An evaluation of Information quality frameworks for the World Wide Web

MB Parker
Faculty of Informatics and Design
Cape Peninsula University of Technology
Cape Town
South Africa
parkerm@cput.ac.za

V Moleshe
Faculty of Informatics and Design
Cape Peninsula University of Technology
Cape Town
South Africa
203071697@cput.ac.za

R De la Harpe
Faculty of Informatics and Design
Cape Peninsula University of Technology
Cape Town
South Africa
delaharper@cput.ac.za

GB Wills
School of Electronics and Computer Science
University of Southampton
Southampton
United Kingdom
gbw@ecs.soton.ac.uk

Abstract

Over the past few years the amount of data immediately available to the consumer has rapidly increased in size. This is due to the growth of the web as an information exchange and creation environment. Data creation on the Internet is increasing as it gives web publishers the opportunity to publish their content without any standards to govern the content. Although the consumer has access to this abundance of information, the lack of standards has lead to various levels of quality problems. There has been much advancement in Search Engine technology to search through these large amounts of content and to retrieve relevant, quality information. However, not all information returned is relevant to its context and it has become more difficult for the consumer to find quality information due to these information quality issues. Barriers that have been identified with regard to the retrieval of relevant information are the problem of too much information and the quality of that information.

This paper therefore address some of the issues of information quality on the web and evaluates a number of frameworks in order to identify common elements, differences and missing elements of such frameworks. A summary of the most common information quality elements is presented as a basis for a more comprehensive view of information quality frameworks available for managing and implementing quality strategies on the web.

Keywords: World Wide Web, information quality, data quality

1. Introduction

The World Wide Web has become an important knowledge and communication resource according to Henzinger & Lawrence (2004). The Internet provides many services to ordinary people and organisations. It is largely used for email, get news, access to government information and to find information on various issues. In the early days of the internet it was mostly used by people with a higher income but now the digital divide is closing up and more people of moderate income are getting online. The internet allows anyone to publish information on it. This directly contributes to the sheer size of the web. Large amounts of information are exchanged over the internet and Zakkon (2005) estimated that there are more than 20,000 new hosts on the internet per day. With this information overload unfortunately there are no regulatory standards for the people producing and publishing the information on the Internet as opposed to information that is printed. Information that is printed can be reviewed or put through refereeing processes before it is actually published and this gives the consumer confidence that they are being exposed to quality information. The Internet as a common vehicle for information distribution has raised the question of the quality of that information and the lack of web publishing standards according to Liu & Huang (2005:99-106).

2. Information quality

Data and information quality has been used interchangeably by various authors (Strong et al, 1997:103-110; Zhu & Gauch, 2001; Kahn et al, 2002:184-193). Information quality has been referred to constantly as information which must satisfy the needs of the user. According to Strong et al (1997:103-110) high quality data is data that is fit for use by the data consumers. The quality or usefulness of data is dependent on the individual who is going to be using it. Good quality data would therefore meet the requirements of its intended use. The concept of quality is therefore relative, depending on the different perceptions and needs of the users of the data. Information is published and used by individuals, corporations, government organisations, educational institutions and many other organisations on the web. Rieh (2002:146-147) identified two methods used by information retrievers as to how they judge and evaluate their information. The first is cognitive authority which is the extent to which users believe that they can trust the information. The other method is predictive judging where the information users have expectations of the information that is returned. When the results returned did not match the expectations of the seeker then they decided to start a new page. It is apparent that information has become and has always been a vital part in the lives of people and it is no surprise that there is a huge emphasis on the quality of it. In the context of the web, the quality of information on the Internet has been described by various authors as being of utmost importance (Liu & Huang, 2005:99-106; Parker, 2004; Rieh, 2002:145-148).

3. Importance of information quality

As mentioned above there is nothing that does not involve the use of information. What may seem like a simple online transaction, for example the purchase of a DVD and using a credit card to pay, involves considerable information processing. For instance information about the customer’s credit details are needed, bank details of the shop that must be

credited and so on. There are financial institutions that deal entirely with the management of the information needed in the above example. This information is vital to the company's existence. Companies like these value their information and quality of it. Large investments are made in the maintenance of the data integrity and the information systems managing the data. For example if the wrong customer was billed for the purchase, a simple letter of apology is not all that is needed to fix the problem. The capturing and use of poor or incorrect data increases operational costs. Time and resources will now have to be re-allocated to detecting the problem and correcting the data. The worse case scenario could lead to the possible loss of the customer who was incorrectly billed. The cost of poor information quality on the internet affects both the consumer and the publishers. Taking the example of the supplier who is selling products over the internet, if they publish incorrect information to the customer, then the customer will make a buying decision based on incorrect information. At a tactical level in an enterprise information is used for decision support and poor quality compromises decision-making according to Redman (1998:79-82).

The Internet as a common vehicle for information distribution has raised the question of the quality of that information according to Liu & Huang (2005:99-106). The internet is a very important information source for many reasons; the booming e-commerce industry where information is exchanged between supplier and customer via the internet. Berland et al (2001:2612-21) stated that an estimated 100 million Americans sought healthcare related information on the Internet and students who utilised it for research and school related work. The same trend is happening in China where the World Wide Web is becoming a pervasive resource for scholars and students according to Liu & Huang (2005:99-106). With this vast amount of data at the disposal of the web user the importance of the quality of information on the web can not be ignored.

4. Identification of data roles

Strong et al (1997:105) suggested that the quality of data should be established during the manufacturing of the data. They identified three roles within the data or information manufacturing cycle. Data evolve through a sequence of stages consisting of data collection, organisation, presentation and application (Liu & Chi, 2002). According to Strong et al (1997) these activities have a direct impact on the quality of the data. The roles include data producers (people, groups or other sources who generate data and are associated with the data-production process), data custodians (people who provide and manage computing resources for storing and processing data and carry responsibility for the security of the data) and data consumers (people or groups who use the data, the people that utilise, aggregate and integrate the data).

Data producers generate data to meet a specification based on the need to represent some aspect of a defined reality. On the web the data producer could be your website author. They should conduct tests to validate the quality and accuracy of information on the web. All information is produced with a purpose and the quality is based on the meeting of that purpose. The data producer will be responsible for the determination of data quality. Strong et al. (1997:106) added that the data custodian should take a broader conceptualisation of data quality. The custodian in the web environment could be the internet hosts of the website. The data consumers are the people visiting the webpage searching for information. Xu et al (2003:51) added a fourth role, data managers, within the data manufacturing cycle. The data managers are responsible for managing data quality in the systems. The manager of the on the webpage can either be the content

publisher or the author of the information on the website. Different data roles might be assigned different priorities to data quality dimensions.

5. Information quality frameworks

Information quality on the internet is defined using a series of dimensions. When defining data quality one will come across words like accuracy, timeliness, completeness, relevance and reliability. These are common dimensions used to define information quality on the internet. Porter (1991:955) described the aim of frameworks is to identify the relevant variables and the questions which should be answered in order to assist the users. Furthermore the framework should seek to help the analyst to make better decisions. It is not enough just defining the quality using the frameworks because they are also dependent on the context in which the data will be used (Shankar & Watts, 2003). Information quality frameworks have been developed over the past few years by various authors in different areas (Zeist & Hendricks, 1996; Strong et al, 1997:103-110; Huang et al, 1999; Leung, 2001:137-152; Zhu & Gauch, 2001; Kahn et al, 2002:184-193; Eppler & Muenzenmayer, 2002; Klein, 2002) that produced common elements of Information quality.

5.1 Frameworks from 1996 to 2000

Zeist & Hendricks (1996:145-160) identified information quality characteristics and sub-characteristics that consisted of functionality, reliability, efficiency, usability, maintainability and portability. Functionality includes suitability, accuracy, interoperability, compliance, security and traceability of information on the webpage. Reliability covers the maturity, recoverability, availability, degradability and fault tolerance of the webpage content. Efficiency of the webpage content investigates the time and resource behaviour. Usability includes the understandability, learnability, operability, luxury, clarity, helpfulness, explicitness, customisability and user-friendliness characteristics of information on the webpage. Maintainability pertains to the analysability, changeability, stability, testability, manageability and the reusability of webpage content. Portability is the adaptability, conformance, replaceability and installability of a webpage.

The authors Strong et al (1997:103-110) stated four information quality areas. The first area covers intrinsic data quality. Intrinsic data quality indicates that information has quality in its own right. It includes: accuracy, objectivity, believability, reputation, pragmatism, usefulness and usability. Accessibility data quality is the second area defined by Strong et al. It emphasises that information on the web must be easily accessible but secure. Accessibility data quality includes: accessibility, access security and shared understanding of data by various social groups. The third area is contextual data quality which is when the information should be provided in time and in appropriate amounts. Contextual data quality includes: relevancy, value-added, timeliness, completeness, amount of data and semantic. Finally representational data quality which comprises aspects related to the format of the information and its meaning. Representational data quality includes: interpretability, ease of understanding, concise representation, and consistent representation and syntactic.

Alexander & Tate (1999) suggested a quality framework for the Web and it included criteria such as authority, accuracy, objectivity, currency, orientation and navigation. Authority is when information is validated and the author of the webpage is visible. Accuracy is when a webpage is reliable and free of any errors. Objectivity is when the

website or WebPages is presented without personal biases. The Currency of the website ensures that the web content is up-to-date. Orientation warrants that there is a clear target audience. Navigation ensures the instinctive design.

Katerattanakul & Siau (1999) described four information quality categories of individual websites adapted from the dimensions by authors Strong et al (1997). The intrinsic category ensures the accuracy and free-of-error of the webpage content. It includes accurate, workable and relevant hyperlinks on the webpage. Contextual category warrants provision of the author's information. Representational information quality refers to the organisation, visual settings, typographical features, consistency, vividness and attractiveness of the webpage. Accessibility ensures the navigational tools used to access and move around on the website.

Shanks & Corbitt (1999) described a semiotic-based framework for the quality of data and it consists of four semiotic levels. Syntactic is when WebPages should be consistent. It should be well-defined and have formal syntax. Semantic ensures that information on the WebPages should be complete and accurate. The information must be comprehensive, unambiguous, meaningful and correct. Pragmatic warrants that the content on the website must be usable and useful. The webpage should be timely displayed, concise, easily accessible and the information must be reputable. The social level ensures shared understanding of meaning and an awareness of biasness on the webpage.

Information quality criteria as mentioned by authors Naumann & Rolker (2000) included subject, object and process criteria. Subject criteria is when the website displays believability, concise representation and understanding of content, interpretability and relevancy of information and the content should add value. Objective criteria ensure that the webpage must be complete, secure, objective, display timeliness and content authors verifiable. Process criteria ensure that information should be accurate, hyperlinks active, available, and consistent representation. The retrieval response time of a webpage forms part of the process criteria.

Zhu & Gauch (2000) described data quality classes of metrics for the retrieval of information on the web. The availability metric is the calculation of the number of broken links divided by total number of links on a webpage. The authority metric is when assigning a score to a reviews website. Currency metric is the time stamp of the last modification of the website. Information-to-noise ratio is the total length of tokens after pre-processing divided by size of webpage. Cohesiveness is how closely related the major topics of the website are. Popularity metric is the number of links to a webpage to determine the popularity of the webpage.

Dedeke (2000) identified a data quality framework that includes ergonomic, accessible, transactional, contextual and representational categories. The ergonomic category is the ease of navigation on the webpage. The accessible category ensures information accessibility, sharing and technical access. Transactional category is the responsiveness of a webpage, controllability, error tolerance, efficiency and adaptability of the content. The contextual category ensures the relevancy, completeness, appropriateness and timeliness of webpage content. Representation is the consistency, conciseness, structure, interpretability, readability and contrast of information on webpage.

5.2 Frameworks from 2001 to 2005

Leung (2001:137-152) defined information quality dimensions as characteristics and sub-characteristics. Functionality characteristic includes suitability, accuracy, interoperability, compliance, security and traceability on webpage content. Reliability characteristic includes maturity, recoverability, availability, degradability and fault tolerance of webpage. The efficiency characteristic includes time and resource behaviour. Usability – This includes understandability, learnability, operability, luxury, clarity, helpfulness, explicitness, customisability and user-friendliness. Maintainability includes analysability, changeability, stability, testability, manageability and reusability. Portability characteristic includes adaptability, conformance, Replaceability and installability.

Information quality were also categorised in the context of the web by authors (Kahn et al, 2002:184-193). These categories included sound information, relevant information, dependable information and usable information. Sound Information includes free-of-error, concise, representation, completeness and consistent representation of information on the webpage. Relevant Information includes appropriate amount, relevancy, understandability, and interpretability, objectivity, accurate and comprehensive webpage content. Dependable Information includes timeliness, security and traceability of WebPages. Usable Information includes believability, accessible, maintainable, reputation, value-added and speed.

Eppler & Muenzenmayer (2002) subdivided their suggested framework into content and media quality. Content quality indicates that the webpage content should include comprehensive, accurate, clear and applicable information. The web authors must ensure that the information on the website should be concise, consistent, correct and current. The content quality is concerned about the quality of the information presented itself on the web. Media quality on the other hand is concerned about the quality of the medium used to deliver the web content. This could include convenience, timeliness, traceability and interaction of the webpage. Other quality criteria are accessibility, security, maintainability and retrieval speed of the webpage.

Klein (2002) identified five key information quality dimensions in the context of the web. Accuracy dimension should ensure that the source and author of the information on the webpage is obtainable. Amount of data ensures that there is not too much or little information on the website or when this information is unavailable. Completeness is when information is missing, lack of depth and website incomplete when compared with other sites. Relevance is when the website purpose is too broad or bias. Timeliness dimension ensures information on a webpage should be current or the date when webpage was published must be known.

Liu & Huang (2005:99-106) recently made mention of the following key dimensions:

- *Source* – The source of the webpage content should be available.
- *Content* – Webpage content is complete.
- *Format and presentation* – The format of the webpage content display consistency.
- *Currency* – Webpage information is current and up to date.
- *Accuracy* – Content is accurate and reliable.

- *Speed* – Webpage is easily downloadable.

According to Eppler & Wittig (2000) an information quality framework should provide a systematic and concise set of criteria to which information can be evaluated. In the context of the World Wide Web the framework should ensure that the webpage content are of a required quality. The frameworks must be able to identify information quality problems on a webpage. It should also provide the basis for information quality measurement and proactive management on the web.

6. Evaluation of information quality frameworks

The information quality frameworks highlighted a number of dimensions that need to be considered to ensure website content quality. Based on the above frameworks we summarised and identified the following common dimensions:

- *Accessibility* – Extent to which the information on the website is readily available and downloadable.
- *Accuracy* – Extent to which the webpage information content are correct and reliable.
- *Appropriateness* – Extent to which the content is appropriate according to what the webpage visitors are requiring.
- *Believability* – The content on the webpage is true and credible.
- *Completeness* – The level to which the web content is not missing and sufficient.
- *Consistency* – All WebPages should be presented in the same format.
- *Ease of Manipulation* – Extent to which the content on the webpage is easy to manipulate.
- *Free-of-Error* – Information on the webpage should be correct and reliable, free of errors.
- *Objectivity* – Webpage content must be unbiased, unprejudiced and impartial.
- *Relevancy* – The webpage content should be applicable, helpful and relevant.
- *Representation* – Extent to which the webpage content is readable, consistent and has formal structure.
- *Reputation* – The information on the webpage is highly regarded with regard to its content.
- *Security* – Extent to which the access to the webpage is restricted appropriately to maintain its security.
- *Speed* – The retrieval or downloadable speed of the webpage content.
- *Timeliness* – Webpage content should be up-to-date.
- *Understandability* – Webpage content should be easily understood or comprehended.
- *Value-added* – Information on the webpage should be beneficial and provides advantages from its use.

Using the above information quality dimensions, the authors summarised, adapted and evaluated the thirteen information quality frameworks for the web in Table 1.

Table 1: Evaluation of information quality frameworks

Information Quality Frameworks														
INFORMATION QUALITY DIMENSIONS	Zeist & Hend-ricks (1996)	Strong et al (1997)	Alexa-nder & Tate (1999)	Kater-attana-kul et al (1999)	Shanks & Corbitt (1999)	Naum-ann & Rolker (2000)	Zhu & Gauch (2000)	Dedeke (2000)	Leung (2001)	Kahn et al (2002)	Eppler & Muenz-enma-yer (2002)	Klein (2002)	Liu & Huang (2005)	**FREQUENCY**
Accessibility	X	X	X	X	X	X	X	X	X	X	X		X	12
Accuracy	X	X	X	X	X	X		X	X		X	X	X	11
Appropriateness	X	X	X					X		X	X	X		7
Believability		X	X	X	X	X	X			X	X			8
Completeness		X			X	X		X		X	X	X	X	8
Consistency		X			X			X		X	X		X	6
Ease of manipulation	X							X	X		X			4
Free-of-error			X	X				X	X	X	X			6
Objectivity		X	X		X	X	X			X	X	X		8
Relevant	X	X	X	X		X	X	X	X			X	X	10
Representation		X	X	X		X				X	X		X	7
Reputation		X	X	X	X								X	5
Security	X	X				X			X		X			5
Source			X	X		X	X			X		X	X	7
Speed						X	X	X		X	X		X	5
Timeliness	X	X	X		X	X	X	X	X	X	X	X	X	12
Understand-ability	X	X				X		X	X	X	X			7
Value-added		X				X								2

X – Information quality dimension exists in framework

An analysis of the above information quality frameworks reveals common dimensions between them. The information quality dimensions that are the most frequent are accessibility and timeliness. The accessibility dimension is concerned about the technical accessibility, data representation issues and data-volume issues. The technical accessibility problem is realised by the website users when security access and permissions of a webpage become barriers of accessibility. The data-volume issue addresses the provision of relevant data that adds value to tasks in a timely manner. When large amounts of data need to be updated to the website it could lead to timeliness problems. This in turn could lead to the problem of accessibility.

The lack of accuracy of the data could lead to poor reputation of the data. This in turn leads to believability problems of the data. The source of the data that causes accuracy, reputation and believability problems could be viewed as adding little value to the website. When the data sources are inaccurate and not believable, the data gradually develops mismatches. The relevancy dimension is another quality dimension that is common amongst most of the above frameworks. The other quality dimensions that had a high frequency number are believability, completeness and objectivity. The quality dimensions with the least occurrences are ease-of-manipulation and value-added. With the value-added quality dimension having a low occurrence it is a clear indication why many individual WebPages lack quality of the information content. There are no web publishing standards involved when publishing content to a website.

The above findings suggest that an information quality framework for the web should at least consist of the following dimensions according to its usage by authors:

- *Accessibility, Timeliness* – occurred 12 times.
- *Accuracy* – occurred 11 times.
- *Relevant* – occurred 10 times.
- *Believability, Completeness, Objectivity* – occurred 8 times.
- *Appropriateness, Representation, Source, Understandability* – occurred 7 times.

The above quality dimensions were chosen based on more than half of the thirteen authors used it in their frameworks.

7. Conclusion

Information quality on the web has proven to be a problem due to the lack of standards for web publishers. This paper summarised thirteen information quality frameworks to identify common dimensions between them. An interesting observation was that none of the authors utilised all the quality dimensions that was identified in the literature. We did however recognise the dimensions that were recommended by most of the authors. These dimensions of *Accessibility, Timeliness*, *Accuracy, Relevant Believability, Completeness, Objectivity Appropriateness, and Representation,* could be useful to be utilised as a basis to manage information quality on the web. Further research should be carried out to investigate how to integrate quality standards into web publishing.

8. References

Alexander, J. and Tate, M. A. 1999. *Web wisdom: How to evaluate and create information on the web*. Mahwah, NJ: Erlbaum.

Berland, G.K., Elliott M. and Morales L.S. 2001. Health information on the Internet: Accessibility, quality and readability in English and Spanish. *Journal of the American Medical Association*, 28(5):2612 – 2621.

Dedeke, A. 2000. A conceptual framework for developing quality measures for information systems. *Proceedings of the 5th International Conference on Information Quality.*

Eppler, M. and Muenzenmayer, P. 2002. Measuring information quality in the web context: A survey of state-of-the-art instruments and an application methodology. *Proceedings of the 7th International Conference on Information Quality (ICQ-02)*.

Eppler, M. and Wittig, D. 2000. A Review of Information Quality Frameworks from the Last Ten years. *Proceedings of the IQ 2000 – The Conference on Information Quality.* Boston, USA, 22-23October 2000.

Henzinger, M. and Lawrence, S. 2004. Extracting knowledge from the World Wide Web. [Online]. Available http://www.pnas.org/cgi/reprint/101/suppl_1/5186.pdf (Accessed 15 May 2006).

Huang, K., Lee, Y. and Wang, R. 1999. *Quality Information and knowledge*. Upper saddle river, NJ: Prentice hall.

Kahn, K., Strong, D. and Wang, R. 2002. Information Quality Benchmarks: Product and Service performance. *Communications of the ACM*, 45(4): 184 – 193.

Katerattanakul, P. and Siau, K. 1999. Measuring information quality of web sites: Development of an instrument. *Proceedings of the 20th international conference on Information Systems*. Charlotte, North Carolina, USA: 279-285.

Klein, B.D. 2002. When do users detect information quality problems on the World Wide Web? *American Conference in Information Systems,*2002.

Leung, H.K.N. 2001. Quality metrics for intranet applications. *Information & Management*, 38(3): 137 – 152.

Liu, L. and Chi, L. 2002. Evolutional Data Quality: A Theory-specific view. *Proceedings of the Seventh International conference on Information Quality (ICQ-02).*

Liu, Z. and Huang, X. 2005. Evaluating the credibility of scholarly information on the web: A cross cultural study. *ScienceDirect*, 37(2):99-106.

Naumann, F. and Rolker, C. 2000. Assessment methods for the information quality criteria. *Proceedings of the 5th International Conference on Information Quality*.

Parker, M. 2004. A generic business intelligence data model to analyse data within a small to medium medical practice (SMMP). Conference Paper. *South African Institute of Computer Scientists and Information Technologists (SAICSIT)*, October 2004. Stellenbosch, South Africa: 111-114.

Porter, M.E. 1991. Towards a dynamic theory of strategy. *Strategic Management Journal,* 12(1):954-1117.

Redman, T. 1998. The impact of poor data quality on the typical enterprise. *Communications of ACM*, 41(2):79-82.

Rieh, S. Y. 2002. Judgment of Information Quality and Cognitive Authority in the Web. *Journal of American Society for Information Science and Technology*, 53(2):145-161.

Strong, D., Lee, Y. and Wang, R. 1997. Data Quality in context. *Communications of the ACM*, 40(5): 103-10.

Shanks, G. & Corbitt, B. 1999. Understanding Data Quality: Social and Cultural Aspects. *Proceedings of the 10th Australasian Conference on Information Systems*.

Shankar, G. and Watts, S. 2003. A relevant, believable approach for data quality assessment. *Proceedings of the 8th International Conference on Information Quality,* 2003: 178-189.

Xu, H., Nord, J., Brown, N. and Nord, G. 2003. Data quality issues in implementing an ERP. *IEEE transactions on knowledge and data engineering*, 1021(1):47-58.

Zakkon, R. 2005. Hobbes' Internet Timeline. [Online]. Available: http://www.zakon.org/robert/internet/timeline/ (Accessed 21 May 2006).

Zeist, R.H.J. and Hendriks, P.R.H. 1996. Specifying software quality with the extended ISO model. *Software Quality Management IV – Improving Quality, BCS*: 145 -160.

Zhu, X. & Gauch, S. 2000. Incorporating quality metrics in centralized/ distributed information retrieval on World Wide Web. *Proceedings of the 23rd international ACM SIGIR conference on research and development in information retrieval,* Athens, Greece.

9. Acknowledgements

The authors would like to acknowledge Cape Peninsula University of Technology, University of Southampton and colleagues for contributions towards this paper.