DEFINING THE INTRINSIC DATA QUALITY FOR WEB PORTALS

Carmen Moraga[1], Angélica Caro[2], Mª Ángeles Moraga[1], Rodrigo Romo Muñoz[3] and Coral Calero[1]
[1]*Alarcos Research Group-Institute of Information Technologies & Systems, University of Castilla-La Mancha, Paseo de la Universidad 4, Ciudad Real, Spain*
[2]*Department of Computer Science and Information Technologies, University of Bio Bio, Chillán, Chile*
[3]*Department of Business Management, University of Bio Bio, Chillán, Chile*

Keywords: Data Quality Characteristics, Web Portal, Data Quality Model.

Abstract: Web portals facilitate access to data on the Internet. Their use and quantity have increased, which has led to the need to satisfy user preferences in order for them to continue on the market. One competition-based strategy is to provide an adequate quality of service, and one of the key factors in this is the quality of the data provided by the Web portal. Bearing this in mind, we have defined a Data Quality (DQ) Model for Web portals. This model is composed of 42 DQ characteristics which are organized into two points of view and four DQ categories. In this paper we present work which is based on the statistical study of users' opinions of one of the characteristics in the model – the Intrinsic DQ category. Our objective is to determine which characteristics in the Intrinsic category will be most relevant as regards the users' gender, age range, level of studies and knowledge of computing, thus allowing us to verify that, according to the aforementioned parameters, the importance of some characteristics varies with regard to others. For example, women place more importance upon Credibility, while men view the Accessibility of the data in Web Portals as being more important.

1 INTRODUCTION

Web portals have been consolidated as an appropriate means to organize and facilitate access to data on the Internet.

As more and more organizations use this means as a way in which to capture followers or 'customers', the competition between them also grows. One basic competition strategy is to provide a quality customer service (Yang et al., 2004), and one of the important factors involved in this is data quality. In the relevant literature, the concept of Information Quality or Data Quality (hereafter DQ) is often defined as ''fitness for use'', i.e., the ability of a collection of data to meet user requirements (Cappiello et al., 2004; Strong et al., 1997). Many DQ models have been proposed in the last few years, and a standard was even proposed in 2008 (ISO/IEC-FDIS-25012, 2008). However, very few works dealing with the DQ in the context of Web portals have appeared. Examples of the few which have are: (Caro et al., 2008; Grigoroudis et al., 2008; Metzger, 2007; Yang et al., 2004).

Our research interest is focused on the creation of a DQ model for Web portals.

In a previous work, we identified a set of DQ characteristics that were organized into four categories (Intrinsic, Operational, Contextual and Representational), and which were divided into two points of view (Inherent and System Dependent). These constitute the SPDQM (SQuaRE-Aligned Portal Data Quality Model).

In order to make SPDQM usable and applicable, we have defined two different approaches: Static and Dynamic. The former is oriented towards determining those characteristics that are relevant to certain types of Web portal users according to their various demographic aspects. That is to say, its intention is to determine which quality characteristics are most important with regard to the users' gender, age range, level of studies, etc. The latter is oriented towards obtaining the level of DQ that exists in each Web portal, at a specific moment, for a given context.

In this paper we present the results of the empirical work carried out for the Static approach in the Intrinsic DQ category of SPDQM.

The remainder of this paper is organized as

Moraga C., Caro A., Moraga M., Romo Muñoz R. and Calero C..
DEFINING THE INTRINSIC DATA QUALITY FOR WEB PORTALS.
DOI: 10.5220/0003929603740379
In *Proceedings of the 8th International Conference on Web Information Systems and Technologies* (WEBIST-2012), pages 374-379
ISBN: 978-989-8565-08-2

follows. Section 2 describes our model SPQDM which was obtained. Section 3 presents the Static and Dynamic approaches. Section 4 focuses on the Static approach towards Intrinsic DQ. In Section 5, the results are shown, and our conclusions and future work are presented in Section 6.

2 SPDQM

Web portals are increasingly more important. However, insufficient attention has been paid to the data quality in them. A Portal Data Quality model has therefore been created to deal with this lack (SPDQM). SPDQM is based on three different elements: PDQM (Portal Data Quality Model) (Caro et al., 2008), a set of DQ characteristics obtained from the relevant literature (Moraga et al., 2009 a) and the ISO/IEC 25012 standard. SPDQM will allow the DQ level in a Web Portal to be discovered, and will guide designers and developers in discovering the most relevant DQ characteristics according to the type of user, which will be determined by various demographic aspects.

Each of the elements used to create our model can be found in (Moraga et al., 2009 b).

2.1 DQ Characteristics for SPDQM

SPDQM contains a set of 42 DQ characteristics that are appropriate to Web portals. The first level corresponds with the two points of view adopted from the ISO/IEC 25012 standard (see Figure 1a).
The second level corresponds with the DQ categories adopted from the PDQM model (Caro et al., 2008) (see Figure 1b):

(a) Intrinsic, which denotes that data have quality in their own right; (b) Operational, which emphasizes the importance of the role of systems, that is, the system must be accessible but secure; (c) Contextual, which highlights the requirement which states that DQ must be considered within the context of the task in hand; and (d) Representational, which denotes that the system must present data in such a way that they are interpretable, easy to understand, and concisely and consistently represented, in a specific context.

The third level corresponds with the set of characteristics in each category (see Figure 1c). Based on the definitions of the PDQM's categories and the ISO/IEC 25012's points of view (see Figure 1a and Figure 1b). Finally, the fourth level contains the sub-characteristics of the characteristics in the previous level (see Figure 1d).

Once SPDQM was defined, the next step was to define how this model could be applied in practice, such that it would be possible to determine the set of DQ characteristics that would be most relevant as regards the user perspective.

This would also assist us in the evaluation and/or improvement of the level of DQ in a Web portal.

The following section describes the two approaches that we were defined in order to apply SPDQM.

3 TWO APPROACHES WITH WHICH TO ADDRESS THE DQ IN WEB PORTALS

This section shows the two approaches used to apply

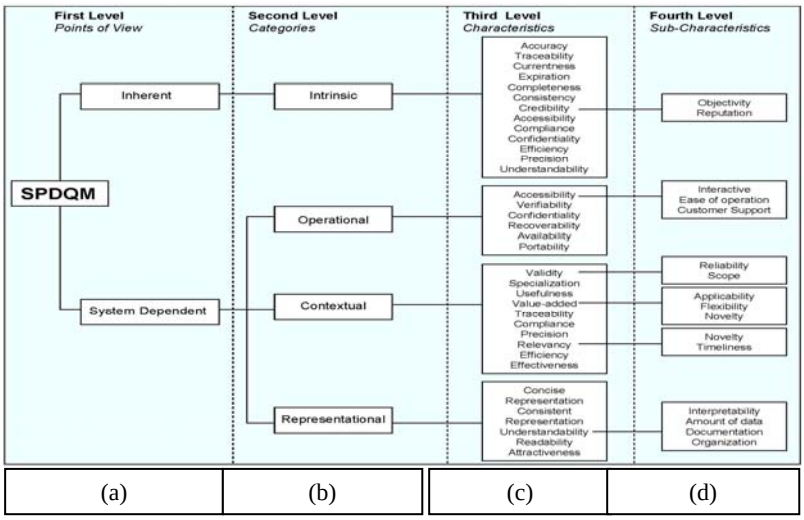

Figure 1: The structure of SPDQM model.

our model.

The Static approach will be carried out only once for each category (Intrinsic and Contextual). This approach consisted of determining the set of characteristics that are most relevant as regards the different types of Web portal users. These types of users will be determined according to different demographic aspects. These demographic aspects are: gender, age range, level of studies, etc, according to the category that will be analysed. The results obtained will be used to create a tool that will allow to guide designers and developers in what the relevant characteristics are for each type of Web portal user. These guidelines will also indicate criteria with which to satisfy user preferences.

A set of measures for each of the characteristics in the Representational and Operational categories will be defined in the Dynamic approach. The objective of this approach is to obtain the level of DQ of each Web portal at a specific moment. The results will vary depending on the Web portal that is being analyzed, and it will not be possible to generalize the results of the analysis. A tool which automates the measures will be also available for this approach. This tool will use the URL in a Web portal to obtain the level of DQ for the characteristics in these categories. This evaluation will then be used as a basis to provide its designers and developers with recommendations for its improvement.

The steps followed in each approach are shown as follows (see Figure 2).

The following steps will be carried out in the Static approach:

- An initial set of questions will be created in relation to the definitions of the characteristics in each category, and to the demographic aspects.
- These questions will be reviewed by a group of experts in statistical analysis.
- The questions will be modified with the feedback obtained from the experts.
- A pilot survey will be sent to a trial group who will fill in the questionnaire. This group will be composed of Web portal users, and will inform us whether the questions are understandable, or whether they will have to be improved.
- The questions in the pilot survey will be reviewed, and the feedback from the trial group will be taken into consideration.
- A questionnaire containing the final set of questions will be created. The questions must be written in such a way that they are understandable to any Web portal user.
- The questionnaires will be distributed, either in printed format or via e-mail, to a heterogeneous group of Web portal users.
- The questionnaires will be collected either by hand or by e-mail.
- The results obtained will be analysed, and any questionnaires containing erroneous data or with questions that had not been answered will be discarded.
- A statistical analysis will be carried out to determine the most relevant DQ characteristics according to the types of users. This will be done by
  - First: The mean value of each characteristic is obtained in order to verify whether all the characteristics are important from the point of view of user preference.
  - Second: A correlation analysis is carried out to verify whether there is a relationship

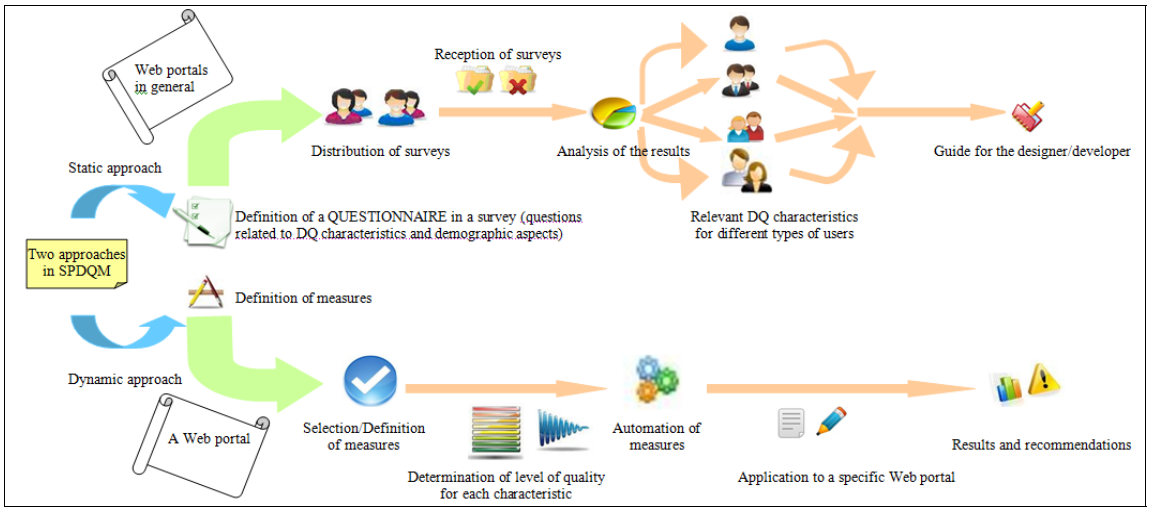

Figure 2: The structure of SPDQM model.

between the category itself and the characteristics associated with that category. This is to reinforce the idea that these characteristics are truly correctly associated with their.

- Third: The set of characteristics is used to create homogeneous groups of characteristics (denominated as factors). These groups are formed of those characteristics which have a considerable amount of correlation with each other. We also attempt to ensure that each group is independent of the others.
- Finally: User groups (denominated as clusters) are created, which are related to the factors. Each user group is formed of different types of users, according to the demographic aspects. The DQ characteristics which are most important to different types of users are therefore grouped together.

- Guidelines containing recommendations for Web portal designers and developers will be created, which would allow them to discover user preferences.

The process used in the case of the Dynamic approach will be as follows:

- A set of measures will be defined for each characteristic.
- The level of quality desired for each characteristic will be determined.
- Recommendations by which to improve the level of data quality in said Web portal will be generated for each DQ characteristic.
- The measures will be automated through the creation of a computerised tool.
- The tool will be made available for direct use by any interested parties in specific Web portals.

In this paper we shall show the application of the Static approach to the Intrinsic DQ category.

4 THE INTRINSIC DQ IN WEB PORTALS

In this section we shall use the Static approach in the Intrinsic DQ category (see Figure 1). The questionnaire in this survey was made up of a total of 20 questions, 16 of which were related to the DQ characteristics in the category (also including a question concerning the definition of the term 'Intrinsic'), and the other 4 of which were related to demographic aspects. We used closed questions, in which only one response was possible.

The 16 questions were created using a 5-point Likert scale, where 1 is "Not at all important" and 5 is "Very important".

The questions related to demographic aspects, on the other hand, allowed us to define the different types of users. These types of users were obtained according to their gender, age range, level of studies and knowledge of computing.

The questionnaire was of the unsupervised type, and was distributed to a heterogeneous group of 150 Web portal users. The results of the questionnaires are analysed in the following section.

5 ANALYSIS OF THE RESULTS

This section shows the study of the results obtained from the questionnaires. Of a total of the 137 questionnaires received, the data from 136 were analysed, since the questions relating to the demographic aspects had not been answered in one of them. This study was carried out by using an SPSS statistical analysis tool.

The objective of our study is, on the one hand to determine whether all the DQ characteristics in the Intrinsic DQ category are important for Web portal users and, on the other, to analyse whether some characteristics are more relevant than others according to the different types of users.

As a starting point, it was necessary to estimate the reliability of the results. This was done by calculating the Cronbach's alpha. The result obtained from this was a value of 0.856, which indicated that the results had good internal consistence. The information is therefore reliable.

A descriptive statistical analysis was then carried out in order to determine whether all the DQ characteristics are important to Web portal users. This analysis allowed us to obtain the central tendency (mean) and the dispersion (typical deviation) for all the variables in the study. The variables correspond with the characteristics that are included in the Intrinsic category, along with the Intrinsic DQ category itself. As a result of this, we observed that the mean value, of all the characteristics, is approximately four. As additional data, we obtained that the characteristics Credibility, Accessibility, Reputation, Consistency and Currentness have a higher mean value. The only characteristic that was below the mean value of four was that of Traceability. Although the value for this category was below four, it had a mean of 3.96, which is very close to four, and it is therefore also considered to be important.

This proximity to the value of four indicates that all the characteristics in the Intrinsic DQ category are, in effect, important to Web portal users.

In order to reinforce the previous idea, we carried out a correlation analysis between the DQ characteristics (which were considered as dependent variables) and the Intrinsic DQ (independent variable). The method used was the Spearman correlation analysis. The results showed that the characteristics are related to the Intrinsic DQ category at an extremely high level.

Having proved that all the characteristics are important and that there is a correlation with the Intrinsic DQ, the next step was to analyse whether some characteristics are more important than others with regard to the different types of users. This was done by carrying out a factorial analysis and a cluster analysis, which are detailed as follows.

5.1 Factorial Analysis

At this point, we created groups of characteristics. Each group was independent of the others and was denominated as a factor. This would allow us to later relate each of these factors to the different types of users.

Table 1 shows the characteristics that correspond with each of the factors obtained.

Table 1: Characteristics in each factor

Factor	Characteristics
Factor 1	Compliance, Confidentiality, Currentness, Understandability, Efficiency, Completeness, Precision
Factor 2	Reputation, Objectivity, Credibility, Accuracy, Consistency
Factor 3	Traceability, Accessibility and Expiration

These three factors combined explain 52.9% of the total variability (which is confirmed by KMO, which has a value of 0.804). Factor 1 accounted for 35.1% of the total variance. Factor 2 accounted for 10.4% of the total variance, and Factor 3 accounted for 7.4%.

All of the DQ characteristics have now been placed in groups. However, our eventual intention was to discover whether there is more relevance between some characteristics and others according to the different types of users. The next step, therefore, was to group these factors by carrying out a cluster analysis whose purpose is shown as follows.

5.2 Cluster Analysis

The results obtained in the factorial analysis were then used to carry out a cluster analysis in order to group the factors by resemblance or similitude. The factors were organised into two clusters, since this is the minimum number of clusters that will allow each factor to be differentiated in a single cluster. In the first cluster, those characteristics located in Factors 1 and 2 are positively valued, whilst cluster 2 positively values the characteristics in Factor 3.

Finally, the clusters were related to the demographic aspects. In this way, it was obtained that Cluster 1 contains users who are principally female, under 25 or over 45, with a basic level of studies or with vocational training, and with a basic or intermediate knowledge of computing. The designers and developers aimed at this type of users should, therefore, bear in mind the characteristics contained in Factors 1 and 2.

The users in Cluster 2 are principally male, between 25 and 45 years of age with vocational training or university studies and with an advanced knowledge of computing. The designers and developers of Web portal aimed at this type of user should therefore bear in mind the characteristics in Factor 3.

The results obtained have been summarized in a table (Table 2) in which it will be observed that all the users consider the characteristics in the Intrinsic DQ category to be relevant (R). It will also be noted that some of these characteristics have been allotted a higher value (R+) and others have been allotted a lower value (R-). Each of the clusters, later gave a greater value to different subsets of characteristics in the Intrinsic DQ category (R+).

Table 2: Relevant characteristics.

Intrinsic Characteristics	Users in General	Cluster 1	Cluster 2
Compliance	R	R+	
Traceability	R-		R+
Reputation	R+	R+	
Objectivity	R	R+	
Credibility	R+	R+	
Accuracy	R	R+	
Consistency	R+	R+	
Accessibility	R+		R+
Confidentiality	R	R+	
Currentness	R+	R+	
Expiration	R		R+
Understandability	R	R+	
Efficiency	R	R+	
Completeness	R	R+	
Precision	R	R+	

6 CONCLUSIONS AND FUTURE WORK

This paper presents a data quality model for Web portals, denominated as SPDQM.

The theoretical definition of the model served as a basis for the definition of two approaches which can be used in the evaluation of the DQ in Web portals. One was a static approach whose objective is to determine those DQ characteristics which are most important for Web portal users. The other was a dynamic approach whose objective is to obtain the DQ level of the data contained in a specific portal at a determined moment. All of this should enable designers and developers to be guided in the construction of portals, in order to make them more appropriate for the users at which they are aimed.

In this paper, we have presented our proposal for the development of the static approach in the Intrinsic DQ category. To do this, we carried out a survey and then analysed the results obtained from the questionnaire by using the SPSS statistical tool. The results obtained have allowed us to determine that all the characteristics are important, and to demonstrate that some characteristics are, in effect, more relevant than others according to the different types of users, who are determined by the following demographic aspects: gender, age range, level of studies and knowledge of computing. Finally, we have indicated the set of characteristics to which more attention should be paid in the DQ of a Web portal according to the user towards whom it is aimed.

Our next step will be to create support guidelines which will be available via a computing tool.

Our short-term future work will be to carry out the static approach for the Contextual characteristics. In the long term we shall develop the dynamic approach for the remaining categories. Our eventual intention is to make our model available to users and developers through a free tool.

ACKNOWLEDGEMENTS

This research has been funded by the following projects: ORIGIN (CDTI-MICINN and FEDER IDI-2010043(1-5)), PEGASO/MAGO project (Ministerio de Ciencia e Innovacion and Fondo Europeo de Desarrollo Regional, TIN2009-13718-C02-01), EECCOO (MICINN TRA2009_0074), and VILMA (JCCM PEII 11-0316-2878).

REFERENCES

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

Caro, A., C. Calero, I. Caballero and M. Piattini (2008). "A proposal for a set of attributes relevant for Web portal data quality." Software Quality Journal 16(4): pp. 513-542.

Grigoroudis, E., C. Litos, V. A. Moustakis, Y. Politis and L. Tsironis (2008). "The assessment of user-perceived web quality: Application of a satisfaction benchmarking approach." European Journal of Operational Research 187(3): pp. 1346-1357.

ISO/IEC-FDIS-25012 (2008). "Software engineering - Software product Quality Requirements and Evaluation (SQuaRE) - Data quality model."

Kitchenham, B. and S. Charters (2007). "Guidelines for performing systematic literature reviews in software engineering." Technical Report EBSE-2007-01, School of Computer Science and Mathematics, Keely University.

Metzger, M. J. (2007). "Making sense of credibility on the web: Models for evaluating online information and recommendations for future research." Journal of the American Society for Information Science and Technology 58(13): pp. 2078-2091.

Moraga, C., M. Moraga, C. Calero and A. Caro (2009) a. Towards the Discovery of Data Quality Attributes for Web Portals. ICWE 2009. LNCS 5648. pp. 251-259

Moraga, C., M. Moraga, A. Caro and C. Calero (2009) b. SQuaRE-Aligned Portal Data Quality Model. Doctoral Consortium. ICWE 2009. (pp. 50-54)

Strong, D., Y. Lee and R. Wang (1997). "Data Quality in Context." Communications of the ACM 40(5): pp. 103-110.

Yang, Z., S. Cai, Z. Zhou and N. Zhou (2004). "Development and validation of an instrument to measure user perceived service quality of information presenting Web portals." Information and Management. Elsevier Science. 42: pp. 575-589.