6th International Conference on the Quality of Information and Communications Technology

September 12-14

12-14 September 2007

LISBON, PORTUGAL

Supported by

Universidade do Minho

FCT
FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Edited by

Ricardo J. Machado

Fernando Brito e Abreu

Paulo Rupino da Cunha

Proceedings

QUATIC 2007

6th International Conference on the Quality of Information and Communications Technology

12-14 September 2007
Lisbon, Portugal

Proceedings

QUATIC 2007

6th International Conference on the Quality of Information and Communications Technology

12-14 September 2007
Lisbon, Portugal

Edited by
Ricardo J. Machado
Fernando Brito e Abreu
Paulo Rupino da Cunha

CPS Conference Publishing Services

Los Alamitos, California

Washington • Tokyo

IEEE computer society

IEEE Computer Society Order Number P2948
ISBN-10: 0-7695-2948-8
ISBN-13: 978-0-7695-2948-6
Library of Congress Number 2007930584

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 Patrick Kellenberger
Cover art production by Joe Daigle/Studio Productions
Printed in the United States of America by Documation LLC

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

QUATIC 2007

6th International Conference on the Quality of Information and Communications Technology

Table of Contents

Invited Papers

Regular Papers

Session on Software Product Quality

Session on Software Methodologies and Quality Assurance

Session on Software Process Improvement Methods

Session on Middleware and Service Architectures

Session on Quality in the Web

Session on Software Process Improvement in Action

Short Papers

SEDES 2007 Workshop

Sixth International Conference on the Quality of Information and Communications Technology

A Probabilistic Approach to Web Portal's Data Quality Evaluation

Angélica Caro[1], Coral Calero[2], Emilia Mendes[3] and Mario Piattini[2]

[1]Dep. Auditoria e Informática,
University of Bio Bio, Chile
mcaro@ubiobio.cl

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

[3]University of Auckland, New Zealand
emilia@cs.auckland.ac.nz

Abstract

Advances in technology and the use of the Internet have favoured the emergence of a large number of Web applications, including Web Portals. Web Portals provide the means to obtain a large amount of information therefore it is crucial that the information provided is of high quality. In recent years, several research projects have investigated Web Data Quality; however none has focused on data quality within the context of Web Portals. Therefore, the contribution of this research is to provide a framework centred on the point of view of data consumers, and that uses a probabilistic approach for Web portal's data quality evaluation. This paper shows the definition of operational model, based in our previous work.

1. Introduction

Web portal is a site that aggregates information from multiple sources on the World Wide Web and organizes this material in an easy user-friendly manner [30]. Over the past decade the number of organizations that provide Web portals has grown dramatically. They provide portals that complement, substitute or extend existing services to their client base [31]. Numerous users worldwide make use of Web portals to inform 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 they need. Likewise, the organizations that provide Web portals need to offer data that meet user requirements and help them achieve their goals. Therefore data quality represents a common interest between data consumers and portals' providers.

In the literature, the concept of Data or Information Quality (DQ hereafter) is often defined as "fitness for use", i.e., the ability of a data collection to meet user requirements [4, 28]. Besides, the terms "data" and "information" are often used as synonyms. In this work we will also use them as synonymous.

In recent years, and due to the specific characteristics of Web applications and their differences from traditional information systems, there research community started to look into the area of DQ on the Web [11]. However, although some studies suggest that DQ is one of the relevant factors when measuring the quality of a portal [20, 31], few address DQ in Web portals. However, as far as the assessment of DQ is concerned, the current view is that it must be focused on the users (data consumers) point of view [3, 4, 9, 13, 29]. The data consumer perspective differs from both the data producer and data custodian perspectives in two important aspects [3]: (1) Data consumers have no control over the quality of available data and (2) the aim of consumers is to find data that match their personal needs, rather than to provide data that meet the needs of others.

Consequently, our research aims to create a Data Quality Model for Web portals focused on the data consumer's perspective. To this end, we divided our work into two parts. The first consisted in the definition of a theoretical model named Portal Data Quality Model, PDQM(t), published on [5]. This resulted in the identification of 33 DQ attributes than can be used to assess a portal's DQ. The second, not finished yet, is to convert PDQM(t) into an operational model, to be used to dynamically assess Web portals' DQ. To reach this goal, the PDQM(t)'s DQ attributes need to be specified in an operational way. This means that we need to define a structure that organizes the DQ attributes, and to which we can associate measures and criteria.

0-7695-2948-8/07 $25.00 © 2007 IEEE
DOI 10.1109/QUATIC.2007.10

Considering the subjectivity of data consumers' perspective and the uncertainty inherent to the quality perception [7], we chose to use a probabilistic approach by means of Bayesian networks to convert PDQM(t) into an operational model. A Bayesian network (BN) is a directed acyclic graph where nodes represent variables (factors) and arcs represent dependence relations between variables. Nodes can embody different types of variables (e.g. observable or latent, categorical, numerical), which do not need to be random. Arcs in a BN connect parent to child nodes, where a child node's probability distribution is conditional on its parent node's distribution. Arcs, nodes and probabilities can be elicited from experts and/or empirical data, and probabilities are conveyed using Node probability tables (NPTs) that are associated to nodes. BNs combine the advantages of an intuitive representation with a sound mathematical basis in Bayesian probability [24]. Building a BN is a three-fold process: first, we need to build the graph structure; second, build the NPTs; and third, to validate both structure and NPTs.

In this paper we will show the creation process of the BN that will support the PDQM, including its validation. This process is an essential aspect in order to achieve the appropriated structure and configuration that allow the assessment of DQ according with the data consumer's point of view. We think that our proposal is an important contribution in an area where as emphasize Gertz et al. in [11] "well-founded and practical approaches to assess or even guarantee a required degree of the quality of data are still missing".

The rest of the paper is organized as follows. Section 2 presents the development process of PDQM, followed by the definition of the theoretical model PDQM(t) in Section 3. Section 4 describes the approach used to convert PDQM(t) to an operational model, the steps developed for the transformation, and the first validation of the model. Finally, conclusions are given in Section 5.

2. Portal Data Quality Model (PDQM)

PDQM is a data quality model for Web portals focused on the data consumer perspective. It is based on three key aspects:

- **Data consumer perspective**. In the late 1990s, the most frequent definition of quality was that of meeting and exceeding customers' expectations [26]. The notion of quality as meeting expectations suggests that quality is defined by conformance to customers' expectations. This situation remains unchanged within the context of data quality; most authors define this concept as "fitness for use" [4, 28], suggesting that quality of data cannot be assessed independently from the users who use data. We employed the DQ expectations of the data consumer on the Internet, proposed in [25] as means to include the data consumers' perspective in our model. These expectations are organized into six categories: Privacy, Content, Quality of values, Presentation, Improvement, and Commitment.
- **Web data quality attributes.** Obtained from DQ frameworks proposed in the literature for different Web contexts. We took advantage of previous work applied to the Web and extended it to Web portals.
- **Web portal functionalities. Under our perspective**, we assume that data consumers judge DQ based on the quality of functionality provided in a Web portal. Therefore we considered the Web portal software functions proposes in [6] as the baseline in our model. These functions are as follows: Data Points and Integration, Taxonomy, Search Capabilities, Help Features, Content Management, Process and Action, Collaboration and Communication, Personalization, Presentation, Administration, and Security.

2.1. The development process of PDQM

To produce the PDQM, we defined a process divided in two parts (see Figure 1). The first part corresponds to the theoretical definition of the model, PDQM(t). Its main goal is to obtain a set of DQ attributes that can be used to evaluate Web portals' DQ, from the data consumers' perspective. The second part consists on the transformation of PDQM(t) into an operational model, to define a structure for organizing DQ attributes by associating measures and criteria with them.

The first part comprises four phases. The first phase compiles from previous literature Web DQ attributes that are in our view pertinent to Web portals. The second phase comprises the building of a matrix to classify the DQ attributes obtained in the previous phase. The third phase uses the matrix created in the second phase to assess the applicability of each Web DQ attribute to a Web portal. Finally, the fourth phase validates the model built in phase 3. The result of the first part of our process is used as an input to the second part (detailed in Section 3), which is used to transform the theoretical PDQM into an operational model. To do so, we decided to use a probabilistic approach (discussed in section 4).

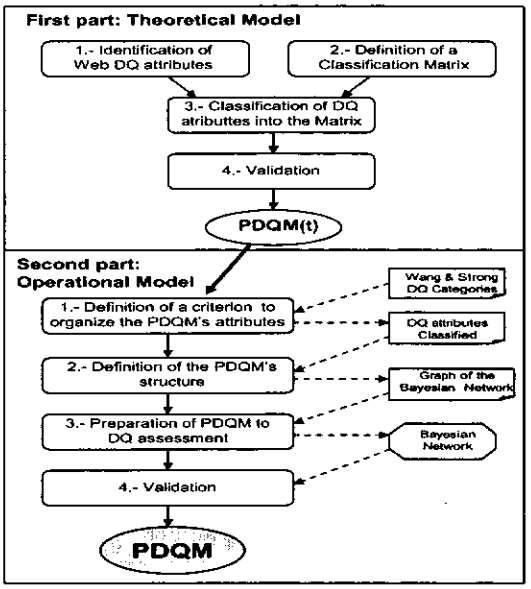


Figure 1. The development process of PDQM.

Four phases comprise this second part. The first phase defines a criterion to organize the PDQM's DQ attributes. The second phase generates the PDQM's graphical structure. The third phase is used to build a model to be used to evaluate the DQ. Finally, the fourth phase validates the model proposed in the third phase.

3. The theoretical model PDQM(t)

Next, we briefly explain each of the phases used to obtain PDQM(t). A more detailed description is given in [5].

3.1. Identification of 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 [16]. Then, we selected previous work proposed for different domains in the Web context (Web sites [8, 14, 21], integration of data [2, 22], e-commerce [15], Web information portals [31], cooperative e-services [10], decision making [12], organizational networks [19] and Data Quality on the Web [11]). As a result of that, and after summarizing the collected initial set of attributes (one hundred), we obtained 41 DQ attributes (see first column in table of Figure 4).

3.2. Definition of a Classification Matrix for Web DQ attributes

In the second phase, we used a matrix to classify the DQ attributes obtained in the previous phase. This matrix associates the two basic aspects considered in our model: the data consumers' perspective by means of their data quality expectations on Internet [25] and the basic functionalities offered in a Web portal [6]. Once the matrix was populated, we ticked the expectations applicable to each of the different functionalities of a Web portal (see Figure 2).

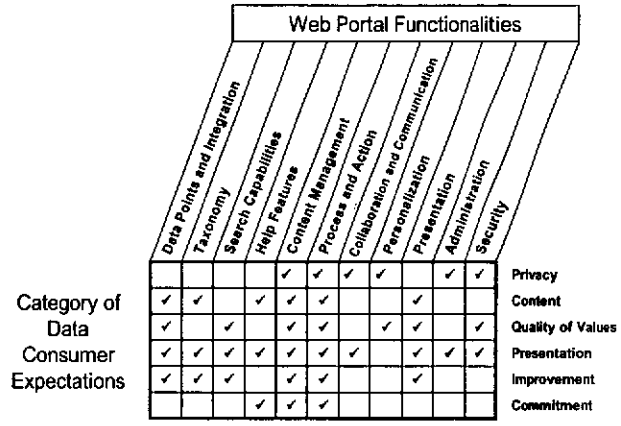


Figure 2. Matrix to classify Web DQ attributes.

3.3. Classification of Web DQ attributes in the Matrix

The third phase used the matrix obtained in the second phase to classify the Web DQ attributes identified in the first phase. Once this classification was complete, we assigned to each relationship between functionality and expectation the DQ attributes that could be used by the data consumer to evaluate the quality of data in a Web portal. This assignment used as basis the appropriateness of each attribute (based on its definition), in relation to the objective of each portal functionality and the user's DQ expectation. Figure 3 shows an example of this classification, and the attributes assigned to each functionality are summarized in the table of Figure 4.

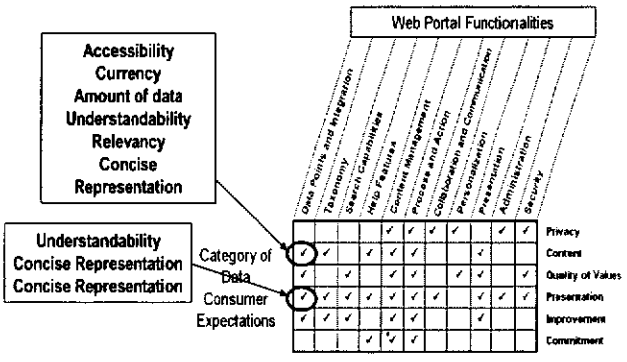


Figure 3. Example of classification of Web DQ attributes into the matrix.

As a result of this phase we identified a set of 34 DQ attributes that can be used for the DQ evaluation in portals, considering the data consumers' perspective.

Functionalities & DQ Attributes	Data Points and Integration	Taxonomy	Search Capabilities	Help Features	Content Management	Process and Action	Collaboration and Communication	Personalization	Presentation	Administration	Security	Number of Functionalities
Accessibility	✓	✓	✓	✓	✓	✓					✓	7
Accuracy	✓					✓		✓			✓	4
Amount of data	✓	✓	✓	✓	✓	✓			✓	✓	✓	9
Applicability					✓	✓						2
Attractiveness									✓			1
Availability					✓	✓	✓					3
Believability	✓	✓	✓		✓	✓			✓			6
Completeness			✓		✓	✓		✓			✓	5
Concise Representation	✓	✓	✓	✓	✓	✓			✓	✓	✓	9
Consistent Representation	✓									✓		2
Cost effectiveness												0
Customer support	✓	✓	✓	✓	✓	✓	✓		✓			8
Currency	✓		✓		✓	✓		✓	✓			6
Documentation					✓							1
Duplicates					✓							1
Ease of operation		✓	✓	✓	✓	✓			✓	✓	✓	8
Expiration	✓					✓		✓	✓			4
Flexibility									✓			1
Granularity												0
Interactivity					✓	✓						2
Internal consistency												0
Interpretability					✓	✓	✓		✓		✓	5
Latency												0
Maintainable												0
Novelty	✓		✓		✓							3
Objectivity					✓	✓						2
Ontology												0
Organization	✓				✓	✓			✓			4
Price												0
Relevancy	✓	✓	✓		✓	✓	✓		✓			7
Reliability	✓	✓	✓	✓	✓	✓			✓			7
Reputation					✓	✓						2
Response time			✓			✓						2
Security					✓	✓	✓			✓	✓	5
Specialization					✓	✓			✓			3
Source's information					✓							1
Timeliness			✓		✓	✓			✓		✓	5
Traceability	✓	✓	✓		✓	✓		✓	✓			7
Understand ability	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	11
Validity	✓	✓		✓	✓	✓		✓	✓		✓	8
Value-added					✓							1
Total of Attributes	16	11	15	8	28	26	6	7	18	6	11	

Figure 4. DQ attributes assigned for functionality

3.4. Validation

The fourth phase comprises the validation of the model obtained in the third phase. To carry out this validation [5], we conducted a survey investigation which purpose was to collect ratings representing the importance of each of the DQ attributes for data consumers. As a result we have obtained a final set of 33 DQ attributes for PDQM(t) (see Table 1).

Table 1
Data quality attributes into PDQM(t)

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

4. Towards PDQM the operational model

So far, we have identified a set of DQ attributes that can be applicable in the Web portals context. However, the definition of a model does not mean that it can be operational, i.e., that it can be used in an assessment process. Indeed, having a set of attributes is not enough to measure/evaluate them or to combine the results of their evaluations to globally assess the quality of the portal data. To reach this goal, we need to find a model that allows (1) taking as input measures collected from the portals, (2) evaluating separately the DQ attributes, and (3) combining these partial evaluations into a global one. In other words, we start by defining a structure to organize the DQ attributes (dependence and definitional relationships). In a second time, we choose/define measures that can be associated to these attributes.

Taking into account all this we believe that BNs fit the explained requirements of our model. To build a BN for PDQM a four-phase process (see Figure 1, second part) has been defined following the stages proposed in [18]: (1) build the graph structure (phase 1 and 2 in Figure 1) and (2) define the node probability tables for each node of the graph as well as measures for the quantifiable variables associated to the each node of the graph (phase 3 in Figure 1). In the rest of this paper the transformation process of PDQM(t) in an operational model will be shown.

4.1. Phase 1: Definition of a criterion to organize PDQM's attributes

As explained in [23], a BN can be built starting

from semantically meaningful units called network fragments. A fragment is a set of related random variables that could be constructed and reasoned about separately from others fragments. Ideally the fragments must make sense to the expert who must be able to supply some underlying motive or reason for them belonging together.

Then, the first phase to build the BN for PDQM, was to define a criterion to create network "fragments". We decided to use the conceptual DQ framework proposed by Wang and Strong [28].

This framework was originally defined for information systems and some aspects inherent to the Web context are not considered, specifically about the role of systems. Then, in our work we have renamed the Accessibility category as Operational category in order to emphasize the importance of the role of systems not only with respect to accessibility and security but also to other aspects as personalization, collaboration, etc; see Table 2, first column.

After this, we classified the DQ attributes of PDQM into these categories; see Table 2, second column. The classification was made considering the literature reviewed, the attributes and categories definitions and the perceptions and experience of the authors. As a result of this phase a BN with two levels and with four network fragments was obtained (one fragment for each one DQ category).

Table 2
Data quality attributes intoPDQM

DQ Category	DQ Attributes
Intrinsic: It denotes that data have quality in their own right.	Accuracy, Objectivity, Believability, Reputation, Currency, Duplicates, Expiration, Traceability
Operational: It emphasizes the importance of the role of systems; that is, the system must be accessible but secure to allow the personalization and collaboration among other aspects.	Accessibility, Security, Interactivity, Availability, Customer support, Ease of operation, Response time
Contextual: It highlights the requirement which states that DQ must be considered in the context of the task in hand.	Applicability, Completeness, Flexibility, Novelty, Reliability, Relevancy, Specialization, Timeliness, Validity, Value-Added
Representational: It denotes that the system must present data in such a way as to be interpretable and easy to understand, as well as concisely and consistently represented.	Interpretability, Understandability, Concise Representation, Consistent Representation, Amount of Data, Attractiveness, Documentation, Organization

4.2. Phase 2: Definition of a structure for PDQM

In this phase, we have generated new levels in the BN based on the relationships of direct influences among the attributes in each category. We used the DQ categories and the DQ attributes definitions, together with our perceptions and experience, to establish these relationships. Our aim was to establish which DQ attribute in a category has direct influence in other attributes in the same category, and eventually in attributes in other category. Each relationship is supported by a premise that represents the direct influence or dependence between an attribute and its parent attribute. In Tables 3, 4, 5 y 6 we summarize the relationships of direct influence, the levels defined to hold the attributes in the BN, and the premise that supports each relationship.

Table 3
Relationships in the intrinsic DQ category

	Relation of Direct Influence		Premise that supports the direct influence relationships
	Level 2	Level 3	
DQ Intrinsic (Level 1)	Accuracy	Duplicates	If a portal delivers duplicates then data cannot be certified free of error. Data consumers can consider not reliable the data delivered.
	Believability	Objectivity	If data are objective (i.e. impartial) then it is more probable that data would be accepted as correct.
		Reputation	If data are trusted in terms of their source or content then it is probable that data and their source can be accepted as correct by data consumers.
		Traceability	If a portal delivers information about the author/owner then data will be more easily traceable. If data are traceable then it is more probable that data can be accepted as correct.
	Currency	Expiration	If the expiry date is known, then it is easier to determine the currency of data.

Table 4
Relationships in the operational DQ category

	Relation of Direct Influence		Premise that supports the direct influence relationships
	Level 2	Level 3	
DQ Operational (Level 1)	Security	-	If data are secure then the security of the system in general will be influenced.
	Accessibility	Interactivity	If data are accessed or retrieved according to one's personal preferences then data are more accessible.
		Ease of Operation	If data can be easily managed and manipulated then data are more accessible.
		Customer Support	If a Web portal provides on-line support then data are more accessible.
	Availability	Response Time	If the response time to obtain data is appropriate then data are considered more available by users.

Table 5
Relationships in the contextual DQ category

	Relation of Direct Influence		Premise that supports the direct influence relationships
	Level 2	Level 3	
DQ Contextual (Level 1)	Validity	Reliability	If users can trust data and their source then this will influence their perception about their validity.
		Completeness	If data are sufficiently wide, deep and concerning the task to be developed, then they can be accepted as valid.
	Value Added	Applicability	If data are specific, useful and easily applicable for the target community then their use will be more beneficial and will provide advantages.
		Flexibility	If data are expandable, adaptable and easily applicable to other needs then they will have more value added for data consumers.
		Novelty	If the data obtained from the portal influence the knowledge and the new decisions then they will have value for users.
	Relevancy	Novelty	If the data obtained from the portal influence the knowledge and the new decisions then they will be relevant.
		Timeliness	If data are available on time then they will be relevant for data consumers.
		Specialization	If the data obtained from the portal are specific for users' interest then they will be relevant

Table 6
Relationships in the representational DQ category

	Relation of Direct Influence		Premise that supports the direct influence relationships
	Level 2	Level 3	
DQ Representational (Level 1)	Concise Representation	-	If data are compactly represented without superfluous elements then they will be better represented.
	Consistent Representation	-	If data are always presented in the same format, are compatible with previous data and consistent with other sources, then they will be better represented.
	Understandability	Interpretability	If data are appropriately presented in language and units for users' capability then they will be understood better.
		Amount of data	If the quantity or volume of data delivered by the portal is appropriate then they will be understood better.
		Documentation	If data have useful documents with meta information then they will be understood better.
		Organization	If data are organized with a consistent combination of visual settings then they will be understood better.
	Attractiveness	Organization	If data are organized with a consistent combination of visual settings then they will be more attractive to data consumers.

As a result of this phase, we obtained the graph of a BN that represents the structure for all the PDQM's DQ attributes, see Figure 5.

4.3. Phase 3: Preparation of PDQM to DQ assessment

Once organized the attributes of PDQM into a BN, the following phase consisted of the preparation of the BN to assess DQ in Web portals. Although our final objective is to create a comprehensive BN model for PDQM, we have decided to work separately with each one of the four fragments in the model. For each one, the following sub-phases may be developed:

a. If necessary, to create synthetic nodes to simplify the sub-network, i.e., to reduce the number of parents for each node.

b. To define quantifiable variables for the entry nodes in the sub-network.

c. To define the node probability tables for each intermediate node in the sub-network. This definition depends on the Web portal domain.

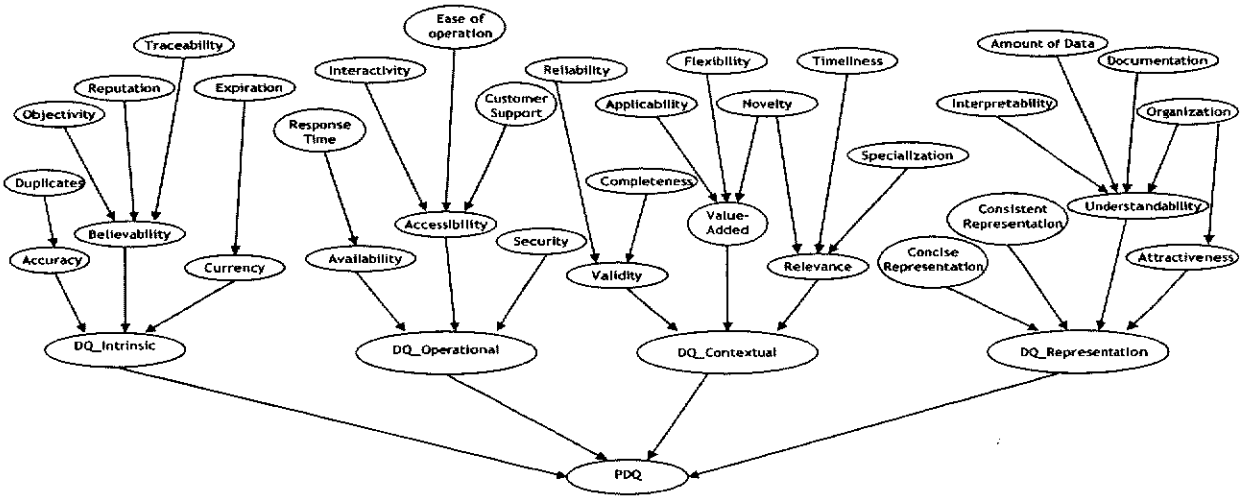


Figure 5. Graph of the BN that represents PDQM

Next, we will develop of sub-phases a, b and c for fragment DQ_Representational and for the domain of university portals.

4.3.1. Simplification of the sub-network (sub-phase a). The original sub-network had two nodes with four parents (Understandability and DQ_Representational) and it was necessary to create two artificial nodes (Representation and Volume of Data) in order to reduce the combinatory explosion in the next step during the preparation of the node probability tables. (See Figure 6).

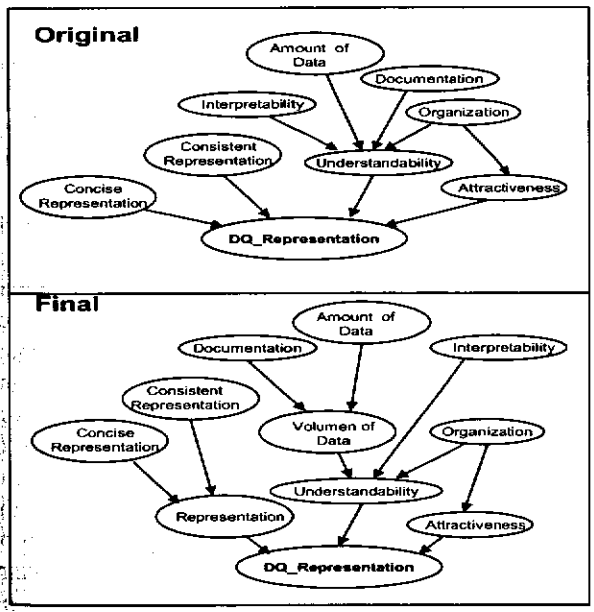


Figure 6. Graph for DQ Representational fragment.

4.3.2. Definition of quantifiable variables (sub-phase b). In this sub-phase measures for the quantifiable variables (input nodes) in the fragment were defined. Thus we have defined an indicator for each input node in the sub-network (see the indicators LCsR, LCcR, LD, LAD, LO and LI in Figure 7, in the last level). To calculate each indicator, we have defined several measures that will be automated and calculated for a given portal. The indicators defined will take a numerical value from between 0 to 1. As the number of possible values for each input node can be infinite, we have transformed them into discrete variables. This is done to ease the definition of probabilities. According to [18] this transformation can be achieved using fuzzy logic. So, in our case we have defined for each indicator a membership function that transforms the value of the indicator into a set of probabilities for each label/class

To show how this process works, we next explain the definition of the LCsR (Level of Consistent Representation) indicator, which generates a measure for the input node Concise Representation. The measures selected for this attribute take as their focal point the consistency of the format and the compatibility between the pages in the portal. This is not only because these aspects are more obvious for data consumers when this attribute is evaluated, but also because they can be measured in an objective way.

We have defined measures based on the use of Style on the pages of the Web portal for this indicator, as well as on the correspondence between a source page and the destination pages.

One type of correspondence measured, for instance, was if the text associated with a link was repeated on a

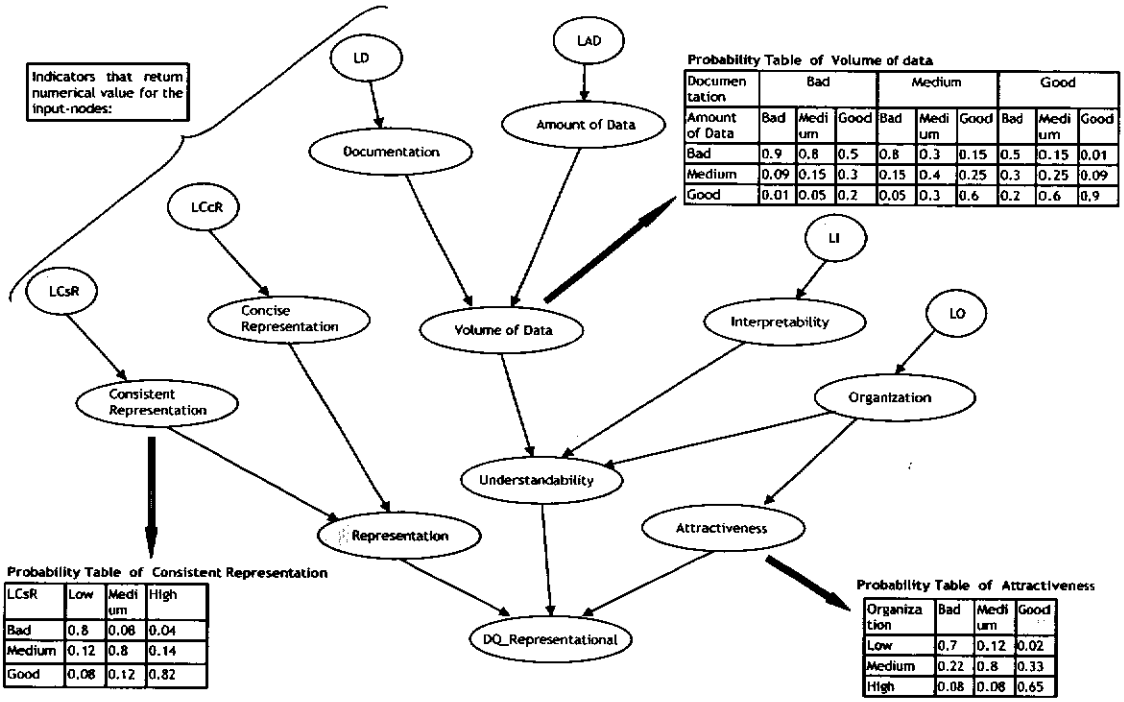
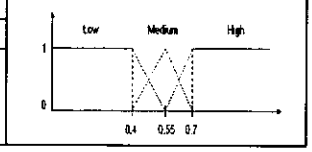


Figure 7. Preparation of fragment DQ_Representational for evaluation process

destination page. A set of base and derived measures had been defined to carry out that task, thus taking these measures as a starting point, we defined their *Analysis Model* [1], which includes an *Equation* that gives a numerical value and a *Decision Criteria* in the form of a membership function; see Table 7.

Table 7
Analysis model for LCsR indicator

LCsR (Level of Consistent Representation)	
Formula	Decision Criteria
LCsR = PSSD*0.5 + SDCD*0.5	
Derived Measures	
PSSD: Pages with the same Style SDCD: Source and Destination Correspondence	

4.3.3. Definition of node probability tables (sub-phase c). We defined a node probability table for each BN node. These probabilities were given by expert judgments and taking account the Web portal domain selected. Figure 7 shows the node probability tables for some nodes in the fragment.

4.4. Phase 4: Validation

There are three different ways which we can use to validate a BN b. First, it can be validated with experts who use different testing scenarios to check if the probabilities provided for selected nodes are suitable; second, it can be validated using data gathered from experiments, where each data point is used as evidence and we assess if the probabilities provided for certain nodes by the BN are suitable; third, we can gather data from experiments to be used to automatically build another BN n and to populate its node probability tables, which are later compared to **b**.

Within the context of this paper we have employed the third option because we wanted to investigate to what extent a BN elicited from experts would be similar to a BN automatically generated from empirical data; however the other two options will also be carried out as part of our future work.

Data was gathered by means of an empirical study where 79 subjects (data consumers) were asked to complete a series of tasks using a real Web portal. These subjects were final year Computer Science undergraduate students enrolled on a BSc Programme at the University of Castilla-La Mancha, Spain. All the subjects had previous experience with using Web portals as data consumers. The instrument used to gather the data comprised a document containing instructions, the motivation for the empirical study, the URL of the Web portal to be used, three tasks to be carried out using this Web portal, and a questionnaire containing nine questions. Once these tasks were completed, subjects were asked to assess, using the questionnaire, the portal's data quality for each of the DQ's attributes previously associated with the DQ's

'Representational' category (see Table 6). They were also asked to provide a global assessment of the Web portal's DQ.

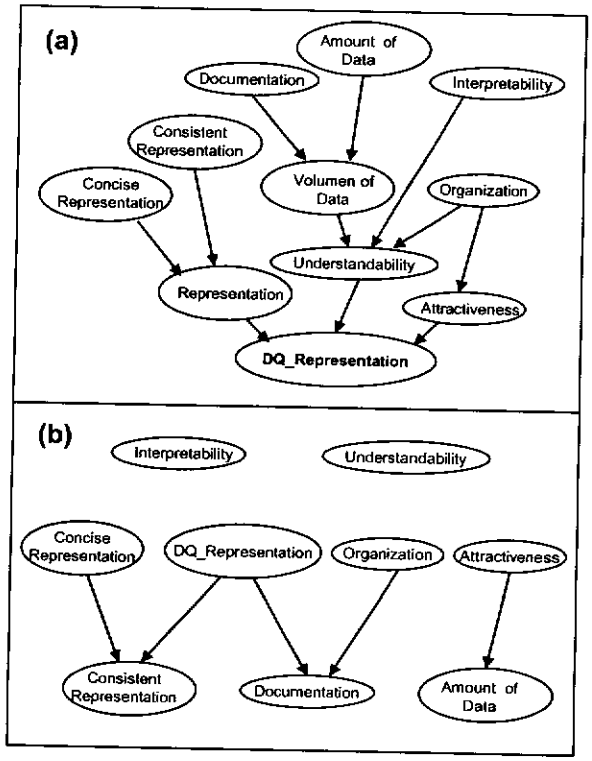


Figure 8. Preparation of fragment DQ_Representational for evaluation process

The data gathered from the empirical study was used as input to a Bayesian Network tool – Hugin Researcher 6.7 from Hugin Expert A/S. There was no need for discretisation of the variables as they were all measured using a three-point ordinal scale. The BN's structure was automatically obtained using the Necessary Path Condition (NPC) algorithm [27]. In addition, prior and conditional probabilities were automatically generated using the EM-learning algorithm [17]. Figure 8 shows that there are clear differences between both BNs, in particular in regard to nodes 'Interpretability' and 'Understandability', which, in the data-based BN (see Figure 8(b)) do not seem to have any relationship with DQ_Representational. The relationship between 'Consistent Representation' and DQ_Representational was the only one kept on both BNs, although with different directions. The node 'Organization' changed from being indirectly related with 'DQ_Representational' (see Figure 8(a)), to having a direct causal effect on 'DQ_Representational'.

These results suggest that the perceptions of users (data consumers) and experts in relation to what DQ attributes affect DQ_Representational and their causal relationships seem to differ. However, to ensure that the trends we have observed persist, we also need to carry out a case study using data consumers other than students, and possibly ask the same students who participated in the case study described in this paper to elicit a DQ_Representational BN. This will enable us to assess if data quality requisites actually change depending on the type of consumers and domain experts.

To what extent were the results presented here influenced by the way in which the HUGIN-based BN was created? HUGIN used well-known standard algorithms for structure building and probabilities generation. However, there are BN tools that use proprietary algorithms which could have led to a BN model slightly different from that obtained using HUGIN. Therefore the choice of BN tool may also be influential upon the results.

5. Conclusions

In this paper we have introduced a DQ model for Web portals (PDQM) centered on the data consumer perspective, and have compared two views of the model, one from experts, and another from data consumers. This model was generated using a two-part process. Part One encompasses the definition of a theoretical model PDQM(t) comprising a set of 33 DQ attributes that can be used to evaluate Web portals' DQ. Part Two encompasses the transformation of PDQM(t) into an operational DQ model, which provides a probabilistic approach by means of a Bayesian Network. We chose to use a Bayesian Network because the causal system relative to Web portal's DQ has an inherently uncertain nature, which accommodates the following requisites:

- Genericity. PDQM must be applicable to any Web portal.
- Adequacy. PDQM is oriented towards the data consumer point of view. It must support the subjective and uncertainty associated with DQ evaluations.
- Flexibility. It must be applicable to different situations (different Web portal domains, different kinds of data consumers, etc.).
- Completeness. The knowledge structure must enable the representation of all the relationships between attributes, e.g., an attribute can simultaneously affect several other attributes.

The PDQM's BN was built using experts' opinion; however to validate this BN we decided to build another BN solely based on data consumer's opinions/perceptions. Therefore a case study was used to obtain data on a Web portal's DQ, used to automatically build a BN to be compared to the BN elicited by experts. Differences between these BNs were identified, and future work is necessary to corroborate or refute these results. This part of our future work will be done by empirical studies with data consumers that allow us to: i) validate the BN obtained by expert's judgements (both the structure and the probability tables). For doing so we are currently implementing a tool to automate the calculation of the quantifiable variables; ii) obtain more data to be used to generate a BN automatically.

The common objective here is to ensure that the structures (BNs) are refined and validated until they are stable. Once we achieve this objective, our next step will be to combine both BNs in order to have a unique BN structure, together with the necessary probability tables that takes the perception of DQ from users and from experts, and that could be used for the assessment of a Web portal's DQ.

6. Acknowledgment

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] Bertoa, M., García, F., y Vallecillo, A. An Ontology for Software Measument, in Ontologies for Software Engineering and Software Technology, Calero, C., Ruiz, F., y Piattini, M., Eds., (2006).

[2] Bouzeghoub, M. y Peralta, V. A Framework for Analysis of data Freshness, International Workshop on Information Quality in Information Systems, (IQIS2004), Paris, France, (2004), pp. 59-67.

[3] Burgess, M., Fiddian, N., y Gray, W. Quality Measures and The Information Consumer, Proceeding of the Ninth International Conference on Information Quality, (2004), pp. 373-388.

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

[5] Caro, A., Calero, C., Caballero, I., y Piattini, M. Defining a Data Quality Model for Web Portals, WISE2006, The 7th International Conference on Web Information Systems Engineering, Wuhan, China, (2006), pp. 363-374.

[6] Collins, H. Corporate Portal Definition and Features: AMACOM, (2001) p.

[7] Eppler, M. Managing Information Quality: Increasing the Value of Information in Knowledge-intensive Products and Processes: Springer, (2003)

[8] Eppler, M., Algesheimer, R., y Dimpfel, M. 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, (2003), pp. 108-120.

[9] Even, A., Shankaranarayanan, G., y Watts, S. Enhancing Decision Making with Process Metadata: Theoretical Framework, Research Tool, and Exploratory Examination, 39th Annual Hawaii International Conference on System Sciences (HICSS'06), (2006), pp. 209a.

[10] Fugini, M., Mecella, M., Plebani, P., Pernici, B., y Scannapieco, M. Data Quality in Cooperative Web Information Systems, 2002.

[11] Gertz, M., Ozsu, T., Saake, G., y Sattler, K.-U. Report on the Dagstuhl Seminar "Data Quality on the Web", SIGMOD Record, vol. 33, N° 1, (2004), pp. 127-132.

[12] Graefe, G. 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, (2003), pp. 133-146.

[13] Herrera-Viedma, E., Pasi, G., y Lopez-Herrera, A. Evaluating the Information Quality of Web Sites: A Quality Methodology Based on Fuzzy Computing with Words, Journal of American Society for Information Science and Technology, 54, (2006), pp. 538-549.

[14] Katerattanakul, P. y Siau, K. Measuring Information Quality of Web Sites: Development of an Instrument, Proceeding of the 20th

International Conference on Information System, (1999), pp. 279-285.
[15] Katerattanakul, P. y Siau, K. Information quality in internet commerce desing, in Information and Database Quality, Piattini, M., Calero, C., y Genero, M., Eds.: Kluwer Academic Publishers, (2001).
[16] Kitchenham, B. Procedures for Performing Systematic Reviews, 0400011T.1, (2004) p.
[17] Lauritzen, S. L. The EM algorithm for graphical association models with missing data, Computational Statistics & Data Analysis, 19, (1995), pp. 191-201.
[18] Malak, G., Sahraoui, H., Badri, L., y Badri, M. Modeling Web-Based Applications Quality: A Probabilistic Approach, 7th International Conference on Web Information Systems Engineering, Wuhan, China, (2006), pp. 398-404.
[19] Melkas, H. Analyzing Information Quality in Virtual service Networks with Qualitative Interview Data, Proceeding of the Ninth International Conference on Information Quality, (2004), pp. 74-88.
[20] Moraga, M. Á., Calero, C., y Piattini, M. Comparing different quality models for portals, Online Information Review., Vol. 30, (2006), pp. 555-568.
[21] Moustakis, V., Litos, C., Dalivigas, A., y Tsironis, L. Website Quality Assesment Criteria, Proceeding of the Ninth International Conference on Information Quality, (2004), pp. 59-73.
[22] Naumann, F. y Rolker, C. Assesment Methods for Information Quality Criteria, Proceeding of the Fifth International Conference on Information Quality, (2000), pp. 148-162.
[23] Neil, M., Fenton, N. E., y Nielsen, L. Building large-scale Bayesian Networks, The Knowledge Engineering Review, 15(3), (2000), pp. 257-284.
[24] Pearl, J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference: Morgan Kaufmann, (1988)
[25] Redman, T. Data Quality: The field guide. Boston: Digital Press, (2000)
[26] Reeves, C. y Bednar, D. Defining quality: Alternatives and implications, Academy of Management Review, 19, (1994), pp. 419-445.
[27] Steck, H. y Tresp, V. Bayesian Belief Networks for Data Mining, Proceedings of The 2nd Workshop on Data Mining und Data Warehousing, Sammelband, (1999).
[28] Strong, D., Lee, Y., y Wang, R. Data Quality in Context, Communications of the ACM, Vol. 40, N° 5, (1997), pp. 103 -110.
[29] Wang, R. y Strong, D. Beyond accuracy: What data quality means to data consumers, Journal of Management Information Systems; Armonk; Spring 1996, 12, (1996), pp. 5-33.
[30] Xiao, L. y Dasgupta, S. User Satisfaction with Web Portals: An empirical Study, in In Web Systems Design and Online Consumer Behavior, Gao, Y., Ed.: Idea Group Publishing, Hershey, PA, (2005), pp. 193-205.
[31] Yang, Z., Cai, S., Zhou, Z., y Zhou, N. Development and validation of an instrument to measure user perceived service quality of information presenting Web portals, Information and Management. Elsevier Science, 42, (2004), pp. 575-589.