# MEASURING INFORMATION QUALITY IN THE WEB CONTEXT: A SURVEY OF STATE-OF-THE-ART INSTRUMENTS AND AN APPLICATION METHODOLOGY

(Practice-Oriented)

**Martin J. Eppler**
University of St. Gallen, Switzerland
Martin.Eppler@unisg.ch

**Peter Muenzenmayer**
Business Media, Switzerland
Peter.Muenzenmayer@bm-ag.com

**Abstract**: Various powerful instruments exist today to evaluate information quality in the web context. They can be categorized into five types of tools, namely performance monitoring systems, site analyzers, traffic analyzers, web mining tools and survey tools (to generate opinion-based user feedback). The combined use of these tools can enable an organization to measure the multiple dimensions of information quality in the Internet or Intranet context. This however requires a clear methodology that is based on systematic sequential steps and on an information quality framework that outlines relevant measurement criteria. In this paper we show which information quality criteria can be measured with the help of these tools and we provide an overview on the most important of these instruments. We present the IQM-methodology to match information quality criteria with adequate measurement tools.

**Key Words**: Information quality audit, IQ criteria, IQ framework, measurement tools, size analyzer, traffic analyzer, web mining, monitoring tools, user surveys, perceived information quality, information quality measurement methodology

## INTRODUCTION: INFORMATION QUALITY PROBLEMS IN THE WEB CONTEXT

One of the most prominent platforms for information provision today is the World Wide Web. The Internet and the intranet have established themselves as the key infrastructures for information administration, exchange, and publication [1]. The rapid proliferation of web-based information applications and services, however, has also led to numerous information quality problems. Typical information quality problems that arise in this context are (see [2], [3], [4]) for example the following:

- A website contains outdated information because its content owners have neglected to update it on a regular basis. The provided information is not *current*.
- The navigation and layout of a website confuse the users. They do not know where to find the information they are looking for. The information is not easily *accessible*.
- The entry page of a portal is contains too many links, references and pieces of information. The

starting page is not *concise* and overloads potential users.

- The website of a company uses an inconsistent style in its various pages. The users do not always know whether they are still in the same domain or not. The information is not presented in a *consistent* manner.
- The users of an intranet cannot access a crucial application, because of frequent intranet down-times. The infrastructure for information provision is not *reliable*.
- The publication process of a news site is sub-optimal leading to a delayed publication of *timely* news.
- The homepage of a company has been inadequately protected. Hackers alter its content because the website's information is not *secure*.
- A website consists of lengthy texts which are difficult to understand. Consequently, users do not return to the site. The information is not *comprehensible*.
- An intranet contains a great number of obsolete links to outside sources that have changed or disappeared. This is frustrating to employees who rely on the pathfinder function of their intranet. The information provided is neither *current* nor *accessible*.
- A website lacks crucial information about a company's products and services. Users must call up the company to find out more about the products. The information provided online is not *complete*.

These problems have great negative consequences for the information consumers. They either cannot find the information they are looking for, they cannot easily interpret and adapt it to their needs or they cannot directly apply it as they would like to.

The responsibility for such shortcoming cannot be traced to one single group of people. It is the collaboration of various information-related functions that leads to such problems [3], namely the work of content producers or authors and content managers (in terms of correctness, conciseness and currency), webmasters (to ensure smooth and consistent publications), IT-support staff (in terms of a reliable and secure infrastructure) and line and product managers (in terms of a website's alignment to information consumer's needs). All of these professional groups can profit from a continuous measurement of information quality to bring problems such as the ones described above to the surface and to devise rapid improvement actions. Information quality measurement can detect (in real-time) whether information is fit for use [5] or not for information consumers, administrators and producers. How this can be done is the described in the next section.

## TOOL CATEGORIES FOR INFORMATION QUALITY MEASUREMENT

With the development of web technology new software has been engineered to help a webmaster in creating, managing and maintaining his or her websites. There is in fact a huge collection of software available that supports the webmaster in many different ways. Starting from freeware, shareware or out-of-the-box tools all the way to powerful (and costly) enterprise software.

There are three main focus areas of such tools. The first one is a very technical one based on *hardware monitoring* and *software testing* known from standard network and server administration. The other two focus areas have different goals in supporting the *maintenance* of websites. The first goal relates to *product* based aspects of a website. Specifically, such tools help to optimize, monitor and *test a website*. The other goal focuses on analyzing the *users' behaviour* on the website, their interaction and their main interests. There are also software tools combining these focus areas, for example integrating information from traffic analyzers with that of legacy systems like ERP[1] or CRM[2] systems or putting them into a data

[1] ERP = Enterprise Resource Planning
[2] CRM = Customer Relationship Management

warehouse for further mining and analysis.

In addition to these continuous, fully-automated measurement tools that collect objective information quality metrics, there are also tools that can gather perceived information quality metrics via surveys. To gather information which is hard to measure technically, there is feedback software to support voting and questioning over the web (on such issues as usability, convenience, completeness, usefulness, relevance, etc.).

Consequently, we can distinguish between the following types of tools that can be used for IQ-measurement in the web context:

a) Performance Monitoring
b) Site Analyzer
c) Traffic Analyzer
d) Web Mining
e) User Feedback

In all five categories there is a large amount of available software, particularly in categories a), b) and c). Categories d) and e) contain the more powerful, but also more expensive tools. Before we provide an overview on leading vendors in all five areas, we briefly describe each tool category.

## a) Performance Monitoring: Server and Network Monitoring Testing

In this category we find tools that observe the availability (e.g., downtime) and performance of servers (e.g., response time) and networks. As this is a well known discipline and just a few web quality criteria are based upon them (such as speed and reliability) we do not describe them extensively in this paper.

## b) Site Analyzer

Site Analyzers help to examine a website based on different quality criteria. Various quality aspects can be examined and represented in an automated and aggregated report. These tools have been developed as websites have grown bigger and more complex and updating them has become more and more challenging (in terms of maintenance to keep the site manageable). Obsolete hyperlinks or missing images have to be found and tested laboriously. The Site Analyzers offer a set of criteria to check for this kind of quality issues. The software of this category reaches from freeware to shareware up to commercial packages.

The tools of this category offer multiple functions. The selection of software is large. Starting with tools that are focused on special aspects or quality criteria like meta information or proper HTML code, there are also suites that take care of a more comprehensive set of quality criteria and represent them in an aggregated report with drill-down possibilities. The standard functionalities focus on identifying:

- broken links and anchors (hyperlinks within a page)
- failures in forms
- orphaned files
- orthography errors
- missing alt tags
- missing or double keywords or page titles
- missing height and width attributes

In addition, these tools examine the following aspects:

- performance and monitoring of servers
- browser compatibility
- site inventory with link structure, used images, types of documents, used image maps, multimedia pages, …
- ratio of old pages versus new pages

Special functionalities that are included in some of the site analyzers relate to:

- corporate identity
- simulation of customer transaction to analyze the performance
- graphical interfaces with drag and drop
- transaction checking
- style sheet independencies
- capacity analyzing and planning
- searchability with the ability to automatically add meta tags

While the site analyzer focuses mainly on the product (e.g., the website). The next group of tool focuses on the users and their behavior.

## c) Traffic Analyzer

The primary purpose of a website is to serve as a communication media for a target group. The usage of the website plays the central role. In order to gain feedback about the traffic and behavior on the website, there are software tools that can collect this data and represent it in reports with insightful diagrams and tables. Besides the websites' integrity and working functionality its actual use is of major interest for strategic and quality issues.

There are two different possibilities to gain data about the user traffic. The first one relies on the server's log-file. The other is a dedicated network collector. The *log-file* is widely used as it is much easier and more economic to implement. The *network collector* allows for a more detailed and precise evaluation as well as a measurement of further IQ-criteria not registered in the log-file. Another advantage of network collectors is the faster evaluation. But a network collector is more expensive both in terms of the required hardware and software.
The standard functionalities of traffic analyzers include:

- page hits, views, visits
- most and least requested files and pages
- information about visitors like geographical segmentation, kind of web browser, installed plug-ins, IP-addresses, …
- graphical evaluation e. g. pie charts, diagrams
- automatic reports trough templates
- different output e. g. html, xml, pdf, etc.
- possibilities to filter information e. g. per day, week, region, browser, tc.
- reverse DNS lookup to show the domain instead of the IP-address
- standard reports e. g. used search engines, keywords and search phrases.

Some, but not all, traffic analyzers also include the following services:

- evaluation of special components e. g. banners, links
- top entry or exit pages
- time spent on pages
- reports on visitor trends

- customer money spent on pages (ROI[3])

As the previous three tools can generate a massive amount of information that has to be analyzed in terms of the underlying quality issues, there is a need for integration. Web Mining tools are one feasible way of integrating IQ-relevant measurement data. They are described in the next paragraph.

## d) Web Mining

In this category we find tools that integrate data from website analyzing, traffic analyzing or legacy systems. With the broader data base more precise analyzes can be performed. Data, for example, can be integrated from a content management system (CMS) with data from traffic analyzers in order to get a better understanding about the costs of maintenance of a website and its return in user traffic. Another example leading to valuable insight about user navigation behavior is the integration of site analyzer data (e. g. the structure of a website) with user traffic data.

## e) User Feedback

Although the tools presented so far are quite powerful, there are still a few quality criteria (such as comprehensiveness, clarity or accuracy) that are hard or even impossible to measure technically or the technical measurement is simply too costly to set up. In those cases a user feedback is the appropriate way to measure information quality criteria.
There are different possibilities to receive user feedback, starting from a simple one or two questions poll, up to fully grown feedback forms with changing question order, changing interview partners names and personalized reports for different user roles. Such user feedback systems typically include the following functionalities:

- graphical representation of the results e. g. pie charts, diagrams
- metric evaluation of the results (e.g. average values)
- graphical front end to build the questionnaire
- templates for layout and context.

Some of the more sophisticated ones offer additional functionalities, such as:

- entire support of the process from creation of the questionnaires, mailing, evaluation of quantitative questions to feedback to the users
- export to different formats and systems
- web based front end
- different possibilities to start the questionnaires e. g. by mail with link or link on web page
- possibilities to filter the results by time, region, …

In the appendix to this paper, we provide an overview of state-of-the-art instruments that cover these functionalities. The appendix lists several tools in each category. It provides the product names, its vendors, and their website-address.

Having given a quick overview on the most important types of tools that can be used for information quality measurement in the Internet and intranet context, we will now give a specific example of how these tools can be combined to measure the various dimensions of information quality.

[3] ROI = Return on Investment

# APPLICATION OF THE TOOLS: THE IQM-METHODOLOGY

In order to use the tools described in the previous section, an organization requires not only relevant hardware (e.g. servers of adequate scale), skills and qualifications (e.g., in handling the software interpreting the results), and resources (in terms of financial scope and time), but also a *measurement methodology*. The information quality measurement (IQM) methodology should ensure that the measurement tools are used correctly, that is to say that they measure the rights things in the right manner. In our view, an information quality measurement methodology consists of two major elements: an *action plan* on how to conduct the measurement (measuring in the right manner), and an information quality *framework* that defines which criteria are worth measuring (measuring the right things). In this section we provide such an action plan and a conceptual framework and we provide an example of applying the methodology.

The action plan or sequence of steps we propose to measure information quality in the web context is outlined in the following table. It consists of four main phases, namely planning the measurement, configuring the measurement tools, conducting the measurement, and following-up on the measurement with corrective actions. It is loosely based on the Deming-cycle of plan-do-check-act (see for example [6]).

1. Measurement Planning
    a) Identification of relevant information quality criteria (adaptation of the IQ-framework) through interviews with stakeholders
    b) Analysis and definition of trade-offs and interdependencies between criteria
    c) Operationalization of the criteria (definition of qualitative and quantitative indicators)
    d) Selection of measurement tools for the required indicators
2. Measurement Configuration
    a) Weighting of the indicators according to strategic priorities
    b) Definition of alert and target values for every indicator
3. Measurement
    a) Data gathering (e.g., monitoring or surveys)
    b) Data analysis (incl. statistical analysis and tests)
    c) Data presentation (aggregation and reporting)
4. Follow-up Activities
    a) Follow-up activities (corrective measures based on alert indicators)
    b) Controlling of activities (e.g., assigning responsibilities)
    c) Adjustment of measurement according to implementation experiences (re-start the cycle at 2. b))

Table 1: The main steps of the IQM (information quality measurement) methodology

For step one (measurement planning) of the methodology, an information quality framework is needed (see step 1a). We use the conceptual information quality framework presented in [2]. It can provide the relevant criteria and indicate possible trade-offs between them. A simplified version of the framework is provided below.

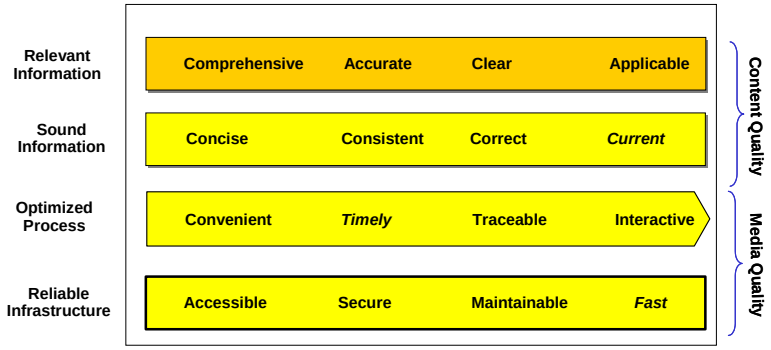


Figure 1: The Conceptual Framework for Information Quality in the Website Context

With these elements – the methodology and its conceptual framework – we can now conduct an information quality audit. A sample audit plan is provided below. It outlines *which* information quality criteria will be measured (they are taken from the framework presented above) and *how* they will be measured with the help of indicators and measurement tools.

| IQ-Criterion | Web-Indicator | Measurement Tool |
|---|---|---|
| 1. Accessibility | # broken links<br># broken anchors | Site Analyzer |
| 2. Consistency | # of pages with style guide deviations | Site Analyzer |
| 3. Timeliness | # of heavy (over-sized) pages/files with long loading times | Site Analyzer |
| 4. Conciseness | # of deep (highly hierarchic) pages | Site Analyzer |
| 5. Maintainability | # of pages with missing meta-information | Site Analyzer |
| 6. Currency | Last mutation > six months | Site Analyzer |
| 7. Applicability | # of orphaned (not visited or linked) pages or user rating | Site Analyzer in combination with Traffic Analyzer, User Surveys |
| 8. Convenience | Difficult navigation paths: # of lost /interrupted navigation trails | Traffic Analyzer, Web Mining Tools |
| 9. Speed | Server and network response time | Server & Network Monitoring Tools, or Site Analyzer |
| 10. Comprehensiveness | User rating | User Surveys |
| 11. Clarity | User rating | User Surveys |
| 12. Accuracy | User rating | User Surveys |
| 13. Traceability | # of pages without author or source | Site Analyzer |
| 14. Security | # of weak log-ins | Site Analyzer/Port scanner |
| 15. Correctness | User ratings | User Surveys |
| 16. Interactivity | # of forms<br># of personalizable pages | Site Analyzer |

Table 2: Measuring IQ-criteria for the website context with relevant indicators and adequate tools

Table 2 shows that different information quality criteria do indeed require different measurement tools and that only a mix of these tools can provide a comprehensive view of information quality on an intranet or Internet website. The table also highlights the fact that some criteria may require more than one indicator. To increase the value and reliability of such a measurement system, combinations of survey-based and automatically generated indicators may be appropriate, as online surveys are subject to several biases (see [7] pp. 222-226) and automatically generated indicators are sometimes difficult to interpret (see [8], p 212). Another important aspect of the measurement process is its consistency over time. Only if the measurement process remains unchanged can the effects of information quality improvements be made visible. Thus, we suggest to leave the steps 1. a.) to 2. a) unchanged as long as the strategic goals have not fundamentally changed.

The main practical advantages of using such a methodology can be summarized as better co-ordination, a greater scope and improved clarity. Below, Gregory Huber, Web Quality Manager at UBS Financial Services Group, outlines his view as a practitioner on the benefits of using such a methodology:

> "The methodology can help to make the information quality management process more efficient and more transparent. It can provide support to make information quality management a regular routine: the people know their responsibilities and their roles. A methodology can also be helpful to identify all relevant stakeholders (e.g., owners, publishers) and their needs with regard to information quality."

In addition to these benefits, the proposed methodology can provide a common terminology or frame of reference for webmasters, users, managers, and IT-staff. It can help to look beyond the measurement *tools* and clarify whether they really measure what is relevant for the stakeholders.

# Conclusion

In this paper we have first given an overview of some typical information quality problems in the Web context. We have shown that they can be related to various information quality criteria. Then, we have given an overview of the state-of-the-art in the area of measurement tools. We have distinguished five groups of such measurement tools and we have provided (in the appendix) various examples of each tool category. We have also outlined the main functionalities of each category. These functionalities are used to measure specific IQ-criteria in a systematic and planned way. Such a systematic way has been proposed with our four step IQM-methodology. The application of the methodology has shown that the proposed tools can be used to measure relevant information quality criteria. A great challenge in this respect is ensuring adequate follow-up activities, so that the measuring process can have a significant impact on the information quality provided on a Internet website or on an entire Intranet. A *continuous* IQ-measurement can reveal whether the implemented activities have improved information quality or not.

# References

[1] Alexander, J. E.; Tate, M. A. (1999) Web wisdom: how to evaluate and create information quality on the web, Mahwah, NJ: Erlbaum.

[2] Eppler, M. (2001) A Generic Framework for Information Quality in Knowledge-intensive Processes, in: Proceedings of the Sixth International Conference on Information Quality, MIT, 2001, pp. 329-346.

[3] Eppler, M., Snoy, R., Mathis, H. (2001) Qualität im Internet. IHA-GfK, Hergiswil.

[4] Eppler, M. (2002) Information Quality in knowledge-intensive Processes. St.Gallen: University of St. Gallen.

[5] Huang, K.-T.; Lee, Y.W.; Wang, R.Y. (1999) Quality Information and Knowledge. New Jersey: Prentice Hall.

[6] English, L. (1999) Improving Data Warehouse and Business Information Quality. Wiley & Sons: New York.

[7] Simsek, Z., Veiga, J. F. (2001) A Primer on Internet Organizational Surveys, in: Organizational Research Methods, Vol. 4 Issue 3, pp. 218-236.

[8] Tierney, P. (2000) Internet-Based Evaluation of Tourism Web Site Effectiveness: Methodological Issues and Survey Results, in: Journal of Travel Research, Vol. 39 Issue 2, pp. 212-220.

# APPENDIX: EXAMPLES OF IQ-MEASUREMENT TOOLS

In this section we provide an overview on specific tools that exist in each tool category.

## Site Analyzers

| Product | Vendor | URL |
|---|---|---|
| Hypertrak Performance Monitor | Trio Networks | www.trionetworks.com |
| Watchfire Enterprise Solution | Watchfire | www.watchfire.com |
| WebAnalyzer 2.0 | InContext | www.incontext.com |
| Webmaster 5.0 | Coast | www.coast.com |

## Traffic Analyzers

| Product | Vendor | URL |
|---|---|---|
| Analog Logfile Analysis 5.03 | University of Cambridge Statistical Laboratory | www.analog.cx |
| Live Stats Web Analytics Server 6 | Deepmetrix | www.deepmetrix.com |
| Nedstat Basic | Nedstat | www.nedstat.com |
| PerfMan for Webservers | ISM | www.perfman.com |
| SiteStat | Nedstat | www.nedstat.com |
| Summary Plus 2.0 | Summary.net | http://summary.net |
| Surfreport 3.0 | netrics.com | www.surfreport.com |
| Urchin Multihome 3 | Quantified Systems | www.urchin.com |
| Website Analysis Suite | Hyperion | www.hyperion.com |
| WebSuxess 4.0 | Exody | www.exody.net |
| Wusage 7.1 | Boutell.com | www.boutell.com |
| Xcavate 1.9 | Expertise | www.exsoft.com |

## Web Mining Tools

| Product | Vendor | URL |
|---|---|---|
| Accrue G2/ Hitlist | Accrue | www.accrue.com |
| C-Insight | Metaedge Corp. | www.metaedge.com |
| Clementine | SPSS | www.spss.com |
| EasyMiner 2 | Mine It | www.mineit.com |
| Synera ePack | Synera | www.synerasystems.com |
| Funnel WebSuite | Quest | www.quest.com |
| Esite | Informatica | www.informatica.com |
| Netgenesis 5 | Netgenesis | www.netgen.com |
| NetTracker eBusiness Solution 5.5 | Sane Solutions | www.sane.com |
| WebAbacus | WebAbacus | www.webabacus.com |
| WebFeedback 3.0 | Liebhart Systems | www.cyberware-neotek.com/WFB |
| Webmaster Pro | Coast | www.coast.com |
| Webmining Genius | Novuweb | www.novuweb.com |
| Webtrends Analysis Suite 7.0 | NetIQ | www.webtrends.com |

## User Feedback Software

| Product | Vendor | URL |
|---|---|---|
| Cont@xt | Information Factory | www.information-factory.com |
| Opinion Poll | Metrix Lab | www.opinionpoll.com |
| Infopoll Business Intelligence Suite | InfoPoll | www.infopoll.com |
| WebSurveyor | WebSurveyor Corp. | www.websurveyor.com |