portals, we have detected certain problems with some measures (for example the number of words per paragraph) which we have been able to solve in an adequate manner.

Finally, we further observed that upon limiting the depth of the links to be analyzed, we also avoided the apparition of outliers in the vast majority of the measurements. We nevertheless decided to maintain the outlier detection strategy, and to eliminate those outliers from the calculations when necessary.

5. Conclusions

This paper explains the refinement actions which were carried out on the PoDQA tool. PoDQA is a tool which implements PDQM, a portal data quality model that uses Bayesian networks. As a result of these refinements, PoDQA is now a stable and efficient tool which accurately calculates the DQ of a portal.

As a future work we will incorporate the rest of the DQ subnetworks into the tool, in order to know the DQ level of all four DQ categories. It will also be necessary to incorporate new domains by defining and including the corresponding probability tables in PoDQA.

6. Acknowledgments

This research is part of the following projects: ESFINGE (TIC2006-15175-C05-05) granted by the Dirección General de Investigación del Ministerio de Ciencia y Tecnología (Spain), CALIPSO (TIN20005-24055-E) supported by the Ministerio de Educación y Ciencia (Spain), DIMENSIONS (PBC-05-012-1) supported by FEDER and by the "Consejería de Educación y Ciencia, Junta de Comunidades de Castilla-La Mancha" (Spain) and COMPETISOFT (506AC0287) financed by CYTED.

7. References

[1] C. Cappiello, C. Francalanci, and B. Pernici. Data quality assessment from the user's perspective. in International Workshop on Information Quality in Information Systems, (IQIS2004). 2004. Paris, Francia: ACM. p. 68-73.

[2] A. Caro, C. Calero, I. Caballero, and M. Piattini. Defining a Data Quality Model for Web Portals. in WISE2006, The 7th International Conference on Web Information Systems Engineering. 2006. Wuhan, China: Springer LNCS 4255. p. 363-374.

[3] M. Eppler and P. Muenzenmayer. Measuring Information Quality in the Web Context: A Survey of State-of-the-Art Instruments and an Aplication Methodology. in Proceeding of the Seventh International Conference on Information Quality. 2002. p. 187-196.

[4] M. Ivory, S. Rashmi, and H. Marti. Empirically Validated Web Page Design Metrics. in SIG-CHI on Human factors in computing systems (SIGCHI'01). 2001. Seattle, WA, USA. p. 53-60.

[5] J. Nielsen, *Designing Web Usability: The Practice of Simplicity*. 2000, Indianapolis: New Riders Publishing.

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

[7] L. Xiao and S. Dasgupta, *User Satisfaction with Web Portals: An empirical Study*, in *In Web Systems Design and Online Consumer Behavior*, Y. Gao, Editor. 2005, Idea Group Publishing, Hershey, PA. p. 193-205.

[8] Z. Yang, S. Cai, Z. Zhou, and N. Zhou, Development and validation of an instrument to measure user perceived service quality of information presenting Web portals. Information and Management. Elsevier Science, 2004. 42: p. 575-589.

IEEE Computer Society Conference Publications Operations Committee

CPOC Chair

Phillip Laplante
Professor, Penn State University

Board Members

Thomas Baldwin, *Manager, Conference Publishing Services* (CPS)
Mike Hinchey, *Director, Software Engineering Lab, NASA Goddard*
Paolo Montuschi, *Professor, Politecnico di Torino*
Linda Shafer, *Professor Emeritus, University of Texas at Austin*
Jeffrey Voas, *Director, Systems Assurance Technologies, SAIC*
Wenping Wang, *Associate Professor, University of Hong Kong*

IEEE Computer Society Executive Staff

Angela Burgess, *Publisher*

IEEE Computer Society Publications

The world-renowned IEEE Computer Society publishes, promotes, and distributes a wide variety of authoritative computer science and engineering texts. These books are available from most retail outlets. Visit the CS Store at *http://www.computer.org/portal/site/store/index.jsp* for a list of products.

IEEE Computer Society *Conference Publishing Services* (CPS)

The IEEE Computer Society produces conference publications for more than 200 acclaimed international conferences each year in a variety of formats, including books, CD-ROMs, USB Drives, and on-line publications. For information about the IEEE Computer Society's *Conference Publishing Services* (CPS), please e-mail: cps@computer.org or telephone +1-714-821-8380. Fax +1-714-761-1784. Additional information about *Conference Publishing Services* (CPS) can be accessed from our web site at: *http://www.computer.org/cps*

IEEE Computer Society / Wiley Partnership

The IEEE Computer Society and Wiley partnership allows the CS Press *Authored Book* program to produce a number of exciting new titles in areas of computer science and engineering with a special focus on software engineering. IEEE Computer Society members continue to receive a 15% discount on these titles when purchased through Wiley or at: *http://wiley.com/ieeecs*. To submit questions about the program or send proposals, please e-mail dplummer@computer.org or telephone +1-714-821-8380. Additional information regarding the Computer Society's authored book program can also be accessed from our web site at: *http://www.computer.org/portal/pages/ieeecs/publications/books/about.html*

Revised: 16 March 2007

New *CPS Online* Workspace

An IEEE Online Collaborative Publishing Environment

CPS Online is a new IEEE online collaborative conference publishing environment designed to speed the delivery of price quotations and provide conferences with real-time access to all of a project's publication materials during production, including the final papers. The ***CPS Online*** workspace gives a conference the opportunity to upload files through any Web browser, check status and scheduling on their project, make changes to the Table of Contents and Front Matter, approve editorial changes and proofs, and communicate with their CPS editor through discussion forums, chat tools, commenting tools and e-mail.

The following is the URL link to the CPS Online Publishing Inquiry Form:
http://www.ieeeconfpublishing.org/cpir/inquiry/cps_inquiry.html