10-2013

HDQM2: Healthcare Data Quality Maturity Model

Javier Mauricio Pinto-Valverde  
*Tecnológico de Monterrey*, mauricio.pinto@xenxar.com

Miguel Ángel Pérez-Guardado  
*Tecnológico de Monterrey*, maperez@itesm.mx

Lorena Gomez-Martinez  
*Tecnológico de Monterrey*, lgomez@itesm.mx

Martha Corrales-Estrada  
*Tecnológico de Monterrey*, mcorrales@itesm.mx

Juan Carlos Lavariega-Jarquín  
*Tecnológico de Monterrey*, lavariega@itesm.mx

Follow this and additional works at: http://scholarworks.wmich.edu/ichita_transactions

Part of the Health Information Technology Commons

WMU ScholarWorks Citation

Pinto-Valverde, Javier Mauricio; Pérez-Guardado, Miguel Ángel; Gomez-Martinez, Lorena; Corrales-Estrada, Martha; and Lavariega-Jarquín, Juan Carlos, "HDQM2: Healthcare Data Quality Maturity Model" (2013). Transactions of the International Conference on Health Information Technology Advancement. 37.  
http://scholarworks.wmich.edu/ichita_transactions/37
HDQM\textsuperscript{2}: Healthcare Data Quality Maturity Model

Javier Mauricio Pinto-Valverde, Miguel Ángel Pérez-Guardado, Lorena Gomez-Martinez, Martha Corrales-Estrada, Juan Carlos Lavriega-Jarquín
Tecnológico de Monterrey
Av. E. Garza Sada 2501, Monterrey NL, Mexico
Telephone Number: +52 (81) 83581400 ext. 5250

mauricio.pinto@xenxar.com, maperez@itesm.mx, lgomez@itesm.mx, mcorrales@itesm.mx, lavariega@itesm.mx

Abstract: Data Quality is a central issue within the development of Healthcare Systems, for both the delivery of the service as well as for the establishment of public policies. Public and private Healthcare Systems in Latin America do not have a framework of reference which allows them to have knowledge of the level of Data Quality (DQ). In other frameworks found within literature the evaluation of DQ has been centered within the dimensions. The intention of this research is to build an evaluation model for DQ based on maturity. This research contributes an innovating and comprehensive model, based on the data life cycle and maturity models, applied to the Health Sector, mainly because it contributes to the switch from the traditional physical mediums to Electronic Medical Records. This article presents the advancements of the research that is being conducted in some hospitals of the Healthcare Cluster of Monterrey, Mexico.

INTRODUCTION

There has been an increase in the interest for Data Quality because during the last fifty years the economy has centered around information as one of its main resources (Eckerson, 2002). According to the TDWI (The Data Warehouse Institute), intellectual capital, which is built from data and information, and know-how are now more important assets than tangible infrastructure is (Eckerson, 2002).

The growth of health organizations has gone hand in hand with the growth in collected, transported and stored data. This has created problems and new challenges which allow securing that the most important asset—the data—has an acceptable level for the delivery of health services. DQ can degrade quickly as its volume grows and moreover with the uncontrolled exchange allowed by the interconnectivity of the Infosphere.

Information is the product which will allow for decision making within a determined context. In other words, both Data as well as Information must satisfy (or exceed) the requirements of those making their use. J. M. Juran states that there could be a high level of DQ "if the same adjusts to what is expected from its use within operations, decision making and planning" (Juran, 1964).

The area intended to be addressed in this research is the Assessment of DQ. Within the revised literature there was a reoccurrence with a key element for the Assessment of DQ: the dimensions of DQ. The dimensions are criteria which qualify an aspect of the group of data to be evaluated. This aspect can be intrinsic and extrinsic to the data and is instrumented through one or more metrics associated with each dimension (Batini & Scannapieco, 2006). In the literature there is an extensive list of dimensions to be considered, however, there is a consensus with regards to the main dimensions and these are: Accuracy, Completeness, Timeliness and Consistency (Batini & Scannapieco, 2006).

DQ is a process of continuous improvement and not a project, and its evaluation must be comprehensive, considering all of the factors affecting it and showing its evolution over time. This is the reason why this research intends to build an assessment model for DQ based in the maturity of its capabilities, given that these models have gotten excellent results in industries such as Software and is widely accepted within the academic community. Philip Crosby was the first to propose a model for maturity management (Crosby, 1980). Afterwards these ideas were adopted by IBM and SEI for quality within Software Engineering, an area where they had wide acceptance and which still continues evolving. Within the literature, there has been found a number of references of maturity models applied to DQ and no reference specifically related to data management within the Healthcare area.

The construction of the maturity model for DQ will be conducted taking as basis models found within literature, considering the peculiarities of the Health Sector and the characteristics of the local context. It is expected that the model, once validated, can serve as a reference and starting point for other DQ evaluation initiatives within national or international Health entities.
This research presents the progress achieved within the Framework Construction for Data Quality Evaluation, based on the Maturity Model approach.

For the construction and validation of this model we will have the participation of the hospitals which form the Healthcare Cluster of the State of Nuevo León in Mexico.

MOTIVATION

Health is a current and critical topic within the development of the communities and constitutes as an indicator of the development of a country (SECRETARIA DE SALUD, 2012).

The low quality of DQ within the health sector has already generated problems (Al-Hakim, 2007) (SECRETARIA DE SALUD, 2012). In developed countries, according to the WHO, one of every ten patients is affected while being hospitalized; it is possible for the medical errors to be related, directly or indirectly, with the level of DQ.

One of the factors which contributes to the mitigation of risks within the Medical Attention Service, is the implementation of Electronic Health Record (EHR), which is the evolution of the traditional medical file in a physical format (SECRETARIA DE SALUD, 2012). The Health System in Mexico, although lagging with regards to the EHR in other developed countries, has made significant efforts within the public and private sector for the implementation of this technology which will noticeably improve the Health service.

There is already a regulation of the Mexican State, through a norm (SECRETARIA DE SALUD, 2012) which pretends to set the foundation for the implementation and standardization of the EHR within public and private health entities. As is mentioned in the justification to the implementation of this norm of EHR, problems have been found in DQ related with Health service processes. These problems would mean a non-quantified expense for medical errors, rework, inefficiency in the delivery of States’ resources to the Health System and degradation within the medical service. Without a doubt this will have an impact over the Quality of Life of all Mexicans.

The norm establishes the DQ dimensions which are expected to be fulfilled by the EHR. The norm is a new official regulation, promulgated in 2010. Given that this is the first time trying to regulate the topic of the EHR, there are several loopholes with regards to the instrumentation of the norm, which only mentions the “what” but does not define the “how”. One of these loopholes is the one regarding the tool that will serve to conduct the DQ evaluation. This lack of instrumentation is an area of important opportunity which this research intends to include with the construction of a Framework, which allows to take cognizance of the level of maturity of the DQ and to size the existent gap in order to mitigate it in the future.

The DQ evaluation has been conducted, until now, in aspects inherent to the data, in other words only considering the qualification criteria of the group and its syntactic and semantic intrinsic characteristics, without considering the level of maturity of the practices guided towards achieving a high level of DQ.

In order to provide an example of the low DQ problem within the Health area, a group of data is presented (see Table 1), which corresponds to some variable taken from the medical records of “Patients”. This example is a typical dataset which can be found in a hospital that has not established an evaluation and improvement mechanism for DQ.

\[
\begin{array}{|c|c|c|c|c|}
\hline
\text{Nº} & \text{Gender} & \text{BirthDate} & \text{BloodGroup} & \text{Gravida} & \text{Allergies} \\
\hline
1 & M & 05/09/1979 & A+ & 0 & PENICILINA \\
2 & M & 04/05/1987 & O- & 0 & NO \\
3 & M & 03/12/1956 & B+ & NINGUNO & NO APICA \\
4 & M & NULL & A Rh+ & 1 & ANESTESICOS \\
5 & M & 05/04/1989 & AB- & NO APICA & NO APICA \\
6 & F & 24/05/1988 & NULL & 1 & NULL \\
7 & M & 19/04/1979 & NULL & 0 & PENCIL \\
8 & NULL & 05-mar-76 & O- & 3 & ESTEROIDES \\
9 & F & 12-34-5678 & B NEG & 2 & NULL \\
10 & NULL & NULL & NULL & 0 & NO \\
\hline
\end{array}
\]

Table 1. Low Data Quality Recordset
From the analysis of this group of data from the perspective of the DQ Dimensions, we can evidence a series of deviations which can mean low DQ. These deviations can go from what is simple: incomplete cells, to what is complex: inconsistent information.

Even though all of the variables of the Table 1 example are critical, some can originate errors that could affect human lives. Such is the case with the "Blood Type" that a case of emergency could result in a delay in the treating of a patient due to not having the specific type, or even worse, administering an incompatible blood type. Both are examples that the level of data within one single variable can result in fatal errors.

Each element which contributes and impacts DQ must be analyzed and placed over a maturity scale. This will help to have a benchmark for each element that comprises the evaluation model for DQ.

This causes motivation to search for a manner to evaluate DQ where the maturity level held by the health institutions within the practices, processes and the application of tools and methodologies for DQ evaluation can be established. The knowledge of DQ maturity level would give certainty with regards to the governing of this data within the health entities. The proposed DQ maturity level starts from the scenario that an information system is composed of basic and fundamental elements for its characterization. These are: People, Processes, Data and Technology (Figure 1).

With this context we can represent that:
- DQ is a current and relevant topic for scientific research.
- A low level of DQ is cause of problems for people, organizations and society.
- There is evidence that some problems within medical attention are related with low DQ level.
- No DQ Evaluation model exists for the Health Sector of Mexico or Latin America.
- According to the revised literature there is a need for DQ evaluation models to be validated in a more rigorous manner and that more study cases be developed in order to give them validity.
- DQ evaluation that considers only characteristics of data is insufficient and there is a need for a holistic vision of the DQ problem.

**Figure 1. Holistic Vision of DQ**

**BACKGROUND**

The concept of Quality has been used in different areas of knowledge in order to assure a better quality of life for society. Norm ISO 9000 defines Quality as “the degree in which a group on inherent characteristics fulfills the requirements”. In order to state that a service or product is of Quality, an evaluation has to be conducted, which allows us to have knowledge of the degree of deviation of the expected characteristics, to later analyze the causes and execute preventive or corrective actions.

Having a deficit within data quality, as with any resource of the company, would provide a reason to not achieve the objectives in a greater or lesser degree. Fisher states that there exists a clear relation between the achievement of the objectives of a company and how they manage their data (Fisher, 2009).

The awareness that the data consists in an important asset for the organization forces the adoption of the necessary measures in order to secure their governance and quality. On the contrary, there would exist an
uncertainty that could be the root of difficult problems to explore and that depending on their urgency could go from simple to fatal.

**Data Quality**

According to Juran: “Data is of high quality if the same fulfills the purpose of its use within operations, decision making and planning” (Juran, 1964).

English, on the other hand, suggests that the fitness for use concept of the date is limited, because the data could be useful not only for the “desired” use, but also for other potential uses (English, 1999).

McGilvray extends this concept and suggests DQ as the level or degree of trust offered by the data, for the use or uses that want to be taken advantage of (McGilvray, 2008). In a simple manner “correct data, at the right place, at the right time, by the right person and which is used for the better performance of the business”.

An important point to highlight is that DQ is related with its use and exploitation. So, as such, it cannot exclude the data consumer, because the same forms part of the evaluation and improvement (Strong, Lee, & Wang, 1997). To this it can be added that the consumer is part of the evaluation and its system, in other words, the organization.

Redman adds that to comply with the quality of fitness for use the data must be accessible, exact, timely, complete and consistent with other relevant and comprehensive sources, they must provide a level of detail that is easy to read and interpret (Redman, 1997).

In this definition, reference has already been made to the multi-dimensionality of DQ, given that there are no sole criteria to be able to define it. Emphasis is made that DQ is more than just accuracy and is not only a technological problem.

With this context we can conclude that DQ:
- Is multi-dimensional.
- Reflects the level of trust attributed to a group of data by its consumers.
- Is the degree of satisfying or exceeding the requirements of a Data consumer.
- Is a holistic concept which not only includes data but also its consumers.
- Is not static and can change through the time and usage of the data.
- Not only relates to accuracy, it must consider the dimensions which are adequate to each specific problem.
- Is not only a technological problem or of processes.
- It is evidenced when somebody accesses data which has a potential use, for a useful purpose.

**Data vs. Information vs. Knowledge**

Data consists on the representation of event of diverse nature such as text, sounds, images, etc. Data represents an abstraction of the real world, limited to the characteristics or properties of interest (Dama International, 2009).

The data on its own is not more than a symbolic representation and in order for the same to have meaning it must have a context. From the data as an atomic unit, Information and Knowledge are built. This hierarchic relation makes DQ to be reflected at higher levels.

Information is the product of data processing within a determined context, turning it into a decision making instrument. Knowledge if Information put into perspective, with the acknowledgement, understanding and familiarity of a situation with its complexity (Dama International, 2009). Knowledge is constructed with basis on the significance of the information and is also a valuable asset and resource for organizations.

**Maturity Models**

The first maturity model was suggested by the quality guru Philip Crosby in his book “Quality of Free” (Crosby, 1980). He declared it as the “Quality Management Maturity Grid” (QMMG). It came from ample experience and livelihood within the implementation of quality management systems. At that time, companies were reluctant to the adoption of quality systems because they perceived it as a pointless expense of resources and in the case of there being any benefits they came in the long term.
One of the contributions of the Crosby model is that it can be used as a self-evaluation of the current situation which will serve as reference to know the gap dimension of the desired maturity state. The possibility of seeing the maturity improvement over time is what gave it its recognition in front of the already known quality systems. This model is a table, as a manner of rubric which describes the five stages of maturity. These are: 1. Uncertainty, 2. Awakening, 3. Enlightenment, 4. Wisdom and 5. Certainty.

This model is generic and is used as the basis for all types of maturity models. Due to its acceptance and usefulness, it is without a doubt the cornerstone for the conduction of this research.

At the end of 1986 the US Government made a request to Software Engineering Institute for the development of a maturity model for software processes. Later this model evolved to the SW-CMM.

The CMM model is founded over a group of key practices and processes, included in Key Process Area (KPA). The KPAs are added within five levels of maturity, which qualify their evolution in the adoption of better practices. The qualification is obtained when demonstrating that the organization has adopted and institutionalized all of the practices of the corresponding level and of the inferior levels.

The levels of the CMM model are: 1. Initial, 2. Repeatable, 3. Defined, 4. Managed and 5. Optimizing.

METHOD

The research will be conducted in three main phases (Figure 2).

Exploratory Phase

The first activity of this phase is the revising of the Literature, which will provide all of the theoretical basis and the foundation for the construction of the model. Although it is located at the beginning, this activity is continuous.

The focus group will involve the personnel of the hospitals in the following areas:

- People: Human Resources, Training.
- Processes: Responsible of Quality and Continuous Improvement
- Medical Area: Medical Director
- Technology: CIO, Responsible of TIC’s.

Another activity that is important is the evaluation of the DQ in datasets obtained from the health entities that participate in the research. There has not been an antecedent methodology detected that allows the measurement of the level of DQ in Mexican hospitals, and this prevents having an objective measurement for the current level of DQ. Knowing the level of DQ in the exploratory phase reinforces the motivation behind this investigation and it helps detect areas of opportunity in the life cycle of Data.

For the purpose of this investigation, information will be taken from demographic data as well as identification of the patients that are considered “Clinical Information” according to the Official Mexican Norm (SECRETARIA DE SALUD, 2012)
Construction of the Model

This is the central part of the project because it will give as a result the model as well as the generation of the theory. For the first activity a hybrid research strategy of Action-Research and Design Science (AR+DS) will be used, inspired in Lee’s framework published in (Kock, 2007).

**Constructs:** Domain vocabulary.
**Model:** Group of prepositions or statements which express the relation between constructs.
**Method:** Group of steps to execute a task.
**Instantiation:** The realization of an artifact within its surroundings.

Form the research activities are expected:
- Process Areas of the Data Quality Maturity Model
- Indicators for each Process Area
- Evaluation criteria for the indicators (Appraisals)

The theory is obtained from the evaluation of the model and the results of the instantiation of the artifact. It is expected that part of the theory be generalized to other areas.

As the second activity, a technic of interviews will be used to conduct the mapping of the model to the criteria of “Value Creation”. The interviews will be conducted to the personnel of strategy and tactics. The goal is to cognizance the impact that maturity has over DQ within value creation for health entities.

Finally, an instantiation will be conducted within a selected unit, where the model will be applied and all of the evidence for the case study will be generated. At the end of the application of the model we will have the maturity evaluation of the entity where the pilot test will be held. This pilot test will be carried out using an audit guide, which will be part of the model as well as a concept test of an artifact of software for the automation of the DQ maturity evaluation.

RESULTS

**Data Quality Assessment**

To identify the dimensions there was a focus group used where the evaluation criteria was defined. Here there was a description of the principal dimensions that appear in DQ literature defined and afterwards there was a tool used to obtain the following dimensions:

- A. Accuracy/Correctness
- B. Completeness
- C. Uniqueness
- D. Duplicates

Although there was success in the consistency of the dimensions selected in the hospitals where the cases were studied, it was noted that in the second case there was a mention of duplicated registered patients as an important problem to detect. This finding was due to empirical and anecdotal evidence of problems in the service due to duplicated records.

In consulting the root cause of these problems, there was no establishment of a unique cause for the problem of evidence to support it. There was also no evidence to be accounted for to confirm and establish the quantity of duplicated records.

In order to evaluate this establishment, clustering techniques were used in the total of the records given, which totaled approximately 600K records. The rate of duplicated records can be confirmed with the records that were analyzed. In presenting the results to the focus group, there was a consensus that the level found was no admissible for the hospital (the confidential agreement does not allow to report the rate of duplicates).

Deviations of other measurements evaluated were also found like:
- Completeness: incomplete e-mail addresses, incomplete addresses, incomplete dates of birth.
- Correctness: incorrect e-mail addresses, incorrect dates of birth, incorrect geographical names

The results were presented to the participating hospitals and there was a general consensus about the results which were also shared with two experts in order to validate them.
Metamodel

From the elements of the Research model, a metamodel was obtained that will help with the conceptualization of the framework, as well as for the construction of the artifact with which concept tests of the Maturity Model will be conducted. This model is formed by the following elements: Levels of Maturity, Practices, Process Areas and Value Creation. The Metamodel was represented using the UML notation (see Figure 3).

For areas of process of the Maturity Model the phases of data life cycle were considered (McGilvray, 2008). The phases were presented in the focus groups and there was a successful consensus in validating the health service delivery. All the phases are presented within the chain value of the health service.

The consensus regarding the maturity model for DQ was that it should be continuous and not staggered due to the fact that it can have different levels of maturity during different phases of the data life cycle.

Focus Group

Very interesting and revealing data was obtained from the focus group during the exploration phase. The most important findings are as follows:

Data
• There is awareness of the value of the data as a main asset.
• There is evidence of low quality data within the health entities.

People
• There is awareness that DQ is not the solely responsible in the Information Technologies area.
• There is awareness that people can contribute to the improvement of DQ.
Processes
• There is no procedure to share data among hospitals.
• There are processes for some of the phases of the Data Life Cycle.

Technology
• Technology helps to reduce the potential errors of DQ.
• There are errors due to processes which are magnified with the use of Information Technologies.

DISCUSSION AND CONCLUSION
The relevance of data for the Health Sector during the advancement of the research has been evident, both for the clinical processes as well as for the establishment of public policies.

The research contributes an innovating and comprehensive model, based on the data life cycle and maturity models, applied to the Health Sector, particularly because the same contributes to the switch from the traditional physical mediums to Electronic Medical Records.

The contribution of this research is the design and validation of metrics within focus groups to calibrate the evaluation process within value creation in terms of compliance with the norms of world class practices, of sustainability of the quality of life within Health, and in the economic benefits derived from the elimination of reprocesses and deviations within clinical diagnostics and in support of decision making through Health Intelligence.

There is an overwhelming amount of existing evidence that DQ in registered patients contains deviations and errors that should be evaluated and improved. If improvement does not happen, these problems could generate adversity in the provided health service.

It was evidenced within the focus groups that there are problems in all perspectives (people, processes, technology and data) which could be the cause for low DQ.

The Health Sector (public and private) does not count with the capabilities for the evaluation of the DQ level with which it operates, so there is uncertainty with regards to the level they have. Because there is no mechanism of evaluation, there will not be one to improve.

LIMITATIONS AND FUTURE WORK
The research is concentrated over some institutions from the Health Cluster of the State of Nuevo Leon; the inclusion of more private and public institutions will contribute with findings which will sustain an external generalization of greater trust.

The Health institutions in Mexico have not adopted a culture in terms of certifying their processes of:
• Data Quality
• Data Life Cycle Maturity
• Holistic Models for DQ Assessment.

The research for the plan of the construction and refining of the model will continue. An instantiation of the artifact to an application to allow for the automation of the evaluation activities and that show the results in a graphic manner will be conducted afterwards.

For the implementation of the evaluation of DQ, identification and demographic data, which partake in the collection of information in a clinic, was taken into account. It is suggested that the research is extended to clinical data that details cases of medical service delivery. It was found that part of this evaluation is performed by the evaluation committee (EHR committee) or the area of health quality.

In fact, due to the volume of the information, there is a need for an analysis to determine what strategy to follow. The restrictions of this suggestion are given by the limited access to these types of data since it is not easy to access all of the EHR databases. To further solidify the evaluation of the capability of maturity of DQ, it is suggested that long-term studies are done to verify the casualties and the impact in the long-run of adopting DQ practices.
REFERENCES


