TOWARDS A DATA QUALITY MODEL FOR WEB PORTALS
Research in Progress

Angélica Caro
Universidad del Bio Bio, Departamento de Auditoria e Informática, La Castilla s/n, Chillán, Chile

Coral Calero, Ismael Caballero, Mario Piattini
ALARCOS Research Group
Information Systems and Technologies Department, UCLM-Soluziona Research and Development Institute
University of Castilla-La Mancha, Paseo de la Universidad, 4 – 13071 Ciudad Real, Spain

Keywords: Data Quality, Information Quality, Web Portals.

Abstract: The technological advances and the use of the internet have favoured the appearance of a great diversity of web applications, among them Web Portals. Through them, organizations develop their businesses in a really competitive environment. A decisive factor for this competitiveness is the assurance of data quality. In the last years, several research works on Web Data Quality have been developed. However, there is a lack of specific proposals for web portals data quality. Our aim is to develop a data quality model for web portals focused -oin three aspects: data quality expectations of data consumer, the software functionality of web portals and the web data quality attributes recompiled from a literature review. In this paper, we will present the first version of our model.

1 INTRODUCTION

In the last years, a growing interest in the subject of Data Quality (DQ) or Information Quality (IQ) has been generated because of the increase of interconnectivity of data producers and data consumers mainly due to the development of the internet and web technologies. The DQ/IQ is often defined as "fitness for use", i.e., the ability of a data collection to meet user requirements (Strong, Lee et al., 1997; Cappiello, Francalanci et al., 2004). Data Quality is a multi-dimensional concept (Cappiello, Francalanci et al., 2004), and in the DQ/IQ literature several frameworks providing categories and dimensions as a way of facing DQ/IQ problems can be found.

Research on DQ/IQ started in the context of information systems (Strong, Lee et al., 1997; Lee, 2002) and it has been extended to contexts such as cooperative systems (Fugini, Mecella et al., 2002; Marchetti, Mecella et al., 2003; Winkler, 2004), data warehouses (Bouzeghoub and Kedad, 2001; Zhu and Buchmann, 2002) or electronic commerce (Aboelmeged, 2000; Katerattanakul and Siau, 2001), among others.

Due to the characteristics of web applications and their differences from the traditional information systems, the community of researchers has recently started to deal with the subject of DQ/IQ on the web (Gertz, Ozsu et al., 2004). However, there are not works on DQ/IQ specifically developed for web portals. As the literature shows that DQ/IQ is very dependent on the context, we have centred our work on the definition of a Data Quality Model for web portals. To do so, we have used some works developed for different contexts on the web but that can be partially applied or adapted to our particular context. For example, we have used the work of Yang et al., (2004) where a quality framework for web portals is proposed including data quality as a part of it.

As the concept of "fitness for use" is widely adopted in the literature (emphasizing the importance of taking into consideration the consumer viewpoint of quality), we have also considered, for the definition of our model, the data consumer viewpoint.

Caro A., Calero C., Caballero I. and Piattini M. (2006).
TOWARDS A DATA QUALITY MODEL FOR WEB PORTALS - Research in Progress.
In *Proceedings of WEBIST 2006 - Second International Conference on Web Information Systems and Technologies - Internet Technology / Web Interface and Applications*, pages 325-331
DOI: 10.5220/0001242703250331

To produce our model, we defined a four-stage process, as set out in figure 1. In the first of these phases, we recompiled web data quality attributes from the literature and which we believe should therefore be applicable to web portals. In the second stage we built a matrix for the classification of the attributes obtained in stage 1. This matrix reflects two basic aspects considered in our model: the data consumer perspective (by means data quality expectations of data consumers on Internet) and the

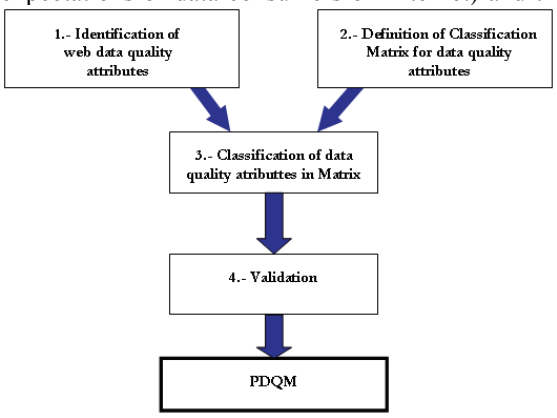

Figure 1: Stages in the development our model.

basic functionalities which a data consumer uses to interact with a Web portal.

Then in our third stage we used the matrix that has been produced, to analyse the applicability of each attribute of Web quality in a Web portal. Finally, in the fourth stage, we will validate our preliminary model, using surveys carried out on the data consumers of a given portal.

In this paper we describe the first version of our model, product of the three first stages of our methodology. The structure of this paper is as follows. In section 2, the components of our model are presented. In section 3, we will deeply describe the first version of our DQ/IQ Web Portal Model. Finally, in section 4 we will conclude with our general remarks and future work.

2 MODEL COMPONENTS

Web Portals are emerging Internet-based applications that enable access to different sources (providers) through a single interface (Mahdavi, Shepherd et al., 2004). The primary objective of a portal software solution is to create a working environment where users can easily navigate in order to find the information they specifically need to perform their operational or strategic functions quickly as well as to make decisions (Collins, 2001), being responsibility of web portals' owners the achievement and maintenance of a high information quality state (Kopcso, Pipino et al., 2000).

In this section, we will present the three basic aspects considerate to define our DQ/IQ model for web portals: the DQ/IQ attributes defined in the web context, the data consumer expectations about data quality, and web portals functionalities.

2.1 Data Consumer Expectations

When data management is conceptualized as a production process (Strong, Lee et al., 1997), we can identify three important roles in this process: (1) data producers (who generate data), (2) data custodians (who provide and manage resources for processing and storing data), and (3) data consumers (who access and use data for their tasks).

As in the context of web-based information systems, roles (1) and (2) can be developed by the same entity (Gertz, Ozsu et al., 2004), for web portals context we identify two roles in the data management process: (1) data producers-custodians, and (2) data consumers.

So far, except for few works in DQ/IQ area, like (Wang and Strong, 1996; Strong, Lee et al., 1997; Burgess, Fiddian et al., 2004; Cappiello, Francalanci et al., 2004), most of the works on the subject have looked at quality from the data producer-custodian perspective. The data consumer's perspective of quality differs from this in two important ways (Burgess, Fiddian et al., 2004):

Data consumer has no control over the quality of available data.

The aim of consumers is to find data that match their personal needs, rather than provide data that meet the needs of others.

Our proposal of a DQ/IQ model for web portals considers the data quality expectations of data consumer because, at the end, it is the consumer who will judge whether a data is fitted for use or not (Wang and Strong, 1996).

We will use the quality expectations of the data consumer on the Internet, proposed in (Redman, 2001). These expectations are organized into six categories: Privacy, Content, Quality of values, Presentation, Improvement, and Commitment.

2.2 Web Portal Functionalities

A web portal is a system of data manufacturing where we can distinguish the two roles established in the previous subsection. Web portals present basic software functionalities to data consumer deploying their tasks and under our perspective, the consumer judges data by using the application functionalities. So, we used the web portal software

functions that Collins proposes in (Collins, 2001) considering them as basics in our model. These functions are as follows: Data Points and

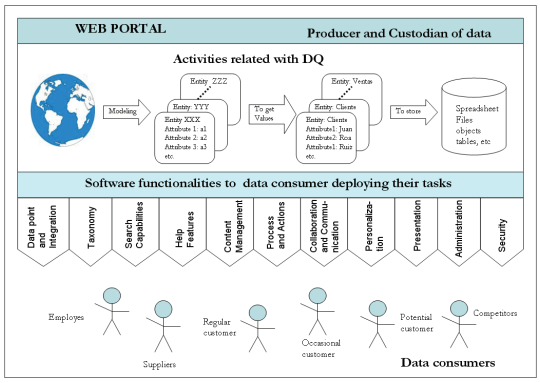

Figure 2: Roles in web portals.

Integration, Taxonomy, Search Capabilities, Help Features, Content Management, Process and Action, Collaboration and Communication, Personalization, Presentation, Administration, and Security. Behind these functions, the web portal encapsulates the producer-custodian role. Figure 2 illustrates this fact.

2.3 Web DQ Revision

By using a DQ/IQ framework, organizations are able to define a model for data, to identify relevant quality attributes, to analyze attributes within both current and future contexts, to provide a guide to improve DQ/IQ and to solve data quality problems (Kerr and Norris, 2004). In the literature, we have found some proposals oriented to DQ/IQ on the web.

Among them, we can highlight those showed in table 1. Related to such proposals, we can conclude that there is no agreement concerning either the set of attributes or, in several cases, their meaning. This situation, probably, is a consequence of the different domains and author's focus of the studied works.

However, from this revision we captured several data quality attributes. The most considered are (we present between brackets different terms used for the same concept): Accuracy (Accurate), in 60% of the works; Completeness, in 50% of the works and Timeliness (Timely), in 40% of the works; Concise (Concise representation), Consistent (Consistent representation), Currency (Current), Interpretability, Relevance, Secure (Security), in 30% of the studies. Accessibility (Accessible), Amount of data (Appropriate amount of information), Availability, Credibility, Objectivity, Reputation, Source Reliability, Traceability (Traceable), Value added are stated in 20% of the works.

Finally, Applicable, Clear, Comprehensive, Confidentiality, Content, Convenient, Correct, Customer Support, Degree of Duplicates, Degree of Granularity, Documentation, Understand ability (Ease of understanding), Expiration, Flexibility, Freshness, Importance, Information value, Maintainable, Novelty, Ontology, Pre-decision availability, Price, Reliability, Response time, Layout and design, Uniqueness, Validity, and Verifiability are only studied in 10 % of the works

Summarizing the above-mentioned attributes, by means of similarity in their names and definitions, we have obtained a set of 28 attributes. Based on these DQ/IQ attributes we will try to identify which ones are applicable to the web portals context by classifying them into the matrix construed by the previous aspects (data consumer expectations x functionalities).

3 RELATIONSHIPS BETWEEN THE COMPONENTS OF THE MODEL

Based on the previous background, we will determine the relationship between the web portal functionalities and the quality expectations of data consumers. Then, we will present the definition of each function according to (Collins, 2001) and we will show their relationships (see figure 3).

Data Points and Integration. They provide the ability to access information from a wide range of internal and external information sources and display the resulting information at the single point-of-access desktop. The expectations applied to this functionality are: *Content* (Consumers need a description of portal areas covered, use of published data, etc.), *Presentation* (formats, language, and others are very important for easy interpretation) and *Improvement* (users want to participate with their opinions in the portal improvements knowing the result of applying them).

Taxonomy. It provides information context (including the organization-specific categories that reflect and support organization's business), we consider that the expectations of data consumer

Table 1: Summary of web DQ/IQ framework in the literature.

Author	Domain	Framework structure
(Katerattanakul and Siau, 1999)	Personal web sites	4 categories and 7 constructors
(Naumann and Rolker, 2000)	Data integration	3 classes and 22 of quality criterion
(Aboelmeged, 2000)	e-commerce	7 stages to modelling DQ problems
(Katerattanakul and Siau, 2001)	e-commerce	4 categories associated with 3 categories of data user requirements.
(Pernici and Scannapieco, 2002)	Web information systems (data evolution)	4 categories, 7 activities of DQ design and architecture to DQ management.
(Fugini, Mecella et al., 2002)	e-service cooperative	8 dimensions
(Graefe, 2003)	Decision making	8 dimensions and 12 aspects related to (providers/consumers)
(Eppler, Algesheimer et al., 2003)	Web sites	4 dimensions and 16 attributes
(Gertz, Ozsu et al., 2004)	DQ on the web	5 dimensions
(Moustakis, Litos et al., 2004)	Web sites	5 categories and 10 sub-categories
(Melkas, 2004)	Organizational networks	6 stages to DQ analysis with several dimensions associated with each one
(Bouzeghoub and Peralta, 2004)	Data integration	2 factors and 4 metrics
(Yang, Cai et al., 2004)	Web information portals	2 dimensions

are: *Content* (consumers need a description of which data are published and how they should be used, 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* (user should expect to convey his/her comments on data in the taxonomy and know the result of improvements).

Search Capabilities. It provides several services for web portal users and needs searches across the enterprise, World Wide Web, and search engine catalogs and indexes. The expectations applied to this functionality are: *Quality of values* (Data consumer should expect that the result of searches is correct, current and complete), *Presentation* (formats and language are important for consumers, for the search and for easy interpretation of results) and *Improvement* (consumer should expect to convey his/her comments on data in the taxonomy and know the result of improvements).

Help Features. They provide help when using the web portal. The expectations applied to this functionality are: *Presentation* (formats, language, and others are very important for easy interpretation of help texts) and *Commitment* (consumer should be easily able to ask and obtain answer to any question regarding the proper use or meaning of data, update schedules, etc.).

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* (it should exist privacy policy for all consumers to manage, to access sources and to guarantee web portals data), *Content* (consumers need a description of data collections, that all data needed for an intended use are provided, etc.), *Quality of values* (consumer should expect that all data values are correct, current and complete, unless otherwise stated), *Presentation* (formats and language should be appropriate for easy interpretation), *Improvement* (consumer should expect to convey his/her comments on contents and their management and know the result of the improvements) and *Commitment* (consumer should be easily able to ask and have any question regarding the proper use or meaning of data, update schedules, etc. answered).

Process and Action. This function enables the web portal user to initiate and participate in a business process of portal owner. The expectations applied to this functionality are: *Privacy* (Data consumer should expect that there is a privacy policy to manage the data about the business on the portal), *Content* (Consumers should expect to find descriptions about the data published for the processes and actions, appropriate and inappropriate uses, that all data needed for the process and actions are provided, etc.), *Quality of values* (that all data associated to this function are correct, current and complete, unless otherwise stated), *Presentation* (formats, language, and others are very important for properly interpret data), *Improvement* (consumer should expect to convey his/her comments on contents and their management and know the result of improvements) and *Commitment* (consumer

should be easily able to ask and to obtain answer to any questions regarding the proper use or meaning of data in a process or action, etc.).

Collaboration and Communication. This function facilitates discussion, locating innovative ideas, and recognizing resourceful solutions. The expectations applied to this functionality are: *Privacy* (consumer should expect privacy policy for all consumers that participate in activities of this function), and *Commitment* (consumer should be easily able to ask and have any questions regarding the proper use or meaning of data for the collaboration and/or communication, etc., answered).

Personalization. This is a critical component to create a working environment that is organized and configured specifically to each user. The expectations applied to this functionality are: *Privacy* (consumer should expect privacy and security about their personalization data, profile, etc.), and *Quality of values* (data about user profile should be correct, current).

Presentation. It provides both the knowledge desktop and the visual experience to the web portal user 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 covered areas , appropriate and inappropriate uses, definitions, information about the sources, etc.), *Quality of values* (the data of this function should be correct, current and complete.), *Presentation* (formats, language, and others are very important for easy interpretation and appropriate use of portals data.) and *Improvement* (consumer should expect to convey his/her comments on contents and their management and know the result of the improvements).

Administration. This function provides service for deploying maintenance activities or tasks associated with the web portal system. The expectations applied to this functionality are: *Privacy* (Data consumers need security for data about the portal administration) and *Quality of values* (Data about tasks or activities of administration should be correct and complete).

Security. It 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* (consumer need privacy policy about the data of the levels of access of data consumers.), *Quality of values* (data about the levels of access should be correct and current.) and *Presentation* (data about security should be in format and language for easy interpretation).

Concerning the relationships established in the matrix of figure 3, we can remark that Presentation is the category of data consumer expectation with more relations. This perfectly fits with the main goal of any web applications, which is to be useful and user-friendly for any kind of user.

The next step is to fill in each cell of the matrix with Web DQ/IQ attributes obtained from the study presented in 2.3. As a result of this, we have a subset of DQ/IQ attributes that can be used in a web portal to evaluate data quality. In table 2, we will show the most relevant attributes for each category of data consumer expectations.

To validate and complete this assignation we plan to work with portal data consumers through surveys and questionnaires. Once the validation is finished, we will reorganize the attributes obtaining the final version of our model.

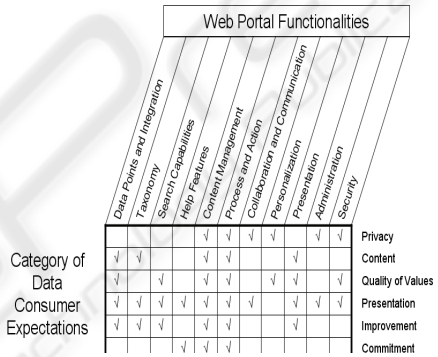

Figure 3: Matrix stating the relationships between data consumer expectations and web portal functionalities.

4 CONCLUSIONS AND FUTURE WORK

The great majority of works found in the literature show that data quality or information quality is very dependent on the context. The increase of the interest in the development of web applications has implied either the appearance of new proposals of frameworks, methodologies and evaluation methods of DQ/IQ or the adaptation of the already-existing ones from other contexts. However, in the web portal context, data quality frameworks do not exist In this paper, we have presented a proposal that combines three aspects: (1) a set of web data quality attributes resulting from a data quality literature survey that can be applicable and useful for a web portal, (2) the data quality expectations of data consumer on the Internet, and (3) the basic functionalities for a web portal. These aspects have been related by obtaining a first set of data quality

Table 2: Web Data Quality attributes applied to web portal functionalities in each category.

Category of Data Consumer Expectations	Web portal functionalities related to each category	Web DQ/IQ attributes applying almost one functionality in each category
Privacy	Content management Process and actions Collaboration and Communication Personalization Administration Security	Security
Content	Data Points and Integration Taxonomy Content management Process and actions Presentation	Accessibility, Currency, Amount of data, Understandability, Relevance, Concise Representation, Validity, Traceability, Completeness, Reliability, Credibility, Timeliness, Availability, Documentation, Specialization, Interpretability, Easy to use
Quality of data	Data Points and Integration Search Capabilities Content management Process and actions Personalization Presentation Security	Accessibility, Currency, Amount of data, Credibility, Understandability, Accuracy, Expiration, Novelty, Relevance, Validity, Concise Representation, Completeness, Reliability, Availability, Documentation, Duplicity, Specialization, Interpretability, Objectivity, Relevance, Reputation, Traceability, Utility, Value-added, Easy to use
Presentation	Data Points and Integration Taxonomy Search Capabilities Help Features Content management Process and actions Collaboration and Communication Presentation Administration Security	Amount of data, Completeness, Understandability, Easy to use, Concise Representation, Consistent Representation, Validity, Relevance, Interpretability, User support, Availability, Specialization, Flexibility
Improvement	Data Points and Integration Taxonomy Search Capabilities Content management Process and actions Presentation	Accessibility, Reliability, Credibility, Understandability, User support, Traceability
Commitment	Help Features Content management Process and actions	Accessibility, Reliability, User support,

attributes for the different data consumer expectations X functionalities.

Our future work, now in progress, consists of validating and refining this model. First of all, it is necessary to check these DQ/IQ attributes with data consumers in a web portal. We plan to make a questionnaire for each web portal functionality. Then, once we have validated the model, we will define a framework including the necessary elements to evaluate a DQ/IQ in a web portal. Our aim is to obtain a flexible framework where the data consumer can select the attributes used to evaluate the quality of data in a web portal, depending on the existing functionalities and their personal data quality expectations.

ACKNOWLEDGEMENTS

This research is part of the following projects: CALIPO (TIC2003-07804-C05-03) supported by the Dirección General de Investigación of the Ministerio de Ciencia y Tecnología (Spain) and DIMENSIONS (PBC-05-012-1) supported by FEDER and by the "Consejería de Educación y Ciencia, Junta de Comunidades de Castilla-La Mancha" (Spain).

REFERENCES

Aboelmeged, M., 2000. A Soft System Perspective on Information Quality in Electronic Commerce. In *Proceeding of the Fifth Conference on Information Quality*,

Bouzeghoub, M. and Z. Kedad, 2001. Quality in Data Warehousing. Information and Database Quality. M. Piattini, C. Calero and M. Genero, Kluwer Academic Publishers.

Bouzeghoub, M. and V. Peralta, 2004. A Framework for Analysis of data Freshness. In *International Workshop on Information Quality in Information Systems, (IQIS2004)*, Paris, France, ACM.

Burgess, M., N. Fiddian, et al., 2004. Quality Measures and The Information Consumer. In *IQ2004*,

Cappiello, C., C. Francalanci, et al., 2004. Data quality assessment from the user´s perspective. In *International Workshop on Information Quality in Information Systems, (IQIS2004)*, Paris, Francia, ACM.

Collins, H., 2001. Corporate Portal Definition and Features. AMACOM

Eppler, M., R. Algesheimer, et al., 2003. Quality Criteria of Content-Driven Websites and Their Influence on Customer Satisfaction and Loyalty: An Empirical Test of an Information Quality Framework. In *Proceeding of the Eighth International Conference on Information Quality*,

Fugini, M., M. Mecella, et al.,2002. Data Quality in Cooperative Web Information Systems.In.

Gertz, M., T. Ozsu, et al.,2004. Report on the Dagstuhl Seminar "Data Quality on the Web".In *SIGMOD Record* vol. 33, Nº 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 *Proceeding of the Eighth International Conference on Information Quality*,

Katerattanakul, P. and K. Siau, 1999. Measuring Information Quality of Web Sites: Development of an Instrument. In *Proceeding of the 20th International Conference on Information System*,

Katerattanakul, P. and K. Siau, 2001. Information quality in internet commerce desing. Information and Database Quality. M. Piattini, C. Calero and M. Genero, Kluwer Academic Publishers.

Kerr, K. and T. Norris, 2004. The Development of a Healthcare Data Quality Framework and Strategy. In *IQ2004*,

Kopcso, D., L. Pipino, et al., 2000. The Assesment of Web Site Quality. In *Proceeding of the Fifth International Conference on Information Quality*,

Lee, Y.,2002. AIMQ: a methodology for information quality assessment.In *Information and Management. Elsevier Science*: 133-146.

Mahdavi, M., J. Shepherd, et al., 2004. A Collaborative Approach for Caching Dynamic Data in Portal Applications. In *Proceedings of the fifteenth conference on Australian database*,

Marchetti, C., M. Mecella, et al., 2003. Enabling Data Quality Notification in Cooperative Information Systems through a Web-service based Architecture. In *Proceeding of the Fourth International Conference on Web Information Systems Engineering*,

Melkas, H., 2004. Analyzing Information Quality in Virtual service Networks with Qualitative Interview Data. In *Proceeding of the Ninth International Conference on Information Quality*,

Moustakis, V., C. Litos, et al., 2004. Website Quality Assesment Criteria. In *Proceeding of the Ninth International Conference on Information Quality*,

Naumann, F. and C. Rolker, 2000. Assesment Methods for Information Quality Criteria. In *Proceeding of the Fifth International Conference on Information Quality*,

Pernici, B. and M. Scannapieco, 2002. Data Quality in Web Information Systems. In *Proceeding of the 21st International Conference on Conceptual Modeling*,

Redman, T., 2001. *Data Quality: The Field Guide*. Digital Press

Strong, D., Y. Lee, et al.,1997. Data Quality in Context.In *Communications of the ACM* Vol. 40, Nº 5: 103 -110.

Wang, R. and D. Strong,1996. Beyond Accuracy: What Data Quality Means to Data Consumer.In *Journal of Management Information Systems* 12(4): 5-33.

Winkler, W.,2004. Methods for evaluating and creating data quality.In *Information Systems* Nº 29: 531-550.

Yang, Z., S. Cai, et al.,2004. Development and validation of an instrument to measure user perceived service quality of information presenting Web portals.In *Information and Management. Elsevier Science* 42: 575-589.

Zhu, Y. and A. Buchmann, 2002. Evaluating and Selecting Web Sources as external Information Resources of a Data Warehouse. In *Proceeding of the 3rd International Conference on Web Information Systems Engineering*,