SPDQM: SQuaRE-Aligned Portal Data Quality Model

Carmen Moraga[1], Mª Ángeles Moraga[1], Angélica Caro[2] and Coral Calero[1],

[1]Alarcos Research Group – Institute of Information Technologies & Systems,
Paseo de la Universidad 4, 13071 Ciudad Real, Spain
Carmen.Moraga@alu.uclm.es, {MariaAngeles.Moraga, Coral.Calero}@uclm.es
[2]Department of Computer Science and Information Technologies,
University of Bio Bio, Chillán, Chile
mcaro@ubiobio.cl

Abstract. Web portals are currently an important means to access Internet information. The use of Web portals permits a vast amount of data to be obtained rapidly. However, the quality of the data recovered by the user is fundamental. We therefore propose a thesis in which a model, denominated as SPDQM (SQuaRE Portal Data Quality Model) will be defined. The proposed model will be based on a previous model, PDQM (Portal Data Quality Model) and the SQuaRE (Software product Quality Requirements and Evaluation) standard. Finally, upon the model‘s completion, an automatic tool with which to asses the data quality (DQ) in Web portals will be developed.

Keywords: Web Portal, data quality, quality model.

1 Introduction

A Web portal is a Website or service that offers a broad array of resources and services for customers and business partners [1]. Many companies currently use Web portals to offer their products, thus providing users with 24-hour access in which to buy them. However, users need to know whether the data offered by Web portals are updated, reliable, correct, and so on. This is not only important for consumers but also for providers. The aforementioned reasons led us to consider the necessity of a Web portal data quality model.

In order to develop this model it is first necessary to study the existing research into data quality in general and the data for Web portals in particular.

With regard to data quality, SQuaRE came into existence some months ago. SQuaRE was selected in this paper, because it is the most recent series of International Standards in which a data quality model is defined. We therefore believe that it is necessary to consider the quality characteristics identified in this standard.

With regard to data quality in Web portals, we should highlight the PDQM model. However, its definition is previous to SQuaRE, it is specific to only one type of Web portal: that of universities, and it has only been partially implemented.

PDQM and SQuaRE will therefore be used as a basis to develop a more complete DQ model for Web portals.

This article is organized as follow. In Section 2, a background is presented, while the thesis proposal is presented in Section 3.

2 Background

This Section presents relevant proposals for our work.

2.1 PDQM (Portal Data Quality Model)

As was previously mentioned, the PDQM (Portal Data Quality Model) model will be used as a starting point. PDQM is focused on the perspective of the data consumer. The development of PDQM was divided into two stages: the theoretical definition and the operational definition of the model [2].

The goal of the theoretical definition was to determine a set of DQ characteristics that are relevant to data consumers when evaluating the DQ of any Web portal. To do this, a set of DQ characteristics proposed in literature was chosen to evaluate DQ in a Web context and a selection of the most relevant characteristics for a Web portal was then defined. This set was empirically validated, resulting in the final set of DQ characteristics for the model. The operational version of PDQM was obtained by first organizing the characteristics into four DQ categories:

- ***Intrinsic,*** which denotes that data have quality in their own right.
- ***Operational,*** which emphasizes the importance of the role of systems; that is, the system must be accessible but secure.
- ***Contextual,*** which highlights the requirement which states that data quality must be considered within the context of the task in hand.
- ***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 each category, influential relationships were then established between the characteristics to determine which characteristics were dependent on other characteristics. As a result of this, a BN (Bayesian Network) was obtained which organizes the 33 DQ characteristics into four network fragments (one for each DQ category). The problem is that PDQM was only created for university Web portals and, up to now, only evaluates the representational category.

2.2 ISO/IEC 25012

SQuaRE is a set of International Standards which consists of different divisions. Our model will be based on the ISO/IEC 25012 which proposes a data quality model that defines fifteen characteristics considered from two points of view: inherent and system dependent [3].

Inherent data quality refers to the degree to which quality characteristics of data have the intrinsic potential to satisfy stated and implied needs when data is used under specified conditions [3]. The characteristics in this set are specifically: “Accuracy”, “Completeness”, “Consistency”, “Credibility” and “Currentness”. System dependent data quality refers to the degree to which data quality is attained and preserved within a computer system when data is used under specified conditions [3]. The characteristics in this set are specifically: “Availability”, “Portability” and “Recoverability”. The set of characteristics for both points of view is composed of “Accessibility”, “Compliance”, “Confidentiality”, “Efficiency”, “Precision”, “Traceability” and “Understandability”.

3 SPDQM (SQuaRE-Aligned Portal Data Quality Model)

This Section presents our thesis proposal:

- ***Problems to solve***: As was explained in the introduction, Internet has increased the amount of data that can be obtained through the network and the ability to provide and obtain information from several sources. Web portals thus serve as an important means to access information. Web portals have therefore undergone an evolution and currently provide a variety of services. One of the aims of many Web portals is to select, organize and distribute content (information or other services and products) in order to satisfy their users/customers [4]. These users wish to recover data with an acceptable level of quality. Web portals developers should therefore take data quality into consideration. Otherwise, if users recover data which are not-up-date or are incorrect, the next time they need information they will probably not access the same Web portal.
- ***Objective of the proposal***: The aim of this proposal is to create a data quality model aligned to SQuaRE and develop a tool which implements SPDQM and evaluates the DQ in a Web portal.
- ***Research methodology***: The research methods used are:
  - Survey: The set of data quality characteristics which are relevant to the existing proposal have been obtained through a survey, based on the steps of Kitchenham [5]. Since PDQM is used as a starting point, our survey covered the time period between 01/01/2006 and 31/12/2008 (the search in PDQM covered the period prior to 2006). As a result, we obtained that 39 characteristics were relevant for the Web portal context.
  - Methodology to develop a quality model: the quality model will be defined by using a methodology which is being developed by colleagues at the University of Castilla-La Mancha and the University of Málaga.
  - Methodology to define the measures: [6] identifies three activities with which to correctly define the measures. These activities are:
    - Measures definition: this is carried out by considering the specific characteristics that we wish to measure and the experience of designers and users.
    - Theoretical validation: this helps us to discover when and how to apply the measures.
    - Empirical validation: this proves the practical utility of the proposed measures.

  This methodology will be used to define the necessary measures.
- ***Model***: The model was developed by using both the set of characteristics proposed in PDQM and ISO/IEC 25012 and those obtained as a result of our survey. Due to the survey selects the DQ characteristics for Web in general, the next step is to refine the set of obtained characteristics and to study the applicability of the

characteristics to the Web portal context. Once the characteristics which are most suitable for Web portals had been chosen it was necessary to resolve conflicts. This was done by detecting both those characteristics which have the same name but a different meaning and those characteristics which have a different name but refer to the same meaning. And when a characteristic has an only subcharacteristic, the subcharacteristic is removed and it is taken into account in the definition of the characteristic. As a final result, we have obtained 42 DQ characteristics (see Table 1).

Table 1: SPDQM

Point of view	Category	Characteristic	Subcharacteristic
Inherent	Intrinsic: it denote that data have quality in their own right	Accuracy	
		Credibility:	Objectivity
			Reputation
		Traceability	
		Currentness	
		Expiration	
		Completeness	
		Consistency	
		Accessibility	
		Compliance	
		Confidentiality	
		Efficiency	
		Precision	
		Understandability	
System Dependent	Operational: it emphasized the importance of the role of Systems; that is, the system must be accessible but secure	Availability	
		Accessibility:	Interactive
			Ease of operation
			Customer Support
		Verifiability	
		Confidentiality	
		Portability	
		Recoverability	
	Contextual: it highlights the requirement which states that data quality must be considered within the context of the task in hand	Validity:	Reliability
			Scope
		Value-added:	Applicability
			Flexibility
			Novelty
		Relevancy:	Novelty
			Timeliness
		Specialization	
		Usefulness	
		Efficiency	
		Effectiveness	
		Traceability	
		Compliance	
		Precision	
	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	Concise Representation	
		Consistent Representation	
		Understandability:	Interpretability
			Amount of data
			Documentation
			Organization
		Attractiveness	
		Readability	

Having identified all the characteristics there are two possibilities to continue with the operationalization of SPDQM:

- We can classify the characteristics according to ISO/IEC 25012 as being "Inherent" and "System Dependent". We would therefore concentrate on both ("Inherent" and "System Dependent"), and would define measures for each of its characteristics. For some characteristics, the measures would be derived from the user's opinions. However, we would like that the vast majority of the measures were automatable to allow us to develop a tool which is capable of evaluating a Web portal. The tool will provide us with values for each indicator that will have been defined (one for each characteristic). The tool will also permit the user to give more or less importance to each indicator. For example, for each indicator it is possible to ask the user if s/he believes it to be very important, not very important or not at all important. For the subjective measures, the tool will ask to the users, their opinion. The tool will use these values to calculate the final value of the characteristics associated with the "Inherent" and the "System Dependent" points of view.

- The second option is to attempt to create an extension of PDQM that is aligned with SQuaRE. To do this we would consider the “inherent” point of view of the ISO/IEC 25012 which corresponds with PDQM’s intrinsic category, and “System Dependent” which corresponds with the other three categories (see Table 1). Therefore, by following the creation of a bayesian network as that of PDQM, we would have to add new characteristics and group them with a maximum of three entries per node. The network would thus have a first and second level with the identified characteristics and the necessary artificial nodes. A third level would contain the PDQM categories, and a fourth level would contain those of ISO/IEC 25012.

We have not yet decided which option is most suitable.

- ***Automatic tool***: We will create an automatic tool to evaluate the DQ in Web portals. This tool will calculate the level of DQ based in measures calculated automatically from the web portal and other measures obtained from user´s input (like a questionnaire). The goal of this automatic tool is to facilitate the developer’s task when incorporating data quality into Web portals and to provide the users with the DQ level of a specific Web portal.
- ***Contributions to Web Engineering***: This model could be used by Web portal designers to improve data quality, promote good practices in DQ on the Web, come up with data cleaning techniques, develop patterns for data refinement or eliminate unnecessary data. Thus, the automatic tool will be developed to provide us with the necessary guidelines to carry out those improvements.

Acknowledgments. This work is part of the INCOME project (PET2006-0682-01) supported by the Spanish Ministerio de Educación y Ciencia, by the IVISCUS project (PAC08-0024-5991) supported by Consejería de Educación y Ciencia (JCCM) and by VIMECUS (TC20080556) supported by the University of Castilla-La Mancha.

PhD Student: Carmen Moraga; **Supervisors**: M^a Ángeles Moraga, Angélica Caro

References

1. Wynn, M., Zhang, S.: Web Portals in SMEs - Two Case Studies. In: Proceedings of the 2008 Third international Conference on Internet and Web Applications and Service (ICIW), pp. 303--308. IEEE Computer Society, Washington, DC (2008)
2. Caro, A., Calero, C., Caballero, I., Piattini, M.: A proposal for a set of attributes relevant for Web portal data quality. Software Quality Journal. 16, 513--542 (2008)
3. [ISO/IEC-FDIS-25012]: Software engineering - Software product Quality Requirements and Evaluation (SQuaRE) - Data quality model. (2008)
4. Domingues, M.A., Soares, C., Jorge, A.M.: A Web-Based System to Monitor the Quality of Meta-Data in Web Portals. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IATW'06), pp. 188--191 (2006)
5. Kitchenham, B., Charters, S.: Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE-2007-01, School of Computer Science and Mathematics, Keely University (2007)
6. Calero, C., Piattini, M., Genero, M.: Method for Obtaining Correct Metrics. In: 3th International Conference on Enterprise Information Systems, pp. 779--784 (2001)