Data Quality: An Assessment of Information Quality Dimensions

Chandrashekar H N [1] and Dr. G Mahadevan[1]

[1]Research Scholar of Ph.D., Department of Computer Science, Rayalaseema University, Kurnool, India
[2]Principal of Annai College of Engineering, Kumbakonam, India

Abstract**:** *Nowadays, activities associated selections creating in an organization are predicated on information and data obtained from data analysis that provides numerous services for constructing reliable and correct method. As information area unit important resources altogether organizations the standard of information is vital for managers and operative processes to spot connected performance problems. Moreover, top quality information will increase opportunity for achieving prime services in a company. However, characteristic numerous aspects of information quality from definition, dimensions, types, strategies, techniques area unit essential to equip strategies and processes for up information. This paper focuses on systematic review of information quality dimensions so as to use at projected framework that combining data processing and applied math techniques to live dependencies among dimensions and illustrate however extracting knowledge will increase method quality.*

Keywords: *Information Quality Dimensions, Data Quality*

1. Introduction

In order to support organization's activity we must always design the method activity befittingly since this involves data. Information is the first foundation in operational, tactical and choices creating activities. As information area unit crucial resources in all organizations, business and governmental application, the quality of knowledge is vital for managers and operational processes to spot connected performance problems.

There are a unit form of information quality problems from definition, measurement, analysis and improvement that area unit essential for making certain high information quality. Because the varied analysis shows if the process's quality similarly as information's inputs are not controlled, once a minute, the degradation of the information quality is obvious. For up process's quality with increased potency in production and administration, mistreatment method style is necessary for automation and management technology. It is well known that each of them gift, services to business and individual user quickly and consistency. Despite availability of huge selection of techniques for accessing and improving information quality like business rules, record linkage and similarity measures, because of rise problem and multiplicity of mistreatment these systems, information quality methodology has outlined and provided. Data quality will offer varied services for an organization similarly as these days, prime quality information will increase chance to realize high services in an organization. Likewise, lack of knowledge quality in organizations can be multiply within the Cooperative system (CIS). In fact, (CIS) is a system with capability to distribute and share general objective between interconnect completely different systems among of varied freelance organization in several geographic area because the information is basic recourses for it. Some researchers have known and tested some effecting factors on information quality within a company with collecting information from survey and interview with senior manager and their results show that management responsibility like commitment up information quality continually, effective communication among neutral and understanding of knowledge quality area unit vital components for influencing information quality in a company. the remainder of this paper discusses on information quality methods and techniques, forms of information, information quality definitions, data quality issues classification and information quality dimensions to provide, basic problems during this field.

2. Data Quality Strategies And Techniques

There are 2 sorts of ways that one tailored for improving knowledge quality particularly data-driven and process-driven, and every strategy employs numerous techniques.

However, up the standard of information is that the aim of every technique.

2.1. Data-Driven

Data-driven is strategy for up the standard of knowledge by modifying the information price directly. Some connected improvement techniques of data-driven are: acquisition of new data, standardization, error localization and correction, record linkage, data and schema integration, supply trait, similarly as price optimization.

2.2. Process- Driven

Process-driven is another strategy that redesigns the process that is created or changed information so as to improve its quality. Process-driven strategy consists of two main techniques: method management and method design. In fact, within the method management information are check and manage among the producing method, whereas within the method redesign the causes of inferiority are eliminated and new method are added in to manufacturing high quality. what is more, adding Associate in Nursing activity that may management format of knowledge before storage is another reality within the method redesign.

However, the benefits of Process-driven are best performing than Data-driven techniques in long amount, because they take away root causes of the standard issues completely. In distinction, Data-driven is dear than Process-driven in long amounting however it's economical in brief period

3. Types Of Data

Data square measure world objects, with ability of storing, retrieving and elaborating through a computer code method and can communicate via a network. Researchers have provided totally different classification for knowledge in several spaces. As implicitly or expressly, 3 sorts of knowledge square measure delineate in the field of DQ. Table 1 present's sorts of knowledge primarily based on this classification.

A second classification of information is predicated on considering data as a product, this model classify knowledge in to a few varieties. Table 2 shows this classification.

Table I: Type of Data as a Implicitly or Explicitly

Types of Data	**Definition**	**Example**
Structured data	Generalization or aggregation of items described by elementary attributes defined within a domain	Relational tables Statistical Data
Unstructured data	A generic sequence of symbols, typically coded in natural language.	Body of an Email Questionnaire with free text answering
Semi structure data	Data that have a structure with some degree of flexibility.	Mark up language, XML

Table II: Type of Data as a Product

Types of Data	Definition
Raw data items	Smaller data unites which are used to create information and component data items
Component data items	Data is constructed from raw data items and stored temporarily until final product is manufactured
Information products	Data ,which is the consequence of performing manufacturing activity on data

Another classification of information is based mostly on strictness to measure and to attain knowledge quality that has 2 category specifically elementary knowledge and mass knowledge. In an organization, knowledge that managed by operational method and represent atomic phenomena of the important world are known as elementary knowledge, (e.g., sex, age), whereas knowledge that are collected from elementary knowledge for applying aggregation function, is named mass knowledge, (e.g., average financial gain that tax money dealer paid in an exceedingly specify city).From purpose of read, data will be classified in numerous sorts supported their usage in sort of field (e.g., network or web).

4. Data Quality Definitions

Data quality has completely different definition on different field and period. Researcher of science and professional created totally different understanding about information quality. In keeping with quality management information quality is acceptable to be used or to satisfy user wants or its quality of information to satisfy client wants. Also, another definition for information quality is fitness to be used. Indeed, quality of data is important for improvement method activity because it is often addressed in numerous field as well as management, medicine, statistics and engineering science. The widespread assortment of definition through information quality might provide chance to higher understand the character of information method.

5. Data Quality Problems Classification

Data quality drawback typically is divided in to two classes that area unit single-source and multi-source drawback. According to some analysis four classes for information quality are known that area unit shown because the following table. As a result, the goal of classifying information quality drawback is illustrating non-standard information and distinctive actual application of knowledge for corresponding necessities.

Table III: Data Quality Problem Classification

Data Quality problem	Category	Definition
Single -source problem	Schema level	Lack of integrity constraints, poor schema designer Uniqueness constraints Referential integrity
	Instance level	Data entry errors Misspelling Redundancy Duplicates Contradictory values
Multi-source problems	Schema level	Heterogeneous data models and schema Design Naming Conflicts
	Instance level	Overlapping contradicting and inconsistence data Inconsistent aggregating Inconsistent timing

6. Data Quality Dimensions And Definition

Table IV illustrate some knowledge quality dimensions and their definition from literature. From the analysis perspective, there is varied numbers of dimensions for data Quality and knowledge quality. In fact, "Data Quality", "Information System" and "accounting and auditing" area unit three initial classes for distinctive correct DQ dimensions. within the field of knowledge Quality ,Wang determined four classes that area unit Intrinsic DQ, Accessibility DQ, Contextual DQ, naturalistic DQ and fifteen dimensions for DQ/IQ (e.g., judgment, credibleness, reputation, worth added). Alternative man of science recognized further dimensions for DQ such as knowledge validation, believability, traceability, convenience for distinctive. Within the space of Information Systems, man of science has known various factors such as reliability, precision, relevancy, usability, and independency. Within the accounting and auditing, researcher explained that accuracy, timeliness and connectedness area unit 3data quality dimensions. Additionally, during this space some scholars explained that control systems want lowest cost and high reliability that refers to some dimensions such as accuracy, frequency and size of knowledge. Base on the ISO customary, quality means that the totality of the characteristics of associate in nursing entity that bear on its ability to satisfy explicit and tacit wants. Knowledge Quality Dimension could be a characteristic or a part of information for classifying data and knowledge requirements. In fact, it offers some way for mensuration and managing knowledge quality yet as data.

So, primary step for understanding knowledge quality dimension will facilitate United States to boost it. Instrument and developer use dimension and taxonomy of separate knowledge via exploitation knowledge quality tools for making and manipulating the knowledge in order to boost data and its method.

Table IV: Data Quality Dimensions

Dimension	Definition
Timeliness	The extent to which age of the data is appropriated for the task at hand. Timeliness refers only to the delay between a change of a real world state and the resulting modification of the information system state. Timeliness has two components: age and volatility. Age or currency is a measure of how old the information is, based on how long age it was recorded. Volatility is a measure of information instability the frequency of change of the value for an entity attribute.
Currency	Currency is the degree to which a datum is up-to-date. A datum value is up-to-date if it is correct is spite of possible discrepancies caused by time-related changes to the correct value. Currency describes when the information was entered in the sources and/or the data warehouse. Volatility describes the time period for which information is valid in the real world.
Consistency	The extent to which data is presented in the same format and compatible with previous data. Refer to the violation of semantic rules defined over the set of data.
Accuracy	Data are accurate when data values stored in the database correspond to real-world. The extent which data is correct, reliable and certified. Accuracy is a measure of the proximity of a data value, v, to some other value, v', that is considered correct. A measure of the correction of the data (which requires an authoritative source of reference to be identified and accessible.
Completeness	The ability of an information system to represent every meaningful state of the represented real world system. The extent to which data are of sufficient breadth, depth and scope for the task at hand. The degree to which values are present in a data collection. Percentage of the real-world information entered in the sources and/or the data warehouse. Information having all having all required parts of an entity's information present. Ratio between the number of non-null values in a source and the size of the universal relation. All values that are supposed to be collected as per a collection theory.
Accessibility	Extent to which information is available, or easily and quickly retrievable.
Duplication	A measure of unwanted duplication existing within or across systems for a particular field, record, or data set.

Data specification	A measure of the existence, completeness, quality and documentation of data standards, data models, business rules, meta data and reference data.
Presentation Quality	A measure of how information is presented to and collected from does how utilize it. Format and appearance support appropriate use of information.
Consistent Representation	To extend to which data is presented in the same format.
Reputation	To extent to which information is highly regarded in terms of source or content.
Safety	It is the capability of the function to achieve acceptable levels of risk of harm to people, process, property or the environment.
Appropriate amount of data	To extend to which data volume of data is appropriate for the task at hand.
Security	Extent to which access to information is restricted appropriately to maintain its security.
Believability	Extent to which information is regarded as true and credible.
Understandability	Extent to which data are clear without ambiguity and easily comprehended. To extend to which data is easily comprehended.
Objectively (Objectivity)	Extent to which information is unbiased, unprejudiced and impartial.
Relevancy	Extent to which information is applicable and helpful for the task at hand.
Effectiveness	It is the capability of the function to enable to users to achieve specified goals with accuracy and completeness in a specified context of use.
Interpretability	To extend to which data is appropriate languages, symbols, and units and the definition are the clear.
Ease of Manipulation	To extend to which data is easy to manipulate and apply to different same format.
Free-of –error	To extend to which data is correct and reliable.
Ease of Use and maintainability	A measure of the degree to which data can be accessed and used and the degree to which data can be updated, maintained, and managed.
Usability	To extent to which information is clear and easily used.
Reliability	Extent to which information is correct and reliable. It is the capability of the function to maintain a specified level of performance when used on specified condition.
Amount of data	To extent to which the quantity or volume of available data is appropriate.
Freshness	Freshness represents a family of quality factors which each one representing some freshness aspect and having on its metrics.
Value added	To extent to which information is beneficial, provides advantages from its use.
Learn ability	It means the capability of the function to enable to user to learn it.
Data Decay	A measure of the rate of negative change to data.
Concise	Extent to which information is compactly represented without being overwhelming (i.e. brief in presentation, yet complete and to the point).

Consistency and Synchronization	A measure of the equivalence of information used in various data stores, applications, and systems, and the processes for making data equivalent.
Data integrity fundamentals	A measure of the existence, validity, structure, content, and other basic characteristics of the data.
Navigation	Extent to which data are easily found and linked to.
Useful	Extent to which information is applicable and helpful for the task at hand.
Efficiency	Extent to which data are able to quickly meet the information needs for the task at hand.
Availability	Extent to which information is physically accessible.
Data Coverage	A measure of the availability and comprehensiveness of data compared to the total data universe or population of interest.
Transactability	A measure of the degree to which data will produce the desired business transaction or outcome .
Timeliness and Availability	A measure of the degree to which data are current and available for use as specified and in the time frame in which they are expected.

7. Discussion

Most people suppose the quality of knowledge is depended solely to its accuracy and that they don't think about and analyze alternative significant dimensions for achieving higher quality. Indeed, quality of knowledge is over considering one dimension thus, the issue of dimensions dependencies is important to improve method quality in numerous domain and applications. Yet, while not knowing the present relations between knowledge quality dimensions, knowledge discovery cannot be effective and comprehensive for decision making method. From previous work detected, not solely dimensions are often powerfully associated with one another but also, knowledge qualities are often supported via the effective dependencies. In fact, choose acceptable dimensions with distinguishing correlation among them will produce high quality knowledge. So as to get dependencies among additional commonly documented dimensions incorporates accuracy, currency, consistency and completeness, we have a tendency to projected framework that combining data processing and applied mathematics techniques to live dependencies among dimensions and illustrate however extracting data will increase method quality. So, supported our hypothesis if there's a correlation between completeness, consistency and accuracy dimensions that thought-about variable and then, think about currency correlation as variable among them, improvement in knowledge quality are going to be happened. Also, reason behind some difficulties on currency dimension the policy is needed.Fig.1 illustrate projected framework for evaluating the effect of freelance dimensions on dependent dimensions.

So, the aim of the projected framework is discovering the dependency structure for the assessed information quality dimensions.

8. Conclusion

From the angle analysis, several students have identified varied methodology and framework for assessing and up knowledge quality through completely different techniques and strategies on the info quality dimensions. They illustrated definitions for dimensions and known additional necessary data quality dimensions. Existing survey known forty knowledge quality dimensions since 1995 until 2015. Since, some dimensions like timeliness, currency, accuracy and completeness area unit additional documented than others, the results of this survey can be wont to realize correlations among knowledge quality dimension supported planned framework with combining data processing and applied mathematics techniques for measurement dependencies among

them and illustrate however method quality can be augmented via the extracting information. Specifically, our future work would be to judge dependency among mentioned knowledge quality dimensions for up method quality.

References

[1] Y. Man, L. Wei, H. Gang, G. Juntao, "A noval data quality controlling and assessing model based on rules," 2010, pp. 29-32.

[2] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations," Communications of the ACM, vol. 39, pp.86-95, 1996.

[3] KQ. Wang, SR. Tong, L. Roucoules, B. Eynard, "Analysis of data quality and information quality problems in digital manufacturing," 2008, pp. 439-443.

[4] M. Heravizadeh, J. Mendling, M. Rosemann, “Dimensions of business processes quality (QoBP)," 2009, pp. 80-91.

[5] D. McGilvray, Executing data quality projects: Ten steps to quality data and trusted information: Morgan Kaufmann, 2008.

[6] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," Journal of management information systems, vol. 12, pp. 5-33, 1996.

[7] M. Bovee, R.P. Srivastava, B. Mak, "A conceptual framework and belief‐function approach to assessing overall information quality," International journal of intelligent systems, vol. 18, pp. 51-74, 2003.

[8] T. C. Redman, Data quality for the information age: Artech House, 1996.

[9] M. Jarke, Fundamentals of data warehouses: Springer Verlag, 2003.

[10] D. P. Ballou and H. L. Pazer, "Modeling data and process quality in multi-input, multi-output information systems," Management science, pp. 150-162, 1985.

[11] F. Naumann, Quality-driven query answering for integrated information systems vol. 2261: Springer Verlag, 2002.

[12] L. Liu and L. Chi, "Evolutionary data quality," 2002.

[13] L. L. Pipino, Y.W. Lee, R.Y. Wang, "Data quality assessment," Communications of the ACM, vol. 45, pp. 211-218, 2002.

[14] S. Knight and J. Burn, "Developing a framework for assessing information quality on the World Wide Web,” Informing Science: International Journal of an Emerging Transdiscipline, vol. 8, pp. 159-172, 2005.

[15] V. Peralta, "Data quality evaluation in data integration systems," Université de Versailles (chair) Raúl RUGGIA Professor, Universidad de la República, Uruguay, 2008.