Selecting Web Functionalities versus Data Quality Dimensions: A First Approach

César Guerra-García[1], Victor Menéndez-Domínguez[2], Ismael Caballero[3], Omar Montaño[1]

[1]Department of Information Technology, Polytechnic University of San Luis Potosí, México
cesar.guerra@upslp.edu.mx
[2]Faculty of Mathematics, Autonomous University of Yucatán, México
mdoming@uady.mx
[3]Information Systems and Technologies Department, University of Castilla-La Mancha, Spain
ismael.caballero@uclm.es

Abstract. Data is considered as fundamental asset for organizations, its strategic value leads to reconsider the importance of maintaining adequate levels of quality in data that is managed by applications. However, within the context of Web applications development, no mechanisms to adequately control Data Quality (DQ) requirements have been still proposed. This proposal is grounded on the idea of anticipating DQ problems that can arise through the functionalities of a Web application. The problems are characterized according to affected DQ dimensions. To do so, our aim is to identify DQ Requirements that will be translated into specific software requirements. By these means, we intend to avoid, or to minimize the effects of the DQ problems on the execution of users task. The main contribution of this paper is a "Model for the selection of DQ requirements according the web functionalities to be implemented".

Keywords. Data quality, Web development, Data quality requirements.

1 Introduction

Organizational performance is seriously affected by data quality problems [1]. The rapid growth of Internet has made that more and more companies rely their Information Systems (IS) on the Web [2]. So, it is possible to state that Web applications have been established itself as an important resource of data that has a strategic value for the organizations. Given this strategic value of data in the running of business processes [3], and taking into account that more frequently organizational data is published through Web applications, organizations need to ensuring acceptable levels of quality. Unfortunately, the Web applications are risked by the known potholes presented in [4], and similar kind of problems to data could arise with the usage of Web application [5]. Considering a systematic review presented in [6], a conclusion was given: neither most of the developers are still familiarized with the underlying concepts, nor do they have available software artifacts to proceed. For the first, Web application developers should first know what is data quality, and interpret how users

understand the concept of data quality for the functionalities they use when managing the Web application [7], and then translate these understanding into convenient software requirement. To better achieve this goal, developers need mechanisms that allow them to represent and include this knowledge in the developing of Web applications as if they were other kind of software requirements. Our research' goal is to provide developers with the adequate mechanism to manage DQ requirements within Web development. The first step was how we could identify which DQ requirements are necessary for each Web application. To answer this question, we realize about the need of identifying the minimal unit of working of a Web application, in order to have a minimal context for analyzing how the various DQ problems could affect them. Given the subjectivity of the concept of data quality (strong dependency on the user's view of the level of quality of the data executing a task), and taking into account that in this sense, the scope of a context could be limited to the execution of a given "part" of the web application, we found that the concept of functionality as generic "part" of a Web Application provided in [7], would let us articulate out research work. In addition, in order to catalogue the DQ problems we found useful the results provided by [4]. With those elements we designed a working strategy whose ultimate goal was to obtain a generic model (MOSCAF) as result of our research. MOSCAF depicts the generic relations between the Web application functionalities and the DQ dimensions which characterize the known DQ problems. The obtained model can be instantiated by any development team, for identifying DQ requirements when developing a specific functionality for a particular Web application. The paper is organized as follows. In section 2 an overview of related areas is provided. The MOSCAF model is described in Section 3. Finally, in section 4 the conclusions are outlined.

2 Related Areas

In order to reduce the negative impact of problems due to inadequate levels of DQ [8], it is paramount that companies can have a quantitative perception of their importance. So, they must assess how good their organizational data resources are for the tasks at hand. Organizations have to deal to the DQ, both in subjective perceptions by individuals that use the data, as objective measures based on a set of data. An assessment of DQ in a subjective way can reflect the needs and experiences of users with a set of data [9, 10]. As mentioned, the most accepted definition for the concept "Data Quality" is "fitness for use" [11]. This means that a user typically evaluates the quality of a set of data for a particular task, which it is done in a specific context, according to a set of criteria or dimensions of DQ. An user performing a role within a IS can specify for a piece of data different DQ software requirements as be necessary, specifying the DQ dimensions that better represent this kind of requirements for a determined functionality. For measuring the level of DQ of a piece of data, it is necessary to identify several DQ dimensions ("DQ model") which can characterize the DQ requirements in a better way [12]. In order to get a broader perspective as possible, we chose the generic DQ model proposed in the standard [13]. As we said previously, is necessary to specify the DQ requirements associated to each one of the functionalities that will be implemented in a Web application. So, we must first reviewing these functionalities described by Collins [7]: Content management, Process and actions,

Search capabilities, Administration, Security, Data points and integration, Communication and collaboration, Presentation, Taxonomy, Personalization and Help features.

3 Model for the selection of DQ requirements according the web functionality to be implemented

Once we mentioned each DQ dimensions and the main functionalities of a Web application, the next step in the research was focused to make an analysis about that DQ dimensions could be part of a DQ requirement at the moment to implement a specific Web functionality. In this respect, it will be possible ensuring that the data that will be used by each functionality have an acceptable level of quality to each user. As initial part of research, we carried out an analysis about which potholes (problems) that normally appear in a IS (defined in [4]) could be in a specific moment related with each one of Web functionalities, as result of this initial phase we got the next matrix of relation, it is showed in Table 1.

Potholes / Web functionalities	Multiple sources	Subjective production	Production errors	Too much information	Distributed systems	Nonnumeric information	Advanced analysis requirements	Changing task needs	Security and privacy requirements	Lack of computing resources
Content Management	√	√	√	√		√	√	√	√	√
Process and actions				√				√		
Search capabilities				√	√	√			√	√
Administration								√		
Security									√	
Data points and integration				√	√				√	
Communication and collaboration				√	√		√			√
Presentation								√		
Taxonomy					√					
Personalization								√	√	
Help features				√						

Table 1. Matrix of relation between Web functionalities and potholes identified.

For sake of space, we only show the first relation between "*Content Management*" and "*Multiple sources*" (significantly, every one of the relation was described as part of our research): *the existence of multiples processes or different sources which generate values of data can cause the problem of not knowing which of these sources really have the major grade of quality. For instance, it generating different values for the "same" data, the choice of the source of information must be done thoroughly, making sure the data are the same*. Taking in account that the two models considered as standard [14] versus international standard [13], introduces different meanings for the DQ dimensions, it is worth making an effort in doing the analogy between the meanings of the same dimension and limiting the scope of each one of them. For that, we studied the meaning of every dimension, showing the results in the following comparative table (see Table 2). The purpose of this comparison was to resolve conflicts in the description of the different DQ dimensions, either the existence of dimensions with the same name and different meanings or dimensions with different names but the same meaning.

Wang & Strong Model	Standard ISO/IEC 25012
Accuracy	Accuracy
Completeness	Completeness
Concise representation	Completeness
Consistent representation	Consistency
Objectivity	Consistency
Beliavability	Credibility
Reputation	Credibility
Timeliness	Currentness
Accessibility	Accessibility
Value-added	Compliance
Security	Confidentiality
Amount of information	Efficiency
Amount of information	Precision
Traceability	Traceability
Easy of understanding	Understandability
Interpretability	Understandability
Variety of data and data sources	Availability
Easy of operation	Portability
Flexibility	Portability
	Recoverability

Table 2. Comparative of DQ Dimensions.

In the last part of analysis, the idea was describing which DQ dimensions could constitute each one of the specific requirements. This required doing an analysis that identifies the Web functionalities with the DQ dimensions from those shown in Table 1. As beginning point, we take in account again the work presented by [4], in which the authors classify with base in their model [14], the DQ dimensions that affect to each one of potholes (see Table 3).

Potholes / DQ dimension	Multiple sources	Subjective production	Production errors	Too much information	Distributed systems	Nonnumeric information	Advanced analysis requirements	Changing task needs	Security and privacy requirements	Lack of computing resources
Consistency	x									
Believability	x	x								
Objectivity		x								
Correctness			x							
Completeness			x					x		
Relevancy			x				x	x		
Concise representation				x		x				
Timeliness				x	x					
Value-added				x	x	x	x	x	x	x
Accessibility				x		x			x	x
Consistent representation					x		x			
Analysis requirements							x			
Security									x	

Table 3. DQ Dimensions that affect to each potholes.

Next, and taking into account that the standard ISO/IEC 25012 is more appropriate for our work, keeping the meaning of dimension, but having in account the change of scope described in Table 2, we rewrite the Table 3 getting the matrix presented in Table 4 in which shows the relationship (indicated by the symbol "√") between Web

functionalities and the DQ dimensions identified by the standard ISO/IEC 25012. In order to complement the results obtained, we decided to conduct a more exhaustive analysis, concerning to other dimensions that might be at one time suspected to be linked with some other functionalities, thus obtaining a set of new relationships (symbol "α"). Table 4 shows the final result "*Model for the selection of DQ requirements according the web functionalities to be implemented - MOSCAF*". In this sense we can say that a DQ requirement may be specified as a subset of each of rows of the matrix, for each functionality to be implemented.

DQ dimensions ISO 25012 / Web functionalities	Accuracy	Completeness	Consistency	Credibility	Currentness	Accesibility	Compliance	Confidentiality	Efficiency	Precision	Traceability	Understandability	Availability	Portability	Recoverability
Content Management	α	√	√	√	√	√	√	√			α	α		α	
Process and actions		√			√	√	√								
Search capabilities		√	√		√	√	√	√			α		α		
Administration		√					√		α	α		α		α	α
Security						√	√	√			α				
Data points and integration		√	√		√	√	√	√					α		
Commuunication and collaboration	α	√	√	α	√	√	√						α		
Presentation		√					√								
Taxonomy	α		√		√		√								
Personalization		√				√	√	√					α		
Help features		√			√	√	√					α			

Table 4. Model for the selection of DQ requirements according the web functionalities.

For a greater understanding of the model (MOSCAF), we describe each one of the different relationships specified ("√", "α"), for sake of space we only show in Table 6 the first subset of relationship about "Content Management".

Accuracy	Completeness	Consistency	Credibility	Currentness	Accessibility
The data managed for their inclusion should represent a correct value according to specific context of use.	All data managed with the application must be complete in each one of its attributes.	All data managed with the application must be coherent in a same environment of use.	The data managed with the application must be credible for users.	Data managed should be updated according to context of its use for each user.	The data managed with the application must always be accessible for users.
Compliance	Confidentiality	Traceability	Understandability	Portability	
The data managed with the application must adhere to laws or standards specified by the Chief of content management.	The data should be classified in different level of confidentiality, it ensuring that data only are accessible by authorized users.	Some information should be provided about "when" and "who" published such data, and "who" will be able to access them.	The data should be managed in an appropriate language, using the symbols or units suitable to be understood by each user.	The data may be installed or moved into any other application in the organization.	

Table 5. Content Management.

4 Conclusions

At present, Information Systems in general and particular Web applications are as important as organization itself. Data are fundamental assets of any organization, and is the raw material of these Information Systems. Therefore, the data must have enough quality to achieve that information systems can satisfy the information needs of users with the adequate quality level. To address this problem, this paper has proposed a "Model for the selection of DQ requirements according the functionalities to be implemented in a Web application (MOSCAF)". With this model, we have tried to facilitate the identification and selection of DQ requirements for a Web application. It can be understood as a way that analysts can follow to write a Requirements Specification Document complemented with management of DQ, always keeping in mind the DQ dimensions that should be implemented for each functionality during the Web applications development.

5 Bibliography

1. Scannapieco, M. and L. Berti-Equille, *Report from the First and Second International Workshops on Information Quality in Information Systems- IQIS 2004 and IQIS 2005 in Conjunction with ACM SIGMOD/PODS Conferences.* SIGMOD RECORD, 2006. **35**(2): p. 50-52.
2. Yang, Z., et al., Development and validation of an instrument to measure user perceived service quality of information presenting Web portals. Information and Management, 2004. 42(4): p. 575-589.
3. Caro, A., et al., A proposal for a set of attributes relevant for Web Portal Data Quality. Software Quality Journal, 2008.
4. Strong, D., Y. Lee, and R. Wang, Ten Potholes in the Road to Information Quality. IEEE Computer, 1997: p. 38-46.
5. Oliveira, P., F.t. Rodrigues, and P. Henriques. A formal Definition of Data Quality Problems. in Tenth International Conference on Information Quality (ICIQ'05). 2005. MIT, Cambridge, MA, USA.
6. Guerra-García, C., I. Caballero, and M. Piattini, A Survey on How to Manage Specific Data Quality Requirements during Information System Development. Communications in Computer and Information Science, 2011. 230(Evaluation of Novel Approaches to Software Engineering): p. 16-30.
7. Collins, H., Corporate Portal Definitions and Features. 2001, New York, NY, USA: Amacom Books.
8. Caballero, I., et al. MMPRO: A Methodology Based on ISO/IEC 15939 to Draw Up Data Quality Measurement Processes. in ICIQ. 2008.
9. Ballou, D.P., R.Y. Wang, and H. Pazer, Modelling Information Manufacturing Systems to Determine Information Product Quality. Management Science, 1998. 44(4): p. 462-484.
10. Wang, R.Y., A Product Perspective on Total Data Quality Management. Communications of the ACM, 1998. 41(2): p. 58-65.
11. Ge, M. and M. Helfert. A Review of Information Quality Research. in International Conference on Information Quality. 2007. MIT, Cambridge, MA, USA.
12. Lee, Y.W., et al., Journey to Data Quality 2006, Cambridge, MA, USA: Massachussets Institute of Technology.
13. ISO-25012, ISO/IEC 25012: Software Engineering-Software product Quality Requirements and Evaluation (SQuaRE)-Data Quality Model. 2008.
14. Wang, R. and D. Strong, Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems; Armonk; Spring 1996, 1996. 12(4): p. 5-33.