西安交通大學 西安交通大學管理学院 THE SCHOOL OF MANAGEMENT XI'AN JIAOTONG UNIVERSITY XJTU-MIT CDSIQ

The 19th International Conference on Information Quality (ICIQ2014)

Big Data: Management & Data Quality

August 1-3, 2014
Xi'an China

19th International Conference on Information Quality (ICIQ2014)

Organizing Committee & Program Committee

Conference Chairs:
Richard Wang, Massachusetts Institute of Technology, USA
Zongben Xu, Xi'an Jiaotong University, China
Wei Huang (Wayne), Ohio University, USA and Xi'an Jiaotong University, China

Program Chairs
Gengzhong Feng, Xi'an Jiaotong University, China
Qin Su, Xi'an Jiaotong University, Xi'an, China
John Talburt, Universityof Arkansas at Little Rock, USA

Advisory Committee:
Yingluo Wang, Xi'an Jiaotong University, China (Co-Chair)
Stuart E. Madnick, Massachusetts Institute of Technology, USA (Co-Chair)
Yang Lee, Northeastern University & MIT, USA
Louiqa Raschid, University of Maryland, USA
Yijun Li，NSFC Management Division
Guoqing Chen，Tsinghua University
Zhiyan Feng, NSFC Management Division

International and Regional Program Committee (sorted by alphabet)
Jacky Akoka, National Conservatory of Arts and Crafts, France
Alexander Borek, University of Cambridge, UK
Jiangjun Cai, Hong Kong University of Science and Technology, China
Jian Chen, Tsinghua University, China
C.H. Cheng, The Chinese University of Hong Kong, China
Waiman Cheung, The Chinese University of Hong Kong, China
Bruce Davidson, Cedars-Sinai Health Systems, USA
Xingchun Diao, PLA University of Science and Technology, China
Jimmy C.H.Fung, Hong Kong University of Science and Technology, China
Jing Gao, Univeristy of South Australia, Australia

Conference Organization Committee(sorted by alphabet)

Feng Wu, Xi'an Jiaotong University, China
En Xie, Xi'an Jiaotong University, China
Xiaotao Yao, Xi'an Jiaotong University, China
Changhong Yuan, Xi'an Jiaotong University, China
Xiaofang Zhou, Suzhou University, China
Guijun Zhuang, Xi'an Jiaotong University, China

Conference Secretary Committee
Hongyun Zhang, Xi'an Jiaotong University, China (Co-Chair)
Jiayin Wang, Washington University, China (Co-Chair)
Yahui Li, Xi'an Jiaotong University, China
Tan Zhang, Xi'an Jiaotong University, China

Sponsorship Committee
Wei Sun, Xi'an Jiaotong University, China (Co-Chair)
Yuewen Liu, Xi'an Jiaotong University, China (Co-Chair)

PREFACE

WELCOME to the **19th *International Conference on Information Quality*** (ICIQ 2014). The conference is hosted by the School of Management, Xi'an Jiaotong University, which is top-level management school in China.

With the arrival of the information era, an increasing amount of significant dicisions are driven by comprehensive and rigorous data analysis. Big Data that has been collected from numerous domains by various technologies presents tremendous opportunities to accelerate the breakthrough discoveries and innovations in different filelds, while making significant, social and economic impacts. Information quality is the foundation of any data analysis. Irregularities are widespread in Big Data and often lead to erronerous conclusions. To assure information accuracy, completeness, relevance, consistency, reliability, and accessibility from large, diverse, distributed and heterogeneous data sets pose many research and practical challenges to scientists.

The purpose of the conference is to bring together researchers and practitioners that are interested in the theory or applications of system-driven information quality enhancement, knowledge discovery, and operation improvement in a wide range of domains and disciplines. It provides an international forum to present original research results, and exchange innovative and practical development experiences. The ICIQ Conference has been instrumental in establishing a premier forum for the community of information quality researchers and practitioners.

We wish to extend our acknowledgement to all conference participants and thank them for their contribution. The Call for Paper response was extraordinary in both quantity and quality. We thank you all for your contribution in establishing Information Quality as a multidisciplinary field and pushing research to be even more useful and practice-oriented. We thank all the members of the program committee who worked diligently reviewing papers and providing valuable feedbacks to the authors. Their rigorous and timely reviews contributed greatly to this exciting program. Without the tireless work of the conference's management team this conference would not be held on time.

Finally, we wish to thank all the authors for their quality submissions and patience with delays and system glitches. We hope you all enjoy the conference and that you are inspired by the discussions through the conference.

The 19th ICIQ Committee

August 1-3, 2014

CONTENTS

Selecting Data Quality Requirements for each different Web Functionalities

Research in Progress

César Guerra-García[1,a],Ismael Caballero[2], Mario Piattini[2]
[1] Department of Information Technology, Polytechnic University of San Luis Potosí, México
[2]Information Systems and Technologies Department, University of Castilla-La Mancha, Spain
[a]cesar.guerra@upslp.edu.mx

Abstract

Data is considered as one of the most important asset for organizations. Its strategic value leads to reconsider the importance of maintaining adequate levels of quality in data that is managed and used by applications, especially in Web Applications as the main organizational showcase. 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 and avoiding DQ problems that can arise through the functionalities of a Web application. The problems are characterized according to the 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 at least to minimize the effects of the DQ problems on the execution of users task. The main contribution of this paper is MOSCAF, a "Model for the selection of DQ requirements according the functionalities to be implemented in a Web application".

Keywords: data quality, web development, data quality requirements specification.

Introduction

Organizational performance is seriously affected by data quality problems (Kahn, Strong et al. 2002, Pipino, Lee et al. 2002, Ballou and Pazer 2003 , Shankaranayanan and Cai 2005, Scannapieco and Berti-Equille 2006). Some facts demonstrates this affirmation (Shankaranayanan and Cai 2005, Cai and Shankaranarayanan 2007): the cost of data quality problems is 8-25% of an organization´s revenue; 40-50% of the budget of companies is dedicated to solving problems associated with a low quality of the handled data. And things could go worse: although the 11% of USA firms recognized to have problems for managing data quality, only the 48% have plans for managing the data quality. The rapid growth of Internet has made that more and more companies rely their Information Systems (IS) on the Web (Yang, Cai et al. 2004). As a consequence, organizational data is or can also be accessible through Web applications. 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 (Mahdavi, Shepherd et al. 2004). Given this strategic value of data in the running of business processes (Yang, Cai et al. 2004, Caro, Calero et al. 2008), and taking into account that more frequently organizational data is published to, and consumed from Web Applications, individuals and organizations need to monitor the quality of data, trying to assure acceptable levels of quality for the applications in use. Unfortunately, the Web Applications are risked by the known potholes presented by Strong et al. in (Strong, Lee et al. 1997), and the similar kind of problems to data could arise with the usage of Web application: missing values, domain violation, existence of synonymous, violation of business domain constraints, semi-empty tuples, inconsistent duplicate tuples, heterogeneity of representation, and so on (Oliveira, Rodrigues et al. 2005). These DQ problems can be characterized by one of several Data Quality Dimensions. For a broader description of the concept of Data Quality dimensions, reader is suggested to consult works in the literature (Lee, Strong et al. 2002, Pipino, Lee et al. 2002, Batini and Scannapieco 2006, Caballero, Verbo et al. 2007). Sadly, most of the current Web applications are not currently data-quality aware. To be so, they would implement some mechanisms being able to avoid DQ problems (Bizer and Cyganiak 2009). From the systematic review that that we have made (Guerra-García, Caballero et al. 2011), we concluded that

neither most of the developers are still familiarized with the underlying concepts, nor do they have available sound 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 (aka which data quality dimensions) for the tasks or functionalities they use when managing the Web application (Collins 2001, Moraga, Calero et al. 2009), and then translate these understanding into convenient software requirement. To better achieve this goal, developers need artifacts or 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. As we have found that no proposals aimed at managing these kinds of specific requirements, we posed that, despite the larger benefit that could be achieved, data quality requirement management is still requiring solutions and, it is worthy to do so. Our research' goal is to provide developers with the adequate mechanism and artifacts to manage DQ requirements within Web application development. Since it is well known that the success in the development and implementation of any application depends heavily on a good elicitation and specification of software requirements by the systems analyst (Pressman 2001, Lowe and Eklund 2002, Nicolás and Toval 2009), our concern was to explore how to capture data quality requirements. We first began by asking what a "DQ requirement" was, and how we could identify which DQ requirements were necessary for each Web application. To answer the latter 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 when playing a role within the organization), 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 by Collins in (Collins 2001), would let us articulate out research work. In addition, in order to catalogue the DQ problems we found useful the results provided by Strong et al. in (Strong, Lee et al. 1997). 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 specific Web application. Then, these DQ Requirement need to be considered as software requirements, some of them would be functional, and other non functional. Anyway, these new requirements would complement the Software Requirements Specification (SRS).

The paper is organized as follows. After this introduction, in section 2 an overview of related pillars is provided. The MOSCAF model is described in Section 3. Section 4 describes the guide of use for this proposed model. Finally, conclusions and future work are outlined in section 5.

Revision of Related Areas

Data Quality

In order to reduce the negative impact of problems (technical, organizational or legal) due to inadequate levels of DQ (Caballero, Calero et al. 2008), it is paramount that companies can have a quantitative perception of their actual 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 (Ballou, Wang et al. 1998, Wang 1998). If the users assess the quality of data as poor, their tasks could be influenced by this assessment (Pipino, Lee et al. 2002). As mentioned, the most accepted definition for the concept "Data Quality" is "*fitness for use*" (Ge and Helfert 2007). 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 Information System 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*. So, *the perception about the DQ level of a set of data could be different for diverse functionalities, even for the same user performing different roles*. For measuring the level of DQ of a piece of data, it is necessary to identify several DQ dimensions (known the set as "*DQ model*") which can characterize the DQ requirements in a better way (Lee, Pipino et al. 2006). In order to get a broader perspective as possible, we chose for our research the generic DQ model

proposed in the international standard ISO/IEC 25012 (ISO-25012 2008). This standard brings together fifteen DQ dimensions from two points of view: 1. *Inherent*: it refers to the extent to which quality characteristics of data have the intrinsic potential to satisfy stated and implied needs when data is used under specified conditions. 2. *System dependent*: it refers to the extent to which data quality is reached and preserved within a computer system when data is used under specified conditions.

Table 1 shows the definition of every DQ dimension described in standard ISO/IEC 25012.

Table 1. DQ dimensions described by standard ISO/IEC 25012

Dimension	Description
Inherent	
Accuracy	The degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use.
Completeness	The degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context or use.
Consistency	The degree to which data has attributes that are free from contradiction and are coherent with other data in a specific context of use.
Credibility	The degree to which data has attributes that are regarded as true and believable by users in a specific context of use.
Currentness	The degree to which data has attributes that is of the right age in a specific context of use.
Inherent and system dependent	
Accessibility	The degree to which data can be accessed in a specific context of use, particularly by people who need supporting technology or special configuration because of some disability.
Compliance	The degree to which data has attributes that adhere to standards, conventions or regulations in force and similar rules relating to data quality in a specific context of use.
Confidentiality	The degree to which data has attributes that ensure that it is only accessible and interpretable by authorized users in a specific context of use.
Efficiency	The degree to which data has attributes that can be processed and provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use.
Precision	The degree to which data has attributes that are exact or that provide discrimination in a specific context of use.
Traceability	The degree to which data has attributes that provide an audit trail of access to the data and of any changes made to the data in a specific context of use.
Understandability	The degree to which data has attributes that enable it to be read and interpreted by users, and are expressed in appropriate languages, symbols and units in a specific context of use.
System dependent	
Availability	The degree to which data has attributes that enable it to be retrieved by authorized users and/or applications in a specific context.
Portability	The degree to which data has attributes that enable it to be installed, replaced or moved from one system to another preserving the existing quality in a specific context of use.
Recoverability	The degree to which data has attributes that enable it to maintain and preserve a specified level of operations and quality, even in the event of failure, in a specific context of use.

Functionalities of a Web Application

As we said previously, and as part of our objectives of research, 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 enumerate and reviewing these functionalities described by Collins in (Collins 2001). At the moment of enumerate these functionalities, these have been reordered using following criteria: "*A major probability of using a Web functionality, a major probability that occur non compliances with the acceptable levels of DQ*". Is worth to highlight that the order proposed has been established taking as base

our experience and knowledge in both areas, Web development as Data Quality. In Table 2 is showed the list of functionalities ordered with a brief description from the developer point of view.

Table 2. Functionalities of a Web Application

Functionalities	Description
Content Management	This function supports the creation, authorization and inclusion (or exclusion) of content. The organizations have information and reference material that is available from the web application.
Process and actions	This function enables Web application users to initiate and participate in business process of organization.
Search capabilities	These provide several search services for Web application users, it supports searches throughout the company Web portal, the World Wide Web, and in search engines of catalogs and indexes.
Administration	This functionality provides two services: the first, a service corresponding to maintenance activities and use of task associated with the Corporate web portal. The second, a service that only can be configured by the administrator, as well as each user through of the personalization.
Security	This provides a description of the levels of access of each user or groups of users for each one of applications and functionalities included in the Web portal.
Data points and integration	These provide the ability to access information from a wide range of information sources both internal as external, and display the resulting information as a single point-of-access in the desktop.
Communication and collaboration	This function facilitates discussion, the location of innovative ideas, and the recognition of resourceful solutions. These elements enables employees to work together of more qualitative form by means of the creation of a virtual environment shared (collaboration), besides supporting the electronic messaging (communication) and adding characteristics of communication and collaboration to the business process (coordination).
Presentation	This provides to Web application users with a visual experience that encapsulates all the functionalities of the portal.
Taxonomy	This provides the context information, including the organization-specific categories that support and reflect the business of it. It provides common terminology used in the organization of fast recognition, and improvement the semantic to the users.
Personalization	This is a critical component in creating a working environment that is organized and configured specifically for each user in the organization. The key is achieve a balance between the information and the attributes needed to be consistent and constant in the personalization of the portal, and the unique characteristics required for each employee.
Help features	These provide help when using the Web application. The aid should be focused both characteristics of the portal as to specific characteristics of the organization.

The Figure 1 shows a typical example of a functionality "*Search Capabilities*" implemented into a Web application, which shows the ability of a Web application to provide search services to the users.

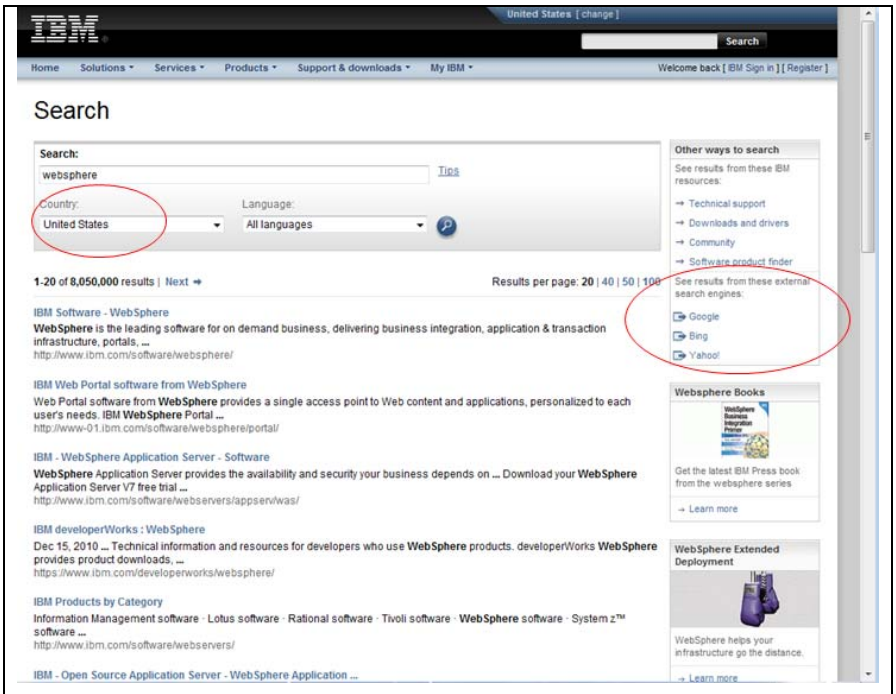

Figure 1. Web functionality "Search Capabilities"

Model for the Selection of DQ Requirements According the Functionalities to be Implemented in a Web Application

Once we described the meaning of each DQ dimensions and enumerated the main functionalities of a Web application, the next step in the research is focused to make an analysis about that DQ dimensions can 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. A complete sketch of all analysis made in the research is presented in Figure 2. Later on, we describe in detail each of the parts of that analysis.

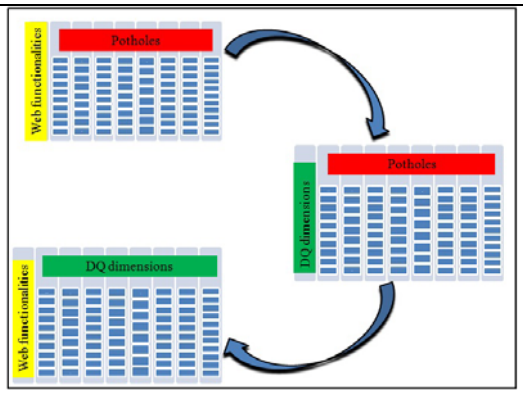

Figure 2. Sketch of the set of analysis carried out

As initial part of research, we carried out an analysis about which potholes or problems that normally appear in a IS (defined by Strong et al. in (Strong, Lee et al. 1997)) could be in a specific moment related or mixed with each one of Web functionalities, as result of this initial phase we got the next matrix of relation, it is showed in Table 3.

Table 3. Matrix of relation between Web functionalities and potholes identified

Potholes / Web functionalities	Multiple sources	Subjective production	Production errors	Too much information	Distributed systems	Nonumeric 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				√						

This Table 3 summarizes the process of analysis carried out; below we described each relation achieved:

Content Management

Multiple sources

To the Content Management, the existence of multiples processes or different sources which generate values of data, it 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, or their combination can lead to higher quality values.

Subjective production

The creation of values for the data in a subjective way or with connotations, it can cause problems when authorizing their inclusion or exclusion in the Web application. The management of all information content should be done through of generation of "rules", in order to reduce the variance in data produced by this subjectivity.

Production errors

When carrying out the creation of collections of values can cause errors in the production of values for the data. A proper management of all collection should cover the process from the generation of data (information source) until its correct storage into the application.

Too much information

Large amounts of stored data can induce the inclusion or exclusion of some reserved information. Also, having too much data stored may cause an exaggerated access time to them, resulting in poor management at the time extracting or summarizing this information.

Nonnumeric information

The management and retrieval of information in image format makes more difficult their managing and indexing. If it is possible to manage content, the information should be handled in a prescribed format and understandable to all processes of the organization (preferably "text"), so that their access, management and storage will be easy within the application.

Advanced analysis requirements

An analysis of content is required in an automatic way to provide to user an added value in data. If it is possible the management of information should carry out in an adequate format (*text*), with the purpose to make different types of analysis of the data in automated manner and in a reasonably short time. Make an information management non-text formats, e.g. images, is extremely complicated and it difficult the analysis and indexing.

Changing task needs

Is necessary accomplishing with the information needs adequate for the different consumers, users or clients, still when these necessities change through the time. A correct management of content showed in the application should provide updated data in the precise moment of perform a task by the users, besides of check periodically the different processes of information.

Security and privacy requirements

The data must be accessible to users; however, we must ensure the privacy and security of the information provided through a proper management, possibly by defining levels of access between users, and / or policies and procedures consistent of security.

Lack of computing resources

The lack of sufficient computing resources to manage content is likely to cause a lack of quality of data managed and stored in the application. Some of the most common problems being the lack of those resources, such as response times too big or data analysis results incomplete.

Process and Actions

Too much information

Large amounts of data can lead to user conflicts, creating miscommunication at the moment to execute a business process with the organization. Providing the right amount of information improves decision making when initiating or participating in a business process.

Changing task needs

It is essential providing adequate and updated data to users, meeting their information needs at the right time requiring it to participate in a business process. Provide outdated data can have a negative impact for the organization, occasioning thus loss of clients with which to start or continue a business process.

Search Capabilities

Too much information

Large amounts of data can do difficult to perform searches inside and outside the application, preventing access to information in a moderate time. Show too much information as a result of a search, may generate dissatisfaction among users of the application, since instead of helping users, they will need for further efforts to understand and discriminate that data do not interest them.

Distributed systems

The data stored in some distributed systems can cause problems with their access, besides an excessive time to locate them. Furthermore some data can be stored in other systems with a different format or representation code. All this results in a search of information with a low quality of data found.

Nonnumeric information

The recovery of data in non-numeric formats is very difficult to index, and consequently more difficult to access it, it resulting in inadequate information retrieval and/or mistaken for its use by users. This can cause user dissatisfaction and a lack of reliability data.

Security and privacy requirements

The users should access only to permissible information, ensuring privacy and security of certain reserved information. Conducting a search of information should be limited only to data of type "public" or "allowed", for the different types of application users.

Lack of computing resources

The lack of sufficient computer resources can cause a limited access to data. Being the main problems facing: the amount of information where the search will be done and the format in which the information is stored. All this leads to get erroneous and incomplete information at the moment to make a search by user.

Administration

Changing task needs

The Administrator of Web application should supply the necessary information to user, doing maintenance into Web application through the functionalities provided for it, in order to meet with the information need requested by each user at the right moment. Likewise, the business processes should be updated into the application, allowing to users perform their tasks with the application, using reliable data.

Security

Security and privacy requirements

Ensure the privacy and confidentiality of data, through different user access levels. If it is possible too, developing policies and procedures of security for each task in the application. The final aim is keeping data accessible only for authorized users.

Data Points and Integration

Too much information

A large amount of data volume can make difficult their access, specially if they are localized in various information systems (e.g. different databases, files with different format of stored, various kind of data warehouses, etc.), thus causing problems when displaying them to the user through a single access point.

Distributed systems

Problems can arise when accessing to data stored in multiple systems or servers in a distributed way, besides that the same data are likely to be stored in different formats or under a different encoding, it causing problems for their indexing when they are displayed to the user through a single access point.

Security and privacy requirements

Easy access to the data from various resources may cause conflict with security and confidentiality requirements of information. To eliminate or at least minimize these problems, we should generate verification requirements and access permissions for each user type in the application, thus protecting the security and confidentiality of data that will be displayed through a single access point.

Communication and Collaboration

Too much information

Large amounts of data may make it harder the efficient transmission of information, resulting in poor communication and collaboration between employees and users of the application, thereby hindering their cooperation to perform or verify a business process.

Distributed systems

Having data distributed across multiple systems and in different formats, can interfere with communication and coordination of employees and users. As in previous relation, a poor information management can generate conflicts between users to access it in a timely manner, damaging the collaboration between them. A typical example is a failed connection when we try to recover data from remote systems, causing delays in the information access.

Advanced analysis requirements

A lack of an automated content analysis, the data will be stored in formats, names and inconsistent definitions, interfering with communication and collaboration of employees and users. One of the most common problems in this case, it refers to the format in which the data are stored (type: int, char, String), the storing data in different format does not permit to carry out analysis of them in an automated way, damaging the communication among users at the moment to make business decisions.

Lack of computing resources

The scarcity of adequate computer equipment and lack of access to information by employees will prevent good communication and collaboration, reducing their productivity. A typical example is to have obsolete computer equipment, insufficient for performing application-specific processes.

Presentation

Changing task needs

It is necessary to provide and present data appropriately updated and in right form to the user, so that meets their specific information needs in all the functionalities available in the application.

Taxonomy

Distributed systems

The data are stored in heterogeneous systems with different values or formats, making it impossible a terminology commonly used in the different categories of the organization. This problem can lead to confusion in the semantics of the concepts used by users, when executing various business processes using the application.

Personalization

Changing task needs

The Web application must provide a work environment organized and configured for each user, according to their specific information needs in a timely manner. Personalization is a very important aspect, because it helps to maintain consistent data for each user.

Security and privacy requirements

It should provide the user the option to customize their work environment in the portal, respecting certain security requirements and privacy information. The user can configure its work environment, showing only the data to which he/she has access according to their role within the application. In addition, this will keep the data in a consistent and steady through an organized work environment.

Help Features

Too much information

Large amounts of stored data will require some help options for the user, in case to perform some specific functionality with a set of data. The help features also will be reflected when showing detailed information related to a specific process of the organization, it explaining how to use and manage certain data of the business.

Taking in account that the two models considered as standard (Wang and Strong 1996) versus international standard ISO/IEC 25012 (ISO-25012 2008), 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 4). 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.

Table 4. Comparative of DQ Dimensions

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

In the last part of analysis, the idea is 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 3. As beginning point, we take in account again the work presented by Strong et al. in (Strong, Lee et al. 1997), in which the authors classify with base in their model (Wang and Strong 1996), the DQ dimensions that affect to each one of potholes (see Table 5).

Table 5. DQ Dimensions that affect to each pothole

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	

Next, and taking into account that the standard ISO/IEC 25012 is more appropriate for the work that we are developing (after all, the Web application can be considered as Information Systems), keeping the meaning of dimension, but having in account the change of scope described in Table 4, we rewrite the Table 5 getting the matrix presented in Table 6, 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 (showed in Table 6 by the symbol "α").

In Table 6 is showed the final result containing the "*Model for the selection of DQ requirements according the functionalities to be implemented in a Web application - MOSCAF*", through this mapping of relationships between the DQ dimensions and the Web functionalities to be implemented into an application. 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 required functionality to be implemented in the application. In this way, we can see, for example, that the Web functionality "*Content Management*" will be related with the DQ dimensions of *Accuracy*, *Completeness*, *Consistency*, *Credibility*, *Currentness*, *Accessibility*, *Compliance*, *Confidentiality*, *Traceability*, *Understandability* and *Portability*. This would mean that if we wanted to

implement this functionality, it would be necessary to consider like a DQ requirement such DQ dimensions related.

Table 6. Model for the selection of DQ requirements according the functionalities to be implemented in a Web application - MOSCAF

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			

For a greater understanding of the model (MOSCAF), we describe each of the different relationships specified ("√", "α") in the next Tables 7 to 17.

	Table 7. Content Management				
Accuracy	*Completeness*	*Consistency*	*Credibility*	*Currentness*	*Accessibility*
The data managed for their inclusion in the application should represent a correct value according to the 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.	The data managed with the application 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 to its management in different level of confidentiality, it ensuring that data only are accessible and interpretable by authorized users.	As part of the management of data, some information should be provided about "when" and "who" published or distributed 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 8. Process and actions			
Completeness	*Currentness*	*Accesibility*	*Compliance*
All data must be complete so that users can carry out a business process with the organization.	Data must be current so that a user can initiate and participate in a business process.	The user can initiate and participate in a business process once the data is accessible for use.	Data must meet certain standards or conventions in force, according to the context in which the user will use data to participate in a business process.

Table 9. Search capabilities			
Completeness	*Consistency*	*Currentness*	*Accesibility*
The data obtained by performing a search should contain values for all their expected attributes.	The data obtained when performing a search should be consistent with other data in a context of use defined.	The data obtained when performing a search must be current to be used in a specific context.	The data should be accessible in the application when performing a search.
Compliance	*Confidentiality*	*Traceability*	*Availability*
The data obtained as a result of a search should be in accordance with the rules or standard established in business processes.	The user will get only the data available for him, according to his level or role in the application.	The system should provide information about the number of times you have been accessed and/or modified a specific data or data set.	The data should be available to be retrieved when running some specific search. This implies that the data should be accessible, that the access path is available and that the user has the appropriate permissions for reading.

Table 10. Administration			
Completeness	*Compliance*	*Efficiency*	*Precision*
The Administrator of the application must ensuring that the data managed in the application remain complete, after doing some activity of maintenance and/or tasks related with the application.	The data must adhere to standards or norms defined by the organization or the Application Administrator.	The Administrator of the application should ensure that data can be accessed with an acceptable level of functioning.	The Administrator of application should ensure that data are defined according to the precision or accuracy required for each specific task.
Understandability	*Portability*	*Recoverability*	
The Administrator should maintain the data in the language, symbol and appropriate unit for each functionality or task of the application.	The data may be installed or replaced by the Administrator during maintenance activities.	The data may be maintained and retrieved by the Administrator during maintenance activities, or in case of system failure.	

Table 11. Security			
Accessibility	*Compliance*	*Confidentiality*	*Traceability*
Data must be classified according to different levels of access, to be accessible to each type of user.	The data should be according to rules or conventions, and so maintain their security policy.	The data will be accessible only by authorized users according to their level of access.	Data should provide information on "*when*" and "*who*" used it.

Table 12. Data points and integration			
Completeness	*Consistency*	*Currentness*	*Accessibility*
All data must be complete for each of its attributes, regardless of the information source from which is obtained.	Data should be consistent no matter the information source from which it is obtained.	The data should be updated regardless of the source of information from which are obtained.	Data should always be accessible to the user, preferably as a single access point in the application.
Compliance	*Confidentiality*	*Availability*	

The data should be in accordance with rules or regulations, regardless of the information source from which are obtained.	The data should be accessible only by authorized users, regardless of the source from which are obtained.	The data should always be available to users at the moment they are needed.	

Table 13. Communication and collaboration			
Accuracy	*Completeness*	*Consistency*	*Credibility*
The data must represent a correct value in the same context by all users, improving with this the communication and performance of them.	Data should be complete in all their attributes, thereby helping the communication and collaboration when users perform their tasks.	The data should not be contradictory for users when executing an activity, it generating a good collaboration between them.	The data should be considered as true by all users, thus improving the collaboration between each them.
Currentness	*Accessibility*	*Compliance*	*Availability*
The data should be updated, thereby helping to make smart solutions and innovative business processes by users.	The data should be available at all times, it enabling to improve communication and collaboration of all users.	Data must conform to actual regulations in the organization, thus creating a better communication between distinct users.	Data should be available to authorized users, this helping them to discuss and resolve problems, besides making decisions.

Table 14. Presentation	
Completeness	*Compliance*
All data should be displayed to the user in full in each of the instances of presentation.	Data displayed to users should comply with the rules and regulations specified in the organization.

Table 15. Taxonomy			
Accuracy	*Consistency*	*Currentness*	Compliance
The data should be syntactically and semantically correct, thus improving the common terminology used in the organization.	Data must be consistently shown in each specific category of the organization, thus improving in a common terminology for application users.	Data should be of the correct period, it helping to improve the company's business in each of their specific categories.	The data must meet specific standards set by the organization, it contributing to a common terminology and improving the semantic and the rapid recognition among users.

Table 16. Personalization				
Completeness	*Accessibility*	*Compliance*	*Confidentiality*	*Availability*
All data should be viewed completely, it aiming to maintain an organized work environment for users. This helps to maintain a consistency in the application customization for each user of the application.	The data may be accessed and configured specifically for each user, thus collaborating to achieve a balance in the work environment of the application.	The data shall be in accordance with certain regulations established in the organization, and their applying in the customization and unique features required by each user of the application.	The data will be presented and configured only for each work environment of each user.	Data should be available to authorized users, this creates a work environment specially configured for each user.

Table 17. Help features				
Completeness	*Currentness*	*Accessibility*	*Compliance*	*Understandability*

All data displayed to user to assist him in the characteristics of the application must be complete in the specific context of use.	The data presented to the user as an aid to perform a specific task should be current.	The data should always be available to be exposed to users to help in any application-specific activity or organization.	The data displayed to help users should to keep existing regulations established by the organization.	The data displayed to the user as assistance of portal should be perfectly read and interpreted in the appropriate language.

Once described the "*Model for the selection of DQ requirements according the functionalities to be implemented in a Web application - MOSCAF*", then we show the guide of use.

Use Guide of Model - MOSCAF

The guide for the use of the MOSCAF model is based on the work of (Franch and Carvallo 2003). It encompasses five steps that are described to follows.

Step 1. UD. Understanding of Domain.

First, it is indispensable describe the application domain, all involved people must have full knowledge both Web functionalities as the DQ dimensions, i.e., they should know and understand the meaning of each one of them. The main input product in this step is to understand both the Web functionalities as the DQ dimensions, in order to generate as output product, a document with a full and detailed description of the domain or environment in which the application will be generated (*DUD*).

Step 2. DDQR. Determining the DQ requirements.

This step consists in selection of DQ dimensions chosen at the moment to develop each one of Web functionalities. In this moment will also be possible considering the selection of one specific DQ dimension, still when does not exists a defined relation between the dimension and the functionality. This could be from a comment or observation given by a person skilled in the application domain. As input product, we must have both the MOSCAF model as the description of domain application (*DUD*), and it getting like an output product a document with the DQ requirements needed (*DDQR*), i.e., the selection of DQ dimensions chosen to each functionality to be implemented.

Step 3. RDQD. Refining the definition of the DQ dimensions selected.

In this step is necessary carrying out the definition of the different DQ dimensions selected according to the context in which they are going to be implemented as part of the distinct Web functionalities. The objective is defining in a clear way the meaning of each DQ dimension according the application domain, so that all people being able to understand the meaning and they are agree with it. The input product of this step is a document with the DQ requirements (*DDQR*), it getting like an output product a refined document with the definition of each DQ dimension chosen (*DDDQR*).

Step 4. HDQD. Defining a hierarchy of DQ dimensions.

Since not all DQ dimensions chosen may be as important, in this step is possible to define a hierarchy, according to the importance degree of the various DQ dimensions that were previously selected, it considering the various comments from experts involved in the definition. The main input product is the document generated in the previous step (*DDDQR*), thus obtaining as a result a document specifying the hierarchy of these DQ dimensions (*DHDQD*).

Step 5. ERDQD. Enunciating the relationships between DQ dimensions.

Specifying the existing relationships (if necessary) between different DQ dimensions previously selected. This is for the purpose of discussing possible relations of collaboration (dependence) or detriment (collision) between DQ dimensions. For instance, if a dimension "*X*" is related to a dimension "*Y*" in a dependent manner, it implies that if the dimension "*X*" is growing in importance, the dimension "*Y*" will grow at the same way. On the other hand, if there is a detrimental relationship between two DQ dimensions, it would mean that implementing one of them in some way could affect the performance of the other one. The main input products for this step are the documents *DDDQR* and *DHDQD*; the output product is a document specifying the existing relationships between the DQ dimensions (*DERDQ*).

Conclusions

At present, Information Systems in general and particular Web applications are as important as organization itself. Data and information 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.

References

Ballou, D. P. and H. L. Pazer (2003). "Modeling Completeness versus Consistency Tradeoffs in Information Decision Contexts " IEEE Transactions on Knowledge and Data Engineering **15** (1): 240-243

Ballou, D. P., R. Y. Wang and H. Pazer (1998). "Modelling Information Manufacturing Systems to Determine Information Product Quality." Management Science 44(4): 462-484.

Batini, C. and M. Scannapieco (2006). Data Quality: Concepts, Methodologies and Techniques. Berlin, Springer-Verlag Berlin Heidelberg.

Bizer, C. and R. Cyganiak (2009). "Quality-driven information filtering using the WIQA policy framework." Web Semant. **7**(1): 1-10.

Caballero, I., C. Calero, M. Piattini and E. M. Verbo (2008). MMPRO: A Methodology Based on ISO/IEC 15939 to Draw Up Data Quality Measurement Processes. ICIQ.

Caballero, I., E. M. Verbo, C. Calero and M. Piattini (2007). A Data Quality Measurement Information Model based on ISO/IEC 15939. 12th International Conference on Information Quality, MIT, Cambridge, MA.

Cai, Y. and G. Shankaranarayanan (2007). "Managing data quality in inter-organisational data networks." International Journal of Information Quality **1**(3): 254 - 271.

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.

Collins, H. (2001). Corporate Portal Definitions and Features. New York, NY, USA, Amacom Books.

Franch, X. and J. P. Carvallo (2003). "Using Quality Models in Software Package Selection." IEEE Software. **20**(1): 34-41.

Ge, M. and M. Helfert (2007). A Review of Information Quality Research. International Conference on Information Quality, MIT, Cambridge, MA, USA.

Guerra-García, C., I. Caballero and M. Piattini (2011). "A Survey on How to Manage Specific Data Quality Requirements during Information System Development." Communications in Computer and Information Science **230**(Evaluation of Novel Approaches to Software Engineering): 16-30.

ISO-25012 (2008). "ISO/IEC 25012: Software Engineering-Software product Quality Requirements and Evaluation (SQuaRE)-Data Quality Model."

Kahn, B. K., D. M. Strong and R. Y. Wang (2002). "Information Quality Benchmarks: Product and Service Performance." Communications of the ACM **45**(4ve): 184-192.

Lee, Y. W., L. L. Pipino, J. D. Funk and R. Y. Wang (2006). Journey to Data Quality Cambridge, MA, USA, Massachussets Institute of Technology.

Lee, Y. W., D. M. Strong, B. K. Kahn and R. Y. Wang (2002). "AIMQ: a methodology for information quality assessment." Information & Management **40**(2): 133-146.

Lowe, D. and J. Eklund (2002). "Client Needs and the Design Process in Web Projects." Journal of Web Engineering, Rinton Press, US **Vol. 1, No. 1**.

Mahdavi, M., J. Shepherd and B. Benatallah (2004). A Collaborative Approach for Caching Dynamic Data in Portal Applications. Proceedings of the fifteenth conference on Australian database.

Moraga, M. A., C. Calero, G. Javier and P. Mario (2009). "Assessment of portlet quality: Collecting real experience." Comput. Stand. Interfaces **31**(2): 336-347.

Nicolás, J. and A. Toval (2009). "On the generation of requirements specifications from software engineering models: A systematic literature review." Inf. Softw. Technol. **51**(9): 1291-1307.

Oliveira, P., F. t. Rodrigues and P. Henriques (2005). A formal Definition of Data Quality Problems. Tenth International Conference on Information Quality (ICIQ'05), MIT, Cambridge, MA, USA.

Pipino, L., Y. Lee and R. Wang (2002). "Data Quality Assessment." Communications of the ACM **45**(4): 211-218.

Pressman, R. (2001). Software Engineering: a Practitioner's Approach. 5/e, McGraw-Hill.

Scannapieco, M. and L. Berti-Equille (2006). "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 **35**(2): 50-52.

Shankaranayanan, G. and Y. Cai (2005). A Web Services Application for the Data Quality Management in the B2B Networked Environment. 38th Hawaii International Conference on System Sciences (HICSS-38 2005), Big Island, HI, USA, IEEE Computer Society.

Strong, D., Y. Lee and R. Wang (1997). "Ten Potholes in the Road to Information Quality." IEEE Computer: 38-46.

Wang, R. and D. Strong (1996). "Beyond accuracy: What data quality means to data consumers." Journal of Management Information Systems; Armonk; Spring 1996 **12**(4): 5-33.

Wang, R. Y. (1998). "A Product Perspective on Total Data Quality Management." Communications of the ACM **41**(2): 58-65.

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 **42**(4): 575-589.