www.ees.uni.opole.pl
ISSN paper version 1642-2597
ISSN electronic version 2081-8319
Economic and Environmental Studies
Vol. 14, No. 4 (32/2014), 361-373, Dec. 2014

A conceptual model for data quality management in a data system on the World Wide Web

Adam CZERWIŃSKI
Opole University, Poland

Abstract: The article presents a conceptual model for data quality management treated as the usability or the compliance of the data product with its specification. The proposed model refers to the well-known TDQM model of Wang based on the Deming's quality improvement cycle. However, the TDQM model does not take into account the impact of the Internet environment on the quality of the data provided by the systems on the Web. The author's model presented in this article takes into account the impact of the Internet on all aspects resulting from data functions in society and organizations. Therefore, it takes into consideration the aspect of promoting data quality management processes, the communication aspect and the aspect of enrichment of individual and collective knowledge. The model also takes into account the fact that the impact of the known properties of the Internet (defined with the acronym MEDIA for example) refers primarily to the contextual quality characteristics of the data on the Web and, only to a small degree, it concerns the internal quality of information pieces described by such features as accuracy, consistency, complexity and precision.

Keywords: data quality, management model, data system, Web service

1. Introduction

There are many reasons, which are strong incentives for development of the data quality management in data systems, and especially of those located on the Web, as the Internet with the World Wide Web is currently a global data system. This also applies to the data resources on the environment, which are the basis for use, protection and development of the environment. Firstly, we can notice a huge increase in the amount of the data collected, processed and made available on Websites. Secondly, the increase of the data amount on the Websites is extremely

Correspondence Address: Adam Czerwiński, Department of Information Economics, Faculty of Economics, Opole University, Ozimska 46a, 45-058 Opole, Poland. Tel. +48 (77) 40 16 862. E-mail: adam.czerwinski@uni.opole.pl.

rapid, which is additionally fueled by globalization processes in the modern economy. In the "global village", in which the differences in time and space are no longer relevant, data can be acquired quickly and easily, but it is of a very poor quality very often. Thirdly, the increase in importance of data and knowledge resources, as production factors within the information economy, known as the knowledge-based economy (KBE), is more and more visible (Skrzypek 2008: 207-214; Czerwiński, 2011: 192-195). It is in the KBE that data as a factor creating knowledge has become the next fundamental economic resource and production factor beside land, labor, capital and entrepreneurship. Fourthly, one may also notice the increase in the importance of data resources for the level of the enterprise competitive potential and the increasing role of the corporate data activities for achieving its competitive edge in the market. As a consequence, it leads to increase of the importance of the so-called data competition as part of traditional industrial economy and post-industrial economy. Fifthly, the role of data as a basis for cooperation and interactions amongst enterprises in supply chains/networks in the digital economy (virtual economy) is growing. Therefore, the data of poor quality, such as inadequate, outdated, inaccurate or incomplete data on the Web is use-less, since it cannot be used to make current business decisions, may not be included in the data resources of various stakeholders creating knowledge structures, may not be used for consumption, and it indeed may mislead. Most of the latest studies related to the assessment of the quality of the data resources are focused on making an overall evaluation of the functionality of web services, i.e. on the assessment of the quality of performance of such services (Czerwiński and Krzesaj, 2014: 82-97). The assessment of the quality of the data provided by web services is treated fragmentary then and omits the impact of the Internet properties on this quality.

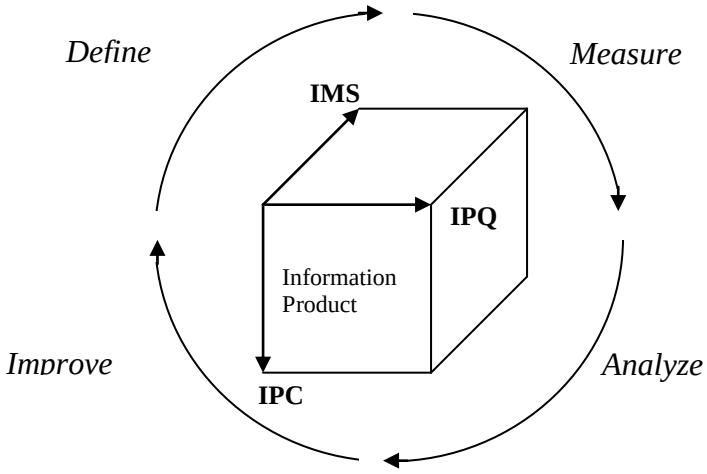
The purpose of this article is to word and to present an original conceptual model for the data quality management in a data system working on the Web.

It is in this article that the data quality management models known from literature are presented. The author's data quality management model in information systems on the Web is described against their background. It is shown how the properties of the network environment affect the quality of the data collected, processed and made available on websites.

2. Data quality management models

The Total Data Quality Management model was presented by R.Y. Wang (1998: 60-65). He has developed the Deming's cycle known from literature and used practical experience in this field. It was in the TDQM model that Wang proposed the data quality management cycle consisting of four consecutive stages: defining, measuring, analyzing and improving data quality: Figure 1.

Figure 1. The TDQM model by R.Y. Wang



Legend:

IPC: Information Product Characteristics
IPQ: Information Product Quality
IMS: Information Manufacturing System

Source: (Wang, 1998: 60).

The TDQM methodology consists of the following four stages (Wang, 1998: 61-65):

1. Defining the data product. At this stage, three tasks are solved:

a) Defining characteristics of the data product. This task is executed at two levels: higher and lower. At the higher, the functionality of the product for the user is interactively defined. At the lower level, the data product structure is defined: its basic units and components as well as their relationships.

b) Defining requirements for the data product quality. This task is carried out from the perspective of suppliers, manufacturers, users and managers of the product by taking into account the relevant quality criteria,
c) Defining the data system generating the product. This task is carried out with a system project, which takes into account the necessary data units and their sources, customers, data streams, information processes and the data bases being applied. While designing data streams, the quality requirements defined at the previous stage are taken into consideration.

2. Measuring the product. It is at this stage that the measures specific for each quality criteria are determined and developed.
3. Analyzing the data product. As part of this stage, it is necessary to examine the main reasons for ongoing problems with the data quality. For those quality characteristics, which have been considered to be too low, the statistical methods for control of processes, recognition of patterns and the Pareto charts analysis are applied.
4. Improving the data product. At this stage, the key areas for improvements should be identified, such as: 1) Matching the data and labor flow to the corresponding data system producing the product, 2) Matching the key characteristics of the data product to business requirements.

A very similar approach to the data quality management is suggested by L. English (2003). The author's method is known as TIQM (Total Information Quality Management) and is composed of six stages:

1. Assessment of the data definition and architecture: defining measures for the quality definition and data models.
2. Assessment of the data quality: defining data quality measures.
3. Assessment of costs and risks related to the use of erroneous data.
4. Designing improvement of the data quality: designing data revisions, reorganization of data processes.
5. Strengthening effects of the improvement in the data quality.
6. Creating the high-quality data environment: implementation of quality management principles; this stage proceeds parallel to other stages and does not have a specific start or end.

Another model for data quality management was proposed by J. Ruževičius and A. Gedminaitė (2007: 22). The authors generalized and combined the TDQM model of Wang with the model of Al Hakim (2004: 170-182) and suggested a supplemented model. It is in this supplemented data quality management model that four new elements, which shape this quality, are added to resources of an organization: information and communication technologies, knowledge and experience of employees dealing with data, instructions and the speed of data creation.

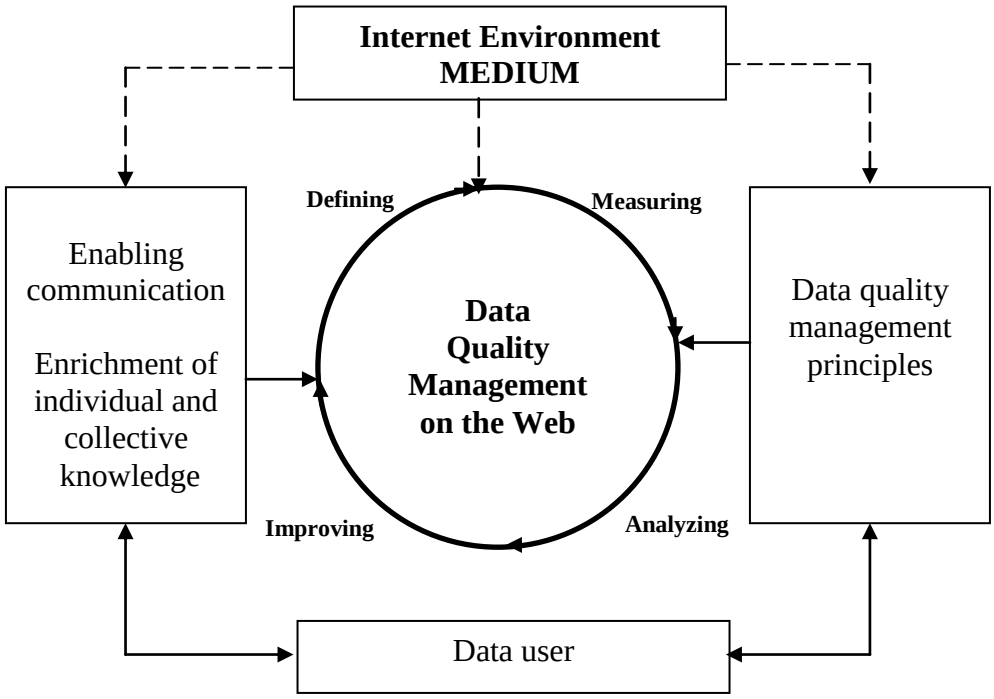
Particularly interesting in this model is taking into consideration information and communication technologies. Modern information and communication technologies greatly facilitate production and distribution of high-quality data. It is directly related to the use of digital data, which are far more resistant to errors at all stages of the production and distribution of data products.

However, none of the presented models takes into account the impact of the Internet environment on the quality of the data provided by the systems on the Web. The author's model presented in this article takes into consideration the impact of the Internet on all aspects resulting from data functions in society and organizations. Hence, it includes the fundamental aspect for supporting management processes in an organization - the communication aspect (enabling communication and establishing relationships amongst participants and organizations) - as well as the aspect of enrichment of individual and collective knowledge.

3. The author's model for data quality management on the Internet

It is in the Figure 2 that the author's model for data quality management in a data system on the World Wide Web is presented.

Figure 2. A conceptual model for data quality management in a data system on the World Wide Web



Source: Author's own elaboration

The Internet environment, in which Web data systems are embedded (Web services, portals, web sites, etc.), is characterized by a set of features that are not applicable to traditional communication media. These advantages have been defined by Ernst & Young in the form of the so-called "the concept of a MEDIUM" (Instytut Logistyki i Magazynowania EAN Polska, 2003: 50-52). It is an acronym created by the first letters of the key Internet features:

- **M**ass
- **E**conomical
- **D**irect
- **I**nteractive
- **U**ltrafast
- **M**easurable

It is in the presented model that the listed features affect the data quality management in the Web data system directly or indirectly (their influence is indicated in the Figure 2 by dotted

arrows). The direct impact concerns individual stages of the data quality management being considered in the TDQM model of Wang.

The impact of the Internet occurs at the beginning of *the stage of defining requirements* for quality of the data product with taking into account data quality characteristics. B. Stvilia et al. (2007: 1723-1724) identified the three following categories of data quality features: internal quality, relational (or contextual) and reputational. The internal data quality includes the features, which can be evaluated by measuring internal data attributes against the benchmark quality in the given culture. Generally speaking, the internal data quality features remain unchanged in time and poorly depend on the context. Therefore, the impact of the Internet on the internal data quality features such as accuracy, consistency, complexity, age of information, naturalness and precision is very low. On the other hand, the relational (or context) data quality features for a data unit (e.g. a message) are not permanent. The context of the use refers here to the context of functioning of the entire data system and can be changed in time and space, in particular under the influence of the Internet environment properties described above. For example, the data and information safety is a feature particularly vulnerable to the effects of various threats from malicious software (viruses and worms, Trojan horses, logic bombs, spyware, etc.) due to the speed and the large-scale of spreading of this type of software and the associated costs related to securing data (Czerwiński, 2005: 303-309). Therefore, while defining requirements for quality of data units in a Web data system, the contextual quality features should be taken into account above all, such as: availability, naturalness, redundancy, accuracy, precision/completeness, safety, semantic compliance, structural compliance, verifiability, and variability/impermanence. It is in the example presented above that it is necessary to determine, to which types of risk the data is exposed and which methods for the protection of the data should be applied in order to guarantee a certain security level.

The reputational data quality is a category of features, which measure the position of the data unit in the cultural or action structure. Therefore, it is very often determined by the origin of the data, which decides on the opinion about the data or the given renown. Hence, the reputation level of the information object within the given community or culture also depends on the key Internet properties.

While defining quality requirements for a data product, limitations associated with available resources have to be taken into account. It is in the model mentioned above that the

time occurs, which is a necessary resource for ensuring the required data quality. J. Ruževičius and A. Gedminaitė (2007) suggest to use a more precise term in place of the time, which means the speed of the data creation (generating). According to them, the faster a data product is produced, the more likely is that it will be of generally poor quality. Despite it, data can be generated in real time in the Internet environment (i.e. without any delays, immediately) and a method to generate it does not need to deteriorate some of its qualitative characteristics. Such a method is syndication, which is a special way to collect and combine original content contained on several web pages in one package, within a special format of network data called RSS (Really Simple Syndication). A data agent can add own data to the newest contents combined in this way (e.g., advertisements, or those, which may arouse interest of a recipient according to him) and redistributes them on-line (Czerwiński and Krzesaj, 2014: 86). It should be noted that the syndication does not affect the internal characteristics of the data quality. However, it can be expected that due to the addition of the agent's data, there will be a change of the contextual data quality features spread under this service. Additionally, it may turn out that the change will be beneficial to the users of the RSS channels: for example the contents will be more relevant for the users and thus will better meet information needs. On the other hand, some of these additional data may be treated as redundant (especially advertisements) and the assessment of the data quality will decrease in this respect.

We also have to deal with properties of the Internet, while *measuring features of the product* supplied by the Web data system. This is due to the fact that the Internet "is measurable" and therefore it enables carrying out fully automated quantitative measurements of traffic on the Web. For example, there are well-known automated techniques for evaluating usefulness of contents contained on Websites[1], which apply special software for assessing the usefulness on the basis of logs into the Web servers. They allow for detailed examination of the activity of users of the given services within a specified period of time. Logs include, among others, data on: the number of references to the server in order to download files and HTML pages, the average length of user sessions, most commonly used path while navigating on the website, which allows for evaluating the usefulness of the contents shared by service. For example, inclusion or omission of specific pages on the website, while users navigate through it, may be a measure for availability of the information published there. The tools providing statistics for Web sites,

[1] Some researchers (Eppler, 2000: 83-96) interpret the term of data quality as a synonym of usability

generated on the basis of a built-in HTML code referring to the service of the service provider offering such statistics, for example Google Analytics, can be an alternative to this type of analysis. The usefulness of the techniques mentioned above for evaluation of data quality on websites has not been fully tested and represents an interesting research problem.

The impact of the Internet properties on data quality on the Web is also visible at *the analysis stage*. As before, it is due to the fact that the Internet "is measurable". In particular, it applies to the possibility of analyzing search log files created by a Web server based on queries asked by users (e.g. the Webalizer tool). It allows for detecting gaps in the quality of the data provided by the web system. Information or data, which are useful, will be appearing in queries and can be found on websites, whereas the ones, which are not useful, will not be appearing in queries but they will be on the websites.

At the stage of *the data quality improvement,* these are interactivity, the massive nature, directness and the speed of the Internet that can be used in the information web-system, while correcting data and reorganizing information processes. The communication of a user with any web information system has a bilateral nature, which causes that control of data and information quality can be performed at the input and output of the system in both feedback cases. The interactive communication is also supported by the massive nature, directness and speed of the Internet (for moderation time). The massive nature of the access to information resources on the Internet gives the possibility of controlling the data quality to everyone, who has access to it. Thanks to this property of the Internet, errors can be quickly detected, corrected or eliminated. The speed of publishing and updating information resources on websites has a similar effect.

It is in the presented model that there are two types of the indirect impact of the Internet properties on data quality:

1. Through the information function in the aspect of communications and enrichment of the knowledge of the users of the information system.
2. Through the data quality management principles.

Ref. 1. The refinement of the information function in terms of communication and enrichment of the knowledge of the users of the information system functioning on the Internet is affected by almost all properties mentioned in the MEDIUM concept. The Internet is an electronic medium enabling omitting unnecessary intermediaries in communication and access to information resources. It is also a very fast medium: sending messages (e.g. per e-mail) to a recipient on

another continent takes just a few seconds. The reduction in the number of links and the speed of transfer in a data channel between a sender and a recipient cause that there are no delays, which cause outdating the information transmitted through it. Overcoming geographical and time barriers extremely facilitates direct communication of users and reaching original sources of information. It allows, among others, for assigning the contents shared on the Internet to the authors (except for anonymous materials). In this way, the assessment of their reliability increases. On the other hand, services available on the Internet (e-mail, instant messaging, chats, forums, and blogs) allow for establishing an effective dialogue amongst parties. It also raises assessment of contextual measures of data quality amongst users. For example, thanks to a multilateral interaction of users of a specialist forum, the relevancy of the contents published on it will be improved, which will result in better meeting the needs of the information users.

Ref. 2. Both information-employees and end-users have their requirements and expectations regarding the information product provided by the web-system. If a mismatch appears between requirements of the both groups, i.e. between providers and consumers of the information, the relevant data quality will not be provided - a quality gap appears. It could be eliminated by implementing detailed rules and principles for data quality management, which should be formalized in the form of instructions and should be applied to every information-worker and end-user of the data. Such instructions on the "framework" quality of a final data product prevent different interpretations of the quality and may contribute to its improvement. The Internet properties such as speed, overcoming geographical and time barriers may facilitate and speed up the dissemination of this kind of quality management principles amongst users of information systems. They can also cheaply and directly reach the persons interested in the data quality assessment.

It is in the presented model that there are also influence measures on the various stages of data quality management resulting from the behavior of users of web-system. In this case, these will be communication ways and forms amongst users described above as well as their individual and collective knowledge and the data quality management principles that will be interacting elements and forces. However, it should be taken into consideration that they are affected both by the Internet environment features mentioned above and the surrounding. The last one has its influence especially in the form of technology and people (which play different roles, not only of users), which manifests itself in the ways of their actions. For example, ruling persons may

establish rules for personnel management, which make people aware of the importance of data quality, for example through training and appropriate motivation and introduction of different consequences (economic and others) for obeying or not-obeying procedures and regulations regarding data quality. An example for an organization, which has created and developed a code of conduct with information on health published on Web pages is *Health on the Net Foundation* (HON). HON is a Swiss non-profit organization founded in 1995, whose aim is to increase knowledge of patients and doctors associated with reliable and useful information on health (HON, 2014). The organization allows everyone, who undertake to comply with this code, to use a special logo, which confirms the high quality of the information available on the websites.

In turn, the development of the technology, in particular of the mobile technology, has caused that the Internet can be accessed from any geographic location. It resulted in creation of new types of information products (e.g. geolocation service), for which the quality assessment is becoming a challenge. The reason is that such services are highly personalized and, therefore, while assessing them, these are contextual quality features above all that should be taken into consideration.

4. Conclusion

The essence of the presented conceptual model for data quality management in a data system on the World Wide Web consists, among others, in:

1. Presenting mechanisms of the Internet impact on various stages of the data quality management on the World Wide Web, starting from defining requirements and ending with improving data quality.
2. Identifying data quality features, which are affected by Internet properties: it has been found that only two of the three categories of quality features, i.e. contextual and reputation features, depend on the indicated Internet features.
3. Showing that the influence of the Internet environment has also an indirect character and is visible in data quality management principles and improving the information function on the Web in terms of communication and enrichment of the knowledge of users. It may result in positive changes of data units in time and space leading to higher users'

evaluation of contextual data quality features (in particular such ones as availability, completeness, and relevancy).

Therefore, the presented conceptual model for data quality management in a data system on the World Wide Web can be used for:

1. Selecting most important quality features in a web data system, which should be analyzed, measured and improved in the first place. These are, above all, contextual quality features such as: availability, naturalness, redundancy, accuracy, precision/completeness, safety, semantic compliance, structural compliance, verifiability, variability/impermanence.
2. Studying directions and strength of the impact of the presented Internet properties on the quality features of the information published on the Internet.

In future, it will allow for determining the scope and the extent of changes under the influence of the Internet, which proceed in the data units embedded in web information systems. Moreover, it will allow for an initial assessment of the risks of using erroneous information on the Internet as well as designing improvement of the quality of the data.

Literature

Al-Hakim, L. (2014). *Information Quality Deployment*. Proceeding of Ninth International Conference on Information Quality: 170-182.

Czerwiński, A. (2011). *Przemiany na rynkach informacji*. Opole: Wydawnictwo Uniwersytetu Opolskiego (in Polish).

Czerwiński, A. (2005). *Niektóre rodzaje przestępczości komputerowej a procesy zarządzania oprogramowaniem w organizacji*. In: *Komputer – przyjaciel czy wróg?*: 303-309. Szczecin: Szczecin University (in Polish).

Czerwiński, A.; Krzesaj, M. (2014). *Wybrane zagadnienia oceny jakości systemu informacyjnego w sieci WWW*. Opole: Wydawnictwo Uniwersytetu Opolskiego (in Polish).

English, L. (2003). Total Information Quality Management: A Complete Methodology for IQ Management. *DM Review* 9: 1-7.

Eppler, M. J. (2000). *Conceptualizing Information Quality: A Review of Information Quality Frameworks from the Last Ten Years*. Proceedings of the 2000 Conference on Information Quality: 83-96. St. Gallen: University of St. Gallen.

HON (2014). Available at: http://www.hon.ch. Accessed 12 July 2014.

Instytut Logistyki i Magazynowania EAN Polska (2003). *Elektroniczna gospodarka w Polsce - Raport 2002* (in Polish).

Ruževičius, J.; Gedminaitė, A. (2007). Business Information Quality and its Assessment. *Engineering Economics* 52 (2): 18-25.

Skrzypek, E. (2008). *Miejsce gospodarki opartej na wiedzy w nowej gospodarce*. In: *Społeczeństwo informacyjne. Stan i kierunki rozwoju w świetle uwarunkowań regionalnych*: 207-214. Rzeszów: Rzeszów University (in Polish).

Stvilia, B. et al. (2007). A Framework for Information Quality Assessment. *Journal of the American Society for Information Science and Technology* 58 (12): 1720-1733.

Wang, R. Y. (1998). A Product Perspective on Total Quality Management. *Communications of the ACM* 41 (2): 60-65.

Konceptualny model zarządzania jakością informacji w systemie informacyjnym w sieci WWW

Streszczenie

W artykule przedstawiono konceptualny model zarządzania jakością informacji traktowanej jako jej użyteczność lub zgodność produktu informacyjnego z jego specyfikacją. Proponowany model nawiązuje do znanego modelu TDQM R.Y. Wanga opartego na cyklu Deminga doskonalenia jakości. Jednakże model TDQM nie uwzględnia wpływu środowiska Internetu na jakość informacji udostępnianej przez systemy informacyjne w sieci WWW. Zaprezentowany w artykule autorski model bierze pod uwagę wpływ właściwości Internetu na wszystkie aspekty wynikające z funkcji informacji w społeczeństwie i w organizacji. Uwzględnia zatem aspekt wspierania procesów zarządzania jakością informacji, aspekt komunikacyjny oraz aspekt wzbogacania wiedzy indywidualnej i zbiorowej. W modelu uwzględniono także fakt, że wpływ znanych właściwości Internetu (określonych np. akronimem MEDIUM) odnosi się przede wszystkim do kontekstowych cech jakości informacji w sieci WWW, a w małym stopniu dotyczy wewnętrznej jakości jednostek informacji opisanych takimi cechami jak np. dokładność, spójność, złożoność czy precyzja.

Słowa kluczowe: jakość informacji, model zarządzania, system informacyjny, serwis WWW