TOWARDS A WEB PORTAL DATA QUALITY MODEL

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

Coral Calero, Ismael Caballero, Mario Piattini
Universidad Castilla-La Mancha, Departamento de Informática, Paseo de la Universidad 4, Ciudad Real, España

Keywords: Data Quality, Information Quality, Web Portals.

Abstract: The technological advance and the internet have favoured the appearance of a great diversity of web applications, one of them are Web Portals. Through this, organizations develop their businesses in a more and more competitive environment. A decisive factor for this competitiveness is the assurance of data quality. In the last years, several research works on Web Data Quality have been developed. However, there is a lack of specific proposals for web portals data quality. In this paper, we will present a proposal for a model data quality for web portals based on some web data quality works.

1 INTRODUCTION

In the last years, a growing interest in the subject of Data Quality (DQ) or Information Quality (IQ) has been generated because of the increase of interconnectivity of data producers and data consumers mainly due to the development of the internet and web technologies. The DQ/IQ is often defined as "fitness for use", i.e., the ability of a data collection to meet user requirements (Strong, Lee et al. 1997; Cappiello, Francalanci et al. 2004). Data Quality is a multi-dimensional concept (Cappiello, Francalanci et al. 2004), and in the DQ/IQ literature several frameworks providing categories and dimensions as a way of facing DQ/IQ problems can be found. One of the most used and referred to is the framework proposed by (Strong, Lee et al. 1997) that states a classification based on the users perspective. Furthermore, methodologies that allow us to manage DQ/IQ within an organization have been proposed. Among them, we can mention: TDQM (Strong, Lee et al. 1997), AIMQ (Lee 2002) and TQdM (English 2001).

Research on DQ/IQ started in the context of information systems (Strong, Lee et al. 1997; Lee 2002) and it has been extended to contexts such as cooperative systems (Fugini, Mecella et al. 2002; Marchetti, Mecella et al. 2003; Winkler 2004), data warehouses (Bouzeghoub and Kedad 2001) and electronic commerce (Aboelmeged 2000; Katerattanakul and Siau 2001), among others. Due to the characteristics of web applications and their differences from the traditional information systems, the community of researchers has recently started to deal with the subject of DQ/IQ on the web with great interest (Gertz, Ozsu et al. 2004).

Related to web, some years ago appeared web portals. However, there are no works on DQ/IQ specifically developed for web portals. As the literature shows that DQ/IQ is very dependent on the context, we centre our work in the definition of a Data Quality Model for web portals. For doing it we use some works developed for other web but that can be applied to our particular context. For example, we use the work of (Yang 2005) where a quality framework for web portals is proposed including the data quality as part of it.

The aim of this paper is to show the result of a review about works in DQ/IQ on the web context and how we have used it for the development of the first version of a web portal DQ model. This paper is organized as follows: in section 2 we show some DQ/IQ issues associated with the web context, DQ/IQ frameworks and methods for evaluation and improvement of DQ/IQ on the web and some web quality characteristics identified from de studied works. In section 3 we present our proposal based on

Caro A., Calero C., Caballero I. and Piattini M. (2006).
TOWARDS A WEB PORTAL DATA QUALITY MODEL.
In *Proceedings of the Eighth International Conference on Enterprise Information Systems - SAIC*, pages 189-192
DOI: 10.5220/0002491501890192

the previously showed review. Finally, section 4 concludes with our remarks and future work.

2 DATA QUALITY ON THE WEB

In this section we review the most important aspects about web DQ/IQ that we founded on the literature.

2.1 Issues About Web Data Quality

Due to the own nature of the web, some concrete issues directly related to the DQ/IQ arise. For example (Eppler and Muenzenmayer 2002), states that typical problems of a web page affect the DQ/IQ as design problems that make it difficult users access to information. Other issue is the integration of structured and non-structured data (Finkelstein and Aiken 1999) and integration of data from different sources (Naumann and Rolker 2000; Angeles and MacKinnon 2004; Bouzeghoub and Peralta 2004; Gertz, Ozsu et al. 2004; Winkler 2004). In both cases the challenger is achieve integrated data that probably do not have the same level of DQ/IQ, and are delivered to the user with a fit level for use. Also, in this context is very important approach the DQ/IQ from users perspective (Angeles and MacKinnon 2004; Cappiello, Francalanci et al. 2004; Gertz, Ozsu et al. 2004) and to help them to understand data and their quality (Finkelstein and Aiken 1999; Gertz, Ozsu et al. 2004).
Finally, the demand for real-time services (Amirijoo, Hansson et al. 2003) and the dynamism on the web, particularly the dynamism with which data, applications and sources change (Pernici and Scannapieco 2002; Gertz, Ozsu et al. 2004), can affect quality.
The identification of these problems, and maybe others that we have not identified still, reveal us the necessity to create specific proposals to this context to approach the DQ/IQ, since otherwise they cannot be considered all their particularities.

2.2 Web Data Quality Frameworks

By using a DQ/IQ framework, organizations are able to define a model for data environment, to identify relevant quality attributes, to analyze attributes within both current and future contexts and to provide a guide to improve DQ/IQ. In the literature, we have found some proposals oriented to DQ/IQ on the web. Among them, we can highlight the ones proposed in (Katerattanakul and Siau 1999; Fugini, Mecella et al. 2002; Pernici and Scannapieco 2002; Eppler, Algesheimer et al. 2003; Graefe 2003; Bouzeghoub and Peralta 2004; Gertz, Ozsu et al. 2004; Melkas 2004; Moustakis, Litos et al. 2004). This summary of the frameworks that we have studied, gives us a global idea of the diversity of focuses that can be considered in the analysis of the DQ/IQ in the web. Through them we can identify the outstanding characteristics of quality to the data in this context.
Once is defined a framework it is possible to approach the quality of the data from the evaluation perspectives and it improves. From the above-mentioned frameworks we can find many methods of DQ evaluation and improvement. These methods can be used by the organizations to help the users and IT managers to capture and analyze the state of DQ in a web application.

2.3 Web Data Quality Characteristic

From de studied works, we have identified 100 data quality characteristic. The most considerate are: Accuracy, in 60% of the works; Completeness, in 50% of the works and Timeliness, in 40% of the works. Concise, Consistent, Currency, Interpretability, Relevance, Secure, in 30% of the studies.
As a conclusion of all these antecedents we think that it is necessary a general quality framework for data on the web that include all the possible characteristics related to them. But not only the quality characteristics are necessary for defining the general DQ/IQ framework, also the problems and different perspectives about the data of the web would be also taken in to account. In particular, the data consumer's expectative.

3 THE DQ/IQ WEB PORTAL FRAMEWORK

There is currently no established conceptual foundation for developing and measuring DQ/IQ of Web applications in general, and Web Portal in particular. Our aim is to create a DQ/IQ framework for Web Portals, considering the next basic considerations:
1. We started with de DQ attributes proposed in the literature. These are the ones identified from the study mentioned previously.
2. We use the data consumer's perspective. We consider the work about quality of data on Internet

of (Redman 2000). He proposes the next categories for the data consumer's expectative: Privacy, Content, Quality of values, Presentation, Improvement and Commitment.
3. We considered the 11 basic functionalities for a Web portal proposed by (Collins 2001): Data Points and Integration, Taxonomy, Search Capabilities, Help Features, Content Management, Process and Action, Collaboration and Communication, Personalization, Presentation, Administration, and Security.
With these elements we first relate the web portal functionalities with data consumer's perspective about data quality. As a result we obtain a matrix (see figure 1). We use the each cell for filling in it those Web DQ/IQ attributes applicable (functionality, expectation).

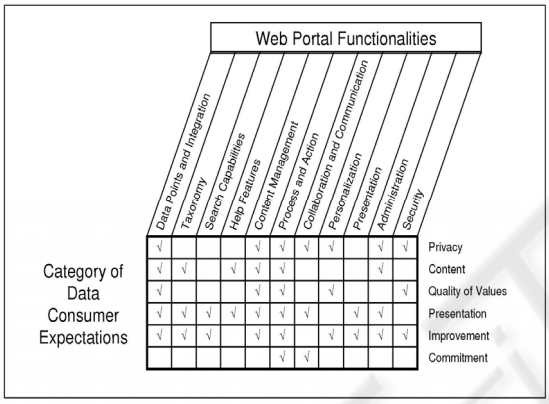

Figure 1: Consumer expectations vs. web portals functionality.

As a matter of example the functionality "Data Point and Integration" connects to: "Privacy" (users expect privacy about the sources and their integration), "Content" (the consumers need a description of portal areas covered, uses of the published data, etc.), "Quality values" (data points correct, current and complete), "Presentation" (formats, language, and others are very important for a easy interpretation) and "Improvement" (user wants to participate with their opinions in the portal improvements knowing the result of applying them). Then we determined for each relation DQ/IQ attributes cab be applied. Table 1 shows the classification done for the "Data Point and Integration". Due to space restrictions in this example we consider only three DQ/IQ attributes.

Table 1: DQ for Data Point and Integration.

Expectation	Dimensions
Privacy	Confidentiality, Maintainable, Verifiability
Content	Accuracy, Concise, Consistent
Quality of Values	Accuracy, Consistent, Interpretability
Presentation	Consistent, Amount of data, Clear
Improvement	Accessibility, Traceability, Clear

However, it will be necessary to validate this assignation. For doping it, we plan to work with user's portals through surveys and questionnaires. Once the validation will be finished, we will reorganize the attributes obtaining the final version of the DQ/IQ web portal matrix.

4 CONCLUSIONS AND FUTURE WORK

The great majority of works found in the literature show that DQ/IQ is very dependent on the context. The increase of the interest in the development of web applications has implied either the appearance of new proposals of frameworks, methodologies and evaluation methods of DQ/IQ or the adaptation of the already-existing ones from other contexts such as Traditional Information Systems, Cooperative Information Systems and Data Warehouses. In this paper, we have described different issues associated with DQ/IQ on the web; many of them are still of great interest for the community of researchers. There are different proposals related to DQ/IQ on the web either to establish quality criteria and/or dimensions relevant to this context or to establish strategies to evaluate and measure these criteria.
We have used the study done for the web DQ/IQ as basis for developing a DQ/IQ model for web portals. In this paper we have showed the first version of this model. The model is composed by two dimensions: the consumer expectations and the web portals functionality. As an example we have classified some web quality characteristics for one of the functionalities and for all the user expectations. As future work, in one hand we must fill in all the cells of the defined matrix with quality characteristics and, on the other hand, to validate the classification in order to prove its correctness.

ACKNOLEDGEMENTS

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

REFERENCES

Aboelmeged, M. (2000). A Soft System Perspective on Information Quality in Electronic Commerce. Proceeding of the Fifth ICIQ.

Amirijoo, M., J. Hansson, et al. (2003). Specification and Management of QoS in Imprecise Real-Time Databases. Proceeding of the Seventh International Database Engineering and Applications Symposium.

Angeles, P. and L. MacKinnon (2004). Detection and Resolution of Data Inconsistences, and Data Integration using Data Quality Criteria. QUATIC'2004.

Bouzeghoub, M. and Z. Kedad (2001). Quality in Data Warehousing. Information and Database Quality.

M.Piattini, C. Calero and M. Genero, Kluwer Academic Publishers.

Bouzeghoub, M. and V. Peralta (2004). A Framework for Analysis of data Freshness. IQIS2004, Paris, France, ACM.

Cappiello, C., C. Francalanci, et al. (2004). Data quality assessment from the user´s perspective. IQIS2004, Paris, Francia, ACM.

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

English, L. (2001). Total Quality data Management (TQdM). Information and Database Quality. M. Piattini, C. Calero and M. Genero, Kluwer Academic Publishers.

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

Eppler, M. and P. Muenzenmayer (2002). Measuring Information Quality in the Web Context: A Survey of State-of-the-Art Instruments and an Aplication Methodology. Proceeding of the Seventh ICIQ.

Finkelstein, C. and P. Aiken (1999). XML and Corporate Portals, Wilshire Conferences.

Fugini, M., M. Mecella, et al. (2002). "Data Quality in Cooperative Web Information Systems."

Gertz, M., T. Ozsu, et al. (2004). "Report on the Dagstuhl Seminar "Data Quality on the Web"." SIGMOD Record vol. 33, Nº 1: 127-132.

Graefe, G. (2003). Incredible Information on the Internet: Biased Information Provision and a Lack of Credibility as a Cause of Insufficient Information Quality. Proc. of the Eighth ICIQ.

Katerattanakul, P. and K. Siau (1999). Measuring Information Quality of Web Sites: Development of an Instrument. Proceeding of the 20th ICIS.

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

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

Marchetti, C., M. Mecella, et al. (2003). Enabling Data Quality Notification in Cooperative Information Systems through a Web-service based Architecture. Proceeding of the Fourth WISE.

Melkas, H. (2004). Analyzing Information Quality in Virtual service Networks with Qualitative Interview Data. Proceeding of the Ninth ICIQ.

Moustakis, V., C. Litos, et al. (2004). Website Quality Assesment Criteria. Proceeding of the Ninth ICIQ.

Naumann, F. and C. Rolker (2000). Assesment Methods for Information Quality Criteria. Proceeding of the Fifth ICIQ.

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

Redman, T. (2000). "Data Quality Guide field."

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

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

Yang, Z. a. C., S. and Zhou, Z. and Zhou, N. (2005). "Development and validation of an instrument to measure user perceived service quality of information presenting Web portals." Information and Management. Elsevier Science 42: 575-589.