World Wide Web (2008) 11:465–484
DOI 10.1007/s11280-008-0048-y

An Applicable Data Quality Model for Web Portal Data Consumers

Coral Calero · Angélica Caro · Mario Piattini

Received: 16 October 2007 / Revised: 10 April 2008 /
Accepted: 30 June 2008 / Published online: 30 July 2008
© Springer Science + Business Media, LLC 2008

Abstract Web portals have emerged as an important means by which to access data on the worldwide. The people that use these applications need to ensure that the data recovered is suitable for the task at hand. That is, they need to know the level of quality of the data obtained. This paper introduces the PoDQA (Portal Data Quality Assessment) tool which implements PDQM, a Portal Data Quality Model, which is centered upon the data consumer perspective. Thus, the measurement of data quality is carried out by using the point of view of data consumers. Our work aims to fill the lack of specific proposals for the DQ evaluation in Web portals and tools that put these proposals into practice. The paper illustrate how PoDQA tool works and how it can be used by data consumers in order to, for example, discover the data quality of a specific web portal. PoDQA also suggests several corrective maintenance activities for users who are interested in the improvement of the data quality of their Web portals.

Keywords web portals · data quality · data quality model · data quality assessment · users · developers

1 Introduction

A Web portal is a site that aggregates information from multiple sources on the Web and organizes this material in an easy user-friendly manner [36]. Over the past decade the number of organizations that provide Web portals has grown dramatically. These

C. Calero (✉) · M. Piattini
Alarcos Research Group, Information Systems and Technologies Department, UCLM—INDRA Research and Development Institute, University of Castilla—La Mancha, Ciudad Real, Spain
e-mail: Coral.Calero@uclm.es

M. Piattini
e-mail: Mario.Piattini@uclm.es

A. Caro
Department of Computer Science and Information Technologies, University of Bio Bio, Chillán, Chile
e-mail: mcaro@ubiobio.cl

organizations provide portals that complement, substitute or extend existing services to their client base [37]. Numerous users worldwide use Web portals to obtain information for their work and to help with decision making. These users, or data consumers, need to ensure that the data obtained are appropriate for their needs. Likewise, the organizations that provide Web portals need to offer data that meet user requirements, thus helping these users to achieve their goals. Therefore data quality represents a common interest between data consumers and portal providers.

Data (or Information) Quality (DQ) is often defined as "fitness for use", i.e., the ability of a collection of data to meet user requirements [3, 33]. This definition and the current view of assessing DQ, involve understanding DQ from the users' point of view [18]. In recent years, several research projects have been conducted on the topic of Web Data Quality. However, there is still a lack of specific proposals for the DQ in Web portals which consider the data consumer's point of view and tools that put these proposals into practice.

In this work we introduce the PoDQA tool, whose aim is to assess the DQ in Web portals. The PoDQA development is part of a greater project on data quality, in which the research focus is to work towards the generation of a generic, adequate, flexible and complete data quality model for Web portals. This project has been introduced in [4] and [5], where the development of PDQM (a data quality model for Web portals) is described. PDQM is centered on the point of view of data consumers and uses a probabilistic approach (based on Bayesian networks) for data quality evaluation.

PoDQA assesses the DQ of a Web portal by using the PDQM model as a basis. The first version of this tool is available in http://podqa.webportalquality.com. This version implements the DQ evaluation for a subpart of PDQM (Representational DQ) as will be explained in this paper.

The organization of this paper is as follows. Section 2 describes the background of our work. Section 3 introduces the PDQM model, emphasizing its approach towards quantifying the DQ. The prototype of PoDQA is described in Section 4. Finally, Section 5 shows our conclusions.

2 Background

DQ is commonly thought of as a multi-dimensional concept [3, 28, 34]. The literature dealing with DQ provides different classifications of the DQ attributes, depending upon the perspective of the authors and the context tackled. On the other hand, in order to assess DQ a growing tendency towards considering the users' point of view exists. In fact, the most common definition of DQ is data that are "fit-for-use", i.e. the ability of a collection of data to meet user requirements [3, 33, 34]. This definition suggests two important ideas. First, that DQ cannot be assessed independently of the people who use the data. And secondly, that DQ is relative and subjective: different users may have diverse opinion about the quality level of the same data.

In [12] DQ assessment is defined as the process of assigning numerical and categorical values (quality scores) to quality criteria in a given data setting. They emphasize that DQ assessment in the Web is a difficult task and that "well-founded and practical approaches to assess or even guarantee a required degree of the quality of data are still missing".

Research on DQ began in the context of information systems [19, 34] and has been extended to contexts such as cooperative systems, data warehouses or e-commerce,

amongst others. Due to the particular characteristics of Web applications and their differences from traditional information systems [29], the research community has begun to deal with the subject of DQ on the Web [12]. In fact, the particular nature of the Web has forced the necessity to pay attention to a series of typical issues in this context which may affect or influence DQ. Among these we might mention: Typical problems of a Web page (un-updated data, publication of inconsistent data, obsolete links, etc.) [9], Integration of structured and non-structured data [10], Integration of data from different sources [1, 2, 12, 24, 35, 38] and Dynamic nature of the Web [12, 27].

Because of the particularities of this context and the necessity of assessing the DQ in the context of its generation [32], in recent years frameworks, models and DQ attributes to deal with DQ in different domains in the Web context have been proposed. Among them, we can highlight those presented in Table 1.

It is important to highlight that many of these are based on well-founded data quality frameworks defined for other fields. One of the most frequently used frameworks is that proposed in the context of information systems by Wang and Strong [34]. This framework establishes four DQ categories in which 15 DQ dimensions are classified (see Table 2).

After studying the Web DQ literature, we have detected a lack of specific proposals of DQ models and/or frameworks for Web portals. Some proposals which tackle the Web portal context, consider DQ as a part of a more general model [22, 37].

Consequently we have developed a DQ model for Web portals, named PDQM, which is centered upon the point of view of data consumers. In order to begin our work, we collected a set of DQ attributes proposed for the Web context, because although Web portals can be considered as an independent category of Web applications [13], a given Web application may belong to more than one category [13]. Thus, the DQ attributes proposed for other Web applications may be useful in the evaluation of the DQ in Web portals. With this strategy in mind, our aim was to take advantage of previous works and to start by using an ample set of DQ attributes as a reference.

Table 1 Web DQ frameworks.

Author	Domain	Model/framework structure
Katerattanakul & Siau 1999 [17]	Personal web sites	4 categories and 7 constructors
Naumann & Rolker 2000 [24]	Data integration	3 classes and 22 quality criteria
Katerattanakul & Siau 2001 [16]	e-commerce	4 categories associated with 3 categories of data user requirements
Pernici & Scannapieco 2002 [27]	Web information systems (data evolution)	4 categories, 7 activities of DQ design and architecture to DQ management
Fugini et al. 2002 [11]	e-service cooperative	8 dimensions
Graefe 2003 [14]	Decision making	8 dimensions and 12 aspects related to (providers/consumers)
Eppler et al. 2003 [8]	Web sites	4 dimensions and 16 attributes
Gertz et al. 2004 [12]	DQ on the web	5 dimensions
Moustakis et al. 2004 [23]	Web sites	5 categories and 10 sub-categories
Melkas 2004 [21]	Organizational networks	6 stages of DQ analysis with several dimensions associated with each one
Bouzeghoub & Peralta 2004 [2]	Data integration	2 factors and 4 metrics
Yang et al. 2004 [37]	Web information portals	2 dimensions and 4 attributes

Table 2 DQ framework of Wang and Strong's.

DQ category	Description	DQ dimensions
Intrinsic	It denotes that data have quality in their own right	Accuracy, objectivity, believability, reputation
Accessibility	It emphasizes the importance of the role of systems; that is, the system must be accessible but secure	Accessibility, security
Contextual	It highlights the requirement which states that data quality must be considered within the context of the task in hand	Relevance, value-added, timeliness, completeness, amount of data
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	Interpretability, ease of understanding, concise representation, consistent representation

3 A data quality model for web portals

PDQM is a data quality model for Web portals which focuses on the data consumer perspective. To produce PDQM, we defined a process which was divided into two parts. The aim of the first part was the definition of a theoretical model based on three key aspects: (1) The data consumer perspective (How data consumers assess DQ, what their expectations of DQ are), (2) Web data quality attributes (proposed in literature for different types of Web applications); and (3) Web portal functionalities (which characterize and distinguish portals from other Web applications). The idea was to obtain a set of DQ attributes that could be used to assess the DQ in Web portals. The first part was, therefore, was made up of four phases:

(1) In the first phase we used previous literature to compile Web DQ attributes that we considered relevant to Web portals. To do this we made a systematic review of the relevant literature and we selected previous work proposed for different domains in the Web context (among them, Web sites, data integration, e-commerce, Web information portals, cooperative e-services, decision making, organizational networks and DQ on the Web). As a result we obtained a set of 100 DQ attributes. We detected certain synonymous amongst the attributes identified. Those attributes were combined along with those which had similar names and meanings, thus obtaining a final set of 41 attributes [4].

(2) In the second phase, we created a matrix with which 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 DQ expectations on the Internet [30] and the basic functionalities offered in a Web portal [6]. Once the matrix was defined, we ticked the expectations applicable to each of the different functionalities of a Web portal.

(3) In the third phase, we used the matrix to analyze the appropriateness of each Web DQ attributes identified in the first phase. This analysis consisted of assigning an expectation related to the DQ attributes that could be used by the data consumer to evaluate the quality of data in a portal to each functionality. For this assignment we used as a 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. As a result of this phase, we obtained a set of 34 DQ attributes through which to evaluate the DQ in Web portals [4]. Figure 1 shows the matrix and an example of the classification of DQ attributes.

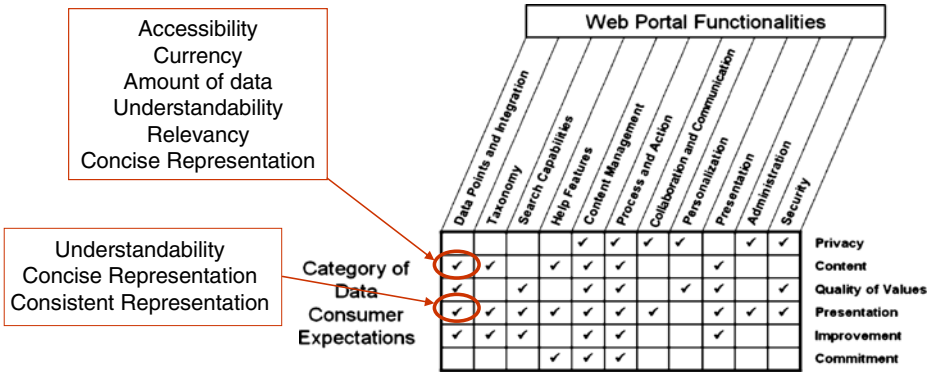


Figure 1 Example of classification of Web DQ attributes in the matrix.

(4) Finally, in the fourth phase we validated the model obtained. To perform this task, we conducted a study by means of a survey. The purpose of this survey was to collect ratings of the importance for data consumers of each of the DQ attributes in the model. As a result of this study, in which one of the consulted attributes was considered to be of less importance by all respondents and in which nobody suggested new attributes, we obtained a final set of 33 DQ attributes (see Table 3). The definition of each DQ attribute of PDQM can be found in "Appendix".

More details of the development of the theoretical version of PDQM can be found in [4].

In the second part, our aim was to convert this theoretical model into an operational one, i.e., one which can, in our case, be used to assess the quality of web portals. In simple terms, this conversion consisted of defining a structure with which to organize DQ attributes and their relationships. Taking into account the intrinsic subjectivity of the data consumer's perspective and the uncertainty inherent in quality perception [7], we decided to use an approach that employs Bayesian networks (BN) and Fuzzy logic [20]. This decision was made by considering a set of properties/requirements that we wished our final model to have. The PDQM must be:

- *Generic.* The PDQM must be applicable to any Web portal.
- *Adequate.* The PDQM is orientated towards the data consumer's point of view. It must support the subjectivity and uncertainty associated with DQ evaluation.

Table 3 DQ Attributes of the theoretical PDQM.

Attractiveness	Documentation	Customer support
Accessibility	Duplicates	Reliability
Accuracy	Ease of operation	Reputation
Amount of data	Expiration	Response time
Applicability	Flexibility	Security
Availability	Interactivity	Specialization
Believability	Interpretability	Timeliness
Completeness	Novelty	Traceability
Concise representation	Objectivity	Understandability
Consistent representation	Organization	Validity
Currency	Relevancy	Value added

- *Flexible.* It must be applicable to different situations. For example, in different Web portal domains, in processes where the model can be used in a partial or complete way or in processes where different kinds of data consumers can be considered. To do this, the structure must support the assignment of different weights to the attributes.
- *Complete.* The structure must allow the representation of all the relationships between the attributes, e.g., an attribute may simultaneously affect several other attributes. In hierarchical models for example, attributes from the same level cannot be related and an attribute cannot affect more than one of the attributes in the upper level.

As a result of the operationalization of the PDQM we have obtained a BN (see Figure 2) which organizes the 33 DQ attributes into four network fragments. A network fragment is a set of related random variables that can be constructed and reasoned on separately from other fragments [25].

In order to create the network fragments, we have used the conceptual DQ framework developed in [33, 34], see Table 2, as a criterion with which to organize the DQ attributes of PDQM. However, in our work we have renamed and redefined the Accessibility DQ category (calling it the Operational DQ category). The idea was to consider aspects which are typical of this context such as personalization, collaboration, etc. Using the definitions of each category and each DQ attribute as a base, we thus classified all the DQ attributes of PDQM into these four categories. After this, relationships of influence between the attributes were established. These relationships were established by using the DQ categories and the DQ attribute definitions, together with our perceptions and experience as a base. Our aim was to establish which DQ attribute in a category has direct influence over other attributes in the same category. As an example, Table 4 shows the relationships defined for the Representational DQ category.

These relations were later confirmed in a validation process of the BN. More details about the generation of the BN for PDQM can be found in [5]. The BN generated for PDQM is shown in Figure 2.

Taking advantage of the possibility of working separately with each fragment of the BN, we decided to start with the Representational DQ. To do this, and in order to complete their operationalization, the following activities were developed:

- Two artificial nodes were created to simplify the fragment network and to reduce the number of parents for each node (node Representation and Volume of Data in

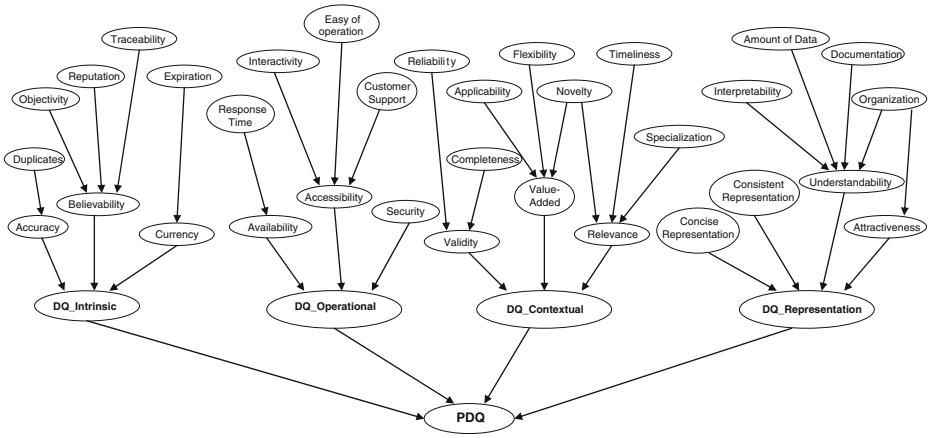


Figure 2 BN graph to represent PDQM.

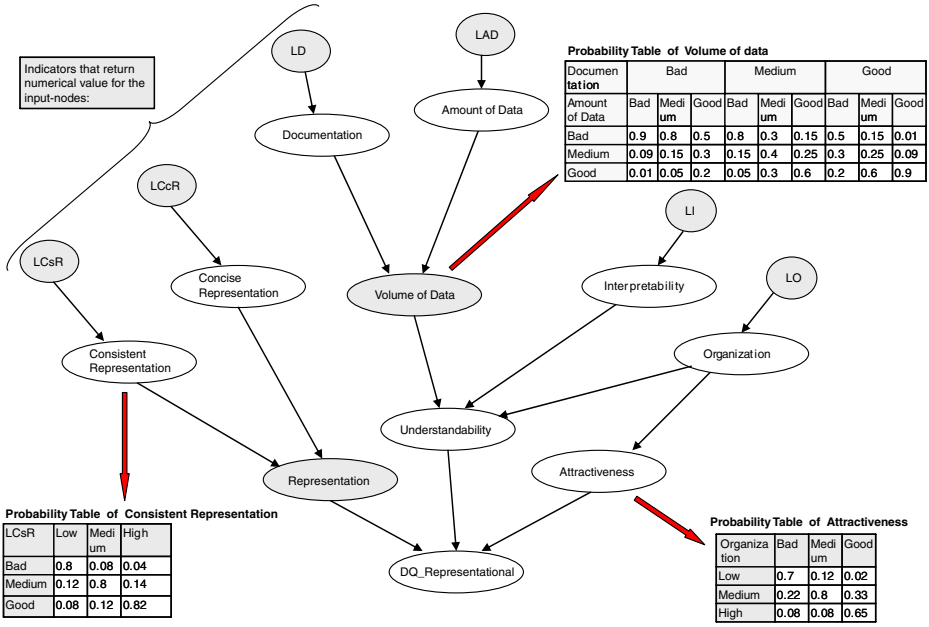
Table 4 Relationships defined for representational DQ category.

	Relation of Direct Influence		Premise that supports the direct influence relationships
DQ representational	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
		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 for data consumers

Figure 3). The aim was to reduce the combinatory explosion in the following step during the preparation of the probability tables.

- Indicators, or quantifiable variables, were defined for each entry node in the fragment (indicators LCsR, LCcR, LD, LAD, LO and LI in Figure 3). The definition of these indicators will be explained in greater depth in the following subsection.
- A probability table was defined for each intermediate node in the fragment (Figure 3 shows the probability table for the Consistent Representation, Volume of Data and Attractiveness nodes). These tables are defined by experts in Web portals. The experts were a group of researchers on DQ and users of Web portals. The procedure used was the following. The experts were presented with a proposal for the probability tables in which the probabilities associated with the various given combinations of parents of each node were already given. The experts analysed and adjusted them in accordance with their knowledge. Case studies then took place in which a group of Web portal users were asked to evaluate a group of portals, after which the users' evaluation was compared with that of the model, and the tables were adjusted in order to obtain a greater coincidence between the users' evaluations and those of the model.

The approach used allows us to represent dependencies among DQ attributes in the form of a BN in which the probability tables of the intermediate nodes are defined in order to capture the characteristics of a specific domain of a web portal. That is, the model can be adjusted, by changing the probability tables, according to the portal domain to be evaluated (for example to the bank domain or the governmental domain).



Documentation	Bad			Medium			Good		
Amount of Data	Bad	Medium	Good	Bad	Medium	Good	Bad	Medium	Good
Bad	0.9	0.8	0.5	0.8	0.3	0.15	0.5	0.15	0.01
Medium	0.09	0.15	0.3	0.15	0.4	0.25	0.3	0.25	0.09
Good	0.01	0.05	0.2	0.05	0.3	0.6	0.2	0.6	0.9

LCsR	Low	Medium	High
Bad	0.8	0.08	0.04
Medium	0.12	0.8	0.14
Good	0.08	0.12	0.82

Organization	Bad	Medium	Good
Low	0.7	0.12	0.02
Medium	0.22	0.8	0.33
High	0.08	0.08	0.65

Figure 3 The network fragment of the representational DQ.

3.1 Quantifying the representational data quality

As we have already indicated, as a part of the operationalization of PDQM it was necessary to define indicators for the entry nodes. This quantification was not easy for several reasons. In the available literature dealing with DQ, relatively few researchers have tackled the difficult task of quantifying DQ [12, 18]. Gertz et al. [12] emphasize that the DQ assessment in the Web is a difficult task because:

- Many criteria are subjective and therefore cannot be assessed automatically.
- Many sources do not publish quality related metadata.
- The sources have a large amount of data so it is therefore difficult to obtain precise quality scores.

After our experiences we can add another reason. A uniform style in the design and construction of Web portals does not exist. This means that it is incredibly difficult to generate general rules which can be applied to the automatic evaluation of DQ. One simple but clear example of this is the lack of uniformity in the handling of the URL with which to make it relatively easy to distinguish whether a portal page is internal or external.

Therefore, one great challenge in our work was to develop measures with which to assess the DQ attributes that are entry nodes in the BN. The rest of the DQ attributes will be calculated on the basis of their causal relationships. The measures have been defined by attempting to generate objective measures, but in some cases this was not possible, and some measures are calculated by using users' valuations.

The representational DQ denotes that a Web portal must present data in such a way that they are interpretable, easy to understand, and concisely and consistently represented [34]. For the definition of measures for this DQ category, we used as a reference the work of

Ivory et al. [15], the Web design recommendations of Nielsen [26] and the assessment methods proposed in [9] and [28].

For this DQ category we have in particular defined measures for six DQ attributes that are entry nodes in the BN (see Figure 3). Five of these can be measured in an objective way and one (Level of Interpretability) is calculated by using a questionnaire that must be answered by the user. In accordance with the definition of the PDQM, the tool will perform measurements for the following indicators:

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

As an example of the measures defined, Table 5 shows the measures for the *Level of Consistency* entry node (and thus the LCsR indicator).

Thus, in order to obtain the score of the Representational DQ in a given Web portal the following steps must be performed. It is first necessary to calculate the measures associated with the indicators: LCsR, LCcR, LD, LAD, LI, LO (the objective measures are calculated

Table 5 Measures for the LCsR indicator in the representational DQ fragment.

Consistent representation attribute	Level of consistent representation (LCsR) indicator
Base measures	LCsR=(PSSD×0.5+SDCD×0.5)
Pages count (PgC)	
Link count (LnC)	
Maximum of pages with the same style (MaSS)	
Link text correspondence (LTC)	
Derived measures	
Source destiny correspondence degree (SDCD): SDCD=LTC/LnC	
Pages with the same style degree (PSSD): PSSD=MaSS/PgC	

automatically and the user's evaluations are requested). Each indicator measured will take a value of between 0 and 1.

Considering that the amount of possible values for an indicator may be infinite, they should be transformed into discrete variables with a limited number of values. In order to carry out this transformation we have used fuzzy logic [31]. The idea of this is that the different values that an indicator may take are replaced with a set of probabilities which represent the degree of membership of each value in various fuzzy labels/classes (for example, "High", "Medium", "Low"). Hence, for each indicator we have defined a membership function that transforms the value of that indicator into a set of probabilities, each of which corresponds to a label/class [20]. A trapezoidal membership function was used for this transformation.

Then, by means a probabilistic classifier (fuzzy logic-based clustering algorithm), the probabilities for each entry node in the BN are calculated. These probabilities are entered in the BN. From each piece of evidence, and by using the corresponding probability table, each node generates a result that is propagated, via a causal link, to the child nodes for the whole network to the level of the Representational DQ. This process is applicable to the whole PDQM model, although only the representational DQ quantifying model has been illustrated here.

The probability tables for the intermediate nodes must be defined by taking the Web portal domain into account. As different communities or social groups may have different viewpoints about the DQ [32], we believe that although it is possible to have generic probability tables it is better to define the probability tables according to the context. For example, in governmental web portals, the influence of the *Amount of Data* in *Understandability* may be different from that of banking web portals. In governmental portals the data consumer might be more tolerant of a large amount of data on the pages whereas in banking portals the data consumer might prefer pages with less information.

In our model, we have solved this difference by changing the probability tables of the intermediate nodes. The idea is to define these specifically for each Web portal domain. In our case, at this moment, we have defined the probability tables for the domain of university portals.

4 A data quality assessment tool

In order to make the PDQM accessible to Web portal users, or data consumers, we decided to implement it. The resulting tool is called PoDQA (Portal Data Quality Assessment) and, at this moment, it implements a sub-part of PDQM, the Representational DQ fragment within the domain of university Web portals.

This tool can be used to achieve three objectives. First, to demonstrate the applicability of PDQM in the DQ evaluation of Web portals. Second, to demonstrate that it is effectively representative of the data consumer perspective. Third, to demonstrate how the PDQM will work and how it could be used by data consumers.

In the following subsections we will describe the main characteristics of the tool, its use, and its application in the DQ assessment in a Web portal.

4.1 The PoDQA tool

The PoDQA tool was built by using a 3-tiered architecture to separate the presentation, application (business), and storage components, using Visual Basic.NET technology (see Figure 4).

By means of the presentation tier the tool provides an interface for the user which allows them to carry out two tasks: users can start an evaluation process and can seek information about the previous evaluations. The application tier is composed of two sub-applications. The first calculates the measures defined in the given portal, stores the results in the database, generates the inputs for the second sub-application and notifies the user when the evaluation process is finished. The second sub-application loads the appropriate BN (corresponding to the Web portal domain), obtains the DQ score and sends the final results to the first sub-application to be stored. Finally, the data tier corresponds to the database in which the results of different evaluations and the tool's management data, are stored.

The main functions of PoDQA are to download a given Web portal, to apply the defined measures to it and to calculate its level of DQ. The objective is to give the user information about the data quality level in a given Web portal. The portal evaluation is made by considering the domain to which it belongs. Thus, for each evaluation the user will have to specify the portal's URL and its domain.

The evaluation process cannot take place in real time because it is necessary to download and analyze all the pages of the Web portal, in order to be able to calculate the defined measures. The tool calculates the measures by using the public information in Web portals. PoDQA stores the results obtained for each portal evaluated. These results will be part of the public information in the Web site of the tool. Thus, any user will be able to obtain the ranking of all Web portals evaluated in each domain. Figure 5 uses a graph to show how PoDQA calculates the DQ level.

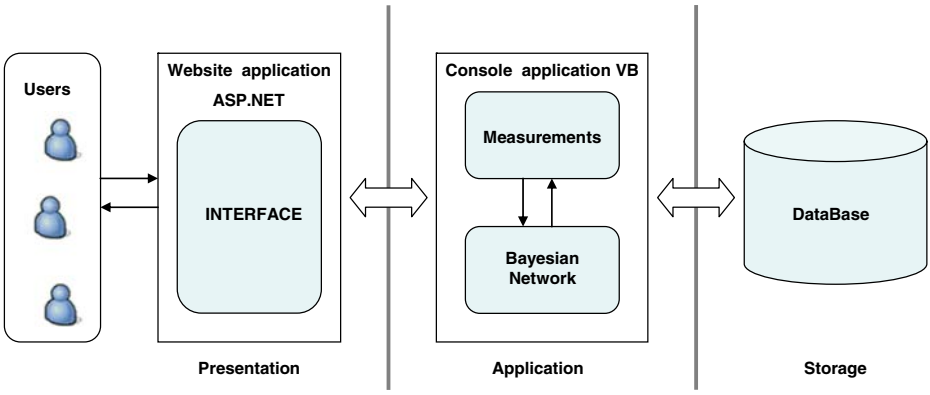


Figure 4 The PoDQA architecture.

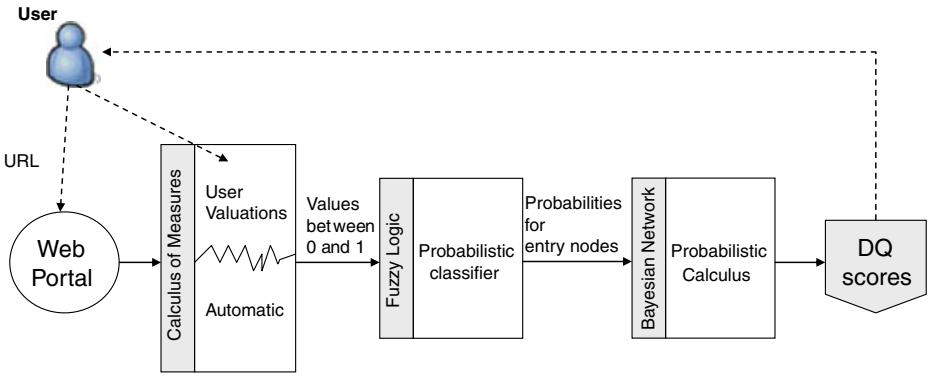


Figure 5 General process to quantify DQ using PoDQA.

Additionally, the tool offers a functionality which is mainly orientated towards Web portal developers and which, based on the results of PoDQA, provides corrective maintenance activities to improve the DQ.

4.2 Considerations in the implementation of the assessment process

When building PoDQA, we have considered certain aspects which are orientated towards improving its efficiency. We have first attempted to reduce the time needed to download and analyse a portal, and the following considerations have been taken into account[1]:

- Only the pages which coincide with the URL entered by the user will be downloaded (those which do not coincide will be considered as external pages).
- If the portal is available in more than one language, then the user will be asked to specify the language which s/he requires.
- Archives which are of zip, doc, txt, jpg, etc formats will not be downloaded.
- The downloading of a portal will be carried out by following all the links from the source page to a depth level of 4.

With respect to the final consideration, we should like to point out the following. Our initial intention was to completely download every portal. However, this implied using a great deal of time and resources. In order to reduce this time, and as we already had the evaluations of one group of portals, we decided to re-evaluate them, but to reduce the depth level of the links which were to be downloaded. As a result of this, we observed that the values calculated by each indicator maintained the tendencies in the values calculated for the whole portal, so we therefore decided to download the portals by following the links to a depth level of 4. Table 6 shows an example of a portal which has been downloaded and measured in its entirety, and the results obtained upon downloading it and measuring it at various levels of depth. As can be seen, the differences between the values obtained from a total download and a download to a depth level of 4 are minimum. The same tendency was repeated in all the Web portals that we tested. This strategy consequently reduces the downloading and measuring time by at least half.

[0] It is important to point out that these considerations do not include aspects related to the "deep web". These aspects are concretely related to the to content which cannot normally be accessed with search engines, such as pages with formats and contents which are adapted to various user profiles, password-protected resources, scripted content or non-HTML content.

Table 6 Values for the indicators for the Web portal evaluated.

Download total	Donwload level 2	Donwload level 3	Donwload level 4
LAD=0.98965	LAD=0.62	LAD=0.9	LAD=0.89
LCcR=0.9919	LCcR=0.82	LCcR=0.9	LCcR=0.9
LCsR=0.12577	LCsR=0.1	LCsR=0.20	LCsR=0.12
LD=0.48772	LD=0.5109	LD=0.52	LD=0.488
LI=not considered because is measured by user valuations			
LO=0.43555	LO=0.38	LO=0.623	LO=0.489

Secondly, in the first evaluations we detected certain problems with some measures, concretely with the extreme values obtained from some indicators. For example, note in Table 4 the values for the LCsR, LCcR and LAD indicators (some values were very close to 1 and others to 0). This situation is reiterative in several of the portals evaluated.

Upon seeking an explanation for this situation, we realised that the origin of these extreme values was in some of the base and derived measures used to calculate each indicator. For example, in order to calculate the Level of Amount of Data (LAD) indicator, it is necessary to know the *distribution of words per page;* and to discover this we need to calculate the *minimum number of words per page* and the *maximum number of words per page*. However, we have found pages which have a minimum of one word and a maximum of 318,823 words; in both cases we have found design problems. Obviously, these values need to be removed from the calculation of the measurement.

For this reason we have refined the calculations made by the tool by detecting and eliminating the outliers in our measures.

4.3 The PoDQA tool usage

The PoDQA tool is a public tool which is available at http://podqa.webportalquality.com. Any user may use it to request the DQ evaluation of a Web portal. The results of the different evaluations will be public. This means that any evaluation will be able to be queried by any user. Users can see both the evaluation that they have requested and also the evaluations which have been asked for by all other users.

The results will be stored in the database. Each time a Web portal is evaluated the new values will also be stored. This allows the user to ask for historical data about the evaluations and to check whether the data quality in the Web portals has being improved. In this way, the user can check the DQ evolution of the Web portal.

When the user decides to start an evaluation process s/he must provide the URL of a Web portal, the Web portal domain, which DQ category s/he wishes to evaluate (at the moment this is only possible for representational DQ), and their e-mail address (see Figure 6). When these data are verified, the process is initiated. If the DQ category to be evaluated includes subjective measures, then a set of questions will be formulated for the user (in the Representational DQ category certain questions will be asked in order to obtain the evaluations for the DQ attribute of Interpretability). Once the calculations are performed the user is contacted (via e-mail) and is invited to visit the PoDQA tool Website again in order to recover the results.

If the user decides to use the PoDQA to ask about previous evaluations s/he can obtain three types of information: (1) the results of evaluations requested by him/her, (2) the results of the previous evaluations of a given Web portal sorted chronologically (requested by any user) and (3) the ranking of the Web portals belonging to a given domain. Figure 7 shows the ranking of a group of Web portals.

Figure 6 A new DQ assessment for a Web portal.

4.4 Example of the DQ assessment of a web portal

To demonstrate how PoDQA works and how it can be used, a DQ assessment process will be developed in which some partial values will be shown, but in a real evaluation the user would not have access to these values. For this evaluation we have used a real university Web portal, but its identity will not be given.

In order to initiate a DQ evaluation process the user executes the PoDQA tool and selects the option "*Assess a portal*". After this, the URL of the Web portal is entered, along with its domain, the user's contact information (name and e-mail) and the DQ category to be evaluated (see Figure 6). Next, PoDQA verifies both the URL and the user's e-mail in order to start the assessment.

The first internal results generated for PoDQA are the values for each indicator. In our example the values obtained are shown in Table 7.

These values are internally transformed into valid entries for the BN (see Figure 8). After this, the BN calculates and propagates the probabilities until the level of the representational DQ in the portal (see the probabilities calculated for each node in Figure 8) is assessed. In this case, as a result we have obtained that the representational DQ level is *Medium* with a probability of 58%, *Good* with a probability of 16% and *Bad* with a probability of 24%.

Once the values for the representational DQ are obtained, they are stored in the PoDQA data base. Next, the tool will send an e-mail to the user to invite him/her to visit the PoDQA website to obtain the results of the DQ evaluation.

When a user accesses the PoDQA website to review the results of the evaluation, s/he will do so under two different sets of circumstances. In the first, the user will be a data consumer who wishes to know results for his/her own personal use. That is to say, s/he wishes to use the results in order to be able to decide whether or not the level of DQ in the portal is appropriate for his/her needs. In this case, the user will receive the results in a language which is easy for him/her to understand. This means that s/he will not only receive numerical results but also an interpretation of them. In our example, it could be said that the representational quality of the data in the portal is medium (because its probability is above 50%).

Figure 7 DQ ranking of a group of Web portals.

In the second set of circumstances, the user will be a Web portal developer who is not only interested in discovering the results of the evaluation, but also in discovering how to improve his/her portal in order to be able to adjust it to the needs of the data consumer. In this case, the user will not only receive the results of the evaluation, but also a list of improvement activities for his/her portal. These activities are directly related to the criteria evaluated through the indicators. For example, in the case of the portal evaluated in this article, some of the improvement activities will be related to the result obtained for the LCsR indicator which obtained quite a low evaluation (see Table 5). Thus in consistence with the definition of the indicator, some of the improvement activities which may be suggested are:

- *Standardising the portal's design through the use of style sheets* [26]. This means that if style sheets are not already being used in the portal, then their design and incorporation is recommended in order to thus ensure coherence and consistency in the presentation of the data. If they are already being used, then their quantity should be reduced to a more appropriate amount in order to present the data in a more uniform and consistent manner.
- *Increasing the correspondence between links and the destination pages.* Review the link texts and check that they are consistent with the content of the destination page.

Table 7 Values for the indicators for the Web portal evaluated.

LCsR	LCcR	LD	LAD	LI	LO
0.12	0.99	0.46	0.99	0.5	0.44

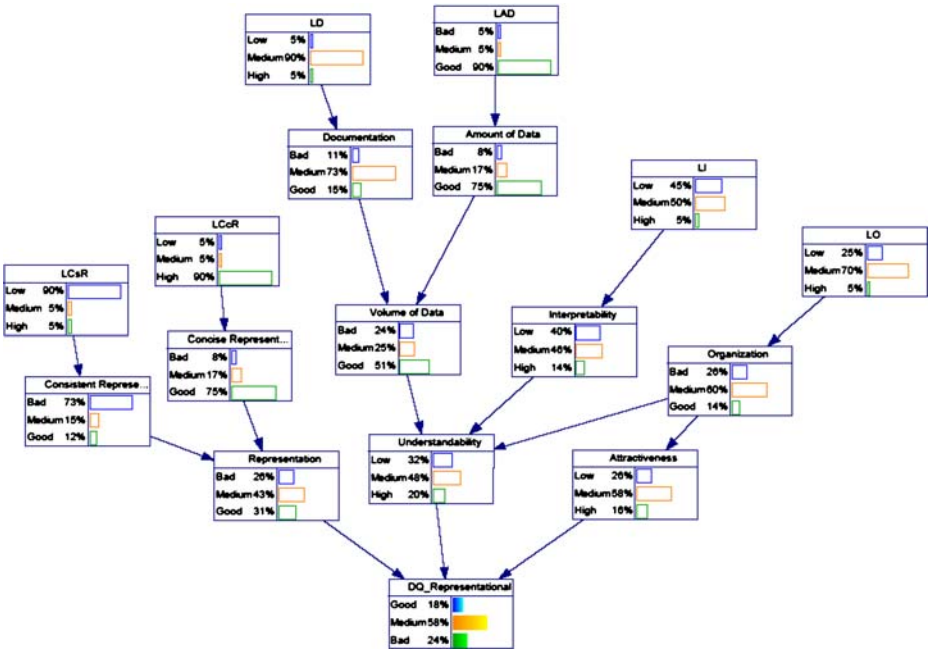


Figure 8 BN calculating the representational DQ level for the Web portal evaluated.

5 Conclusions

Nowadays, numerous users worldwide make use of Web portals to seek information for their work. These users, or data consumers, need to ensure that the data obtained are appropriate to their needs. In recent years, several research projects have been conducted on the topic of Web Data Quality. However, there is still a lack both of specific and practical proposals for the DQ in Web portals, and which consider the data consumer's point of view. In this work we have presented the PoDQA tool, which is based on a data quality model called PDQM, and is used to assess the DQ in Web portals.

PDQM is a data quality model for Web portals which is composed of 33 DQ attributes grouped into four DQ categories. The method defined to evaluate the DQ is based on the use of a Bayesian networks. This method allows us to adjust the DQ model according to the Web portal domain to be evaluated. The model can be applied to assess the DQ in a specific DQ category or to assess the DQ in all DQ categories. Together with the BN, the method is also accompanied by various measures which are used to calculate a set of indicators that generate input values for the BN. Through this approach our intention has been to acknowledge the user's subjectivity in the DQ evaluation. This subjectivity is represented by means of probability tables derived from expert's opinions.

At this moment, PoDQA implements the DQ evaluation for the Representational DQ category in the university Web portal domain. The main functionalities of the PoDQA are: the level of representational DQ is calculated for a given Web portal and the data quality ranking for the Web portals evaluated is shown for a specific Web portal domain. As has previously been mentioned, PoDQA is available at http://podqa.webportalquality.com.

Our future work is to extend the tool to the whole PDQM. That is, we will implement the measures for the rest of the DQ categories of the model. The PoDQA will consequently

be able to offer users the possibility of evaluating the data quality in a Web portal in either only one of the four DQ categories, or in all of them at the same time. In this last case, if there are conflicting reports from different categories, in the sense that the DQ levels are different, this conflict will be solved by adjusting the corresponding probability table.

Another task for the future is the adaptation of the BN to other Web portal domains. Thus, PoDQA will allow the users to select between several Web portal domains and will assess the DQ of each one, based on their specific characteristics which will be represented by means of their probability tables.

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

Appendix

Table 8 Data quality attributes of PDQM.

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 accepted 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 representation	The extent to which data are always presented in the same format, are compatible with previous data and consistent 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 on-line 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, aggregated, 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

Table 8 (continued).

Attribute	Definition
Interactivity	The extent to which the way which data are accessed or retrieved can be adapted to one's personal preferences through interactive elements
Interpretability	The extent to which data are in language and units that are appropriates for consumer capability
Novelty	The extent to which data obtained from the portal influence knowledge and new decisions
Objectivity	The extent to which data are unbiased and impartial
Organization	The organization, visual settings or typographical features (colour, text, font, images, etc.) and the consistent combinations of these various components
Relevancy	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 trusted or highly regarded in terms of their source or content
Response time	Amount of time until complete response reaches the user
Security	Degree to which information is passed privately from user to information source and back
Specialization	Specificity of data contained and delivered for a Web portal
Timeliness	The availability of data "on time", that is, within the time constraints specified by the destination organization
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

References

1. Angeles, P., MacKinnon, L.: Detection and resolution of data inconsistences, and data integration using data quality criteria. In: Proceedings of the QUATIC'2004, pp. 87–93 (2004)
2. Bouzeghoub, M., Peralta, V.: A framework for analysis of data freshness. In: Proceedings of the International Workshop on Information Quality in Information Systems (IQIS2004), pp. 59–67. ACM, Paris, France (2004)
3. Cappiello, C., Francalanci, C., Pernici, B.: Data quality assessment from the user's perspective. In: Proceedings of the International Workshop on Information Quality in Information Systems (IQIS2004), pp. 68–73. ACM, Paris, Francia (2004)
4. Caro, A., Calero, C., Caballero, I., Piattini, M.: Defining a Data Quality Model for Web Portals. In: Proceedings of the WISE2006, The 7th International Conference on Web Information Systems Engineering, pp. 363–374. Springer LNCS 4255, Wuhan, China (2006)
5. Caro, A., Calero, C., Piattini, M.: Development process of the operational version of PDQM. In: Proceedings of the WISE2007, The 8th International Conference on Web Information Systems Engineering pp. 436–448. Springer LNCS, Nancy, Francia (2007)
6. Collins, H.: Corporate Portal Definition and Features. AMACOM (2001)
7. Eppler, M.: Managing Information Quality: Increasing the Value of Information in Knowledge-intensive Products and Processes. Springer, New York (2003)
8. Eppler, M., Algesheimer, R., Dimpfel, M.: Quality Criteria of Content-Driven Websites and Their Influence on Customer Satisfaction and Loyalty: An Empirical Test of an Information Quality

Framework. In: Proceedings of the Eighth International Conference on Information Quality, pp. 108–120 (2003)
9. Eppler, M., Muenzenmayer, P.: Measuring Information Quality in the Web Context: A Survey of State-of-the-Art Instruments and an Application Methodology. In: Proceedings of the Seventh International Conference on Information Quality, pp. 187–196 (2002)
10. Finkelstein, C., Aiken, P.: XML and Corporate Portals. Access in http://www.wilshireconferences.com/xml/paper/xml-portals.htm (1999)
11. Fugini, M., Mecella, M., Plebani P., Pernici B., Scannapieco M.: Data Quality in Cooperative Web Information Systems. Personal Communication. citeseer.ist.psu.edu/fugini02data.html (2002)
12. Gertz, M., Ozsu, T., Saake, G., Sattler, K.U.: Report on the Dagstuhl Seminar "Data Quality on the Web". SIGMOD Rec. **33**(1), 127–132 (2004)
13. Ginige, A., Murugesan, S.: Web Engineering: An Introduction. IEEE Multimed **8**(1), 14–17 (2001)
14. Graefe, G.: Incredible Information on the Internet: Biased Information Provision and a Lack of Credibility as a Cause of Insufficient Information Quality. In: Proceedings of the 8th International Conference on Information Quality, pp. 133–146 (2003)
15. Ivory, M., Rashmi, S., Marti, H.: Empirically Validated Web Page Design Metrics. In: Proceedings of the SIG-CHI on Human factors in computing systems (SIGCHI'01), pp. 53–60. Seattle, WA, USA (2001)
16. Katerattanakul, P., Siau, K.: Information quality in internet commerce desing. In: PiattiniCaleroGenero, M.C.M. (ed.) Information and Database Quality, pp. 45–56. Kluwer Academic, Dordrecht (2001)
17. Katerattanakul, P., Siau, K.: Measuring Information Quality of Web Sites: Development of an Instrument. In: Proceedings of the 20th International Conference on Information System, pp. 279–285 (1999)
18. Knight, S.A., Burn, J.M.: Developing a Framework for Assessing Information Quality on the World Wide Web. Inf. Sc. J. **8**, 159–172 (2005)
19. Lee, Y.: AIMQ: a methodology for information quality assessment. Inf. Manage. **40**, 133–146 (2002), Elsevier Science
20. Malak, G., Sahraoui, H., Badri, L., Badri, M.: Modeling web-based applications quality: a probabilistic approach. In: Proceedings of the 7th International Conference on Web Information Systems Engineering, pp. 398–404. Springer LNCS, Wuhan, China (2006)
21. Melkas, H.: Analyzing Information Quality in Virtual service Networks with Qualitative Interview Data. In: Proceedings of the Ninth International Conference on Information Quality, pp. 74–88 (2004)
22. Moraga, M.Á., Calero, C., Piattini, M.: Comparing different quality models for portals. Online Inf. Rev. **30**(5), 555–568 (2006)
23. Moustakis, V., Litos, C., Dalivigas, A., Tsironis, L.: Website Quality Assessment Criteria. In: Proceedings of the Ninth International Conference on Information Quality, pp. 59–73 (2004)
24. Naumann, F., Rolker, C.: Assesment Methods for Information Quality Criteria. In: Proceedings of the Fifth International Conference on Information Quality, pp. 148–162 (2000)
25. Neil, M., Fenton, N.E., Nielsen, L.: Building large-scale Bayesian Networks. Knowl. Eng. Rev. **15**(3), 257–284 (2000)
26. Nielsen, J.: Designing Web Usability: The Practice of Simplicity. New Riders, Thousand Oaks, CA (2000)
27. Pernici, B., Scannapieco, M.: Data Quality in Web Information Systems. In: Proceedings of the 21st International Conference on Conceptual Modeling, pp. 397–413 (2002)
28. Pipino, L., Lee, Y., Wang, R.: Data quality assessment. Commun. ACM. **45**(4), 211–218 (2002)
29. Pressman, R.: Software Engineering: a Practitioner's Approach, 5th edn. McGraw-Hill, New York (2001)
30. Redman, T.: Data quality: The field guide. Digital, Boston (2000)
31. Sahraou,i H., Boukadoum, M., Chawiche, H.M., Mai, G., Serhani, M.A.: A fuzzy logic framework to improve the performance and interpretation of rule-based quality prediction models for object-oriented software. In Proceedings of the 26th Computer Software and Applications Conference (COMPSAC02) (2002)
32. Shanks, G., Corbitt, B.: Understanding Data Quality: Social and Cultural Aspects. In: Proceedings of the 10th Australasian Conference on Information Systems, pp. 785–797. Wellington, New Zealand (1999)
33. Strong, D., Lee, Y., Wang, R.: Data quality in context. Commun. ACM. **40**(5), 103–110 (1997)
34. Wang, R., Strong, D.: Beyond accuracy: What data quality means to data consumers. J. Manage. Inf. Syst. **12**(4), 5–33 (1996) Armonk; Spring
35. Winkler, W.: Methods for evaluating and creating data quality. Inf. Syst. **29**(7), 531–550 (2004)
36. Xiao, L., Dasgupta, S.: User Satisfaction with Web Portals: An empirical Study. In: Gao, Y. (eds.) Web Systems Design and Online Consumer Behavior, pp. 193–205. Idea Group, Hershey, PA (2005)

37. Yang, Z., Cai, S., Zhou, Z., Zhou, N.: Development and validation of an instrument to measure user perceived service quality of information presenting Web portals. Inf. Manage. **42**, 575–589 (2004), Elsevier Science
38. Zhu, Y., Buchmann, A.: Evaluating and Selecting Web Sources as external Information Resources of a Data Warehouse. In: Proceedings of the 3rd International Conference on Web Information Systems Engineering, pp. 149–160 (2002)