Part I

Information Quality and the Web

Chapter 2

Information Quality

The availability of reliable, accurate, and up-to-date information is crucial for any decision making. Compared to concepts like data integrity and security which have been studied in detail since the introduction of relational database technology, the notion of information quality is relatively young and its general conceptualization as well as the methods developed to assess and improve information quality are very diverse. Information quality shows a steadily growing interest amongst practitioners and researchers [GzSS03] as information is increasingly seen as the most valuable asset of an organization and dealing with information quality problems can be very expensive and time consuming [Red98]. A further reason which drives the work on information quality is the increasing interconnectivity amongst information producers, mainly spurred through the development of the Internet and web-based information systems.

2.1 Related Work

Information quality has been researched in various application contexts. Within Information Systems research, the different approaches can be classified into two general groups:

One group of authors is focusing on improving the information quality within information production processes inside organizations. These approaches are influenced by the Total Quality Management philosophy [Jur74, Dal03, Bec01, WZL00] and are thus analyzing information quality from a management perspective, aiming at installing appropriate workflows and organizational regulations to ensure information quality. Information quality management is seen as a continuous cycle involving four major steps: the definition of information quality requirements, measuring these requirements, analyzing the results, and taking the appropriate steps for improving information quality based on the evaluation results [Wan98].

Prominent representatives of this group are Richard Wang, who is leading the Total Data Quality Program[1] at the Massachusetts Institute of Technology [Wan98, WS96, WZL00], and Thomas Redman whose work aims at providing managers with practical guidelines to analyze and improve information quality within business processes [Red96, Red98, Red01]. Information quality has also been studied in the context of data warehousing projects. Jarke and Vassiliou [JV97] as well as Ballou and Tari [BT99] analyze the role of information quality in the design, operation and evolution of data warehouses. Their work has a management perspective as they perceive the quality of information sources as influenceable.

A second group of authors are investigating information quality in situations where the information providers are autonomous and cannot be directly managed as in an organizational setting. The problem of assuring information quality in the context of web-based information systems falls into this category, since the Web is an open information space consisting of information from autonomous information providers. Due to the lack of manageability, the authors focus on assessing information quality in order to support information consumers in their decision whether to use certain information or information sources for accomplishing specific tasks. Prominent representatives of this group are Felix Naumann [Nau02] and Chen et al. [CZW98] who developed approaches to quality-driven source selection and query planning; Michael Gertz [Ger96], and Mecella et al. [MSV^{+}02] who are researching information quality assessment in the context of federated databases; and Eppler and Muenzenmayer [EM02] as well as Alexander and Tate [AT99] who are proposing methods to assess the quality of websites.

2.2 Terminology

The concept of information quality is a domain-specific subconcept of the general concept of quality. As with all abstract concepts, there are various definitions for quality. A widely accepted definition is provided by the International Organization for Standardization (ISO), which standardized the concept of quality as a basis for the ISO 9000 family of quality management standards [ISO05]. The ISO definition has evolved over time. ISO 8402, which was published in 1986, defines quality as "the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs." [ISO86]. This definition already implies the

[1] http://web.mit.edu/tdqm/www/ (retrieved 09/25/2006)

idea that quality is determined by various characteristics and that quality is not a general property of a product or service but depends on needs to be satisfied. The ISO 8402 definition is constricted to products and services. This constriction is lifted by the ISO 9000 standard which supersedes ISO 8402. The latest ISO 9000 revision IS9000:2005 defines quality as: "Degree to which a set of inherent characteristics fulfills requirements" [ISO05]. A characteristic is defined as "distinguishing feature" [ISO05]. Requirement is defined as "need or expectation that is stated, generally implied or obligatory" [ISO05]. This definition is more general and can be applied to any entity. Quality is again defined as a relation between requirements and features. A second, very popular and even shorter definition is given by Joseph Juran, a prominent author and practitioner of the Total Quality Management movement [Jur74, Dal03, Bec01]. Juran defines quality simply as "fitness for use" [Jur74].

Juran's definition has been adopted by most authors working on information quality. [WS96, SLW97, TB98, Nau02, EM02, KB05] commonly define information quality as the fitness for use of information. This definition implies two important aspects:

- Information quality is task-dependent. A user might consider the quality of a piece of information appropriate for one task but not sufficient for another task.

- Information quality is subjective, as a second less quality-concerned user might consider the quality of the same piece of information appropriate for both tasks.

Information quality is commonly conceived as a multidimensional construct [WS96, Red96, BWPT98, Nau02, PS03], as the "fitness for use" may depend on various factors such as accuracy, timeliness, relevancy, completeness, consistency, or interpretability. Information quality dimensions are not independent of each other and typically only a subset of the dimensions is relevant in a specific situation. Which quality dimensions are relevant and which levels of quality are required for each dimension is determined by the specific task at hand and the subject preferences of the information consumer [Nau02, WS96]. For instance, for the task of deciding when to sell a stock, it is essential to have the latest market information. As for the task of getting an overview of a market, timeliness is less relevant but completeness is essential.

In [Wan98], Richard Wang proposes an extended entity-relationship diagram [Che76] notation for visualizing which quality dimensions are regarded as relevant for specific attributes. Figure 2.1 shows an example diagram for a stock information systems used by a bank to consult its clients [Wan98]. Within the diagram, quality dimensions are attached to the ovals which represent the attributes of an entity.

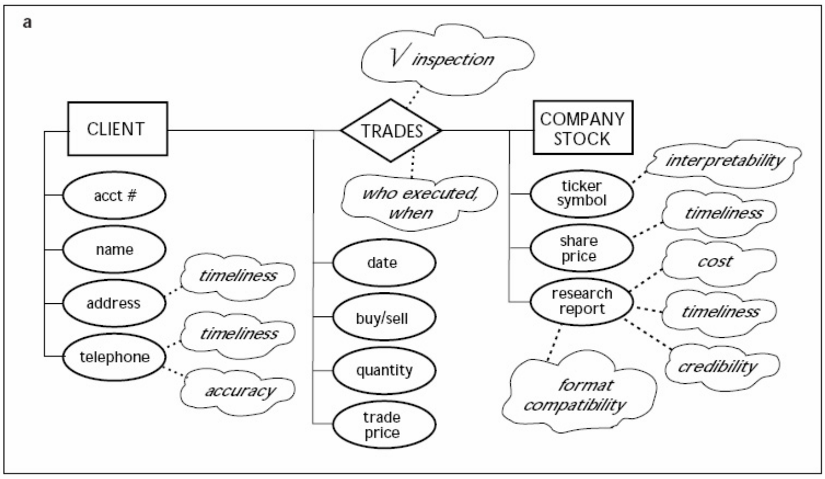

Figure 2.1: Extended entity-relationship diagram visualizing which quality dimensions are regarded as relevant for specific attributes [Wan98].

2.3 Information Quality Dimensions

Several research efforts have put together catalogs of information quality dimensions [WS96, Red96, JV97, CZW98, MSV^{+}02, Ger96, EM02]. Others have compiled multiple catalogs [Nau02, EW00, KB05]. In order to compare the different approaches, the proposed quality dimensions are grouped into four categories, following a categorization schema introduced by Richard Wang [WS96]:

Intrinsic Dimensions denote that information has quality in its own right. Intrinsic dimensions are independent of the user's context. They are capturing whether information correctly represents the real world and whether information is logically consistent in itself.

Contextual Dimensions highlight the requirement that information quality must be considered within the context of the task at hand and the subjective preferences of the user. In contrast to the intrinsic dimensions, contextual dimensions cannot be assessed in a general fashion but need to be assessed based on the user's context and subjective preferences.

Category	*Dimension*	*Wang* [WS96]	*Redman* [Red96]	*Jarke* [JV97]
Intristic Dimensions	Accuracy	√	√	√
	Consistency			√
	Objectivity	√		
	Timeliness	√	√	√
Contextual Dimensions	Believability	√		√
	Completeness	√	√	√
	Understandability	√		
	Relevancy	√	√	√
	Reputation	√		
	Verifiability		√	
	Amount of Data	√	√	
Representational Dimensions	Interpretability	√	√	√
	Rep. Conciseness	√	√	
	Rep. Consistency	√	√	√
Accessibility Dimensions	Availability	√	√	√
	Response Time			
	Security	√		

Table 2.1: Information quality dimension catalogs proposed by authors with a management perspective.

Representational Dimensions capture aspects relating to the representation of information within information systems.

Accessibility Dimensions capture aspects involved in accessing information.

Table 2.1 shows an overview of the quality dimension catalogs proposed by authors who have a management perspective on information quality. The most comprehensive and most widely cited catalog from this group has been proposed by Wang and Strong [WS96]. The catalog was created empirically by asking information consumers which dimensions they consider most relevant. The catalog has successfully been used within information quality management projects in industry and government [WS96]. Jarke and Vassiliou's catalog was created by specializing Wang and Strong's dimensions for use within data warehousing projects [JV97]. The catalog therefore emphasises aggregated information. The other authors do not describe the methodology used to create their catalogs and it seems likely that they were created in an ad-hoc manner by listing dimensions which seem relevant in the context on which the authors focused.

Category	*Dimension*	*Chen*	*Mecella*	*Gertz*	*Eppler*
		[CZW98]	[MSV$^+$02]	[Ger96]	[EM02]
Intristic Dimensions	Accuracy	√	√	√	√
	Consistency		√		
	Objectivity				
	Timeliness	√	√	√	√
Contextual Dimensions	Believability				
	Completeness	√	√	√	
	Understandability				√
	Relevancy	√			√
	Reputation				
	Verifiability				√
	Amount of Data	√		√	
Representational Dimensions	Interpretability				
	Rep. Conciseness				√
	Rep. Consistency				√
Accessibility Dimensions	Availability			√	√
	Response Time	√			√
	Security				√

Table 2.2: Information quality dimension catalogs proposed for autonomous information sources.

Table 2.2 shows the quality dimension catalogs proposed by authors with a focus on assessing the quality of autonomous information sources. Compared to the catalogs proposed by authors from the first group, the catalogs from the second group contain less quality dimensions. The catalogs focus on intristic and contextual dimensions. Representational and accessibility dimensions are less prominent than in the first group of catalogs. A potential explanation for this observation might be that the authors consider representational and accessibility issues as solved by the assumed technical setting.

Table 2.3 shows the distribution of the proposed dimensions. Accuracy and timeliness are the most popular dimensions appearing in all catalogs. Completeness appears in six catalogs; relevancy and availability in five catalogs; all other dimensions in less than five.

The comparison of the different catalogs shows that the sets of quality dimension strongly reflect the application domain. The same is true for the definitions of specific dimensions [KB05]. Thus Richard Wang's finding from back in 1995, that "there is a lack of consensus, both on what constitutes a set of good data quality dimensions, and on what an appropriate definition

Dimension	*Count*
Accuracy	7
Timeliness	7
Completeness	6
Relevancy	5
Availability	5
Rep. Consistency	4
Amount of Data	4
Interpretability	3
Rep. Conciseness	3
Security	2
Objectivity	2
Believability	2
Understandability	2
Verifiability	2
Response Time	2
Consistency	2
Reputation	1

Table 2.3: Distribution of information quality dimensions.

is for each" [WSF95] still seems to be valid today.

The following sections refer to a definition from literature for each quality dimension and discuss the dimensions in relation to web-based information systems.

2.3.1 Intrinsic Dimensions

Intrinsic dimensions are independent of the user's context. They are capturing whether information correctly represents the real world and whether information is logically consistent in itself.

Accuracy is the degree of correctness and precision with which information in an information system represents states of the real world [WW96]. Accuracy is a very important quality dimension and early information quality research was mainly focused on this dimension [WS96]. Within information production processes inside organizations, accuracy can be improved by installing organizational procedures, like having information double checked by two independent people, or by technical means, like calibrating sensors or verifying customer contact information received through a website against an address database. The autonomy of information providers on the Web prevents the installation of binding procedures for assuring the accuracy of information. Instead of centralized procedures, the focus shifts to information providers convincing information consumers that their information is accurate by publishing information together with meta-information about the circumstances in which information was created [PS03]. Notice, that the concept of accuracy implies the assumption that information can be captured in an objective fashion. Thus, accuracy is not applicable to subjective information, like consumer's tastes or political views.

Timeliness is the degree to which information is up-to-date [KSW02]. Timeliness can be seen in an objective fashion, meaning that information represents the current state of the real world [Nau02]. Timeliness can also be seen as task-dependent, meaning that the information is timely enough to be used for a specific task [PLW02]. Timeliness is arguably one of the most important quality dimensions for Web information, for providing new information instantly is a major success factor of the Web, compared to newspapers or journals [Nau02]. The importance of timeliness led to the incorporation of timestamps into web documents [ISO03a] and the base protocol of the Web: HTTP [GBL99].

Consistency implies that two or more values do not conflict with each other [MSV$^+$02]. Information on the Web is likely to be inconsistent as it is provided by multiple information providers, which might use different procedures to capture information, have different levels of knowledge and different views of the world.

Objectivity is the extent to which information is unbiased, unprejudiced and impartial [PLW02]. The applicability of the objectivity dimension depends on the type of information. For instance, the height of a building can be measured objectively. The objectivity of other types of information, like product descriptions, might be influenced by the information provider's preferences or intentions. Objectivity overlaps with the concept of accuracy.

2.3.2 Contextual Dimensions

Contextual dimensions reflect the requirement that information quality must be considered within the context of the task at hand and the subjective preferences of the information consumer.

Relevancy is the extent to which information is applicable and helpful for the task at hand [PLW02]. Relevancy is an important quality dimension in the context of web-based systems, as information consumers are often confronted with an overflow of potentially relevant information. Approaches to assess the relevancy of web documents are used within web search engines, which sort documents according to their relevancy for a given query using a combination of hyperlink analysis [PBMW98] and information retrieval methods [Fer03].

Completeness is the degree to which information is not missing. Pipino et al. [PLW02] collect three different definitions for the completeness dimension: Schema completeness which is the degree to which entities and attributes are not missing in a schema; column completeness which is a function of the missing values in a column; and population completeness which refers to the ratio of entities represented in an information system to the complete population, e.g. "If a column should contain all 50 states of the US and it contains only 43, than we have population incompleteness" [PLW02]. Completeness is context dependent. For instance, a list of all German stocks will be regarded as complete by an investor focusing on Germany, but will not be regarded as complete by another investor who wants to get an overview of all European stocks.

Amount of Data is the extent to which the volume of data is appropriate for the task at hand [PLW02]. This means that information is of sufficient breadth and depth for the task of the information consumer. On the other hand, it also means that the information consumer is not overwhelmed by too much detail.

Understandability is the extent to which data is easily comprehended by the information consumer [PLW02]. Understandability is related to interpretability. Interpretability refers to technical aspects, for instance, whether information is represented using an appropriate notation, while understandability refers to the subjective capability of the information consumer to comprehend information.

Believability. The extent to which information is regarded as true and credible [PLW02]. Believability can be seen as expected accuracy [Nau02]. While accuracy refers to the verifiable precision with which information about the real world is captured by an information system, believability refers to trusting information without checking. Believability is, for instance, an important quality dimension for forecasts, which by nature cannot be verified in advance. Believability is subjective as there are different levels of cautiousness and different subjective procedures for deciding which information to believe.

Verifiability is the degree and ease with which the information can be checked for correctness [Nau02]. Related concepts are traceability and provability. For information which might be biased, verifiability plays an important role in the information consumer's decision whether to accept information [Nau02].

Offensiveness. A further subjective quality dimension that is important in the context of web-based systems but has not been considered in the catalogs discussed above, is the offensiveness of information. Information consumers can consider web content offensive for moral, religious, or political reasons. Out of the motivation to protect children from offensive web content, various approaches have been developed to restrict the access to web content regarded as sexually explicit or violent. Some of the approaches rely on information providers adding meta-information, like Internet Content Rating Association labels [Int05], to their content. Other approaches rely on analyzing content itself or use central authorities which maintain black lists of offensive websites.

2.3.3 Representational Dimensions

Representational information quality dimensions capture aspects like the conciseness, consistency, and interpretability of information.

Representational Consistency is the extent to which information is represented in the same format [PLW02]. Information found on the Web is in general not very consistently represented, as the web architecture does not restrict the representations of resources to certain formats and as the decentralization of the Web leads to numerous communities using different formats [JW04]. An example for the inconsistent representation of information on the Web is the parallel use of text formats like HTML, PDF, and Microsoft Word. The problem of inconsistent representation can be handled by employing wrappers and mediators to translate between different representations [Nau02]. An example of successful format mediation is Google's ability to index various text formats.

Representational Conciseness is the extent to which information is compactly represented [PLW02]. Information representation formats for the Web, like XHTML [Ste02], XML [BPSMM00] and RDF [Bec04b], do not emphasize conciseness but are rather verbose in order to be self-descriptive and unambiguous.

Interpretability is the extent to which information is in appropriate languages, symbols, and units, and the definitions are clear [PLW02]. Web technologies try to improve the interpretability of information by using self-descriptive formats, identifying objects and the terms used to describe these objects with globally unique identifiers, and by employing various schema languages to provide definitions for terms [JW04].

2.3.4 Accessibility Dimensions

Accessibility dimensions capture aspects involved in accessing information.

Accessibility is the extent to which information is available, or easily and quickly retrievable [PLW02]. Providing access to a huge number of information sources is the main success factor of the Web and improving the accessibility of information is the main motivator behind the standardization of web technologies. The accessibility of information is influenced by practical factors like time-of-the-day and week-dependent network congestion, worldwide distribution of servers, highly concurrent usage, denial-of-service-attacks, or planned maintenance interruptions [Nau02]. These practical accessibility problems can be tackled by replicating or caching information.

Response time measures the delay between submission of a request by the user and reception of the response from the system. The response time depends on the type and complexity of the request. In a Web setting, the response time also depends on unpredictable factors, such as network traffic, server workload, etc. As with accessibility, response times can be improved by locally replicating or caching information.

2.4 Summary

This chapter established the notion of information quality. Information quality is commonly seen as the fitness for use of information. This definition implies that information quality is task dependent and subjective. Information quality is a multi-dimensional concept. Which dimensions are relevant and which quality levels are required is determined by the task at hand and the subjective preferences of the information consumer.

The chapter presented seven catalogs of information quality dimensions from literature. Due to the different research context of their authors, the catalogs differ widely. Accuracy, timeliness, completeness, relevancy, and availability were identified as the most popular dimensions appearing in at least five catalogs. Afterwards, the different dimensions were discussed in relation to web-based information systems.

The next chapter gives an overview of different metrics that can be used to assess information quality dimensions. Afterwards, the concept of quality-based information filtering policies is introduced.