A Bayesian network to represent a data quality model

Article *in* International Journal of Information Quality · January 2007

DOI: 10.1504/IJIQ.2007.016392 · Source: DBLP

CITATIONS
3

READS
257

4 authors, including:

Angelica Caro
University of Bío-Bío
53 PUBLICATIONS **400** CITATIONS
SEE PROFILE

Coral Calero
University of Castilla-La Mancha
193 PUBLICATIONS **2,799** CITATIONS
SEE PROFILE

Mario Piattini
University of Castilla-La Mancha
1,123 PUBLICATIONS **13,225** CITATIONS
SEE PROFILE

A Bayesian network to represent a data quality model

Angélica Caro*

Department of Computer Science and Information Technologies,
University of Bio Bio, Chillán, Chile
Fax: 34 926 295 354 E-mail: mcaro@ubiobio.cl
*Corresponding author

Coral Calero

Information Systems and Technologies Department,
Alarcos Research Group,
UCLM-INDRA Research and Development Institute,
University of Castilla-La Mancha,
Paseo de la Universidad, 4, Ciudad Real, Spain
Fax: 34 926 295 354 E-mail: Coral.Calero@uclm.es

Houari A. Sahraoui

Department d'Informatique et de Recherche Opérationnelle,
Université de Montreal, CP 6128 succ, Centre Ville,
Montréal, QC H3C 3J7, Canada
E-mail: sahraouh@iro.umontreal.ca

Mario Piattini

Information Systems and Technologies Department,
Alarcos Research Group,
UCLM-INDRA Research and Development Institute,
University of Castilla-La Mancha,
Paseo de la Universidad, 4, Ciudad Real, Spain
Fax: 34 926 295 354 E-mail: Mario.Piattini@uclm.es

Abstract: Web portals provide data to many people where data consumers need to assess Data Quality (DQ). In our previous work a Portal Data Quality Model (PDQM) was developed. PDQM is focused on data consumers' perspective and is composed by 33 attributes appropriate for DQ evaluation. Now, we have organised these attributes into a generic and operational structure. Considering the uncertainty inherent in perception of quality, we decided to use a probabilistic approach, using Bayesian Networks (BNs). This paper, explains the definition of the BN structure that supports PDQM.

Within it, users can navigate with ease to find the information they specifically need to perform their operational or strategic tasks quickly and to make speedy decisions (Collins, 2001). Data consumers who use the data offered by these applications must ensure that these are appropriate for the use they need to put them to. It is of fundamental importance, then, to assess the quality of data.

In the related 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 and their differences from the traditional information systems (Pressman, 2001), the research community has begun to study the subject of data quality on the web (Gertz et al., 2004).

However, in a systematic review of the literature (Caro et al., 2005), we have found no works on (DQ) that address the particular context of web portals, in spite of the fact that some works do indeed highlight DQ as one of the relevant factors affecting the quality of a portal (Moraga et al., 2004; Yang et al., 2004). On the other hand, DQ research has gradually begun to approach the DQ from the data consumers' perspective and not only from that of the data producers or data custodians (Wang and Strong, 1996; Burgess et al., 2004; Cappiello et al., 2004; Shankaranarayanan and Cai, 2005; Even et al., 2006; Herrera-Viedma et al., 2006). The data consumers' perspective differs from the other two perspectives in two important aspects (Burgess et al., 2004):

- data consumers have no control over the quality of available data
- 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, in a previous work we defined a Portal Data Quality Model (PDQM) which focused on the perspective of the data consumer (Caro et al., 2006). To produce the PDQM, we have considered three key elements:

- DQ expectations for web data consumers
- a set of web DQ attributes identified in the literature
- The functionalities that a web portal may offer to its users.

Based on these elements, we have defined a model composed of 33 DQ attributes that can be applied when assessing DQ in a web portal.

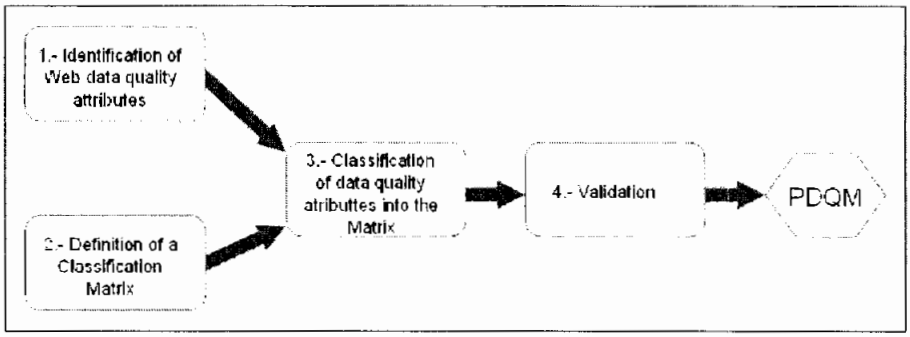
However, defining a model does not necessarily mean that it can be operational. In other words, it must not be taken for granted that it can be actually used to assess DQ of web portals. We thus need to produce an operational version of PDQM that can be used in assessment processes. To this end, we need to define a structure where DQ attributes can be organised, and in which to associate sub-criteria and measures to those attributes.

To map PDQM into an operational model, we decided to use an approach based on Bayesian network. This choice is motivated by the fact that, as remarked in Eppler (2003), any approach to quality (and hence also data quality) has not only an objective component, but also a subjective one. The uncertainty derived from the

- *Web portal functionalities.* Web portals present basic software functionalities so that data consumers can perform their tasks. As we see it, the consumer judges data while actually using the application functionalities. So we used the web portal software functions proposed in Collins (2001) considering these as basic in our model. These functions are the following: Data Points and Integration, Taxonomy, Search Capabilities, Help Features, Content Management, Process and Action, Collaboration and Communication, Personalisation, Presentation, Administration, and Security. Behind these functions, the web portal is really playing the twin role of data producer and data custodian. These functionalities were what we used to analyse the applicability in a web portal of the web DQ attributes gathered from the literature.

Based on these elements, we have developed a four-phase process for producing the PDQM (see Figure 1).

Figure 1 Phases of PDQM development



During the first phase, we gathered web DQ attributes from the literature. They were those that, in our opinion, are pertinent to web portals. In the second phase, we built a matrix for the classification of the DQ attributes obtained in the previous phase. This matrix reflects two basic aspects considered in our model: the data consumer perspective and the basic functionalities which a data consumer uses to interact with a web portal. In our third phase, we used the matrix obtained to analyse the applicability of each web DQ attribute in a web portal. Finally, the fourth phase corresponds to the validation of the model. In the next subsections we will explain each phase.

2.1 Identification of web DQ attributes

This phase consisted of gathering web DQ attributes from the literature. To do so, we carried out a systematic review of the relevant literature (Kitchenham, 2004). We selected works proposed for different domains in the web context (websites (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), organisational networks (Melkas, 2004) and DQ on the web (Gertz et al., 2004)). As a result of this selection and after summarising the initial set of attributes collected, we obtained 41 DQ attributes.

Table 2 Categories of data consumer expectations about the DQ on the internet proposed by Redman (continued)

Category	*Description: 'The data consumer'*
Presentation	Should expect data formats to convey the data properly and that they be easy to read. Unless a format is straightforward, the Consumer should expect to find instructions on reading the data. The Publisher's choice of language will be clear and any technical terms used will be fully defined. In addition, the consumer should expect to be able to interpret data properly if he or she follows instructions
Improvements	*Should expect* to have a means of conveying his or her comments about data, be they good or bad, to the Publisher and that these comments will be acted upon in a responsible manner. The consumer should also expect to be provided with useful summaries of actual quality levels of the data he or she is using and they will be notified if recently published data are abnormally deficient. There should be a summary of performance measurements, indicating the results of improvements
Commitments	*Should expect* to be able to ask any questions regarding the proper use or meaning of data, update schedules, etc., easily and have them answered
	The Data Publisher will be fair and honest and will give him/her the answer to any query. The Consumer should also expect the Data Publisher to adhere to its published policies

Table 3 Web portal functionalities proposed by Collins

Portal functionality	*Description*
Data points and integration	They provide the ability to access information from a wide range of internal and external information sources and display the resulting information at the single point-of-access desktop
Taxonomy	It provides information context (including the organisation-specific categories that reflect and support the organisation's business)
Search capabilities	This provides several services for web portal users and different needs in order to support searches across the company, the World Wide Web, and in search engine catalogues and indexes
Help features	These provide help when using the web portal
Content management	This function supports content creation, authorisation, and inclusion in (or exclusion from) web portal collections
Process and action	This function enables the web portal user to initiate and participate in a business process of a portal owner
Collaboration and communication	This function facilitates discussion, locating innovative ideas, and recognising resourceful solutions
Personalisation	This is a critical component in creating a working environment that is organised and configured specifically for each user
Presentation	It provides the web portal user with both the knowledge desktop and the visual experience that embraces all of the portal's functionality
Administration	This function provides a service for deploying maintenance activities or tasks associated with the web portal system
Security	This provides a description of the levels of access that each user (or groups of users) is allowed for each portal application and software function included in the web portal

each portal functionality and to the DQ expectation of each one of the 38 relationships identified in the matrix. For instance, Figure 3 shows the DQ attributes assigned for the relationships (Data Points and Integration, Content) and (Data Points and Integration, Content). Due to restrictions on space, we do not show the DQ attributes classified in each relationship. We have summarised the attributes assigned by functionality; see Table 4.

Figure 3 Example of classification of web DQ attributes in the matrix

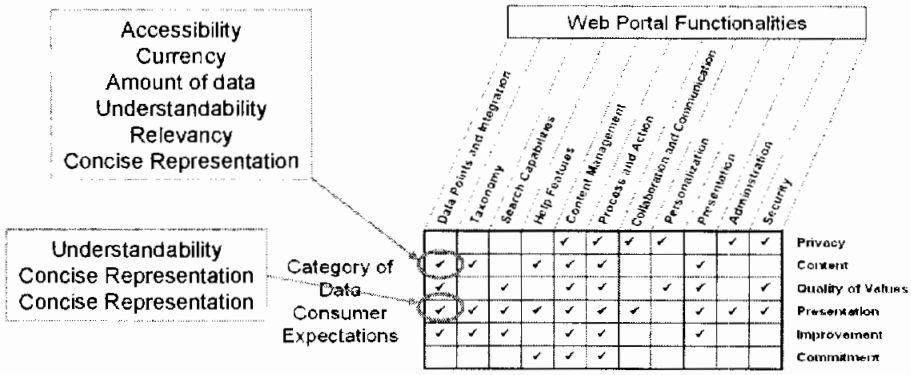


Table 4 Data quality attributes assigned for each functionality

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	Interactivity	Internal consistency	Interpretability	Latency	Maintainable	Novelty	Objectivity	Ontology	Organization	Price	Relevancy	Reliability	Reputation	Response time	Security	Specialization	Source's information	Timeliness	Traceability	Understandability	Validity	Value-added	Total of Attributes
Data Points and Integration	✓	✓	✓				✓		✓	✓		✓	✓				✓								✓			✓		✓	✓						✓	✓	✓		16
Taxonomy	✓		✓				✓		✓			✓				✓														✓	✓						✓	✓	✓		11
Search Capabilities	✓		✓				✓	✓	✓			✓	✓			✓									✓					✓	✓		✓				✓	✓			15
Help Features	✓		✓						✓			✓				✓														✓								✓	✓		5
Content Management	✓		✓	✓		✓	✓	✓	✓			✓	✓	✓	✓	✓				✓		✓			✓	✓		✓		✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	25
Process and Action	✓	✓	✓	✓		✓	✓	✓	✓			✓	✓			✓	✓			✓		✓				✓		✓		✓	✓	✓	✓	✓	✓		✓	✓	✓		25
Collaboration and Communication						✓						✓					✓					✓								✓				✓				✓			8
Personalization		✓						✓					✓				✓																			✓	✓	✓			7
Presentation			✓		✓		✓		✓			✓	✓			✓	✓	✓				✓						✓		✓	✓				✓		✓	✓	✓		16
Administration			✓						✓	✓						✓																	✓				✓				6
Security	✓		✓					✓	✓	✓						✓						✓											✓			✓		✓	✓		11
Number of Functionalities	7	4	9	2	1	3	6	6	9	2	0	8	8	1	1	8	4	1	0	2	0	5	0	0	3	2	0	4	6	7	7	3	2	5	3	1	5	7	11	8	1	

Some attributes have not been assigned, because in our analysis they were found to be non-important or non-visible for the data consumer. Thus the first version of PDQM has 34 DQ attributes (see their definitions on Appendix A).

2.4 Validation

The fourth phase consisted in the validation of the model obtained (third phase). To perform this task, we decided to conduct a study by means of a survey. The purpose of this survey was to collect ratings of the importance of each of the DQ attributes.

- to take measures collected from the portals as input
- to evaluate the DQ attributes separately
- to combine these partial evaluations into an overall one.

In other words, we start by defining a structure for organising the DQ attributes (dependence and definitional relationships). The second thing to for us do is to choose/define measures that can be associated with these attributes.

The definition of our structure has been driven by a set of properties/requirements that the final model must satisfy. The most important are the following:

- *Genericity*. PDQM must be applicable to any web portal.
- *Adequacy*. PDQM is oriented towards the point of view of the data consumer. It must support the subjectivity and uncertainty associated with DQ evaluation.
- *Flexibility*. It must be applicable to different situations. This would include, for example, in different web portal domains, or in processes where the model can be used partially or completely. It might also find application in processes where different kinds of data consumers can be considered. To have this breadth of application, the structure must support the assignation of different weights to the attributes.
- *Completeness*. The structure must allow the representation of all the relationships between the attributes, e.g., an attribute can affect several other attributes simultaneously. In hierarchical models for example, attributes from the same level cannot be related and an attribute cannot affect more than one attribute in the upper level.

A consideration of all these requirements leads us to believe that Bayesian networks are good tools for supporting our model structure.

3.1 A Bayesian Network (BN) for the structuring of the DQ attributes

A Bayesian Network (BN) is a directed a cyclic graph, whose nodes are the uncertain variables and whose edges are the causal or influential links between variables. A conditional probability function models the uncertain relationship between each node and its parents (Neil et al., 2000). BN provides a graphical and intuitive method for capturing the relationships between attributes in a task or domain. In addition, the most distinctive features of BN are their ability to represent changing configurations and respond to them.

In our context, BNs offer an interesting framework with which it is possible to:

- Represent the relationships between DQ attributes intuitively and explicitly, by connecting influencing factors to influenced ones. Such a representation facilitates the comprehension of the model, its validation, its evolution and its exploitation.
- Deal with subjectivity and uncertainty using probabilities.
- Use the network obtained to predict/estimate the data quality of a web portal.
- Isolate factors responsible for when there is low data quality.

Table 6 DQ categories of Wang and Strong's framework

DQ category	*Description*
Intrinsic	It denotes that data have quality in their own right
Accessibility	It emphasises the importance of the role of systems; that is, the system must be accessible but secure
Contextual	It highlights the requirement which states that data quality must be considered within the context of the task in hand
Representational	It denotes that the system must present data in such a way that they are interpretable, easy to understand, and concisely and consistently represented

Keeping in mind the definition of each DQ category, we have classified all data quality attributes of PDQM into these four categories, as shown in Table 7. This classification is consistent with other classifications proposed in the literature, in particular with the one by Wang and Strong (1996).

As a result of this phase we have a two-level BN as a first structure for the PDQM attributes. In the following phase more levels will be generated, determining for each category the relationships between attributes.

Table 7 Classification of DQ attributes of PDQM into the four DQ categories

DQ category	*Dimensions*
Intrinsic	Accuracy, objectivity, believability, reputation, currency, duplicates, expiration traceability
Operational	Accessibility, security, interactivity, availability, customer support, ease of operation, response time
Contextual	Applicability, completeness, flexibility, novelty, reliability, relevancy, specialisation, timeliness, validity, value-added
Representational	Interpretability, understandability, concise representation, consistent representation, amount of data, attractiveness, documentation, organisation

4.2 Phase 2: Building a graph that represents PDQM

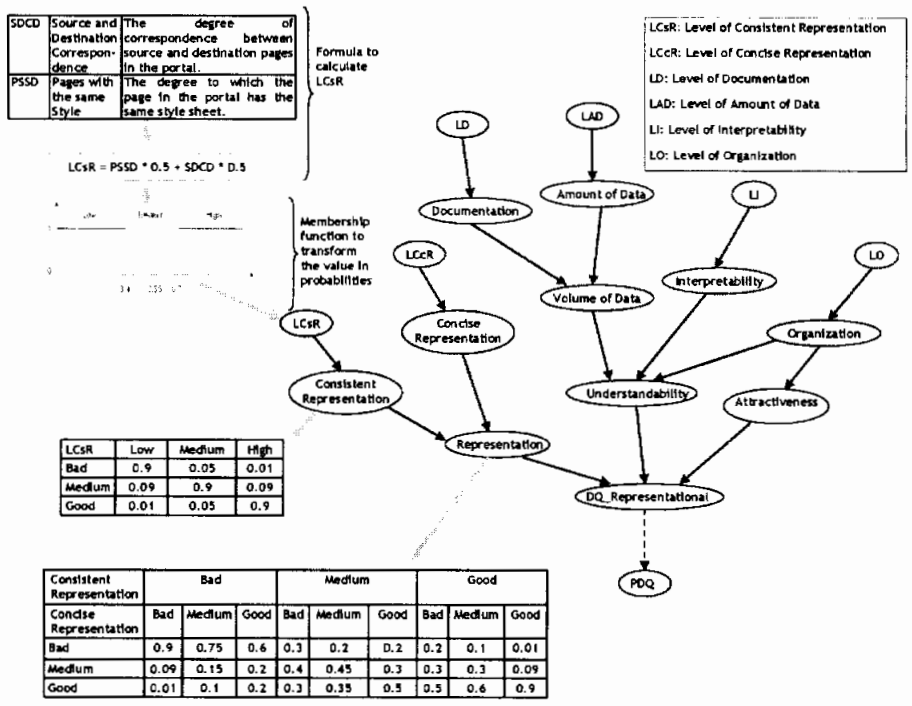
In this phase, we generated new levels in the BN, based on the relationships of direct influences between the attributes in each category. We used the DQ categories and the DQ attribute definitions, together with our perceptions and experience, to establish these relationships. Our aim was to establish which DQ attribute in a category has direct influence on other attributes in the same category, and eventually on attributes in another category. Each relationship is supported by a premise that represents the direct influence or dependence between an attribute and its parent attribute. In Table 8, we will summarise the relationships of direct influence, the levels defined for holding the attributes in the BN, along with the premise that supports each relationship.

Table 8 Relationships of direct influence between the DQ attributes of PDQM (continued)

	Relation of direct influence		
	Level 1	*Level 2*	*Premise that supports the direct influence relationships*
DQ contextual (Level 1)	Validity	Reliability	If users can trust data and their source, this will then influence their perception about the validity of said data
		Completeness	If the data are of sufficient breadth and depth and if they are to do with the task to be undertaken, 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 have more added value for users
		Novelty	If the data obtained from the portal influence knowledge and new decisions, then they will have value for users
	Relevancy	Novelty	If the data obtained from the portal influence knowledge and new decisions then they will be relevant
		Timeliness	If data are available on time then they will be relevant for the data consumer
		Specialisation	If the data obtained from the portal are specific for the users' interest then they will be relevant
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, compatible with previous data and consistent with other sources, then they will be represented better
	Understandability	Interpretability	If data are appropriately presented in language and in units that are appropriate to user 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
		Organisation	If data are organised with a consistent combination of visual settings then they will be understood better
	Attractiveness	Organisation	If data are organised with a consistent combination of visual settings then they will be more attractive for data consumers

quantifiable variable LCsR (Level of Consistent Representation), the formula for calculating it and the membership function for transforming it are shown.

Figure 5 Illustration of how we can use the sub network DQ Representational



After this, two types of probabilities must be defined. These are, in the first place *Input-node probabilities*, which represent attributes that can be directly measured and produced by a transformation of numerical-value measures into probabilities (Figure 5, probability table for Consistent Representation node) and *Intermediate-node probabilities*. The latter are obtained through tables that define conditional probabilities of the different values that can be taken by quality characteristics of the node which knows the values of the characteristics of the parent nodes (Figure 5, probability table for Representation node). These tables are defined using the judgment afforded by experts.

Once this process is complete, the BN will be ready for use in an evaluation process. To show how this process works, we will now explain it for a sub-part of the sub-network (Figure 5, the area marked).

The process is as follows, therefore. For a given web portal we calculate the PSSD and SDCD measures associated with the indicator LCsR. When we apply the formula defined we will obtain a value between 0 and 1. Using the membership function we will transform this value in a set of probabilities for the labels 'Low', 'Medium' and 'High'.

7 Conclusions

In this paper, we have in the first place presented our previous work, the development of a DQ model for web portals. The main contribution of that work is the identification of a set of 33 DQ attributes which are based on the perspective of the data consumer and that can be used for DQ evaluation in web portals. Secondly, we have shown how we have started to transform our model into a DQ evaluation framework, using a BN. The choice of this approach for generating the framework comes from its ability to get around many issues in quality assessment: threshold value definition, measure combination, and uncertainty.

The most important contribution of this work is the organisation of the PDQM attributes into a BN that represents a generic structure, which can be used in different evaluation contexts.

One of the advantages of our framework will be its flexibility. Indeed, the idea is to develop a global framework that can be adapted to both the goal and the context of evaluation. From the point of view of the goal, the user can choose the sub-network that evaluates the characteristics he is interested in. From the contextual point of view, the parameters (probabilities) can be changed in such a way as to consider the specific context of the portal being evaluated. This operation can be performed using historical data available from the organisation.

As part of our project we are also developing a DQ evaluation tool. With this, any data consumer can obtain a DQ evaluation of a web portal. For the time being, it is being implemented for the DQ Representational but the idea is to incorporate the complete PDQM. This will be a good opportunity for the data consumer to know the DQ of web portals that he or she uses and to have some way of to seeing what the differences are between them.

Likewise, the designer/developer could use this model (and the tool) to determine the level of DQ in a web portal. They could identify specific aspects of DQ that are not appropriate for data consumers (detecting, for example, a low DQ level in a branch of some of the sub-networks in the model) and thus improve their portals.

Acknowledgements

This research is part of the following projects: ESFINGE (TIC2006-15175-C05-05), 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.

This work was performed during the stay of Houari Sahraoui at the University of Castilla-La Mancha under the "Programa Nacional De Ayudas Para La Movilidad de Profesores en Régimen de año sabático", from the Spanish Ministerio de Educación y Ciencia, REF: 2004-0161.

Katerattanakul, P. and Siau, K. (2001) 'Information quality in internet commerce design', in Piattini, M., Calero, C. and Genero, M. (Eds.): *Information and Database Quality*, Kluwer Academic Publishers, pp.45–56.

Kitchenham, B. (2004) *Procedures for Performing Systematic Reviews*, Joint Technical Report Software Engineering Group, Department of Computer Science, Keele University, Keele, Staffs ST5 5BG, UK.

Mahdavi, M., Shepherd, J. and Benatallah, B. (2004) 'A collaborative approach for caching dynamic data in portal applications', *Proceedings of the Fifteenth Conference on Australian Database*, Dunedin, New Zealand, Vol. 27, pp.181–188.

Malak, G., Sahraoui, H., Badri, L. and Badri, M. (2006) 'Modeling web-based applications quality: a probabilistic approach', *7th International Conference on Web Information Systems Engineering*, Springer LNCS, Wuhan, China, Vol. 4255, pp.398–404.

Melkas, H. (2004) 'Analysing information quality in virtual service networks with qualitative interview data', *Proceeding of the Ninth International Conference on Information Quality*, Cambridge, Massachusetts, USA, pp.74–88.

Moraga, M.Á., Calero, C. and Piattini, M. (2004) 'A first proposal of a portal quality model', *IADIS International Conference, International Association for Development of the Information Society (IADIS) E-society*, Ávila, Spain, Vol. 1, pp.630–638.

Moustakis, V., Litos, C., Dalivigas, A. and Tsironis, L. (2004) 'Website quality assesment criteria', *Proceeding of the Ninth International Conference on Information Quality*, Cambridge, Massachusetts, USA, pp.59–73.

Naumann, F. and Rolker, C. (2000) 'Assesment methods for information quality criteria', *Proceeding of the Fifth International Conference on Information Quality*, Cambridge, Massachusetts, USA, pp.148–162.

Neil, M., Fenton, N.E. and Nielsen, L. (2000) 'Building large-scale Bayesian Networks', *The Knowledge Engineering Review*, Vol. 15, No. 3, pp.257–284.

Pressman, R. (2001) *Software Engineering: A Practitioner's Approach*, 5/e, McGraw-Hill, New York, USA.

Redman, T. (2000) *Data Quality: The Field Guide*, Digital Press, Boston.

Shankar, G. and Watts, S. (2003) 'A relevant, believable approach for data quality assessment', *Eighth International Conference on Information Quality (IQ2003)*, Cambridge Massachusetts, USA, pp.178–189.

Shankaranarayanan, G. and Cai, Y. (2005) 'A web services application for the data quality management in the B2B networked environment', *38th Hawaii International Conference on System Sciences (HICSS-38 2005*, IEEE Computer Society, Big Island, HI, USA.

Strong, D., Lee, Y. and Wang, R. (1997) 'Data quality in context', *Communications of the ACM*, Vol. 40, No. 5, pp.103–110.

Wang, R. and Strong, D. (1996) 'Beyond accuracy: what data quality means to data consumers', *Journal of Management Information Systems; Armonk*, Vol. 12, Spring, pp.5–33.

Yang, Z., Cai, S., Zhou, Z. and Zhou, N. (2004) 'Development and validation of an instrument to measure user perceived service quality of information presenting web portals', *Information and Management*, Elsevier Science, Vol. 42, pp.575–589.

Appendix A Data quality attributes (continued)

Attribute	*Definition*
Security	Degree to which information is passed privately from user to information source and back
Specialisation	Specificity of data contained and delivered for a web portal
Source information	The extent to which information about the author/owner of web portal is delivered to the data consumers
Timeliness	The availability of data 'on time', that is, within the time constraints specified by the destination organisation
Traceability	The extent to which data are well-documented, verifiable, and easily attributed to a source
Understandability	The extent to which data are clear, without ambiguity, and easily comprehensible
Validity	The extent to which users can judge and comprehend data delivered by the portal
Value added	The extent to which data are beneficial and provide advantages from their use

Appendix A Data quality attributes

Attribute	*Definition*
Accessibility	The extent to which the web portal provides enough navigation mechanisms for visitors to reach their desired data faster and easier
Accuracy	The extent to which data are correct, reliable, and certificated to be free of error
Amount of data	The extent to which the quantity or volume of data delivered by the portal is appropriate
Applicability	The extent to which data are specific, useful and easy applicable for the target community
Attractiveness	The extent to which the web portal is attractive for its visitors
Availability	The extent to which data are available by means of the portal
Believability	The extent to which data and their source are aecepted as correct
Completeness	The extent to which the data, provided by a web portal are of sufficient breadth, depth, and scope for the task at hand
Concise representation	The extent to which data are compactly represented without superfluous or non-related elements
Consistent reprcsentation	The extent to which data are always presented in the same format, are eompatible with previous data and eonsistent with other sources
Currency	The extent to which the web portal provides non-obsolete data
Customer support	The extent to which the web portal provides online support by means of text, e-mail, telephone, etc.
Documentation	Amount and usefulness of documents with meta information
Duplicates	The extent to which data delivered for the portal contains duplicates
Ease of operation	The extent to which data are easily managed and handled (i.e., updated, moved, aggrcgated, etc.)
Expiration	The extent to which the date until which data remain current is known
Flexibility	The extent to which data are expandable, adaptable, and easily applied to other needs
Interactivity	The extent to which the way which data are accessed or retrieved can bc adapted to one's personal preferences through interactive elements
Interpretability	The extent to which data are in language and units that are appropriate for consumer capability
Novelty	The extent to which data obtained from the portal influence knowledge and new deeisions
Objectivity	The extent to which data are unbiased and impartial
Organisation	The organisation, visual settings or typographieal features (colour, text, font, images, etc.) and the consistent combinations of these various components
Relevaney	The extent to which data are applicable and helpful for users' needs
Reliability	The extent to which users can trust the data and their source
Reputation	The extent to which data are trustcd or highly regarded in terms of their source or content
Response time	Amount of time until complete response reaches the user

References

Baldi, P., Frasconi, P. and Smyth, P. (2003) *Modeling the Internet and the Web; Probabilistic Methods and Algorithms*, Wiley, New York, USA.

Bouzeghoub, M. and Peralta, V. (2004) 'A framework for analysis of data freshness', *International Workshop on Information Quality in Information Systems, (IQIS2004)* ACM, Paris, France, pp.59–67.

Burgess, M., Fiddian, N. and Gray, W. (2004) 'Quality measures and the information consumer', *Proceeding of the Ninth International Conference on Information Quality*, Cambridge Massachusetts, USA, pp.373–388.

Cappiello, C., Francalanci, C. and Pernici, B. (2004) 'Data quality assessment from the user's perspective', *International Workshop on Information Quality in Information Systems, (IQIS2004)*, ACM, Paris, Francia, pp.68–73.

Caro, A., Calero, C., Caballero, I. and Piattini, M. (2005) 'Data quality in web applications: a state of the art', in Isaias, P. and Nunes, M.B. (Eds.): *IADIS International Conference WWW/Internet 2005*, Lisboa-Portugal, Vol. 2, pp.364–368.

Caro, A., Calero, C., Caballero, I. and Piattini, M. (2006) 'Defining a Data Quality model for web portals', *WISE2006, The 7th International Conference on Web Information Systems Engineering*, Springer LNCS 4255, Wuhan, China, pp.363–374.

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

Eppler, M. (2003) *Managing Information Quality: Increasing the Value of Information in Knowledge-intensive Products and Processes*, Springer, Berlin, Germany.

Eppler, M., Algesheimer, R. and Dimpfel, M. (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*, Cambridge Massachusetts, USA, pp.108–120.

Even, A., Shankaranarayanan, G. and Watts, S. (2006) 'Enhancing decision making with process metadata: theoretical framework, research tool, and exploratory examination', *39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, Washington DC, USA, Vol. 8, p.209a.

Fugini, M., Mecella, M., Plebani, P., Pernici, B. and Scannapieco, M. (2002) 'Data quality in cooperative web information systems', *Personal Communication*, citeseer.ist.psu.edu /fugini02data.html.

Gertz, M., Ozsu, T., Saake, G. and Sattler, K-U. (2004) 'Report on the Dagstuhl Seminar data quality on the web', *SIGMOD Record*, Vol. 33, No. 1, pp.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*, Cambridge, Massachusetts, USA, pp.133–146.

Herrera-Viedma, E., Pasi, G. and Lopez-Herrera, A. (2006) '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*, Vol. 54, pp.538–549.

Katerattanakul, P. and Siau, K. (1999) 'measuring information quality of web sites: development of an instrument', *Proceeding of the 20th International Conference on Information System*, Charlotte, North Carolina, USA, pp.279–285.

With this value, and using the probability table of the Consistent Representation node, we can derive the value of the Consistent Representation. (For example, the probability of having a consistent representation 'Good', when LCsR takes a 'Good' value, is 0.9). So at this point we will have 'evidence' that will be propagated via causal link to the child nodes of Consistent Representation- in this case the node Representation.

This process is similar for the rest of the input nodes in the sub-network. In general, the idea is that values for input nodes are measured directly for a given web portal. These values are transformed into a set of probabilities corresponding each to a label/class. Those values calculated for each input node are known as 'evidence'. This evidence propagates through the BN via causal links until the level of representational DQ in the web portal is obtained.

6 Future work

The definition of the graph structure for our model is a great advance in our work. At this stage, the evaluation of its application in a specific context is the next step to develop. To do that, we will perform the activities highlighted in the previous section and do so for the whole BN. What we will then have is a BN that is prepared to assess DQ in a web portal context. An empirical evaluation of our model will be needed in order to prove its validity.

We are also working on the automatic calculation of the measures for input nodes and the implementation of the BN structure, in order to assess the DQ of a given portal.

We can highlight the following limitations of our work, which we will deal with in our future projects.

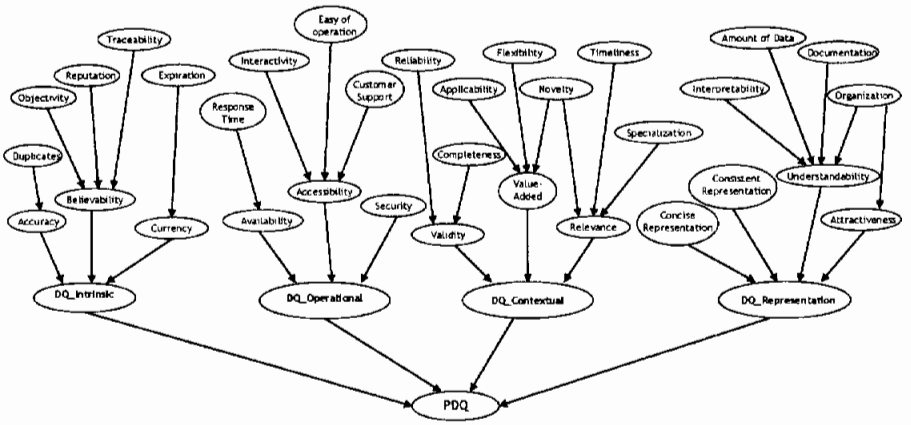
Correctness of the measures and definitions of node probability tables

The measures defined for the quantifiable variable may not be correct, in the sense that they may not represent the point of view of the data consumer. On the other hand, the definition of the node probability tables could be deficient. To improve this, we will carry out 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 PDQM. Moreover, the data obtained will be used for the refinement of the tables by the self-learning mechanism of the BN.

Generation of BN based on judgment given by experts

The BN presented in this work has been generated by experts who have taken into account the point of view of the data consumers. In an attempt to improve and/or verify it, we will develop an alternative line of research. In this new line, the BN will be generated from data automatically. We plan to compare both models (the one from experts and the one obtained automatically), our goal being to validate our present model or to obtain a better model for PDQM (as a combination of both proposals).

Figure 4 Graph of the BN that represents PDQM



As can be seen in the resulting graph, some relationships are more complex than others. For example: the variables Accessibility and Understandability have three and four parents, respectively. With the aim of facilitating the use of the BN, new nodes (synthetic nodes) need to be created which permit the reduction of the number of parents for each node.

5 Using the BN created for PDQM

The goal of this paper is to show how we have defined the structure for PDQM. That constitutes a first step in making PDQM operational. Nevertheless, and by way of example, we will give a preview of how we intend to complete this model and to prepare it for use in assessing the DQ in web portals. We have selected the DQ_Representational sub-network for this purpose.

5.1 Preparing the BN of DQ_Representational for use in DQ assessment

In order to simplify the BN and so as to reduce a future combinatory explosion, we have created two synthetic nodes: Representation and Volume of Data. The incorporation of these nodes allowed the reduction of the number of parents for the Understandability and DQ_Representational nodes (see the original sub-network in Figure 4 and the new sub-network in Figure 5).

We then defined a quantifiable variable for each input node in the sub-network. See Figure 5 – the nodes in the last level. To calculate each quantifiable variable, several measures will be defined. Each quantifiable variable will take a numerical value between 0 and 1. As the number of possible values for each input node can be infinite, we will transform them into discrete variables. According to Malak et al. (2006) this transformation can be achieved using fuzzy logic. So for each quantifiable variable we will define a membership function that transforms the value of the indicator into a set of probabilities, each corresponding to a label/class. For example, in Figure 5 the

We have built the graph of the BN which represents PDQM, (shown in Figure 4) with Table 8 as our starting point. In this BN we can observe the following levels:

- Level 0, where the PDQ is the node that represents data quality in the whole portal.
- Level 1, where the node represents the DQ in each DQ category in a portal. Obviously, the node PDQ is defined in terms of the other four nodes.
- Level 2, where nodes represent the DQ attributes with a direct influence over each one of the DQ categories.
- Level 3, where nodes represent the DQ attributes with a direct influence over each one of the DQ attributes in Level 2.

Table 8 Relationships of direct influence between the DQ attributes of PDQM

	Relation of direct influence		
	Level 1	*Level 2*	*Premise that supports the direct influence relationships*
DQ intrinsic (Level 1)	Accuracy	–	Accuracy is an attribute that has a direct influence on the DQ intrinsic in a Web portal. Accuracy is an important DQ indicator for the data consumer (Wang and Strong, 1996)
	Believability	Objectivity	If data are objective (i.e., impartial) then data are more likely to 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 the data consumer
		Traceability	If the portal delivers information about the author/owner then data will be more easily traceable
			If data are traceable then data are more likely to be accepted as correct
	Currency	–	Currency is an attribute that has a direct influence on the DQ intrinsic in a web portal. Currency is an important DQ indicator for the data consumer (Wang and Strong, 1996)
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 they are more accessible
		Ease of operation	If data can be easily managed and manipulated then they are more accessible
		Customer support	If the Web portal provides online support then data are more accessible
	Availability	Response time	If the response time for obtaining data is appropriate, then data are considered to be more available to users

- Another interesting property of the Bayesian approach is the fact that it considers 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.

The next section will show the process followed in structuring DQ attributes of PDQM in a BN.

4 Structuring DQ attributes in a Bayesian Network (BN)

An appropriate grouping of attributes is a very relevant factor in achieving the requirements set out in the previous section. In this section, we will show the process developed for building the graph structure that represents the 33 DQ attributes of PDQM.

This process has been performed in two phases. First, we have defined the criterion for organising DQ attributes hierarchically. Secondly, we have built a graph to represent PDQM; that is, the DQ attributes organised in the previous phase will be represented in the form of a BN. In the next subsections we will explain each phase of the generation of the BN.

4.1 Phase 1: Defining a hierarchical structure for PDQM

To organise the data quality attributes of PDQM into a hierarchical structure we have used the conceptual DQ framework developed in Wang and Strong (1996) as the criterion for classification. We have decided to use this framework for the following reasons:

Wang and Strong's framework focuses on the point of view of the data consumer in the same way as our own seeks to do.

All the dimensions (attributes) considered in this framework are also in the set of attributes of PDQM.

The framework has a simple structure, with two levels that give us freedom to organise the attributes of PDQM in the most appropriate way, according to our given aim.

Wang and Strong's framework considers four DQ categories (see Table 6). This framework was defined for information systems; some aspects inherent to the web context are not considered, specifically about the role of systems. In our work, therefore, we have renamed the Accessibility category as operational category. With this new name, our intention is to emphasise the importance of the role of systems, not only with respect to accessibility and security, but also in terms of personalisation, collaboration, etc.

So for our structure we have held on to the intrinsic, contextual and representational categories and we have incorporated the operational category.

The survey questionnaire was composed of 34 questions, one for each DQ attribute. Each question was measured by using a 5-point Lickert scale where 1 means 'Not Important' and 5 'Very Important'.

We used a sample of student subjects (convenience sampling) for our survey. A group of 70 Master students in the final-year (fifth) of Computer Science was enrolled (from a software engineering class). The total effective sample was 54, or 77% of the subjects that had initially been enrolled.

We decided that DQ attributes that had a mean of three or more (considering the choices 'moderately important', 'important' and 'very important') would be kept in the PDQM. All the others are rejected. This first simple filtering would be followed by a refinement and validation phase.

Regarding the results of the survey, 33 DQ attributes obtained a mean of three or more (97 %). These 33 attributes made up the new version of PDQM. Table 5 shows the retained DQ attribute list and descriptive statistics about them.

Table 5 Final set of DQ attributes of PDQM

Attribute	*Mean*	*Min*	*Max*	*Attribute*	*Mean*	*Min*	*Max*
Attractiveness	4.06	2	5	Interactivity	3.19	1	5
Accessibility	4.52	3	5	Interpretability	3.87	2	5
Accuracy	4.28	2	5	Novelty	3.67	2	5
Amount of data	3.96	2	5	Objectivity	3.50	1	5
Applicability	4.00	2	5	Organisation	3.94	2	5
Availability	4.60	3	5	Relevancy	4.09	2	5
Believability	4.15	2	5	Reliability	4.15	2	5
Completeness	3.85	2	5	Reputation	3.46	2	5
Concise representation	3.63	2	5	Response time	4.30	2	5
Consistent representation	3.63	2	5	Security	4.22	2	5
Currency	4.54	3	5	Source's information	2.56	1	5
Customer support	3.54	1	5	Specialisation	3.61	2	5
Documentation	3.31	1	5	Timeliness	4.06	2	5
Duplicates	3.00	1	5	Traceability	3.63	1	5
Ease of operation	3.72	2	5	Understandability	4.02	2	5
Expiration	3.28	1	5	Validity	3.57	1	5
Flexibility	3.26	2	5	Value added	3.98	1	5

3 A probabilistic approach for the structuring of PDQM

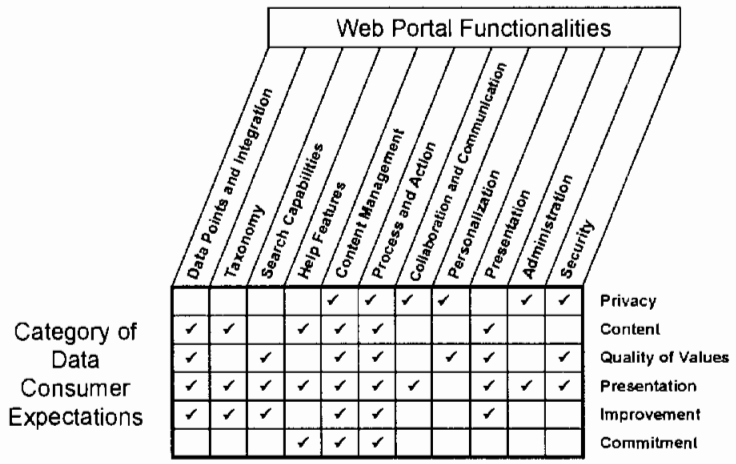
So far, we have identified a set of DQ attributes that can be applicable in the web portal context. The definition of a model does not mean that it can be operational, however, i.e., that it can be used in an assessment process. Indeed, just having a set of attributes is not enough to be able to then go on to measure/evaluate them. Neither is combining the results of their evaluations enough to give an overall assessment of the quality of the portal data. To reach this goal, we need to find a model that allows us

With the matrix created, based on the definitions of expectations and functionalities as well as on our perceptions, we carried out an analysis of what expectations are applicable to each particular web portal functionality. For example, regarding the Data Points and Integration functionality we have the following analysis:

> Based on their definition, we have considered that this functionality is related to four categories of DQ expectations: Content, Quality of value, Presentation and Improvement. The reasoning used was:
>
> - Content, because consumers need a description of portal areas covered, use of published data, etc.
> - Quality of value, because the data consumer should expect the result of searches to be correct, up-to-date and complete.
> - Presentation, because formats, language and other aspects are very important for easy interpretation.
> - Improvement, because users could wish to participate with their opinions in the portal improvements, as well as to know what the results of applying those improvements are.

The analysis carried out for each functionality is more detailed in Caro et al. (2006). Figure 2 shows all relationships established in the matrix. Each relationship is represented by a '√' mark.

Figure 2 Matrix for classifying web DQ attributes



2.3 Classification of web DQ attributes in the matrix

In the third phase, we used the matrix obtained to classify the web DQ attributes identified in phase 1. So for each relationship between functionality and expectation, we assigned the web DQ attributes that could be used by the data consumer to evaluate DQ in a portal. We did this intuitively, by studying what was appropriate for each attribute (based on its definition). This assignation was carried out in relation to the objective of

Table 1 shows these attributes, pointing out for each of them the work where they were put forward, as well as the total number of pieces of work where they can be found referred to. In addition, the symbols × and ⊗ were used to represent how they were combined (× indicates the same name and meaning and ⊗ marks the fact that only the meaning is the same).

Table 1 Web data quality attributes 1–41

Author	Year	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	Understandability	Validity	Value-added	Number of Attributes
Nauman and Rolker	2000		×	×			×	×	×	×	×		×		×								×	×			×			×	×	×	×	×	×		×	⊗	×		×	22
Katerattanakul and Siau	1999, 2001	×	⊗			×																							×									⊗			6
Eppler	2001	×	×		×				×	⊗	⊗			×			⊗				×	⊗		×								⊗	×			×	×	⊗			16
Fugini et al.	2002		⊗		×			⊗	×													×								⊗			⊗			×					8
Pernici and Scannapieco	2002		×						×									×													⊗										4
Graefe	2003	⊗						⊗	×														×			×				⊗									⊗	⊗	6
Bouzeghoub and Peralta	2004													×																						×					2
Gertz	2004								×					×		×				×							×														5
Mecikas	2004	×	×	⊗					×	×	×	×					×		×				×			×				×		×				×	×	⊗		×	20
Moustakis	2004				⊗				⊗																					⊗					×					4
Yang et al.	2004		×						⊗	×				×																⊗											5
Number of references		4	7	2	3	1	1	6	7	3	3	1	1	4	1	1	2	1	1	1	1	2	3	1	1	1	2	1	1	6	2	2	3	4	1	1	5	3	4	1	3	

2.2 *Definition of a classification matrix*

In the second phase, we have built a matrix for the classification of the DQ attributes obtained in the previous phase. This matrix relates two basic aspects considered in our model: the data consumer perspective, as seen by their expectations of DQ on the internet (Redman, 2000) and the basic functionalities in a web portal (Collins, 2001), defined in Tables 2 and 3 respectively.

Table 2 Categories of data consumer expectations about the DQ on the internet proposed by Redman

Category	*Description: 'The data consumer'*
Privacy	Should expect the Publisher to state explicitly and follow both its consumer privacy policy and its privacy policy regarding others (other consumers, individuals, organisations, and so forth)
Content	Should expect the Publisher to be explicit in: describing what data are published and how they should be used. The publisher is also to be explicit in describing appropriate and inappropriate uses of the data published; all data needed for an intended use will be provided (unless otherwise stated). Easy-to-understand definitions of every important term will be clearly stated, along with all original sources of data
Quality of values	Should expect all published data to be correct and that the Publisher will give a guarantee on the correctness of data published, or that the Publisher will state its policy regarding incorrect data. He or she should also expect data values to be current, unless otherwise informed by the Publisher – all relevant data will be published, unless otherwise stated

subjectivity can be dealt with by BNs. The subjectivity is present in, among other places, the judgement of data consumers, some DQ attributes (for example, interpretability and understandability), the measurement of some DQ attributes such as attractiveness and relevancy, as well as in the calculus of DQ and the ranking of it.

We use a Bayesian Network (BN) to structure, refine, and represent PDQM for the perspective of DQ evaluation in web portals. The construction of a BN for a particular quality model can be done in two phases (Malak et al., 2006). First, we build the graph structure. This structure is essential for capturing the appropriate relationships between DQ attributes. Secondly, we define the node probability tables for each node of the graph. The assignment of probabilities must be done according to the context of evaluation (Shankar and Watts, 2003).

This paper addresses the first phase of the BN definition, essentially. In particular, we show how we built a generic structure for the PDQM attributes and explain how we plan to use this for concrete assessment.

The rest of the paper is organised as follows. Section 2 presents a summary of the PDQM model. The approach for structuring DQ attributes of PDQM is presented in Section 3. Section 4 describes, step by step, the process that led to the generation of the targeted structure. An example of how we use the generic structure is presented in Section 5. Section 6 is given over to the forthcoming steps towards the production of a fully operational DQ model. Finally, Section 7 summarises and concludes the paper.

2 PDQM

PDQM is a data quality model for web portals which takes the data consumer perspective as its central focus. To generate this model we have analysed how the data consumer evaluates DQ and what aspects are relevant for him or her. Following this approach, we have based our model on three key elements:

- *The data consumer perspective.* The currently accepted view of assessing DQ involves understanding it from the users' point of view. Strong et al. (1997) suggest that quality of data cannot be assessed independently of the people who use them. That is why in our model we have started from the DQ expectations of the data consumer on the internet in our consideration of the data consumer perspective, as proposed in Redman (2000). These expectations are organised into six categories: privacy, content, quality of values, presentation, improvement, commitment.
- *Web data quality attributes.* In a systematic review of the literature we have found some DQ frameworks and models proposed in different domains in the context of the web. From these frameworks and models we have gathered together DQ attributes that have been proposed for application in the web context. The idea was to take advantage of work that had already been carried out in the web context and to apply it to web portals.

Keywords: Data Quality; DQ; information quality; web portals; Bayesian Network; BN; data quality model; data quality evaluation.

Reference to this paper should be made as follows: Caro, A., Calero, C., Sahraoui, H.A. and Piattini, M. (2007) 'A Bayesian network to represent a data quality model', *Int. J. Information Quality*, Vol. 1, No. 3, pp.272–294.

Biographical notes: Angélica Caro has a PhD in Computer Science from the Castilla-La Mancha University in Ciudad Real, Spain. She is an Assistant Professor at the Department of Computer Science and Information Technologies of the Bio Bio University in Chillán, Chile. Her research interests include: data quality, web portals, data quality in web portals and data quality measures. She is author of papers in national and international conferences on this subject.

Coral Calero has a PhD in Computer Science and is an Associate Professor at the Escuela Superior de Informatica of the Castilla-La Mancha University in Ciudad Real, Spain. She is a member of the Alarcos Research Group, in the same University, specialised in Information Systems, Databases and Software Engineering. Her research interests inelude: databases design, web quality software quality, software measurement. She has published, among others, in *Information Systems Journal*, *Software Quality Journal*, *Information and Software Technology Journal* and has served as reviewer of several conferences and journals. She is also author and editor of books on quality and databases.

Houari A. Sahraoui is an Associate Professor at the Department of Computer Science and Operations Research (Software Engineering Group), University of Montreal. He held the position of lead researcher of the Software Engineering group at CRIM (Research Center on Computer Science, Montreal). PhD in Computer Science, Pierre and Marie Curie University, his research interests include: artificial intelligence techniques applied to SE, object-oriented metrics, software quality, software visualisation, and software reverse and reengineering. He has published around 80 papers and edited two books. He served as steering, program and organisation committee member in several major conferences and journals.

Mario Piattini is a Full Professor at the Castilla-La Mancha University in Ciudad Real, Spain. His research interests include software quality, metrics and maintenance. He holds a PhD in Computer Science from the Technical University of Madrid, and leads the Alarcos Research Group at the UCLM. He is CISA and CISM by ISACA. He leads the Joint INDRA–UCLM Software Research and Development Center. He is member of ACM and the IEEE Computer Society. Furthermore, he is the author of several books and papers on databases, software engineering and information systems.

1 Introduction

During the past decade, an increasing number of organisations have established web portals to complement, substitute or widen already-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 online information and information-related services (Yang et al., 2004). Moreover, they create a beneficial working environment.