An Applicable Data Quality Model for Web Portal Data Consumers

Coral Calero • Angélica Caro • Mario Piattini

Received: 16 October 2007 / Revised: 10 April 2008 / Accepted: 30 June 2008 / Published online: 30 July 2008 © Springer Science + Business Media, LLC 2008

Abstract Web portals have emerged as an important means by which to access data on the worldwide. The people that use these applications need to ensure that the data recovered is suitable for the task at hand. That is, they need to know the level of quality of the data obtained. This paper introduces the PoDQA (Portal Data Quality Assessment) tool which implements PDQM, a Portal Data Quality Model, which is centered upon the data consumer perspective. Thus, the measurement of data quality is carried out by using the point of view of data consumers. Our work aims to fill the lack of specific proposals for the DQ evaluation in Web portals and tools that put these proposals into practice. The paper illustrate how PoDQA tool works and how it can be used by data consumers in order to, for example, discover the data quality of a specific web portal. PoDQA also suggests several corrective maintenance activities for users who are interested in the improvement of the data quality of their Web portals.

Keywords web portals • data quality • data quality model • data quality assessment • users • developers

1 Introduction

A Web portal is a site that aggregates information from multiple sources on the Web and organizes this material in an easy user-friendly manner [36]. Over the past decade the number of organizations that provide Web portals has grown dramatically. These
organizations provide portals that complement, substitute or extend existing services to their client base [37]. Numerous users worldwide use Web portals to obtain information for their work and to help with decision making. These users, or data consumers, need to ensure that the data obtained are appropriate for their needs. Likewise, the organizations that provide Web portals need to offer data that meet user requirements, thus helping these users to achieve their goals. Therefore data quality represents a common interest between data consumers and portal providers.

Data (or Information) Quality (DQ) is often defined as “fitness for use”, i.e., the ability of a collection of data to meet user requirements [3, 33]. This definition and the current view of assessing DQ, involve understanding DQ from the users’ point of view [18]. In recent years, several research projects have been conducted on the topic of Web Data Quality. However, there is still a lack of specific proposals for the DQ in Web portals which consider the data consumer’s point of view and tools that put these proposals into practice.

In this work we introduce the PoDQA tool, whose aim is to assess the DQ in Web portals. The PoDQA development is part of a greater project on data quality, in which the research focus is to work towards the generation of a generic, adequate, flexible and complete data quality model for Web portals. This project has been introduced in [4] and [5], where the development of PDQM (a data quality model for Web portals) is described. PDQM is centered on the point of view of data consumers and uses a probabilistic approach (based on Bayesian networks) for data quality evaluation.

PoDQA assesses the DQ of a Web portal by using the PDQM model as a basis. The first version of this tool is available in http://podaq.webportalquality.com. This version implements the DQ evaluation for a subpart of PDQM (Representational DQ) as will be explained in this paper.

The organization of this paper is as follows. Section 2 describes the background of our work. Section 3 introduces the PDQM model, emphasizing its approach towards quantifying the DQ. The prototype of PoDQA is described in Section 4. Finally, Section 5 shows our conclusions.

## 2 Background

DQ is commonly thought of as a multi-dimensional concept [3, 28, 34]. The literature dealing with DQ provides different classifications of the DQ attributes, depending upon the perspective of the authors and the context tackled. On the other hand, in order to assess DQ, a growing tendency towards considering the users’ point of view exists. In fact, the most common definition of DQ is data that are “fit-for-use”, i.e., the ability of a collection of data to meet user requirements [3, 33, 34]. This definition suggests two important ideas. First, that DQ cannot be assessed independently of the people who use the data. And secondly, that DQ is relative and subjective: different users may have diverse opinions about the quality level of the same data.

In [12] DQ assessment is defined as the process of assigning numerical and categorical values (quality scores) to quality criteria in a given data setting. They emphasize that DQ assessment in the Web is a difficult task and that “well-founded and practical approaches to assess or even guarantee a required degree of the quality of data are still missing”.

Research on DQ began in the context of information systems [19, 34] and has been extended to contexts such as cooperative systems, data warehouses or e-commerce, amongst others. Due to the particular characteristics of Web applications and their differences from traditional information systems [29], the research community has begun to deal with the subject of DQ on the Web [12]. In fact, the particular nature of the Web has forced the necessity to pay attention to a series of typical issues in this context which may affect or influence DQ. Among these we might mention: Typical problems of a Web page (an-updated data, publication of inconsistent data, obsolete links, etc.) [9], Integration of structured and non-structured data [10], Integration of data from different sources [1, 2, 12, 24, 35, 38] and Dynamic nature of the Web [12, 27].

Because of the particularities of this context and the necessity of assessing the DQ in the context of its generation [32], in recent years frameworks, models and DQ attributes to deal with DQ in different domains in the Web context have been proposed. Among them, we can highlight those presented in Table 1.

It is important to highlight that many of these are based on well-founded data quality frameworks defined for other fields. One of the most frequently used frameworks is that proposed in the context of information systems by Wang and Strong [34]. This framework establishes four DQ categories in which 15 DQ dimensions are classified (see Table 2).

After studying the Web DQ literature, we have detected a lack of specific proposals of DQ models and/or frameworks for Web portals. Some proposals which tackle the Web portal context consider DQ as a part of a more general model [22, 37]. Consequently we have developed a DQ model for Web portals, named PDQM, which is centered upon the point of view of data consumers. In order to begin our work, we collected a set of DQ attributes proposed for the Web context, because although Web portals can be considered as an independent category of Web applications [13], a given Web application may belong to more than one category [13]. Thus, the DQ attributes proposed for other Web applications may be useful in the evaluation of the DQ in Web portals. With this strategy in mind, our aim was to take advantage of previous works and to start by using an ample set of DQ attributes as a reference.

### Table 1. Web DQ frameworks.

<table>
<thead>
<tr>
<th>Author</th>
<th>Domain</th>
<th>Model/Framework structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kaarntanen &amp; Siau 1999 [17]</td>
<td>Personal web sites</td>
<td>4 categories and 7 constructs</td>
</tr>
<tr>
<td>Naarmann &amp; Rolker 2000 [34]</td>
<td>Data integration</td>
<td>3 classes and 22 quality criteria</td>
</tr>
<tr>
<td>Kaarntanen &amp; Siau 2001 [16]</td>
<td>e-commerce</td>
<td>4 categories associated with 3 categories of data user requirements</td>
</tr>
<tr>
<td>Pernici &amp; Scannapieco 2002 [27]</td>
<td>Web information systems (data evolution)</td>
<td>4 categories, 7 activities of DQ design and architecture to DQ management</td>
</tr>
<tr>
<td>Graefe 2005 [14]</td>
<td>Decision making</td>
<td>8 dimensions and 12 aspects related to (providers/consumers)</td>
</tr>
<tr>
<td>Eppler et al. 2003 [8]</td>
<td>Web sites</td>
<td>4 dimensions and 16 attributes</td>
</tr>
<tr>
<td>Goret et al. 2004 [12]</td>
<td>DQ on the web</td>
<td>5 dimensions</td>
</tr>
<tr>
<td>Moukas et al. 2004 [23]</td>
<td>Web sites</td>
<td>3 categories and 10 sub-categories</td>
</tr>
<tr>
<td>Melkas 2004 [21]</td>
<td>Organizational networks</td>
<td>6 stages of DQ analysis with several dimensions associated with each one</td>
</tr>
<tr>
<td>Yong et al. 2004 [17]</td>
<td>Web information portals</td>
<td>2 dimensions and 4 attributes</td>
</tr>
</tbody>
</table>
The diagram shows the relationship between different components and their interactions. It seems to be a conceptual model for understanding and analyzing a complex system.

The text on the page is a list of items, possibly headings or topics, organized in a way that suggests they are related to the diagram. The content is too fragmented to provide a coherent translation, but it appears to be discussing various elements and their connections within a larger framework.
- *Flexible.* It must be applicable to different situations. For example, in different Web portal domains, in processes where the model can be used in a partial or complete way or in processes where different kinds of data consumers can be considered. To do this, the structure must support the assignment of different weights to the attributes.

- *Complete.* The structure must allow the representation of all the relationships between the attributes, e.g., an attribute may simultaneously affect several other attributes. In hierarchical models for example, attributes from the same level cannot be related and an attribute cannot affect more than one of the attributes in the upper level.

As a result of the operationalization of the PDQM we have obtained a BN (see Figure 2) which organizes the 33 DQ attributes into four network fragments. A network fragment is a set of related random variables that can be constructed and reasoned on separately from other fragments [25].

In order to create the network fragments, we have used the conceptual DQ framework developed in [33, 34], see Table 2, as a criterion with which to organize the DQ attributes of PDQM. However, in our work we have renamed and redefined the Accessibility DQ category (calling it the Operational DQ category). The idea was to consider aspects which are typical of this context such as personalization, collaboration, etc. Using the definitions of each category and each DQ attribute as a base, we thus classified all the DQ attributes of PDQM into these four categories. After this, relationships of influence between the attributes were established. These relationships were established by using the DQ categories and the DQ attribute definitions, together with our perceptions and experience as a base. Our aim was to establish which DQ attribute in a category has direct influence over other attributes in the same category. As an example, Table 4 shows the relationships defined for the Representational DQ category. These relations were later confirmed in a validation process of the BN. More details about the generation of the BN for PDQM can be found in [9]. The BN generated for PDQM is shown in Figure 2.

Taking advantage of the possibility of working separately with each fragment of the BN, we decided to start with the Representational DQ. To do this, and in order to complete their operationalization, the following activities were developed:

- Two artificial nodes were created to simplify the fragment network and to reduce the number of parents for each node (node Representation and Volume of Data in

Figure 2). The aim was to reduce the combinatorial explosion in the following step during the preparation of the probability tables.

- Indicators or quantifiable variables, were defined for each entry node in the fragment (indicators LCSR, LCSR, LD, LD, LC, LO, and LI in Figure 3). The definition of these indicators will be explained in greater detail in the following subsection.

A probability table was defined for each intermediate node in the fragment (Figure 3 shows the probability table for the Consistent Representation, Volume of Data and Attractiveness nodes). These tables are defined by experts in Web portals. The experts were a group of researchers on DQ and users of Web portals. The procedure used was the following. The experts were presented with a proposal for the probability tables in which the probabilities associated with the various given combinations of parents of each node were already given. The experts analysed and adjusted them in accordance with their knowledge. Case studies then took place in which a group of Web portal users were asked to evaluate a group of portals, after which the users' evaluation was compared with that of the model, and the tables were adjusted in order to obtain a greater coincidence between the users' evaluations and those of the model.

The approach used allows us to represent dependencies among DQ attributes in the form of a BN in which the probability tables of the intermediate nodes are defined in order to capture the characteristics of a specific domain of a web portal. That is, the model can be adjusted, by changing the probability tables, according to the portal domain to be evaluated (for example to the bank domain or the governmental domain).
show how PQDs can aid the DPGF local to obtain the initial or final weather products evaluated in each domain. Figure 3 uses a graph to represent how the information in the model is compiled into the DPGF. The values in the graph are asymptotic, or the final product is considered the information in the DPGF. The DPGF contains the current state of the model and its internal state. The DPGF is represented as a representation of the model. The DPGF can be used to evaluate the DPGF's performance and its ability to provide current weather information.
The FODCA tool is a public tool which is established at the Ministry of Housing and Urbanization.

4.2. Conditions in the implementation of the assessment process

The conditions in the implementation of the assessment process are influenced by a variety of factors. It is necessary to know the strategy of wards and the integrated approach to deal with the complex nature of urban and rural areas. These conditions are influenced by the interaction of various factors, such as land use, socioeconomic conditions, and physical characteristics. In addition, the conditions in the implementation of the assessment process are influenced by the interaction of various factors, such as land use, socioeconomic conditions, and physical characteristics. In addition, the conditions in the implementation of the assessment process are influenced by the interaction of various factors, such as land use, socioeconomic conditions, and physical characteristics. In addition, the conditions in the implementation of the assessment process are influenced by the interaction of various factors, such as land use, socioeconomic conditions, and physical characteristics. In addition, the conditions in the implementation of the assessment process are influenced by the interaction of various factors, such as land use, socioeconomic conditions, and physical characteristics.
Figure 7: DO matrix of a group of web portals

Table 7: Value for the variables for the web portal environment

The table shows the correspondence between the variables and the distribution of the criteria. The table includes information on the importance of the criteria and the frequency of their occurrence. This information is used to assess the performance of the web portals. The table is divided into two sections: the left section shows the criteria and the right section shows the frequency of their occurrence.

Figure 8: new DO assessment for web portal

The graph shows the results of the DO assessment for the web portal. The x-axis represents the criteria, and the y-axis represents the frequency of their occurrence. The graph includes a legend that explains the colors used to represent the different categories of occurrence (e.g., high, medium, low). The graph is generated using a statistical software tool. The tool is used to analyze the data and generate the graph. The tool includes features for data manipulation, analysis, and visualization. The tool is user-friendly and includes a graphical user interface (GUI) that allows users to interact with the data.
5 Conclusions

By means of the previous analysis, it is possible to draw some important conclusions about the role of the PDD in terms of its potential to influence decision-makers. The PDD is a powerful tool for communication, education, and awareness-raising, and can be used in a variety of contexts to promote sustainable development. However, it is also important to recognize that the PDD is not a panacea and that it must be used in conjunction with other forms of communication and education to achieve its full potential. In conclusion, the PDD is a valuable tool for promoting sustainability, but it must be used in a responsible and strategic manner to ensure that it achieves its full potential.