Software Qual J (2008) 16:513–542
DOI 10.1007/s11219-008-9046-7

A proposal for a set of attributes relevant for Web portal data quality

Angélica Caro · Coral Calero · Ismael Caballero · Mario Piattini

Published online: 15 March 2008

Abstract Data Quality is a critical issue in today's interconnected society. Advances in technology are making the use of the Internet an ever-growing phenomenon and we are witnessing the creation of a great variety of applications such as Web Portals. These applications are important data sources and/or means of accessing information which many people use to make decisions or to carry out tasks. Quality is a very important factor in any software product and also in data. As quality is a wide concept, quality models are usually used to assess the quality of a software product. From the software point of view there is a widely accepted standard proposed by ISO/IEC (the ISO/IEC 9126) which proposes a quality model for software products. However, until now a similar proposal for data quality has not existed. Although we have found some proposals of data quality models, some of them working as "de facto" standards, none of them focus specifically on web portal data quality and the user's perspective. In this paper, we propose a set of 33 attributes which are relevant for portal data quality. These have been obtained from a revision of literature and a validation process carried out by means of a survey. Although these attributes do not conform to a usable model, we think that it might be considered as a good starting point for constructing one.

Keywords Data quality · Web portal · Data quality attributes · Data consumer

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

C. Calero · I. Caballero · M. Piattini
Alarcos Research Group. Information Systems and Technologies Department, University of Castilla-La Mancha, Paseo de la Universidad 4, Ciudad Real, Spain
e-mail: Coral.Calero@uclm.es

I. Caballero
e-mail: Ismael.Caballero@uclm.es

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

1 Introduction

Over the past decade, the number of organizations which have Web portals has grown considerably. Some of these organizations have established portals to complement, substitute, or widen existing services for their clients (among these we might mention: bank portals, university portals, cultural portals, commercial portals, etc.). In general, portals provide users with access to different data sources (providers) (Mahdavi et al. 2004), as well as to on-line information and information-related services (Yang et al. 2004). They also create a working environment which is easy for users to navigate to find the data they need in order to perform their operational or strategic functions speedily and/or to make decisions quickly (Collins 2001). In this context, organizations must face the challenge of achieving and maintaining a state of high data quality (Kopcso et al. 2000) because this aspect is a key factor for both them and their customers. Obviously, the higher the quality of portal data is, the more likely it is that users will return to the portal. Moreover, it is fundamental that Web portal users be able to evaluate the quality of the data obtained from a portal so that they can make decisions such as the choice of the best portal from various different ones.

In the relevant literature, the concept of Information Quality or Data Quality (hereafter DQ)[1] 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). Research on DQ began in the context of information systems (Strong et al. 1997; Lee 2002) and has been extended to contexts such as cooperative systems (Fugini et al. 2002; Marchetti et al. 2003; Winkler 2004), data warehouses (Bouzeghoub and Kedad 2001; Zhu and Buchmann 2002) or e-commerce (Aboelmeged 2000; Katerattanakul and Siau 2001), among others. Due to the particular characteristics of Web applications and their differences from the traditional information systems (Pressman 2001), the research community has started to deal with the subject of DQ on the Web (Gertz et al. 2004). However, in a systematic review of 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 highlight DQ as being one of the most relevant factors in the quality of a Web portal (Yang et al. 2004; Moraga et al. 2006).

It is important to note that, as opposed to software products in which the ISO/IEC 9126 standard with the definition of a general model for software quality is available, no such standard exists for data quality. Perhaps this is due to the fact that data may be considered as simply another software product and that the ISO/IEC 9126 can therefore be applied (or tailored) to this context. However, experience has demonstrated that this is not true and that data have some peculiarities that are not shared with a general software product.

In fact ISO/IEC is now working on the SQUARE (Software Quality Requirements) family of standards. SQUARE will contain the ISO/IEC 25010 that will provide a model for software product quality, defining software product quality characteristics and how they are decomposed into sub-characteristics, but it will also contain the ISO/IEC 25012 in which a model composed by a set of DQ characteristics will be included.

It is well known that a normal manner in which to work with (general) quality models is to tailor them to a specific domain. In the case of the ISO/IEC 9126 there are several proposals that do this. Obviously, this is not the case for DQ because still there is no standard which can be used to tailor, in our case, the data portal domain.

[1] As with much of the research into DQ, in this paper we will use the terms information and data as being synonymous.

The aim of our research, therefore, is the identification of a set of attributes which are relevant for the assessment of Web portal data quality, conforming to a good basis for further tailoring process for different portal contexts such as, for example, bank, university, business etc. portals. Along with the aforementioned aims, our objective is to do this from the point of view of the data consumer. This focus differs from the data producer's or data custodian's perspective in two important aspects (Burgess et al. 2004): (1) Data consumers have no control over the quality of available data and (2) the aim of consumers is to find data that match their personal needs, rather than to provide data that meet the needs of others.

The idea is, then, to offer a general set of DQ attributes which are relevant from the data consumer point of view in such a manner that they can be used in the definition of a portal data quality assessment process. We believe that this proposal might even be useful for ISO/IEC because this set could be studied for its inclusion (total or partial) in the new data quality standard.

We have used three basic elements to identify this set of attributes: (1) a set of Web DQ attributes identified in literature, (2) the expectations with regard to DQ according to data consumers on the Internet, as described by Redman in (Redman 2000), and (3) the functions that a Web portal may offer its users (Collins 2001).

The rest of the paper is organized as follows. In Sect. 2, the basic elements used to identify the set of attributes are presented. Section 3 shows how we have combined them. In Sect. 4 we present the validation of the set of attributes obtained. Section 5 presents an initial phase with which to organize these attributes within a structure. Finally, Sect. 6 describes our conclusions and future work.

2 Basic elements of our study

As we have already mentioned, our objective is to identify a set of attributes which are relevant for web portal data quality from the data consumer point of view. Taking all these considerations into account, we have worked with three basic elements for the purpose of identifying data quality attributes: (1) The data consumer perspective, (2) A set of Web DQ attributes, and (3) The basic Web portal functions. The following subsections will describe each one of these briefly.

2.1 Data consumer perspective

In the late 1990s, the most frequent definition of quality was that of meeting and exceeding customers' expectations (Reeves and Bednar 1994). The notion of quality as meeting expectations suggests that quality is defined by conformance to customer expectations. These may relate to excellence, value or to other salient attributes that are relevant to consumers in shaping their perceptions of quality (Nelson et al. 2005). This situation is not different in the context of data quality; most authors define this concept as "fitness for use" (Strong et al. 1997; Cappiello et al., 2004). Moreover, the view of assessing DQ which is currently accepted involves understanding it from the point of view of the user (or data consumer) (Knight and Burn 2005).

Taking the above into account, we decided to focus our work on the perspective of the data consumer, this being the first basic element of our study. To represent this perspective we have used a study developed by Redman (2000), in which he established the DQ

expectations of data consumers on the Internet. We have used Redman's definitions of each of these categories as a basis through which to identify which of the DQ attributes are appropriate when considering the DQ expectations of the data users or consumers of a Web portal.

These expectations are grouped into six categories: Privacy, Content, Quality of Values, Presentation, Improvement and Commitment (See definitions in Appendix A).

2.2 Web data quality attributes

The notion of DQ has been widely studied in literature and is commonly approached as a multi-dimensional concept (Wang and Strong 1996; Redman 2000; Cappiello et al. 2004; Gertz et al. 2004). We can, furthermore, observe that various DQ attributes have been proposed, according to an author's philosophical view-point (Knight and Burn 2005) and the context studied. With the idea of taking advantage of work already carried out and applying it to Web portals, we also decided to recompile DQ attributes proposed in literature for Web and/or the context of Web portals. Next, and by following the methodology proposed in (Kitchenham 2004), we carried out a systematic review of the relevant literature and selected those works in which DQ attributes which were applicable to our particular context were proposed. Works for different domains in the Web context were selected. Among these were: data integration (Naumann and Rolker 2000; Bouzeghoub and Peralta 2004), e-commerce (Katerattanakul and Siau 2001), Web information portals (Yang et al. 2004), cooperative e-services (Fugini et al. 2002), decision making (Graefe 2003), organizational networks (Melkas 2004) and data quality on the Web (Katerattanakul and Siau 1999; Eppler et al. 2003; Gertz et al. 2004; Moustakis et al. 2004).

As a result of this review, it was possible to define a basic set of one hundred DQ attributes proposed for different domains in the Web. Table 1 shows the research works used as sources of Web DQ attributes; the author, the Web domain and the number of DQ attributes obtained from the model/framework are shown for each of them.

Table 1 Research works used as source of Web data quality attributes

Autor	Domain	No. of DQ attributes obtained from the model/framework
Katerattanakul and Siau (1999)	Personal web sites	6 DQ attributes
Katerattanakul and Siau (2001)	e-Commerce	
Naumann and Rolker (2000)	Data integration	22 DQ attributes
Pernici and Scannapieco (2002)	Web information systems (data evolution)	4 DQ attributes
Fugini et al. (2002)	e-Service cooperative	8 DQ attributes
Graefe (2003)	Decision making	8 DQ attributes
Eppler et al. (2003)	Web sites	16 DQ attributes
Gertz et al. (2004)	DQ on the web	5 DQ attributes
Moustakis et al. (2004)	Web sites	4 DQ attributes
Melkas (2004)	Organizational networks	20 DQ attributes
Bouzeghoub and Peralta (2004)	Data integration	2 DQ attributes
Yang et al. (2004)	Web information portals	5 DQ attributes

It is interesting to note the lack of consensus between the researchers in this area with regard to the terminology used to refer to the DQ attributes (we can found terms as: attribute, dimension, characteristic, factor or criterion are used to indicate the same concepts) and the definitions of them.

Although not all of the DQ attributes obtained from literature were proposed to be used in the evaluation of DQ from the data consumer's point of view, their relevance and importance (visibility) for the data consumer were analyzed before their selection.

2.3 Web portal functionalities

DQ needs to be assessed within the context of its generation and intended use (Katerattanakul and Siau 1999; Knight and Burn 2005). As a result of their research, Strong et al. (1997), conclude that in order to study DQ it is necessary to incorporate the task context and the process by which users access and manipulate data to meet their task requirements. Therefore, in order to represent the context of Web portals, we have selected a set of basic software functions which represent the basic services that a Web portal might offer its users in order for them to access and manipulate data. The process presented in this paper assumes that a data consumer judges the quality of data when he or she carries out his/her task using these functionalities or services.

As basic functions of a Web portal we have used those proposed by Collins in (Collins 2001). These functions are as follows: *Data Points and Integration, Taxonomy, Search Capabilities, Help Features, Content Management, Process and Action, Collaboration and Communication, Personalization, Presentation, Administration*, and *Security*.

3 Combining the elements

Having described our three basic elements, we shall now describe the four-phase process defined to identify the set of DQ attributes with which to evaluate the DQ in web portals using the data consumer's perspective (see Fig. 1).

During the first phase, the Web DQ attributes which we believe may be applicable to Web portals are identified in the pertinent literature. In the second phase, a matrix for the classification of the attributes obtained in the previous phase is built.

In the third phase, the matrix obtained is used to analyse the applicability of each Web DQ attribute in the Web portal DQ evaluation. Finally, in the fourth phase, the model is validated by means of a survey performed with a group of Web portal data consumers. In the following subsections we will explain each phase.

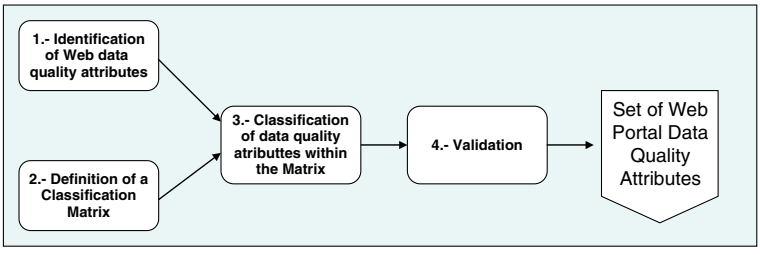

Fig. 1 Development process

3.1 Identification of web data quality attributes

To carry out this phase we used a set of one hundred Web DQ attributes obtained from literature (see subsection 2.2). On analyzing the names and definitions of the DQ attributes, we realized that we could reduce the amount of attributes. So, having detected synonyms and homonyms amongst the set of attributes, we combined these obtaining a final set of 41 Web DQ attributes.

Table 2 uses columns to show the final set of attributes, and rows to show the works in which they were proposed. It also shows the total number of works that make reference to each attribute. In addition, the symbols $\times$ and $\otimes$ are used to represent how they were combined: $\times$ indicates the same name and meaning while $\otimes$ shows that only the meaning is similar.

3.2 Definition of a classification matrix

Once we had obtained an initial set of DQ attributes which were potentially applicable to the evaluation of DQ in the context of a Web portal, we next defined a matrix which would allow us to analyze the relevance of each DQ attribute. The matrix defined represents two of the three basic elements of our study: the data consumer perspective and the basic Web portal functionalities.

The idea is to reflect the fact that a data consumer assesses the DQ in a Web portal when he or she uses the different functions (or services) which it offers. Our aim, therefore, is to form a relationship between the functions of a Web portal and the DQ expectations of data consumers. In other words, we shall use these relationships to attempt to determine what the data consumer expects as regards data content, or what is delivered by a Web portal, in deciding whether it is "fit for use".

In order to build the matrix, we have used the basic functions in a Web portal (Collins 2001) (presented in subsection 2.3) for one dimension of the matrix and the DQ expectations of the data consumer on the Internet (Redman 2000) (presented in subsection 2.1) for the other dimension. By using the definitions of the functions and expectations as a base, we next analyzed which expectations were related to which portal functions. The result is represented in Fig. 2, where "$\surd$" shows that the interaction between a pair (function, expectation) makes sense.

In the following paragraphs we shall explain the relationship (function, expectation) established in the matrix for each function. We shall first define the function and then the expectations that were related to it.

- Data points and integration. *These provide the ability to access information from a wide range of internal and external information sources and to display the resulting information on the single point-of-access desktop*. The expectations applied to this function are: *Content* (the consumers need a description of the portal areas covered, use of published data, etc.), *Quality of value* (the data consumer should expect the results of searches to be correct, up-to-date and complete), *Presentation* (formats, language, and other aspects are very important for easy interpretation) and *Improvement* (users not only wish to participate in portal improvements with their opinions, but also to know what the results of applying them are).
- Taxonomy. *This provides the context of information (including the organization-specific categories that reflect and support the organization's business)*. In our opinion,

Table 2 Web data quality attributes 1–41

Author	Year	Accessibility	Accuracy	Amount of data	Applicability	Attractiveness	Availability	Believability	Completeness	Concise representation	Consistent representation	Ccost effectiveness	Customer support	Currency	Documentation	Duplicates	Ease of operation	Expiration	Flexibility	Granularity	Interactive	Internal consistency	Interpretability	Latency	Maintainable	Novelty	Objectivity	Ontology	Organization	Price	Relevancy	Reliability	Reputation	Response time	Security	Specialization	Source's information	Timeliness	Traceability	Understand ability	Validity	Value-added	Number of attributes
Nauman and Rolker	2000		x	x			x	x	x	x	x		x		x								x	x			x			x	x	x	x	x	x			x	⊗	x		x	22
Katerat-tanakul and Siau	1999, 2001	x	⊗			x																							x							x			⊗			6
Eppler	2001	x	x		x				x	⊗	⊗			x			⊗				x	⊗			x								⊗	x			x	x	⊗				16
Fugini et al.	2002		⊗		x			⊗	x													x									⊗			⊗			x					8
Pernici and Scann-apieco	2002		x						x									x														⊗										4
Graefe	2003	⊗					⊗	x															x			x					⊗								⊗	⊗		8
Bouzeg-houb and Peralta	2004													x																							x						2
Gertz	2004								x					x		x				x							x																5
Melkas	2004	x	x	⊗					x	x	x	x	x				x		x				x				x				x		x	⊗	x			x	x	⊗		x	20
Moustakis	2004				⊗				⊗																						⊗				x							4
Yang et al.	2004		x						⊗	x				x																	⊗											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	1	6	2	2	3	4	1	1	5	3	4	1	3	

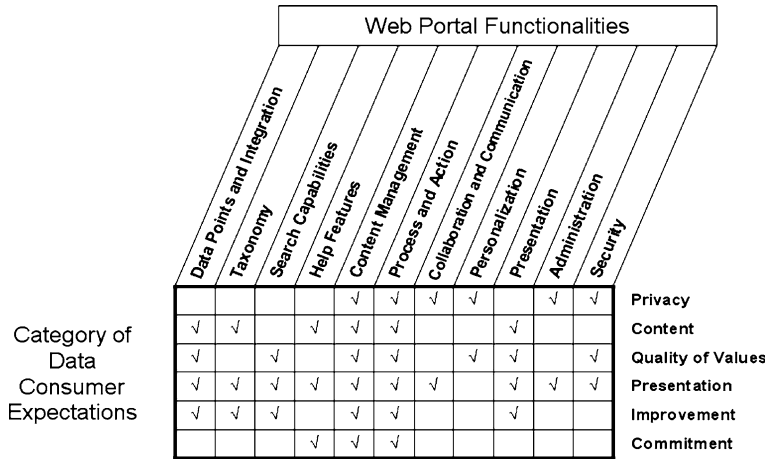

Fig. 2 Matrix for the Web data quality attributes classification

the expectations of the data consumer are: *Content* (consumers need a description of which data are published and how they should be used, as well as easy-to-understand definitions of every important term, etc.), *Presentation* (formats and language in the taxonomy are very important for easy interpretation; users should expect to find instructions when reading the data), and *Improvement* (the user should expect to be able to convey his/her comments on data in the taxonomy and to know what the result of improvements are).

- Search capabilities. *These provide several services for Web portal users and their needs, supporting searches throughout the company, the World Wide Web, and in search engine catalogs and indexes*. The expectations applied to this functionality are: *Quality of values* (the data consumer should expect the results of searches to be correct, up-to-date and complete), *Presentation* (formats and language are important for consumers, both in searching and in easy interpretation of results) and *Improvement* (consumers should expect to be able to convey their comments and suggestions about the data connected with the search capacity of the portal. They would also expect to be made aware of the results of improvements).
- Help features. *These provide help when using the Web portal*. The expectations applied to this functionality are: *Presentation* (formats, language, and other aspects are very important for the easy interpretation of help texts) and *Commitment* (the consumer should be able to ask and obtain answers to any question regarding the proper use or meaning of data, update schedules, etc, with ease.).
- Content management. *This function supports content creation, authorization, and inclusion in (or exclusion from) Web portal collections*. The expectations applied to this functionality are: *Privacy* (a privacy policy which can be used by all data consumers to manage access to sources and guarantee Web portal data should exist), *Content* (consumers need a description of data collections and to see that all the data needed for an intended use are provided, etc.), *Quality of values* (a consumer should expect all data values to be correct, up-to-date and complete, unless otherwise stated), *Presentation* (formats and language should be appropriate for easy interpretation), *Improvement* (consumers should expect to be able to convey their comments on content and its management and to be made aware of the results of any improvements) and

Commitment (consumers should find it easy to ask any questions regarding the proper use or meaning of data, the updating of schedules, and so on, and to have them answered).

- Process and action. *This function enables Web portal users to initiate and participate in a business process of a portal owner*. The expectations applied to this functionality are: *Privacy* (data consumers should expect there to be a privacy policy with which to manage data about the business on the portal), *Content* (consumers should expect to find descriptions about data related to processes and actions, along with appropriate and inappropriate uses. They would also expect all the data needed for the process and actions to be provided, etc.), *Quality of values* (that all data associated with this function are correct, up-to-date and complete, unless otherwise stated), *Presentation* (formats, language, and other aspects are very important in the correct interpretation of data), *Improvement* (a consumer should expect to be able to convey his/her comments on contents and their management and to know the results of improvements) and *Commitment* (the consumer should be able to ask and obtain an answer to any question regarding the proper use or meaning of data in a process or action, etc., with ease.).
- Collaboration and communication. *This function facilitates discussion, the location of innovative ideas, and the recognition of resourceful solutions*. The expectations applied to this functionality are: *Privacy* (the consumer should expect a privacy policy for all consumers who participate in the activities of this function), and *Commitment* (a consumer should be able to ask any questions connected with the proper use or meaning of data for collaboration and/or communication, etc. and to have these questions answered, with ease).
- Personalization. *This is a critical component in creating a working environment that is organized and configured specifically for each user*. The expectations applied to this functionality are: *Privacy* (the consumer should expect privacy and security with regard to their personalized data, profile, etc.), and *Quality of values* (data about the user profile should be correct and up-to-date).
- Presentation. *This provides Web portal users with both the knowledge desktop and the visual experience that encapsulates all of the portal's functionality*. The expectations applied to this functionality are: *Content* (the presentation of a Web portal should include data about areas covered, appropriate and inappropriate uses, definitions, information about sources, etc.), *Quality of values* (the data of this function should be correct, up-to-date and complete.), *Presentation* (formats, language, and other aspects are very important for the easy interpretation and appropriate use of portal data.) and *Improvement* (the consumer should expect to be able to convey his/her comments on contents and their management and to become aware of the results of any improvements).
- Administration. *This function provides a service for deploying maintenance activities or tasks associated with the Web portal system*. The expectations applied to this function are: *Privacy* (Data consumers need security for data about the portal administration) and *Quality of values* (Data connected with administrative tasks or activities should be correct and complete).
- Security. *This provides a description of the levels of access that each user or groups of users are allowed for each portal application and software function included in the Web portal*. The expectations applied to this functionality are: *Privacy* (consumers need a privacy policy with regard to the data of the levels of access of data consumers).*Quality of values* (data about the levels of access should be correct and up-to-date) and *Presentation* (data about security should be in a format and language which are easy to interpret).

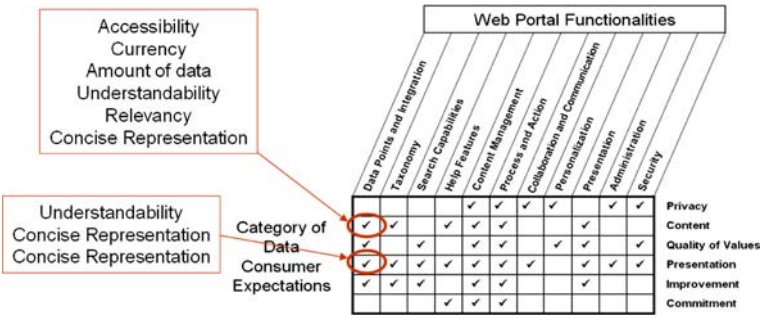

Fig. 3 Classification of Web DQ attributes within the matrix

3.3 Classification of web data quality attributes in the matrix

The third phase in the identification of the DQ attributes (see Fig. 1), consisted of the classification of the Web DQ attributes (shown in subsection 3.1) within the relationships (functionality, expectation) established in the classification matrix (see previous subsection). The idea was to assign the appropriate attributes in order to assess DQ (see Fig. 3), taking into account the DQ expectation of data consumers by functionality. For example, Fig. 3 shows the DQ attributes that have been identified for two relationships of the Data Points and Integration functionality. Table 3 summarizes the DQ attributes for each functionality. The last line of the matrix shows the total number of functionalities to which each attribute is applicable. The number of attributes which are applicable to a specific functionality are shown in the final column.

As can be seen in Table 3, there are seven DQ attributes which were not classified in the matrix: Cost effectiveness, Granularity, Internal consistency, Latency, Maintainable, Ontology and Price. This is basically due to the fact that, in our opinion, these attributes are not applicable to any relationship established in the matrix such as, for example, Granularity and Internal Consistency. To be specific, Granularity is defined as "the degree of granularity in the sources, which allows us to describe the properties of data in more detail" (Gertz et al. 2004). We consider that for data consumers, it might be very difficult to know the level of granularity of data sources, so they cannot therefore evaluate whether the granularity presented is that which is most appropriate for the right use. With regard to the Internal Consistency attribute, this is defined as "the degree to which the values of the attributes of an instance of a schema element satisfy the specific set of semantic rules defined on the schema element" (Fugini et al. 2002). This attribute obviously reflects internal aspects of data which are not accessible to data consumers.

As a result of this phase we have obtained a set of 34 DQ attributes which can be used to assess the DQ in Web portals. These attributes are outlined in Table 4. Their definitions can be seen in Appendix B.

4 Validation of the set of attributes

As is outlined in the previous section, up to this point, we have established relationships between the functionalities of a Web portal and the DQ expectations of the data consumer on the Internet. On the basis of these relationships we have intuitively identified the Web DQ attributes which could be used to evaluate the quality of data in a Web portal.

Table 3 Data quality attributes for functionality

Functionalities	Accessibility	Accuracy	Amount of data	Applicability	Attractiveness	Availability	Believability	Completeness	Concise representation	Consistent representation	Cost effectiveness	Customer support	Currency	Documentation	Duplicates	Ease of operation	Expiration	Flexibility	Granularity	Interactive	Internal consistency	Interpretability	Latency	Maintainable	Novelty	Objectivity	Ontology	Organization	Price	Relevancy	Reliability	Reputation	Response time	Security	Specialization	Source's information	Timeliness	Traceability	Understand ability	Validity	Value-added	Number of attributes
	7	4	9	2	1	3	6	5	9	2	0	8	6	1	1	8	4	1	0	2	0	5	0	0	3	2	0	4	0	7	7	2	2	5	3	1	5	7	11	8	1	
Data points and integration	✓		✓				✓		✓	✓		✓	✓				✓								✓			✓		✓	✓								✓	✓	✓		16
Taxonomy	✓		✓		✓		✓		✓			✓	✓			✓													✓	✓								✓	✓			11
Search capabilities	✓		✓				✓	✓	✓			✓	✓			✓									✓					✓		✓	✓				✓	✓				15
Help features	✓		✓						✓			✓				✓									✓					✓	✓						✓	✓	✓			8
Content management	✓		✓		✓	✓	✓	✓	✓			✓	✓	✓	✓	✓				✓		✓			✓			✓		✓	✓			✓	✓		✓	✓		✓		28
Process and action	✓		✓		✓	✓	✓	✓	✓			✓	✓			✓	✓			✓		✓			✓			✓		✓	✓			✓	✓		✓	✓				26
Collaboration and communication						✓						✓										✓								✓				✓				✓				6
Personalization	✓							✓					✓				✓																				✓	✓				7
Presentation			✓		✓		✓		✓			✓	✓			✓	✓	✓				✓						✓		✓				✓			✓	✓	✓			18
Administration			✓						✓	✓						✓																	✓					✓				6
Security	✓		✓					✓	✓							✓						✓						✓					✓			✓	✓	✓			11
Number of functionalities																																									

Table 4 Data quality attributes

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

In line with the method we have defined, the last phase is its validation (see Fig. 1, first part, phase 4). The objective is to validate the set of DQ attributes and, if necessary, to eliminate those DQ attributes which are considered less important from the data consumer's point of view or to add new attributes, if the users indicate them.

4.1 Method

In order to perform this phase, we decided to carry out a study by conducting a survey. To carry out this survey, "the principles of survey research" proposed in (Pfleeger and Kitchenham 2001; Kitchenham and Pfleeger 2002a, b, c, d, 2003) were used. These principles state that a survey is a comprehensive system for collecting information to describe, compare or explain knowledge, attitudes and behaviour (Pfleeger and Kitchenham 2001). The survey instrument is thus part of a larger survey process with clearly-defined activities; see Table 5.

4.2 Activity 1. Setting specific, measurable objectives

The objective of our survey was defined as: "*To obtain the opinion of Web portal data consumers with respect to the importance of each of the DQ attributes selected in the previous phases of our method*".

4.3 Activity 2. Planning and scheduling the survey

Having taken the aforementioned objective into account, our survey was targetted towards Web portal data consumers. Bearing in mind that working with students implies certain advantages, such as the fact that these subjects' knowledge tends to be homogenous and that it is possible to count on a high number of subjects at the same time, we decided to carry out the survey with a group of students who have experience as Web portal users and who therefore comply with the role of data consumers. The students' knowledge is, in this case, considered to be homogenous because they were all experienced in the use of Web portals as data consumers, and were therefore capable of distinguishing between one portal and another with regard to the data quality that could be obtained from them. In our opinion, this previous experience favours the fact that they had different points of view

Table 5 Main activities of a survey process

Activity	Description
1. Setting specific, measurable objectives	The objectives are essential for all subsequent survey process activities, in particular to instrument design and development, and must be clear and measurable. Each objective is a statement of the survey's expected outcomes
2. Planning and scheduling the survey	The target population must be determined, based on the objectives. The best way to obtain the information needed has to be decided. It is also necessary to determine factors such as an appropriate sample size and the largest possible response rate
3. Ensuring that appropriate resources are available	To undertake a survey, resources to support the survey design are required. For instance, resources to build the instrument and to distribute it
4. Designing the survey	Depending on the survey's objectives and the resources available, the survey design must be selected. It may be a descriptive or experimental design
5. Preparing the data collection instrument	The survey instrument is usually a questionnaire. The appropriate questions must be selected when constructing the questionnaire. In addition, the appropriate type and number of questions, the questionnaire format, etc must be determined
6. Validating the instrument	The instrument must be evaluated, i.e., survey reliability (how reproducible a survey's data is) must be assessed, along with survey validity (how well a survey instrument measures what it sets out to measure). The two most common ways to organize an evaluation are focus groups and pilot studies
7. Selecting participants	Once a target population has been defined, a sampling method must be selected to obtain the sample for the survey. The two most common sampling methods are the probabilistic and non-probabilistic methods
8. Administering and Scoring the instrument	These activities consist of the application of the survey instrument to the sample defined
9. Analyzing the data	This corresponds to the analysis of the data collected. The data must first of all be validated in order to eliminate the inconsistent and incomplete responses. Many analysis techniques may then be used to analyze the data
10. Reporting the results	In the research community, conferences and academic journals and are common places to publish the results of a survey

both about the DQ in Web portals and about the importance of each DQ attribute within this context.

Having considered the previously mentioned advantages and the fact that the tasks in question were simple, we decided that it would be appropriate to carry out the survey with students.

As regards sample size, this was the total number of students in the software engineering class in the final-year (fifth) of Computer Science and was made up of 70 subjects.

4.4 Activity 3. Resources available in developing the survey

We were able to create the instrument, to access a group of subjects, to administrate the survey and to analyze the results. Consequently, we believe that we did indeed obtain all the resources necessary to perform the survey.

4.5 Activity 4. Designing the survey

With the survey objective in mind, a descriptive design was selected. In our case, we wish to describe the importance given by data consumers to the set of DQ attributes previously identified. So, we think that this type of design is appropriate for describing the phenomenon we are interested in (Kitchenham and Pfleeger 2002a).

4.6 Preparing the data collection instrument

To prepare the questionnaire, the questions were chosen by bearing in mind the purpose and goal of the survey (Kitchenham and Pfleeger 2002c) mentioned previously. Hence, we chose one closed question to ask about the importance of each attribute selected and one open question to ask about any other aspect that was important to the data consumer but which had not been considered in the questionnaire for considering its inclusion in the set of attributes. The questions were created by using conventional language and expressing simple ideas. Negative questions were not included. In each one of the first 34 questions, the attribute asked for was formulated in terms of its definition. In an effort to reduce the time used to complete the survey, we standardized the response format (Kitchenham and Pfleeger 2002c). All closed questions were measured using a 5-point Likert scale anchored by "1" as "Not Important" to "5" being "Very Important" (See Appendix C).

4.7 Validating the instrument

The instrument was validated by following two strategies. First, as a pilot study, we used a survey previously developed for a subset of DQ attributes, specifically for the attributes classified for the Web portal function "Data Points and Integration" (In this pilot survey the subjects were contacted by e-mail, and a total of 69 effective responses were received; more details of this work are available in (Caro et al. 2006)). The purpose of this was to use this experience to prepare a better questionnaire. For example, in this survey we also asked about the importance of the DQ attributes, but gave only the name of each attribute. The result, however, was that many respondents reported that it was not easy to give a reply to this question on these terms. In the new questionnaire, the questions contained the complete description of each attribute.

Secondly, a pre-test of the questionnaire was given to 10 respondents (all PhD students with experience as Web portals users). The purpose of the pre-test was to improve the questions that a majority of respondents did not understand or considered confusing. As a result, two questions were modified to achieve better understanding on the part of the respondents.

4.8 Selecting participants

To choose the subjects for the survey, the non-probabilistic method of "convenience sampling" was used (Kitchenham and Pfleeger 2002d). The sample consisted of 70 subjects (all students in a software engineering class).

4.9 Administering and scoring the instrument

The questionnaire was delivered directly to the subjects, in printed format, and the nature of the study, the time given for completing the survey (less than twenty-five minutes), and its importance were explained.

4.10 Analyzing the data

We performed the survey with an expected sample of 70 subjects (see subsection 4.8). In practice, the questionnaire was answered by 54 subjects, as the remaining students did not attend, so our response rate was, therefore, 77 %. After data screening, we found one questionnaire with one question unanswered. Taking into account the independence of each question, we did not eliminate this particular questionnaire and it was only this single response that was not considered.

Descriptive statistics of the 34 DQ attributes are presented in Table 6. Most of the 34 DQ attributes had a range of 2 to 5 (21 of them). The exceptions were *Accessibility, Currency* and *Availability* which had the smallest range with values of between 3 and 5; these attributes were considered the most important for the respondents. Other exceptions were: *Documentation, Duplicates, Expiration, Source Information, Interactivity, Objectivity, Customer support, Traceability, Validity and Value Added*, which had a full range of 1–5. Among them, we can find the attributes with less importance for the respondents. On the other hand, most of the 34 DQ attributes had means that were higher than 3; that is, most of the DQ attributes surveyed were considered to be moderately important, or of a greater level of importance.

The last question (number 35), which dealt with new attributes not included in the questionnaire, was not answered by any of the subjects.

Table 6 Descriptive statistics of DQ attributes

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	Organization	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	Specialization	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

These results led us to decide that those DQ attributes which had a mean of 3 or more would be conserved; we similarly rejected those attributes that did not fulfil these conditions. Thus, the DQ attribute "Source Information" was eliminated. We are consicious that the difference between the mean of the eliminated attribute and the means of the other attributes is not very large. It is, however the only one which is under 3, which is to say that it is below the mean concept of importance with which participants were asked to evaluate the attributes in the survey. Therefore, if we look further than the numerical difference, it seemed to be conceptually relevant to eliminate it.

As a consequence of this validation, the final set is composed of a set of 33 DQ attributes. This number of attributes might appear to be very high, but this is not a problem as it allows us to ensure that all aspects which are relevant to users are taken into account. What is more, it will be possible to adjust this set of attributes once the model has been adapted to a specific domain.

4.11 Threats to validity

As is usually the case, different threats can affect the validity of the results of an experiment. In this section, and by following the framework proposed in (Wohlin et al. 2000), we discuss some threats that affect the following types of validity-construct, internal, external and conclusion.

Construct validity: This threat is concerned with the relationship between theory and observation. We used a 5-point Likert scale, which allows respondents to express a numerical opinion on a scale of 'not important' to 'very important'. We considered that this scale was efficient enough to gather the opinion of the subjects. We are, however, conscious of the fact that certain studies consider ratings to be less reliable than rankings (Krosnick 1999) and do not, therefore, dismiss the idea of carrying out new experiments with which to evaluate the model generated at a future date.

Internal validity: Internal validity is the degree to which conclusions about the causal effect of independent variables on dependent variables can be drawn. A lack of internal validity could lead to results that are not derived from causal relationships. With regard to internal validity, we considered the following issues carefully:

- *Differences between subjects*: All the subjects had the same profile (students enrolled in a software engineering class), thus reducing the variability between and among subjects.
- *Problems with the language*: We used subjects from Spain and the survey was written in Spanish, so no problems arose in this respect.
- *Task complexity*: The tasks were the same for all the subjects, so this aspect was not considered to be threat.
- *Persistence effect*: This was not present, as the subjects had never participated in a similar survey.
- *Learning effects*: This aspect was not considered to be a threat as the survey was only applied once and therefore no learning took place.
- *Fatigue effects*: The time taken in completing the survey was around 20 min. Considering the type of responses expected and received, we can reasonably consider that the effect of fatigue is minimal.
- *Subject motivation*: The subjects were volunteers and were convinced that their contribution was very important for research in the field of Data Quality. Participation in the experiment was on a voluntary basis and the experiment was not part of the

students' formal assignments. Since "Quality on the Web" was a topic which the subjects had already grasped, they were interested in participating. It is thus reasonable to claim that the subjects were sufficiently motivated.
- *Other factors*: Plagiarism and the subjects' influence on each other were controlled. They were informed that they should not talk to other subjects or share answers with them.

External validity: External validity is the degree to which the results of the research can be generalised to the population under study and to other research settings. The greater the external validity, the more the results of an empirical study can be generalised to actual software engineering practice. If external validity is not assured, the empirical results cannot be generalized to the population. As far as external validity is concerned, the following issue was considered:

- *Material and task used*: The material delivered was a printed questionnaire of the survey, and the subjects did not carry out any previous tasks in order to answer the survey.
- *Subjects*: Due the difficulty of gathering a group of subjects wishing to participate in the survey together in a short amount of time, this survey was conducted by using students. We justified the decision to use students by the following two reasons. First, many authors agree that, for many phenomena, using students has little impact on the external validity of a study (Höst et al. 2000; Carver et al. 2003) and the results of a study. Second, in our particular case, the survey needed the respondents to have experience in the use of Web portals and not skills or knowledge in any technical aspects. So, the computer science students are appropriate for our study, basically because the students had experience as Web portal users, were data consumers of these applications and clearly represent the population under study.

Conclusion validity: Conclusion validity defines the extent to which conclusions are statistically valid. The only issue that might affect the statistical validity of this study is the size of the sample (54 subjects). We are not concerned about this, because we consider that the 33 attributes identified at this stage is a high enough quantity for assessing the DQ in Web portals. These attributes have, moreover, previously been used in contexts which are similar to those of Web portals and have, in the majority of cases, also been validated.

5 Structuring the DQ attributes

In order to advance towards our final objective (the definition of a DQ model for web portals) we have decided to organize the set of DQ attributes identified within a structure. For this, we have used as basis one of the DQ models which is most frequently used and referenced in literature, the Wang and Strong model [43]. In the following subsections we will explain how we have defined an initial structure for the DQ attributes.

5.1 Wang & Strong model

The most widely known model amongst those which are currently available within the DQ field, and which is used as a standard "de facto", is the hierarchical framework created by the Wang and Strong model [43]. This model contains fifteen attributes of data quality, and is organized in four categories, see Table 7.

Table 7 Wang and Strong model

DQ Category	Description	DQ attributes
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	Access, Security
Contextual	It highlights the requirement which states that data quality must be considered within the context of the task in hand	Amount of Information, Completeness, Relevancy, Timeliness, Value-Added
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, Easy of Understanding, Concise Representation, Consistent Representation

We have therefore used this model as basis from which to generate an initial structure for our DQ portal model.

5.2 Tailoring the Wang & Strong model

By bearing in mind the definition of each DQ category and the definition of each of the DQ attributes identified, we have attempted to classify each attribute in one of four categories, as is shown in Table 8.

However, as Table 8 shows, certain attributes were not assigned to any of the categories. This was due to the fact that, given the definitions of these attributes, it was not possible for them to be directly assigned. Because of this, we decided to adapt one of the categories and redefine it in order to adjust it to the context of Web portals, thus allowing these attributes admitted. In particular, the Accessibility category has been retitled as Operational category. With this new name, our intention is to emphasize the importance of the role of systems, not only with respect to accessibility and security, but also in terms of personalization, collaboration, etc.

5.3 Final structure for Web portal data quality attributes

As a result of the previous classification, we have obtained a hierarchical structure of the attributes, which can be seen in Fig. 4.

We believe that, in spite of the hierarchical nature of this structure and the fact that it only has two levels, it is sufficiently flexible to subsequently be used to refine and modify it in function of its application to DQ evaluation.

5.4 The DQ model proposed and ISO/IEC 25012 draft

Given that ISO/IEC 25012 (N3792 - 2007) is still not a definitive model we shall only comment upon certain of its relevant aspects which will allow us to compare it with the model proposed in this paper.

Table 8 Classifying the DQ attributes in the four DQ categories of the Wang and Strong model

DQ Categories	Accessibility	Accuracy	Amount of data	Applicability	Attractiveness	Availability	Believability	Completeness	Concise representation	Consistent representation	Customer support	Currency	Documentation	Duplicates	Ease of operation	Expiration	Flexibility	Interactive	Interpretability	Novelty	Objectivity	Organization	Relevancy	Reliability	Reputation	Response time	Security	Specialization	Timeliness	Traceability	Understand ability	Validity	Value-added
DQ Intrinsic		✓					✓					✓		✓		✓					✓				✓					✓			
DQ Accessibility	✓																										✓					
DQ Contextual				✓				✓									✓			✓			✓	✓				✓			✓	✓
DQ Representation		✓	✓		✓				✓	✓			✓						✓			✓									✓	

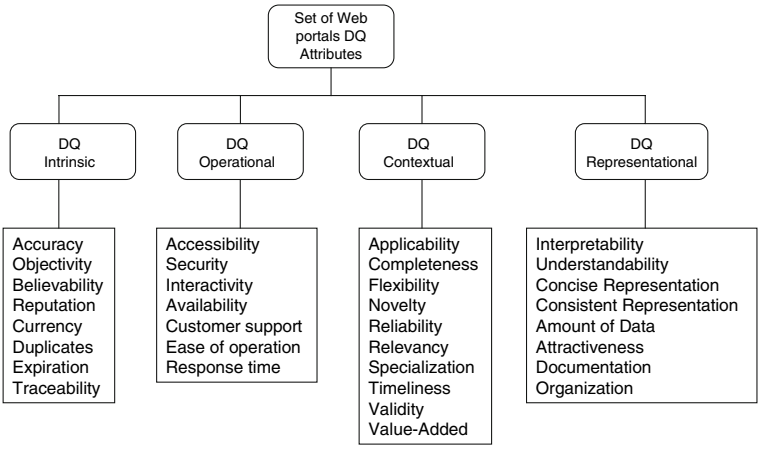

Fig. 4 Final structure of the Web portal DQ attributes

We could specifically mention that ISO/IEC proposes a model made up of 16 DQ attributes as opposed to ours which proposes 33 DQ attributes which are grouped into four categories. ISO/IEC considers that it is possible to tackle the measurement of these attributes from two points of view: Inherent and Extended. The difference between these points of view centres upon whether the evaluation of DQ will be made independently of the technological system to which the data are associated or whether, on the contrary, it will be made by considering them. We tackle the evaluation from the point of view of the data consumer, and therefore from a point of view which goes further than that of technological aspects. We also tackle aspects which are relative to usability and quality in the use of the data. Our focus includes aspects of the DQ which are related to data representation, its organization, its attractiveness, the value-added, etc., which do not appear to have been considered in ISO/IEC 25012.

We do not plan to carry out a more detailed comparison at present owing to the fact that the ISO/IEC 25012 norm is still a rough draft and is therefore subject to changes which may make it invalid.

6 A case study for validating the proposed model

A case study of the domain of bank portals has been carried out with the objective of checking our model's validity within a specific Web portal domain. This was done by carrying out a survey upon a group of bank portal data consumers, each of whom was asked to evaluate the importance that each of the attributes in our model had in their domain. There now follows a description of this case study and its results.

6.1 Method

We conducted a survey to collect data on the importance of each of attribute of our model to data consumers in the bank portal domain. The questionnaire asked the

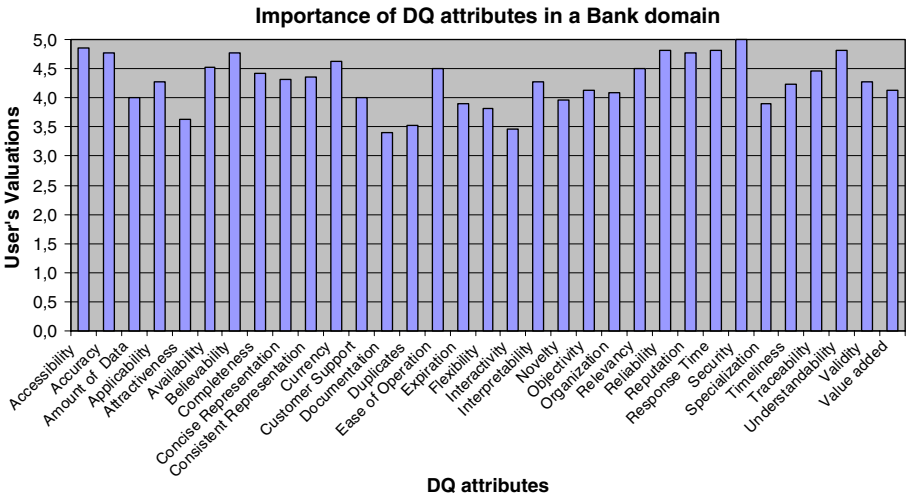

Fig. 5 Results of the case study

respondents (in 33 closed questions) to rate the importance of each DQ attribute for their domain on a scale of 1–5, where 1 was not important and 5 extremely important. We also included an open question which asked them to indicate whether, in their opinion, it was necessary to consider others attributes. Finally, the questionnaire was administered to 22 respondents: all bank executives and data consumers of bank portals.

6.2 Results

As a result of this case study it is possible to say that all of the attributes in the model are highly pertinant to the domain studied. As the graph in Fig. 5 shows, the mean number of the evaluations carried out by the participants in the survey upon the importance of each attribute in its domain is, in all cases, superior to 3. Furthermore, no new attributes were suggested. We can also highlight that within this specific domain the most important attributes were considered to be: Availability, Currency, Accuracy, Believability, Reputation, Reliability, Response Time, Understandability, Accessibility and Security, all of which obtained a mean evaluation of over 4.5.

It is probable that these valorations might change in a different domain, and that the most relevant attributes would be others, and it is for this reason that our model should contemplate using a measurement process which recognizes and manages the differences between these domains.

Finally, this case study has demonstrated that, from a user perspective, the set of attributes in our model as a whole is pertinent and is, in principle, complete.

7 Conclusions and future work

Web portals are applications which have, over the last decade, established their position as information sources and/or as a means of accessing information. Of course those who

look for information by means of these portals need some means by which to ensure that this information is indeed suitable for the use that they require. In other words, they really need to assess the level of the quality of the data obtained. However, within the literature studied we have found no specific proposals for data quality models for Web portals.

Moreover, although it is possible to find a standard for software product quality, there is no similar proposal for data quality. Although the new SQUARE family of standards will include a specific proposal of DQ standard, at this moment there is nothing which can be used to tailor it to a specific domain (in our case web portal data).

In this paper, we have presented the development of a process through which to obtain of a set of attributes for web portal DQ, which focuses on the data consumer's point of view. The process has been built by using three basic elements: a set of Web DQ attributes found in the relevant literature, DQ expectations of data consumers on the Internet, and the functionalities which a Web portal may offer its users.

At this moment our set is composed of 33 DQ attributes which, from the data consumer's perspective, can be used to assess the DQ in Web portals. We consider the proposed set of attributes to be an important step towards achieving a DQ model for Web portals, as they can be used in the definition of a quality assessment process. We believe that this proposal would perhaps even be useful for ISO/IEC since this set could be studied for its inclusion (total or partial) in the new data quality standard. A case study has been carried out to show the validity of these attributes within a concrete domain of Web portals (the bank portal domain), thus demonstrating that the set of DQ attributes is correct and complete.

In order to advance towards our final objective (that of defining a DQ model for Web portals) we have organized the 33 DQ attributes in a hierarchical structure. This has been done by using the Wang and Strong model [43], one of the models which is most frequently used and referenced in literature.

As future work, we plan to continue working with our set of attributes, attempting to use them in the assessment of the quality of the data in a web portal. Portal users will, therefore, have a model with which they can discover the DQ level of the portals which they use. This will also be of use to designers who will be able to evaluate whether their portal is appropriate to their users' needs. Our idea is to use an approach which will allow us to assess DQ automatically, analysing the html code of web portals, and simplifying its use for the data consumer. We plan to define measures for the DQ attributes in PDQM based on the measures proposed in literature (Eppler and Muenzenmayer 2002) and others which are necessary for the context and type of attribute measured. Within the evaluation of DQ, we shall also attempt to consider the subjectivity associated with the point of view of data consumers.

We also plan to study the alignment of our proposal with the ISO/IEC DQ standard (when it will be ready) in order to adapt it, if necessary.

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.

Appendix A

Categories of data consumer expectations concerning the DQ on the internet

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, organizations, and so forth)
Content	*should expect* the Publisher to be explicit in: describing what data are published and how they should be used, to describe appropriate and inappropriate uses of the published data; all data needed for an intended use will be provided (unless otherwise stated); easy-to-understand definitions of every important term and all original sources of data will be clearly stated
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 it will state its policy regarding incorrect data. He/she should also expect data values to be current, unless otherwise informed by the Publisher- all relevant data will be published, unless otherwise stated
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, he/she should expect to be able to interpret data properly if he/she follows instructions
Improvements	*should expect to* have a means to convey his/her comments about data, be they good or bad, to the Publisher and that these comments will be acted upon in a responsible manner. He/she should also expect to be provided with useful summaries of actual quality levels of the data he/she is using and 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

Appendix B

Attribute	Definition
Data quality attributes considered in the model	
Accessibility	The extent to which the Web portal provides sufficient navigation mechanisms for visitors to reach their desired data faster and easier
Accuracy	The extent to which data are correct, reliable, and certified as being 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 easily 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 through the portal
Believability	The extent to which data and their sources are accepted as correct

Appendix B continued

Attribute	Definition
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 by 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
Interactivity	The extent to which the way in 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 appropriate 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 sources
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	Degree of specificity of data/information contained in and delivered by the Web application, i.e. it should incorporate all details which might be seen by its visitors
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 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, non-ambiguous, 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
Data quality attributes discarded from the model descartados del modelo	
Cost effectiveness	The degree to which the cost of collecting appropriate data/information is reasonable
Granularity	The degree of granularity in the sources, which allows us to describe the properties of data in more detail

Appendix B continued

Attribute	Definition
Internal consistency	The degree to which the values of the attributes of an instance of a schema element satisfy the specific set of semantic rules defined in the schema element
Latency	Amount of time needed for user to obtain the first data/information from the Web application
Maintainability	The degree to which data/information can also be easily accessed in the future
Ontology	The degree to which an ontology exists which centres upon the description of the schemes of the sources (Knowledge of this structure scheme is extremely important in the management of the integration process)
Price	Monetary charge per consultation

Appendix C. Survey Tool[2]

Survey on Web Portal Data Quality

Thank you for taking part in this survey!

As you answer this questionnaire, bear in mind your experience as a Web portal user (university portals such as those of the University of Castilla la Mancha, or from banks, businesses, and so on). In each of the questions in the survey, an aspect of Web Portal data quality will be described. You will be asked to assess how important these features are, in your view. This evaluation of the relative importance of each aspect should be shown by assigning to it a numerical value between 1 and 5, where the meaning of each value is the following: This aspect, in your opinion is

1. not important.
2. of little importance.
3. moderately important.
4. important.
5. very important.

1. That the data contained in, and delivered by, a Web portal is set out attractively, in an eye-catching and pleasant form , is, in your opinion:

Not important	1	2	3	4	5	Very important

2. That a Web portal provides sufficient navigation mechanisms for the data to be accessed speedily and with ease is, in your opinion:

Not important	1	2	3	4	5	Very important

3. That the data contained in, and delivered by, a Web portal is up-to-date is, in your opinion:

Not important	1	2	3	4	5	Very important

4. That the data contained in, and delivered by, a Web portal is specific, useful and easy to apply to your needs, is, in your opinion:

Not important	1	2	3	4	5	Very important

5. That the quantity of data delivered by a Web portal (in each link, as the result of a search on each page,etc.) is appropriate, is, in your opinion:

Not important	1	2	3	4	5	Very important

6. That the data contained in, and delivered by, a Web portal is wide-reaching enough, as well as of sufficient relevance and depth for the task in hand is, in your opinion:

Not important	1	2	3	4	5	Very important

[2] In this paper we have included an English version of the survey. The original survey was presented in Spanish so that the students could understand it as well as possible.

7. Being able to trust that the data contained in, and delivered by, a Web portal are appropriate, is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

8. The availability of certain data from a Web portal when you need to obtain them is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

9. That the information documentation concerning the data contained in, and delivered by, a Web portal, (the support offered in relation to that data) is of the right quantity and quality is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

10. That the information provided by a web portal does not contain duplicate data when accessed, is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

11. That the data contained in, and delivered by, a Web portal (in each link, as the result of a search on each page,etc.) is clear, unambiguous and easy to understand , is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

12. That the data contained in, and delivered by, a Web portal is detailed enough is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

13. Being able to believe that the data and their source/s are correct is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

14. That the data contained in, and delivered by, a Web portal is precise, correct and guaranteed to be free of mistakes, is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

15. The abilty to know how long the data contained in, and delivered by, a Web portal can be considered up-to-date, is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

16. That the data contained in, and delivered by, a Web portal is easy to handle and control in the tasks you perform is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

17. That the data contained in, and delivered by, a Web portal is adaptable and applicable to different requirements, is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

18. That a Web portal gives information about the author and/or the source of the data contained in the portal, is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

19. That the way in which the data contained in, and delivered by, a Web portal is adaptable to your personal preferences through interactive elements, is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

20. That the data contained in, and delivered by, a Web portal is presented in a language that is appropriate and easy to interpret, is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

21. That the data contained in, and delivered by, a Web portal is novel and that it has an influence on your knowledge and the decisions you have to take is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

22. That the data contained in, and delivered by, a Web portal is impartial and bias-free is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

23. That a Web portal give you the data you need within a time period that suits your needs is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

24. That the data contained in, and delivered by, a Web portal is organized according to certain criteria and that it uses a consistent combination of visual controls and/or stylistic devices (colours, text, different types and sizes of font, pictures, and so on), is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

25. That the data contained in, and delivered by, a Web portal is applicable to the tasks you need to carry out and/or is suitable to meet your needs and that they are useful in these areas, is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

26. That the data contained in, and delivered by, a Web portal is presented in a form that is compact, free of superfluous data and of elements that are not at all pertinent, is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

27. That the data contained in, and delivered by, a Web portal is presented in a form that is consistent with other sources, using the same, or compatible, formats on the various different pages is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

28. That the data contained in, and delivered by, a Web portal is worthy of great respect as regards content and sources is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

29. That the data contained in, and delivered by, a Web portal is protected from unauthorised access and manipulation is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

30. That a Web portal provides on-line support in attending user queries on data contained in the portals by means of e-mail, 'phone, etc is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

31. That the time taken between a request for information (entry to a page, or through a query, etc.), and the satisfactory and complete answering of that request is appropriate to your needs is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

32. That the data contained in, and delivered by, a Web is well documented, verifiable and easy to attribute to a single source, is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

33. That the data contained in, and delivered by, a Web portal can be judged and seen to be valid from the users' point of view is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

34. That the data contained in, and delivered by, a Web portal provides those using it with advantages and benefits, is, in your opinion:

Not important | 1 | 2 | 3 | 4 | 5 | Very important

If there are any other aspects which are relevant to the quality of data that a web portal may contain and which have not been reflected in the questions in this survey, please add your remarks in the space provided below.

References

Aboelmeged, M. (2000). A soft system perspective on information quality in electronic commerce. In *Fifth Conference on Information Quality* (pp. 318–319).

Bouzeghoub, M., & Kedad, Z. (2001) Quality in data warehousing. In M. Piattini, C. Calero, & M. Genero (Eds.), *Information and Database Quality*. Kluwer Academic Publishers.

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

Burgess, M., Fiddian, N., & Gray, W. (2004). Quality Measures and The Information Consumer. In *Ninth International Conference on Information Quality* (pp. 373–388).

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

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

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

Carver, J., Jaccheri, L., Morasca, S., & Shull, F. (2003). Issues in Using Students in Empirical Studies in Software Engineering Education. In *9th International Software Metrics Symposium (METRICS'03)* (239 pp). IEEE Computer Society, Los Alamitos, CA, USA.

Collins, H. (2001). *Corporate portal definition and features*. AMACOM.

Eppler, M. (2001). A generic framework for information quality in knowledge-intensive processes. In Proc. sixth international conference on information quality. pp. 329–346.

Eppler, M., Algesheimer, R., & 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. In *Eighth International Conference on Information Quality* (pp. 108–120).

Eppler, M., & Muenzenmayer, P. (2002). Measuring information quality in the web context: A survey of state-of-the-art instruments and an aplication methodology. In *Seventh International Conference on Information Quality* (pp. 187–196).

Fugini, M., Mecella, M., Plebani, P., Pernici, B., & Scannapieco, M. (2002). Data Quality in Cooperative Web Information Systems.

Gertz, M., Ozsu, T., Saake, G., & Sattler, K.-U. (2004). Report on the dagstuhl seminar "Data Quality on the Web". *SIGMOD Record, 33*(1), 127–132.

Graefe, G. (2003). Incredible Information on the Internet: Biased information provision and a lack of credibility as a cause of insufficient information quality. In *Eighth International Conference on Information Quality* (pp. 133–146).

Höst, M., Regnell, B., & Wohlin, C. (2000). Using students as subjects—a comparative study of students and professionals in lead-time impact assessment. *Empirical Software Engineering, 5*, 201–214.

Katerattanakul, P., & Siau, K. (1999). Measuring information quality of web sites: Development of an instrument. In *20th International Conference on Information System* (pp. 279–285).

Katerattanakul, P., & Siau, K. (2001). Information quality in internet commerce desing. In M. Piattini, C. Calero, & M. Genero (Eds.), *Information and database quality* (pp. 45–56). Kluwer Academic Publishers.

Kitchenham, B. (2004). Procedures for Performing Systematic Reviews.

Kitchenham, B., & Pfleeger, S. L. (2002a). Principles of survey research part 2: Designing a survey. In *SIGSOFT Softw. Eng. Notes* (Vol. 27, pp. 18–20). ACM Press.

Kitchenham, B., & Pfleeger, S. L. (2002b). Principles of survey research part 4: Questionnaire evaluation. In *SIGSOFT Softw. Eng. Notes* (Vol. 27, pp. 20–23). ACM Press.

Kitchenham, B., & Pfleeger, S. L. (2002c). Principles of survey research: Part 3: Constructing a survey instrument. In *SIGSOFT Softw. Eng. Notes* (Vol. 27, pp. 20–24). ACM Press.

Kitchenham, B., & Pfleeger, S. L. (2002d). Principles of survey research: Part 5: Populations and samples. In *SIGSOFT Softw. Eng. Notes* (Vol. 27, pp. 17–20). ACM Press.

Kitchenham, B., & Pfleeger, S. L. (2003). Principles of survey research part 6: Data analysis. In *SIGSOFT Softw. Eng. Notes* (Vol. 28, pp. 24–27). ACM Press.

Knight, S. A., & Burn, J. M. (2005). developing a framework for assessing information quality on the world wide web. *Informing Science Journal, 8*, 159–172.

Kopcso, D., Pipino, L., & Rybolt, W. (2000). The assesment of web site quality. In *Fifth International Conference on Information Quality* (pp. 97–108).
Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*, 537–567.
Lee, Y. (2002). AIMQ: A methodology for information quality assessment. *Information and Management, 40*(2), 133–146.
Mahdavi, M., Shepherd, J., & Benatallah, B. (2004). A collaborative approach for caching dynamic data in portal applications. In *Proceedings of the fifteenth conference on Australian database* (Vol. 27, pp. 181–188).
Marchetti, C., Mecella, M., Scannapieco, M., & Virgillito, A. (2003). Enabling data quality notification in cooperative information systems through a web-service based architecture. In *Fourth International Conference on Web Information Systems Engineering* (pp. 329–332).
Melkas, H. (2004). Analyzing information quality in virtual service networks with qualitative interview data. In *Ninth International Conference on Information Quality* (pp. 74–88).
Moraga, M. Á., Calero, C., & Piattini, M. (2006). Comparing different quality models for portals. *Online Information Review, 30*, 555–568.
Moustakis, V., Litos, C., Dalivigas, A., & Tsironis, L. (2004). Website quality assesment criteria. In *Ninth International Conference on Information Quality* (pp. 59–73).
Naumann, F., & Rolker, C. (2000). Assesment methods for information quality criteria. In *Fifth International Conference on Information Quality* (pp. 148–162).
Nelson, R., Todd, P., & Wixom, B. (2005) Antecedents of information and system quality: An empirical examination within the context of data warehouse. *Journal of Management Information Systems, 21*, 199–235.
Pernici, B., & Scannapieco, M. (2002). Data quality in web information systems. In *21st International Conference on Conceptual Modeling* (pp. 397–413).
Pfleeger, S. L., & Kitchenham, B. (2001). Principles of survey research: Part 1: Turning lemons into lemonade. *SIGSOFT Softw. Eng. Notes* (Vol. 26, pp. 16–18). ACM Press.
Pressman, R. (2001). *Software Engineering: A Practitioner's Approach. 5/e*. McGraw-Hill.
Redman, T. (2000). *Data quality: The field guide*. Boston: Digital Press.
Reeves, C., & Bednar, D. (1994). Defining quality: Alternatives and implications. *Academy of Management Review, 19*, 419–445.
Strong, D., Lee, Y., & Wang, R. (1997). Data quality in context. *Communications of the ACM, 40*(5), 103–110.
Wang, R., & Strong, D. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*, 5–33.
Winkler, W. (2004). Methods for evaluating and creating data quality. *Information Systems, 29*, 531–550.
Wohlin, C., Runeson, P., Höst, M., Ohlson, M., Regnell, B., & A., W. (2000). *Experimentation in Software Engineering: An Introduction*, Kluwer Academic Publishers, 2000.
Yang, Z., Cai, S., Zhou, Z., & Zhou, N. (2004). Development and validation of an instrument to measure user perceived service quality of information presenting Web portals. *Information and Management, 42*, 575–589.
Zhu, Y., & Buchmann, A. (2002). Evaluating and selecting web sources as external information resources of a data warehouse. In *3rd International Conference on Web Information Systems Engineering* (pp. 149–160).

Author Biographies

Angélica Caro has a PhD in Computer Science and is 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 Associate Professor at the Escuela Superior de Informatica of the Castilla-La Mancha University in Ciudad Real. She is a member of the Alarcos Research Group, in the same University, specialized in Information Systems, Databases and Software Engineering. Her research interests include: advanced databases design, database quality, software metrics, database metrics. She is author of papers in national and international conferences on this subject. She has published in Information Systems Journal, Software Quality Journal, Information and Software Technology Journal and SIGMOD Record Journal. She has organized the web services quality workshop (WISE Conference, Rome 2003) and Database Maintenance and Reengineering workshop (ICSM Conference, Montreal 2002).

Ismael Caballero has an MSc and PhD in Computer Science from the Escuela Superior de Informática of the Castilla-La Mancha University in Ciudad Real. He actually works as an assistant professor in the Department of Information Systems and Technologies at the University of Castilla-La Mancha, and he has also been working in the R&D Department of Indra Sistemas since 2006. His research interests are focused on information quality management, information quality in SOA, and Global Software Development.

Mario Piattini has an MSc and a PhD in Computer Science (Politechnical University of Madrid) and a MSc in Psychology (UNED.). He is also a Certified Information System Auditor and a Certified information System Manager by ISACA (Information System Audit and Control Association) as well as a Full Professor in the Department of Computer Science at the University of Castilla-La Mancha, in Ciudad Real, Spain. Furthermore, he is the author of several books and papers on databases, software engineering and information systems. He is a coeditor of several international books: "Advanced Databases Technology and Design", 2000, Artech House, UK; "Information and database quality", 2002, Kluwer Academic Publishers, Norwell, USA; "Component-based software quality: methods and techniques", 2004, Springer, Germany; "Conceptual Software Metrics", Imperial College Press, UK, 2005. He leads the ALARCOS research group of the Department of Computer Science at the University of Castilla-La Mancha, in Ciudad Real, Spain. His research interests are: advanced databases, database quality, software metrics, security and audit, software maintenance.