MMPRO: A METHODOLOGY BASED ON ISO/IEC 15939 TO DRAW UP DATA QUALITY MEASUREMENT PROCESSES

(Research-in-Progress)

Ismael Caballero, Coral Calero, Mario Piattini
Department of Information Technologies and Systems (UCLM)
Indra-UCLM Research and Development Institute
Paseo de la Universidad 4 – 13071 Ciudad Real, Spain
{Ismael.Caballero, Coral.Calero, Mario.Piattini}@uclm.es

Eugenio Verbo
Department of Research & Development (Indra Software Labs, S.L.U.)
Indra-UCLM Research and Development Institute
Ronda de Toledo s/n – 13003 Ciudad Real, Spain
{emverbo}@indra.es

Abstract: Nowadays, data plays a key role in organizations, and management of its quality is becoming an essential activity. As part of such required management, organizations need to draw up processes for measuring the data quality (DQ) levels of their organizational units, taking into account the particularities of different scenarios, available resources, and characteristics of the data used in them. Given that there are not many works in the literature related to this objective, this paper proposes a methodology -abbreviated MMPRO- to develop processes for measuring DQ. MMPRO is based on ISO/IEC 15939. Despite being a standard of quality software, we believe it can be successfully applied in this context because of the similarities between software and data. The proposed methodology consists of four activities: (1) Establish and sustain the DQ measurement commitment, (2) Plan the DQ Measurement Process, (3) Perform the DQ Measurement Process, and (4) Evaluate the DQ Measurement Process. These four activities are divided into tasks. For each task, input and output products are listed, as well as a set of useful techniques and tools, many of them borrowed from the Software Engineering field.

KeyWords: Data Quality Measurement Plans, Data Quality Measurement Information Model, ISO/IEC 15939, DQMIM.

1. INTRODUCTION

Batini et al. [2] mention situations in which data with inadequate levels of quality originate problems that negatively affect information systems and, therefore, organizational performance. The most common sources of these inadequate levels are some potholes that can be found during the life cycle of data, such as those described by Strong et al. in [34]. Problems caused by this lack of quality can be classified at different levels according to their nature: technical (such as those relating to the implementation of data warehouses [25]), organizational (such as loss of customers [30], large financial losses [7, 13, 23] or even dissatisfaction of workers [11, 34]) and legal, because of privacy and/or national legislation (like in the Spanish Organic Law for Data Protection from 1999).

In order to minimize the negative impact of these problems on the execution of their activities, it is essential that organizations can assess whether the DQ level of their data is adequate. This involves firstly the definition of DQ measures on the data, and secondly the establishment of some valid acceptance ranges for the values obtained via measurement procedures. In this sense, and applying classical techniques and tools for quality management, it will be possible to more efficiently locate data with inadequate levels of DQ and its causes.

However, due to the very nature of data, it can be quite difficult to make a proper definition of the measures from the point of view of classical quality management foundations. Although DQ literature shows a lot of proposals for defining measures, they usually lack of an organizational focus. Thus, it is necessary to find an approach that enables a perception of measurement of DQ within an organizational context, taking into account the proper resources of the organization in order to develop DQ Measurement Processes which are integrated together with the remaining processes of the organization in the Process Model. An approach that can make easier the definition of DQ measurement processes is the one proposed by the TDQM Program at MIT [32, 38], which provides a view from classical theories of quality, in which data can be considered as raw material for the manufacturing process, being the resulting product an Information Product (IP).

Precisely, the special characteristics of data make it difficult to define DQ measures. Some of the methodologies proposed in works [3, 11, 18, 19, 24, 31] put forward several of the aforementioned characteristics, and typically, their corresponding authors provide solutions to their particular contexts. In spite of having identified a set of common concepts about DQ measuring, researchers normally use different terms for the concepts, making it difficult to translate their results among scenarios different from theirs. To attempt to bring together the used concepts with a set of unified terms, Caballero et al. in [4] have analyzed the most referenced works, and they have proposed a unified terminology according to the information model provided by ISO/IEC 15939 [21], but extended with the vocabulary corresponding to the specific characteristics of the DQ field. The result is a **Data Quality Measurement Information Model** (DQMIM), which is going to be used as a basis for this paper.

An added value of ISO/IEC 15939 is that it also provides a methodology for defining software measurement processes, bearing in mind organizational aspects and the possibility of using feedback about the process to improve it. We have tailored this methodology with the specifics of the DQ in order to fill the previously mentioned gap in the field of DQ. Hence, the main contribution of this article to the field of DQ is MMPRO, a methodology based on that proposed by ISO/IEC 15939 for planning DQ measurement processes, which are going to be integrated into the organizational process model.

The definition of the methodology does not catalogue data quality measures, nor does it provide a recommended set of measures to apply to data quality improvement projects. It does, however, involve not only identifying the activities and tasks, but also the identification of input and output products for each one. In addition, it identifies, from those used in the field of Software Engineering, some techniques and tools usable to transform inputs into outputs. We believe that the structure of the methodology will facilitate the incorporation of data quality issues to good practices of software quality as an effort to strive for the recognition of data quality in organizational environments.

The literature already describes many generic methodologies for assessing DQ, as surveyed by Batini et al. in [2]. For the purpose of this work, we understand that the fundamental difference between measurement and assessment is the specific focus of the latter to determine the validity and usefulness of data within a context, while the measurement is only aimed at obtaining values to be used in the assessment, without giving support to any judgment [1]. Although it may seem that this simplification makes our proposal less important, in the literature there are not any similar initiatives with sufficient generality, as it is a challenging research issue [1]. The great difference between the proposed methodology, MMPRO, and other proposals in the literature is that our work takes into account both the special characteristics of the data and also the inherent aspects of the DQ measurement within the organizational context. We consider that this is the main reason for which MMPRO can be not considered as yet another methodology for measuring DQ.

The remainder of the paper is structured as follows: Section 2 provides a brief presentation of the terminology proposed in DQMIM which is going to be used throughout this paper. Section 3 describes the methodology itself. Section 4 outlines some conclusions and future work; and finally, acknowledgements and references.

2. TERMINOLOGY

This section presents some of the terms which are going to be used in the methodology described in this paper. These terms belong to the DQMIM, which, as previously said, is an attempt to bring together the different DQ terminology used by the most referenced authors in the literature. Albeit Table 1 shows a mapping between DQ measurement concepts and the unified terminology provided by DQMIM, some rationale is introduced in the following paragraphs for a better understanding of the paper.

In spite of the existence of the various definitions for DQ [1], most authors agree that a piece of data (namely an organizational unit –portion of the organization-, organizational data bases, relational tuples, and so on), can be said to be of quality if it is valid for the purpose for which a user wants to use it ("*fitness for use*" [1, 35]). In order to gain generality in our explanations, let us extend the term "**user**" by means of the term "**stakeholder**" to make reference to any agent involved in the use of data [38].

As is known, this intuitive definition of quality, based on fitness for use, has two major implications: multi-dimensional perception of the quality, and dependence on context. These two implications give rise to adaptations to the characteristics of data and their context in order to adequately define metrics.

For any DQ measurement process (the set of activities and resources aimed at achieving values about the DQ level of data of an entity within the limits of an organizational unit), stakeholders must firstly identify the reasons for which they want to measure the level of DQ of the pieces of data used in any of the underlying processes of the organization. According to DQMIM, one can say that those stakeholders would want to meet certain **information needs**, which need some information products containing data that must be interpreted in order to satisfy the information need. It is important to highlight a difference: the information products are the outcome of the measurement process, whereas for DQ researchers, the concept of Information Product (IP) is the outcome of a data manufacturing process [38].

One of the strategies used to face the study of the multi-dimensional perception of the DQ of the entities having data is to divide the quality into smaller qualities, as ISO / IEC 9126 does for the software. These features are called DQ Dimensions [1, 22] although based on ISO/IEC15939, DQMIM proposes the term "**measurable concept**" instead. In this paper, we are going to align our explanations to the terms provided by the DQMIM. So, we are not going to use the terms "*DQ dimensions*" or "*metrics*" any longer, but **measurable concept** and **measure** respectively. In addition, it can be said that for each one of the information needs, it would be possible to identify one or various measurable concepts in order to make an interpretation about the level of the quality of the entities having data. So, for each scenario in which a DQ measure of the data used is needed, the most appropriate measurable concepts must be chosen according to the users' DQ requirements. The set of usable DQ measurable concepts is known as a **DQ model**. Eligible DQ models can come from either the Measurement Experience Base of the organization or from literature, in which some specific DQ models are referenced, like those for Web [8, 12], to name a few; it is also important to point out that ISO is currently working on the standard ISO / IEC 25012 [20], as part of the SQUARE family. This standard will propose a DQ model for IS. In any case, the classification proposed by Strong et al. in [35] is currently the most widely used.

DQ Measurements are defined in relation to the measurable attributes of the entities whose DQ level is required to be assessed. Typically, and in order to make the interpretation of the results easier, an indicator (a type of measurement according to DQMIM) must be outlined. An indicator consists of two main components: an analysis model (which can imply the use of other measures) and a decision criterion (which allows combining measures corresponding to different measurable concepts in order to achieve an overall perception of the DQ level of the entities being observed).

Typically, the combined DQ measurements are derived measures, which require a measurement function. For such kinds of functions, values for variables come from base measures. These base measures need measurement methods which can simply consist of counting the overall number of attributes, or counting only those which satisfy a condition. To complete the definition of a base measure, an expected range and the data type of values to be collected must be provided. Therefore, **measuring consists of collecting data values about a measurable attribute according to a measurement method**.

ISO 15939 Concept	Meaning in ISO 15939	DQ Field
Information need	An insight necessary to manage objectives, goals, risks and problems.	IQ Assessment Objectives [11], "Problem" [22], fundamental projects [18]
Measure (verb)	To make a measurement.	Metrics [11, 22, 23], Measures[11, 30]
Measure (noun)	A quantitative or categorical representation of attributes.	
Measurement	A set of operations having the objective of determining a value of a measure.	
Measurable Concept	A concept whose measurement satisfies different information needs.	DQ dimensions [11, 22, 23], IQ Criterion [24]
Measurable Attribute	A property or characteristic of an entity that can be distinguished quantitatively or qualitatively by human or automated means.	Information Group for Assessment [11]
Data Store	An organized and persistent collection of data that allows its retrieval.	Relational Database, Object Relational, XML, Spreadsheet.
Entity	An object that is to be measured.	Data Models, Data Values, Data Policies.
Stakeholder	An individual or organization that sponsors measurements and provides data or is a user of the measurement results.	People creating or updating a group of data[11]; Collector, custodians, consumer [22]; Data Customer, Data Manager and Data Manufacturer, Data Supplier [38]
Base measure	An attribute and the method for quantifying it. It is functionally independent of other measures.	Metric, Measure
Derived measure	A measure defined as a function of base measures.	Metric, Measure
Decision criteria	Numerical threshold or targets used to determine the need for action or further investigation, or to describe the level of confidence in a given result.	Indicator [37]
Function	An algorithm or calculation performed to combine two or more base measures.	-
Indicator	Measure that provides an estimate or evaluation of specified attributes derived from a model with respect to defined information needs.	Indicator [37]
Measure	A quantitative or categorical representation of attributes.	Metric, Measure
Measurement	A set of operations to determine a value of a measure.	-
Measurement method	Logical sequence of operations, described generically, used in quantifying an attribute with respect to a specified scale.	-
Measurement procedure	Set of operations, described specifically, used in the performance of a particular measurement according to a given method.	-
Measurement Process	The process for establishing, planning, performing and evaluating data quality measurement within an overall project or organizational measurement structure.	
Model	An algorithm or calculation combining measures (base or derived) associated to a decision criteria.	-
Observation	An instance of applying a measurement procedure to produce a value for a base measure.	-
Scale	An ordered set of values, continuous or discrete, or a set of categories to which the attribute is mapped. It can be of one of the following types: Nominal, Ordinal, Interval, Ratio.	Scale [36]
Type of method	Two types of methods can be distinguished depending on the nature of the operations used to quantify an attribute: Subjective and objective.	-
Unit of measurement	A particular quantity defined and adopted by convention, with which other quantities of the same kind are compared in order to express their magnitude relative to that quantity.	-
Value	A numerical or categorical result assigned to a base measure, derived measure or indicator - a statistic.	Metric, Measure.

Table 1. Mapping between some concepts of DQ found in literature and DQMIM proposed in [4]

Sometimes, to be able to collect values, the meaning of pieces of data must be completed with metadata in accordance with the measurable concept to be measured. So, the measurement procedure needs, apart from value for data and value for the metadata, a business rule that describes how to use the value for metadata, for example, to decide whether a piece of data is or not sound, if the associated metadata value belongs to a specific valid domain or not. To make measurement processes repeatable, it is necessary that the value of metadata remains attached to the piece of data that is completed. Wang et al. in [37] propose a solution to the relational model by labeling data as if they were common relational attributes. Caballero et al. in [4] propose an XML schema called DQXSD that allows adding tags to XML files. There are even some proposals, like [6, 29] suggesting the use of Semantic Technologies to perform this attachment, and thus enabling Web agents to process tasks related to DQ measurement.

Another important consideration that must be taken into account is the need for selecting among all, only a representative set of pieces of data to be inspected when collecting data for a base measure. This need can be encouraged, for example, as an attempt to minimize the computational effort invested in the measurement in order to not consume resources that would be entirely devoted to the processing of data. The number of pieces of data and its distribution will depend on the nature of information needs. In addition, and to not interfere in the measurement process, it is necessary to choose in the life cycle of the data which is the best moment or the best time interval to carry out the measurement method [31].

As we have seen throughout this section, managing all these concepts can be complicated within an organizational context. Therefore, we understand the need for some guidance in defining DQ Measurement processes. In order to provide such guidance, in the following section, we present MMPRO, which is the main contribution of this paper.

3. THE METHODOLOGY MMPRO

The aim of this section is to briefly describe the most interesting characteristics of MMPRO in order to guide stakeholders when measuring the levels of DQ within an organizational context. Intended stakeholders are commonly data quality managers or even executives within organizations who are interested in performing management and/or technical actions, primarily for improvement purposes.

ISO/IEC 15939 defines the scope of a measurement process to the extent of an organizational unit. Based on this idea, the main aim of this methodology is to define DQ Measurement Processes for the identifiable organizational units in order to collect, analyze and report information products related to the levels of data quality of information systems or any of their components, typically data stores, to support effective management of the processes, and to objectively demonstrate the DQ of the Information Products [21]. From this point onwards, we are going to use the term **scenario** to refer to that situation in which an organizational unit is used.

The methodology, which is “information need”-driven, comprises of four activities which are divided into several tasks. As a part of the tailoring work of ISO/IEC 15939, for each proposed task we decided to introduce a list with the main input and output products which are going to be presented by means of tables throughout the section. The use of specific techniques and tools (some borrowed directly from the Software Engineering field) is also proposed, although the final choice depends on the preferences and possibilities of each organization. In order to make the methodology more usable, the paper also identifies those stakeholders who should participate in the implementation of each task.

3.1. COM. Establish and sustain the DQ measurement commitment.

Experts in quality agree on the need to delegate responsibility for quality management to a reduced and skilled group of people committed to the organization [10]. In addition, this group must be able to access all of the elements belonging to the organizational units which need to be measured.

To give the necessary support to this need, the first goal of this methodology is to outline a multidisciplinary team of workers, who can take responsibility for planning and executing the process for measuring the DQ level of the entities.

To meet the target proposed in this activity, the following tasks must be carried out:

3.1.1. COM.1. Identify a Team for the Management of the DQ Measurement Plan (TM-DQMP).

This team will be responsible for coordinating efforts and organizational resources in order to adequately plan the DQ Measurement Process. The TM-DQMP should be composed of people being able to perform roles with direct responsibility over data and its integrity, throughout the life cycle of the data within the Information System, so that it can be known who is using the data and for what purpose, so as to assess the nature and extent of any deficiencies that may exist as well as assessing the impact that problems related to the DQ can cause. This team should be also multidisciplinary, cohesive, its members must be complemented in knowledge and skills, and be able to criticize their own work. In addition, the team must have a working environment where they identify a working method with effective ways of communication, as well as having available catalogues of techniques and tools usable for each task. Table 2 shows the main artifacts for this activity.

Products	Input	• List of required roles. • List of scenarios with their corresponding organizational units.
	Output	• Proposal for a TM-DQMP.
Tools and techniques		• Interviews. • Work Sessions. • Typical Human Resources tools and techniques.
Stakeholders		• Organizational experts in data and DQ. • Organizational managers of process and business models.

Table 2. Artifacts for COM.1.

3.1.2. COM.2. Establish and Communicate the Commitments of TM-DQMP members.

Once the necessary roles for TM-DQMP have been identified, it is time to assign concrete capable people to the candidates for roles of TM-DQMP. It would be interesting to conduct interviews and meetings with candidates to check their interest in the job, and their suitability and, if necessary, their commitment to developing the DQ Measurement Process. Table 3 gives more details about the artifacts of this activity.

Products	Input	• Proposal for a TM-DQMP. • Organizational chart.
	Output	• Commitment of the organizational components to participate in the measurement plan elaboration. • Work load documents.
Tools and techniques		• Interviews, Work Sessions. • Typical Human Resources tools and techniques. • Organizational Communication Methods. • Interpersonal Negotiation Techniques.
Stakeholders		• DQ experts. • Organization managers. • TM-DQMP.

Table 3. Artifacts for COM.2.

3.1.3. COM.3. Assign the Human Resources to the TM-DQMP.

After achieving the commitment of the components of the proposed TM-DQMP to the DQ Measurement Process, their availability and time constraints must be studied and tailored, if possible, in order to determine the feasibility of the measurement process. To do so, somebody should conduct interviews with those responsible for the organization to identify potential overlaps among other projects being performed, in order to avoid affecting the measurement plan. Some useful tools to achieve this goal are Gantt or PERT charts. Moreover, it is necessary to classify capabilities and skills of the components of TM-DQMP and, depending on these and concordance to what is required to measure, assign tasks to each participant in order to get their best performance. Table 4 shows artifacts for this task.

Products	Input	• TM-DQMP components. • List of tasks of the DQ Measurement Process.
	Output	• Final composition of the TM-DQMP. • List with assignable people to tasks of the DQ Measurement Process.
Tools and techniques		• Work sessions. • Interviews. • Planning tools like Gantt or PERT diagrams. • Time and Budget Estimation tools and techniques. • Cost/benefit analysis.
Stakeholders		• Organization managers. • TM-DQMP components. • People willing to participate in the measurement process.

Table 4. Artifacts for COM.3.

3.2. PMP. Plan the DQ Measurement Process.

The main goal of this second activity is to properly outline the DQ Measurement Process. In order to document the plan, the terms provided by the terminology DQMIM will be used. As a result, a document containing the plan for executing the DQ Measurement Process is obtained. This activity consists of the following tasks:

3.2.1. PMP.1. Characterize Organizational Units and Scenarios.

Provided that the scenarios with organizational units are the context for measurement, it is really important to fully describe these organizational units as better as possible, since the DQ Measurement Process is going to be attached to them. As a first step, it is necessary to identify the class of resources which are susceptible to be measured for data quality. The most generic unit is the whole organization itself, but if we considered it like this the problem would become difficult to tackle. Instead, we propose taking a smaller perspective centered on the information manufacturing process ([38]) or on the information management process ([5]). Batini and Scannapieco in [1] classify the collections of data belonging to the organizations into two main blocks: Internal group of data (organizational database collections) and external sources of data (data providers). The information manufacturing/management processes address databases. For these databases, Even and Shankaranarayanan in [14], consider a hierarchical decomposition into datasets (like relational tables, or xml or rdf files), which comprises of data records (like relational tuples, or elements in xml files); each of these data records is formed by data items (values corresponding to relational attributes; or values corresponding to xml attributes). In order to complete the characterization of the scenario, it is also necessary to identify both the role of the stakeholders performing any kind of task with the organizational units and their requirements, which are expected to be met by the working organizational unit.

In order to achieve better results, it would be very helpful to have a graphical representation of the information process for a better identification and characterization of the scenario and the organizational unit. Table 5 shows the main artifacts of this task.

Products	Input	• Organizational maps of processes. • Data Model. • Process Model.
	Output	• Characterized Organizational Units. • Stakeholders using data. • Stakeholders' need.
Tools and techniques		• IPMAP[32], BPMN[26], SPEM [27], Activity Diagrams of UML 2.0 • Work Session. • ORME-DQ Matrices by [2]
Stakeholders		• TM-DQMP components.

Table 5. Artifacts for PMP.1

3.2.2. PMP.2. Identify and prioritize information needs.

For each organizational unit, besides technical and managerial requirements, stakeholders are expected to have some DQ requirements. These kinds of requirements allow MMPRO users to identify what is needed to measure, that is to say, the information needs according to nomenclature of DQMIM. One of the most common information needs is to understand the anatomy of DQ problems and their patterns, as Lee et al. state in [22]. So, users of MMPRO could be interested in achieving sufficient evidence (a.k.a. measures) to identify from manifestations of DQ problems how important (the magnitude) each one is, and what are their sources in order to plan improvement actions. Strong et al. in [34] identify several potholes and their warning signs in information systems, which allows concretion of the information needs. Indeed, the provided schema can be used as a guide for establishing the measurement process, since it also identifies the measurable concepts that are of interest for the information need. So, when non-conformities are reported, then TM-DQMP must inspect the characterized organizational units in the scenario in order to go into the root of DQ problems. Other examples of possible information needs are the interest of stakeholder in quantifying the DQ Risk as Batini et al. propose in [2], or the technical problems identified by Oliveira et al. in [25] for data stores. But as not all information needs are of equal importance it is necessary to prioritize them, taking into account their relative level of criticality for the organization.

Since the information needs do not only concern technical aspects, but also organizational and managerial ones, the business relationships among the main users of Information Systems should be taken into account and properly gathered, because they can affect the plan of the DQ Measurement Process. These relationships must be communicated to the stakeholders who are going to be involved in the measurement process, in order to provide as much information as possible regarding their DQ requirements and their perception of how to measure the levels of quality. Table 6 shows some artifacts for this task.

Products	Input	• DQ requirements specification. • Organizational activity reports of IS.
	Output	• List of prioritized information needs.
Tools and techniques		• Quality Identification Tools, like brainstorming, adjacency matrices, Interviews, … • DQ Risk Identification by [2] • Communication techniques.
Stakeholders		• TM-DQMP managers. • Organization managers.

Table 6. Artifacts for PMP.2

3.2.3. PMP.3. Identify DQ Measurable Concepts

Measurable concepts (aka DQ Dimensions) are rational criteria that represent user requirements for judging the DQ of the organizational units in the scenario. The TM-DQMP is responsible for selecting those measurable concepts that best meet the information needs. As a basis for carrying out such a task, DQ models that best fit to DQ requirements of the organizational unit must be used. To select the best fitting DQ measurable concepts, some tools and techniques can be applied, such as brainstorming, working sessions, Goal-Question-Metrics (GQM), interviews or Delphi method. The methodology described by Franch and Carvalho in [15] could also be used but tailored and extended with the specific characteristics of DQ. For example, the aforementioned proposal by Strong et al. [34] could be used as a decision tree in which, for each kind of pothole, being the subject of the study through the information needs, the involved measurable concepts must be identified.

It is very important to bear in mind that, sometimes, the chosen DQ measurable concepts can be not orthogonal, and there may be dependencies among them, as [9, 16, 17] have analyzed. This fact entails modeling how they can affect themselves and how they can affect to the DQ Measurement Process.

In Table 7, a summary of the main artifacts for this task is shown.

Products	Input	Information needs. DQ Requirement Specification (DQ-USR). DQ measurable concepts catalogue.
	Output	Relevant DQ measurable concepts list. List of entities to be measured.
Tools and techniques		Goal-Question-Metrics [33]. Franch and Carvalho's methodology [15] Cataloguing.
Stakeholders		TM-DQMP components. Stakeholders affected for the DQ Measurement Process

Table 7. Artifacts for PMP.3.

3.2.4. PMP.4. Define or Identify DQ Measures.

Once DQ measurable concepts have been identified for each information need, it is time to define the measures associated to the organizational units. According to the DQMIM, the measure can be of one of the following types: **base measure** (defined without taking into account any other measure: it is defined by means of a measurement procedure), **derived measure** (which takes into account any other base or derived measure: it is defined by means of a measurement function) and **indicator** (which needs a decision criteria and an analysis model). Typically, and in order to better assess the DQ level, an indicator might be provided, with decision criteria based on ranges of values for acceptance defined for each measurable concept (and likely for each combination of measurable concepts) within each scenario. So, measures can become useful to better satisfy the information needs through the chosen corresponding measurable concept(s). ISO/IEC 15939 proposes defining a set of candidate measures, and then selecting the best fitting ones. In this sense, the Goal Question Metric methodology defined by [33], can be also used as a guide to obtain measures by progressively refining questions (information needs) and responses (measurable concepts) to identify the measurable attributes by which measures are going to be defined.

In addition, for each DQ Measure, it is necessary to identify:

- A Unit of Measurement.
- A Scale (nominal, ordinal, interval, or ratio). Pipino et al. in [28] provide some of the most used measurable concepts and analyze how to choose a scale for them.
- A Formal Definition of the DQ Measure by means of a measurement procedure according to the type of the measure. This is probably the most challenging research line in the DQ field, because of the wide variety of influential issues. The fourth chapter by [22] and the second one by [1] present interesting formulae which can be used as measurement functions for derived measures. On the other hand, a formal definition could also include some discussion on specific issues, such as those proposed by Even and Shankaranarayanan in [14], which consists of the introduction of some modifiers for the values taken as impartial measures aimed at modeling and introducing contextual issues like those corresponding to organizational units for a utility-driven assessment into the impartial measurement.

The DQ Measurement Process would originate changes to the technology of the organizational unit. So, as a part of the definition of the measurement procedures, the definition of a DQ measurement process may cause changes in both the process model and the data model of IS. For example, some of the defined measurement procedures can require the addition of metadata to complete the meaning of the data for measuring a specific measurable concept (e.g. the timestamp for timeliness).

Together with the definition of the measurement procedures, a data collection procedure must be outlined, which is used to provide data belonging to the organizational unit for the corresponding measurement procedure. In order to make either the data collection procedure or the measurement procedure more reliable, the idea of automating them is interesting. However, it is also important to appropriately describe the manner in which the measurement results will be reported as part of DQ Measurement Process.

For all issues dealt with in this section, it is assumed that DQ measurements are being defined from scratch. In addition, these measures might be defined as generically as possible in order to enable their re-use. Some organizations could have a Measurement Experience Base containing parameterized DQ measurements which have been previously developed. The parameters can be the following: the organizational unit to measure, how many entities must be taken into account, and when to collect the data and apply the measurement methods. In situation, the main aim of this task is to identify the best fitting DQ measurement from among those stored in the Measurement Experience Base, and in the following task, give adequate values to the parameters in order to particularize the measures for a specific scenario. This task must be repeated for those measurable concepts that are involved in the definition of the measure for the information need.

Finally, and given that information needs may address technical and managerial issues, it is important to highlight that the TM-DQMP must be multidisciplinary and skilled enough to face this activity, which can be regarded as one of the most important of the methodology presented.

Table 8 shows the main artifacts for this task.

Products	Input	• Information needs. • Relevant measurable concepts. • Measurement Experience Base. • Data and process model.
	Output	• List of measures to be applied to each measurable concept.
Tools and techniques		• GQM [33]. .
Stakeholders		• TM-DQMP components.

Table 8. Artifacts for PMP.3.

3.2.5. PMP.5. Locate the instances of the Entities which need to be measured.

As previously said, one of the parameters of the DQ Measurement Procedure is the organizational unit which has data whose DQ levels need to be measured. This data can be found in data stores of various kinds, such as relational databases or semi-structured, sequential access files, XML documents or spreadsheets, or even documents of Semantic Web, and also patterns of data, securities data, data domains, business rules, or user interfaces [4]. The nature of the data stores can affect the used technology, if the corresponding measurement and data collection procedures can have been automated.
The most useful techniques for locating these entities are working sessions with those responsible for the data resources of the organization. Table 9 summarizes the main artifacts for this task.

Products	Input	• List of entities to be measured. • Organizational data model.
	Output	• List of the location of data repositories whose DQ is to be measured.
Tools and techniques		• Inspections. • Work sessions.
Stakeholders		• TM-DQMP components. • Data Resources managers.

Table 9. Artifacts for PMP.4.

3.2.6. PMP.6. Determine the amount of data to Measure: sampling data.

Depending on the purpose of measurement or the need not to sacrifice performance of the IS with the processing of the DQ measurement, it may be necessary to delimit the number of entities that must be taken into account when measuring in order for a sample to be statistically meaningful. In that case, one must select a representative sample of the entire set of entities and then extrapolate the results. The parameters of the sample (type of sampling, acceptable ratio of invalid data, sample size, minimum and maximum values for acceptance or rejection) can be calculated according to standards such as ISO 2859, provided that the entities having data are to satisfy the conditions and constraints imposed by these standards. In addition, Lee et al. in chapter 5 of [22] provide further information about sampling.

For a list of artifacts of EMP.5 see Table 10.

Products	Input	• Information needs. • Data repositories to be measured. • Computational cost of measuring the whole data collection. • Effort-cost relation of measuring the whole data collection. • DQ Measurement Method. • Data Collection procedure.
	Output	• Viability study of the measurement of the whole data collection. • Parameters of sampling. • Data Collection Method review.
Tools and techniques		• UNE-EN-ISO66020 / ISO 2859.
Stakeholders		• TM-DQMP components.

Table 10. Artifacts for PMP.5.

3.2.7. PMP.7. Do a temporal plan for data collecting.

The last of the parameters is related to temporal issues. Data collection procedures could not be a punctual activity, but their implementation could require a time interval. In addition, it may happen that somebody could be interested in studying the temporal evolution of the DQ level of an entity [30]. It is also essential to consider the temporary human assignment which was carried out during the task COM.3 so as to avoid collisions between the measurement activities of TM-DQMP and other work they can do within the organization apart from measurement planning.

For all these reasons, it is necessary to perform a temporal plan of the measurement so that the DQ-MPMT can obtain the most significant values, without hindering the work of the rest of the workers of the organization. Some of the tools that can be used for timing are Gantt charts or even if we are dealing with certain parts of the business process, we could use some process modeling notations, such as BPMN [26], SPEM [27], or IPMAP, specific DQ notation for representing Information Manufacturing Processes developed by Shankaranarayan et al. [32].

Table 11 contains the main artifacts for this task.

Products	Input	• List of DQ measures. • List of entities to be measured. • Data quantity to be measured for each entity. • Specification and temporal restrictions for data collection. • Information Product life cycle. • Data Model and Process Model.
	Output	• Data collection Plan for each measure.
Tools and techniques		• Gantt diagrams. • BPMN, IPMAP, SPEM, Activity Diagrams of UML 2.0.
Stakeholders		• TM-DQMP components.

Table 11. Artifacts for PMP.6.

3.3. *PeMP. Perform the DQ Measurement Process.*

3.3.1. PeMP.1. Integrate measurement procedures into organizational units.

Once the DQ Measurement Process has been fully defined, their measurement methods must be fitted into the organizational units, in order to obtain meaningful results within organizational contexts. If possible, the measurement methods must be automated as much as possible by means of software routines according to the previously developed plan (in PMP.7). This fact probably implies modifying both the process model and the data model of the organizational unit because of the integration of the software routines aimed at implementing the measurement methods according to the type of measurement. The changes must be communicated to the people responsible for the organizational units.

Artifacts for this task can be seen in Table 12.

Products	Input	• Data Model. • Process Model. • DQ Measurement Process.
	Output	• Modified Data Model. • Modified Process Model.
Tools and techniques		• Reengineering.
Stakeholders		• TM-DQMP components.

Table 12. Artifacts for PeMP.1.

3.3.2. PeMP.2. Collect and validate data for measurement procedures.

This is the central task of the methodology, which consists of running the DQ measurement process in order to obtain values by means of the corresponding measurement methods associated to each measure. If sampling has been planned, then it must be done prior to computing any of the measurement methods. For this task, it is demanded that stakeholders can access the data which is to be measured. Obtained values must be stored together the contextual information that enables further understanding of the data within its context. After collecting values, in order to obtain representative values, it would be necessary to analyze the validity of the data. Table 13 shows the artifacts for this task.

Products	Input	• DQ Measurement Process.
	Output	• Values for the DQ Measures contained in the DQ Measurement Process.
Tools and techniques		• Those required for the corresponding DQ measurement procedures.
Stakeholders		• People responsible for organizational units. • TM-DQMP components.

Table 13. Artifacts for PeMP.2.

3.3.3. PeMP.3. Develop objective reports to enable further analysis of the DQ level.

It is necessary to process data (e.g. aggregate or transform by means of statistical analysis) in order to develop objective reports, bearing in mind that released results must meet information needs. These reports are to be delivered to people being able to analyze and interpret the results in order to perform an assessment and to propose DQ improvements for organizational units. Together with the reports, contextual information is required to be also supplied to DQ analysts and managers. See Table 14 for artifacts.

Products	Input	• Validated data regarding DQ measures.
	Output	• DQ Measurement Reports.
Tools and techniques		• Classical Quality Tools, like histograms, Pareto and Ishikawa Diagrams, …
Stakeholders		• TM-DQMP components.

Table 14. Artifacts for PeMP.3.

3.3.4. PeMP.4. Communicate and share the results.

Once reports about the results of the measurement process have been made, it is necessary to communicate the results to whom it may concern, namely DQ analysts and managers. See Table 15 contains the main artifact of this task.

Products	Input	• DQ Measurement reports. • List of stakeholders requiring DQ Measurement reports.
	Output	• Success Reports of Delivering. • Corresponding feedbacks.
Tools and techniques		• Any typical communication tools, like email, web, etc.
Stakeholders		• TM-DQMP components.

Table 15. Artifacts for PeMP.4.

3.4. ERMP. Evaluate DQ Measurement Process.

3.4.1. ERMP.1. Evaluate DQ Measurement Process and standardize lesson learned.

This task is aimed at evaluating if the DQ measurement process really has met the information needs. It is not about whether obtained values for DQ measurement are good according to DQ requirements. An analysis must be performed for identifying the weaknesses and the strengths of the DQ measurement process. From this analysis, some lessons are expected to be derived, and in order to avoid future similar weaknesses, they must be standardized and added to the Measurement Experience Base. The main artifacts are shown in Table 16.

Products	Input	• Reports containing results of measurements. • List with information needs. • Measurement Experience Base.
	Output	• Report with adequacy of the DQ Measurement Process. • Lessons learned. • Revised Measurement Experience Base.
Tools and techniques		• Typical Quality tools.
Stakeholders		• TM-DQMP components.

Table 16. Artifacts for ERMP.1.

3.4.2. ERMP.2. Identify potential improvements to DQ Measurement Process.

The weakness of a report containing the results of a DQ Measurement process can probably be due to errors when creating the report or to the weakness of the DQ Measurement process. From the report with the adequacy of the DQ Measurement Process, it is necessary to isolate the possible causes of the weaknesses, and modify either the measurement procedure or the procedure to elaborate the report with the results of the measurements. Table 17 shows the main artifacts for this task.

Products	Input	• Lessons learned. • Measurement Experience Base.
	Output	• Extended Measurement Experience Base .
Tools and techniques		• Classical Quality Tools, Deming's PDCA
Stakeholders		• TM-DQMP components.

Table 17. Artifacts for ERMP.2.

4. Conclusions and Future Work

Measurement has been recognized as a key activity within any quality management field. Since organizations have begun to realize that some of their current problems can be due to inadequate levels of DQ, they have begun to devote resources and efforts to managing the quality of the data used in their business processes.

In this regard, researchers in the DQ field are committed to developing artifacts that help organizations to measure their DQ. As the DQ field is still very young, it is necessary to rely on other more established fields such as the Software Engineering one. So research efforts can be based on solid foundations which are familiar to software developers and managers. As an example of this kind of foundation, we can mention the existing standards for software measurement and software quality. We have chosen one of these standards, ISO / IEC 15939, as the basis for the methodology MMPRO presented in this paper.

Our proposal aims to fill the gap and also to attempt to complement the existing DQ assessment frameworks in the literature with organizational issues. The main contribution of this paper is not the methodology itself and the survey of the literature attached to each task, but the advantage of being able to use it as an efficient guide that takes into account the aforementioned characteristics of data in planning the measurement of the DQ organizational resources containing relevant data which is going to be used in business processes.

MMPRO consists of four activities with their corresponding tasks. To make it more usable as a possible methodology, artifacts have been identified for each of these tasks. Since the methodology has been adapted from standard ISO/IEC 15939, a well-known one in the field of measurement software, any professional who is familiar with it could easily apply the methodology in its own context and begin to introduce concepts about DQ into his/her catalogue of good practices.

We have also realized the importance of having reliable and automatable DQ measures. Accordingly, our current line of work focuses on two aspects: on the one hand, to validate the methodology through its application to real cases, with the aim of identifying measurable concepts that could accept automatable measurement procedures, and on the other hand, to develop a set of tools for automating the process of developing DQ Measurement plans by using Semantic Technologies.

5. ACKNOWLEDGEMENTS

This research is part of the projects HERMES (TSI-020100-2008-155) supported by the Spanish Ministerio of Industria, Turismo y Comercio through AVANZA I+D Program, and the projects ESFINGE (TIN2006-15175-C05-05) and CALIPSO (TIN 2005-24055-E), both supported by the Spanish Ministerio of Educación y Ciencia.

6. REFERENCES

[1] Batini, C. and Scannapieco, M., *Data Quality: Concepts, Methodologies and Techniques*. Data-Centric Systems and Applications. Springer-Verlag Berlin Heidelberg Berlin, 2006.

[2] Batini, C., Barone, D., Mastrella, M., Maurino, A., and Ruffini, C. *A Framework and a Methodology for Data Quality Assessment and Monitoring*. in *12th International Conference on Information Quality*. 2007. MIT, Cambridge, MA.

[3] Burgess, M.S.E., Gray, W.A., and Fiddian, N.J. *Quality Measures and the Information Consumer*. in *Ninth International Conference on Information Quality (ICIQ'04)*. 2004. MIT, Cambridge, MA, USA.

[4] Caballero, I., Verbo, E.M., Calero, C., and Piattini, M. *A Data Quality Measurement Information Model based on ISO/IEC 15939*. in *12th International Conference on Information Quality*. 2007. MIT, Cambridge, MA.

[5] Caballero, I., Caro, A., Calero, C., and Piattini, M., "IQM3: Information Quality Maturity Model". *Journal of Universal Computer Science*, 14. 2008. p. 1-29.

[6] Caballero, I., Verbo, E.M., Calero, C., and Piattini, M. *DQRDFS:Towards a Semantic Web Enhanced with Data Quality*. in *Web Information Systems and Technologies*. 2008. Funchal, Madeira, Portugal.

[7] Cai, Y. and Shankaranarayanan, G., "Managing data quality in inter-organisational data networks". *International Journal of Information Quality*, 1 (3). 2007. p. 254 - 271.

[8] Caro, A., Calero, C., Caballero, I., and Piattini, M., "A proposal for a set of attributes relevant for Web Portal Data Quality". *Software Quality Journal*. 2008.

[9] DeAmicis, F., Barone, D., and Batini, C. *An Analytical Framework to analyze Dependencies among data Quality Dimensions*. in *ICIQ'06*. 2006. MIT, Cambridge, MA, USA.

[10] Deming, W.E., *Out of Crisis*. MIT Center for Advanced Engineering Study Cambridge: MA, 1986.

[11] English, L., *Improving Data Warehouse and Business Information Quality: Methods for reducing costs and increasing Profits*. Willey & Sons New York, NY, USA, 1999.

[12] Eppler, M. and Muenzenmayer, P. *Measuring Information Quality in the Web Context: A Survey of State-of-the-Art Instruments and an Application Methodology*. in *Proceeding of the Seventh International Conference on Information Quality*. 2002.

[13] Eppler, M. and Helfert, M. *A Classification and Analysis of Data Quality Costs*. in *International Conference on Information Quality*. 2004. MIT, Cambridge, MA, USA.

[14] Even, A. and Shankaranarayanan, G., "Utility-driven assessment of data quality". *SIGMIS Database*, 38 (2). 2007. p. 75-93.

[15] Franch, X. and Carvallo, J.P., "Using Quality Models in Software Package Selection". *IEEE Software.*, 20 (1). 2003. p. 34-41.

[16] Gackowski, Z., "A formal definition of operation quality of factors: a focus on data and information". *International Journal on Information Quality*, 1 (2). 2007. p. 225 - 249.

[17] Ge, M. and Helfert, M. *A Review of Information Quality Research*. in *Interantional Conference on Information Quality*. 2007. MIT, Cambridge, MA, USA.

[18] Gebauer, M., Caspers, P., and Weigel, N. *Reproducible Measurement of Data Quality Field*. in *Tenth International Conference on Information Quality (ICIQ'05)*. 2005. MIT, Cambridge, MA, USA.

[19] Gustavsson, M. *Information Quality Measurement*. in *International Conference on Information Quality*. 2006. MIT, Cambdrige, MA, USA.

[20] ISO-25012, "ISO/IEC 25012: Software Engineering - Software Quality Requirements and Evaluation (SQuaRE) - Data Quality Model (Draft)". 2006.

[21] ISO/IEC, *FDIS 15939: Software Engineering-Software Measurement Process*. 2002.

[22] Lee, Y.W., Pipino, L.L., Funk, J.D., and Wang, R.Y., *Journey to Data Quality*. Massachusetts Institute of Technology Cambridge, MA, USA, 2006.

[23] Loshin, D., *Enterprises Knowledgement Management: The Data Quality Approach*. Morgan Kauffman San Francisco, CA, USA, 2001.

[24] Naumann, F. and Rolker, C. *Assessment Methods for Information Quality Criteria*. in *Fifth International Conference on Information Quality (ICIQ'2000)*. 2000. MIT, Cambridge, MA, USA.

[25] Oliveira, P., Rodrigues, F.t., Henriques, P., and Galhardas, H. *A Taxonomy of Data Quality Problems*. in *Second International Workshop on Data and Information Quality (in conjunction with CAISE'05)*. 2005. Porto, Portugal.

[26] OMG, *Business Process Model and Notation 2*. 2008, Object Management Group.

[27] OMG, *Software Process Engineering Meta-Model, version 2.0- http://www.omg.org/technology/documents/formal/spem.htm*. 2008, Object Management Group.

[28] Pipino, L.L., Wang, R.Y., Kopcso, D., and Rybolt, W., *Developing Measurement Scales for Data-Quality Dimensions*, in *Information Quality*, R.Y. Wang, et al., Editors. 2005, ME Sharpe: Armonk, NY, USA. p. 37-51.

[29] Preece, A., Missier, P., Embury, S., Jin, B., and Greenwood, M., "An ontology-based approach to handling information quality in e-Science". *Concurrency and Computation: Practice and Experience*, 20 (3). 2008. p. 253-264.

[30] Redman, T., *Data Quality: The field guide*. Digital Press Boston, 2000.

[31] Redman, T.C., *Data Quality for the Information Age*. Artech House Publishers Boston, MA, USA, 1996.

[32] Shankaranarayanan, G., Wang, R.Y., and Ziad, M. *IP-MAP: Representing the Manufacture of an Information Product*. in *Fifth International Conference on Information Quality (ICIQ'2000)*. 2000. MIT, Cambridge, MA, USA.

[33] Solingen, R.v., Latum, F.v., Oivo, M., and Berghout, E.W. *Application of software measurement at Schlumberger RPS: towards enhancing GQM*. in *Proceedings of the 6th European Software Control and Metrics (ESCOM) conference*. 1995. The Netherlands.

[34] Strong, D., Lee, Y., and Wang, R., "Ten Potholes in the Road to Information Quality". *IEEE Computer*. 1997. p. 38-46.

[35] Strong, D.M., Lee, Y.W., and Wang, R.Y., "Data Quality in Context". *Communications of the ACM*, 40 (5). 1997. p. 103-110.

[36] Wang, R., Pierce, E.M., Madnick, S., and Fisher, C., eds. *Information Quality*. Advances in Management Information Systems, ed. V. Zwass. 2005, M.E. Sharpe: Saddle River, NJ.

[37] Wang, R.Y., Reddy, M., and Kon, H., "Towards quality data: An attribute-based approach". *Journal of Decision Support Systems*, 13 (3-4). 1995. p. 349-372.

[38] Wang, R.Y., "A Product Perspective on Total Data Quality Management". *Communications of the ACM*, 41 (2). 1998. p. 58-65.