CALYDAT: A METHODOLOGY FOR EVALUATING DATA QUALITY DIMENSIONS BASED ON DATA PROFILING TECHNIQUES

(Research-in-progress)

Yonelbys Iznaga
Universidad de las Ciencias Informáticas, Cuba
yiznaga@uci.cu

César Guerra
Universidad Politécnica San Luis Potosí, México
cguerra74@gmail.com

Ismael Caballero
Insituto de Tecnologías y Sistemas de Información, Universidad de Castilla-La Mancha, España
Ismael.Caballero@uclm.es

Abstract: Any organization that needs to satisfy their business objectives and uses data to implement organizational processes, must have knowledge of how these data satisfy the preset quality requirements. These requirements are expressed by means of certain data quality dimensions. In some contexts, models and methodologies of data quality assessment require of mechanisms to control and monitor the level of quality of data. Proposing a methodology with a qualitative diagnosis of the data quality dimensions and using data profiling techniques to measure some of these dimensions, will have a significant impact on the processes of appropriate use of the data. The main contribution of this paper is a methodology that assesses the data quality, by diagnosing its dimensions through surveys and data profiling techniques. The paper also presents the results obtained in a real case study, which served to validate the methodology.

Key words: Data quality, data quality dimensions, data profiling, methodology.

INTRODUCTION

During decades, data management has acquired a growing significance in companies, because data constitute the blood of the organization, and without them, corporations cannot align with their organizational strategy [7]. In 2002, only in the United States of America, the annual expenditure of poor data quality for enterprises was six billion dollars, according to estimations of TDWI (The Data Warehouse Institute) [20]. Because electronic data are so pervasive, data quality (hereafter DQ) plays a critical role in all business and governmental applications [1] and it is recognized as a relevant performance issue of operating processes [3]. Companies that decide to implement complex information systems such as Decision Support System (DSS), Executive Support Systems (ESS) or Enterprise Resource Planning (ERP), among others, should understand that the success of these systems also depends largely on their data. According to ISO / IEC 25012, DQ is defined as "*the degree to which the characteristics of the data are suggested conditions and needs when used under specific conditions*." [9].

Therefore, before any operation, it is important to assess the suitability degree of the use of data involved in the task, according to the context in which they are. Data profiling is one of the techniques that helps diagnose the DQ in specific contexts, which is the "*data analysis systems to understand its content, structure, quality and dependencies*" [4]. Indeed, doing data profiling and monitoring the defects of data, are useful activities for assessing DQ in specific contexts.

Although, nowadays there exists some models, methodologies and tools to carry out the data profiling processes, in our context, it is possible to find some needs such as: assessment of some characteristics of DQ using techniques and tools of data profiling, definition of roles and responsibilities for the DQ control, organization of the process through the use of artifacts and documents and the frequent reporting to the organization of the DQ diagnosis, depending on user types and level securities, in order to involve and engage members and roles that interacts with these data.

This paper is organized in five main sections. Section II describes existing methodologies for DQ control and assessment, and data profiling models that currently exists. Section III presents CALYDAT, the proposed methodology and a description of their characteristics, principles, scope, processes, activities and people in charge. Section IV presents the obtained results from applying the proposed methodology in a real context. Finally, Section V presents the conclusions and the main intentions for future work.

BACKGROUND

Methodologies for DQ assessment

Many authors have made contributions for DQ. Several of these have offered the most relevant categorizations of DQ dimensions, such as in [10, 14, 17, 21, 22, 25]**.** This research was based on the data quality characteristics introduced in ISO/IEC 25012, which are: accuracy, completeness, consistency, credibility, timeliness, accessibility, compliance, confidentiality, efficiency, traceability, portability, understandability, availability and recoverability [9].

For making a comparative study of existing methodologies for DQ assessment, we consider several aspects, including: dimensions used, cost and types of data and information systems involved. The methodologies for DQ assessment and improvement have been classified in four categories [1]:

- *complete methodologies*, which provide support to both the assessment and improvement phases, and address both technical and economic issues;
- *audit methodologies*, which focus on the assessment phase and provide limited support to the improvement phase;
- *operational methodologies,* which focus on the technical issues of both the assessment and improvement phases, but do not address economic issues;
- *economic related methodologies:* which focus on the evaluation of costs.

This research is based on audit methodologies. Some of these methodologies are AIMQ [13], CIHI [26], AMEQ [20] and IQM [5]:

AIMQ Methodology: (*A Methodology for Information Quality Assessment* [13]): It is the only methodology of information quality based on benchmarking. It draws heavily on the PSP/IQ model (Table 1), which classifies the DQ dimensions according to the interest and priority of users and administrators. **AIMQ** has four classifications for DQ: comprehensive, reliable, useful and usable, into which DQ dimensions fall. It uses questionnaires for the identification and diagnosis of both DQ dimensions and measures of information quality.

	Conforms to specifications	***Meets or exceeds the customer expectations***
Product Quality	*Sound information*	*Useful information*
Service Quality	*Dependable information*	*Usable information*

Table 1. The PSP/IQ model.

CIHI methodology (*Canadian Institute for Health Information* **[26]): CIHI** focus on the control of DQ of data stored in the Canadian Institute of Health Information, specifically in the monitoring of the size, heterogeneity and quality of the stored data. Data quality evaluation is based on a four-level hierarchical model. At the first level, 86 basic quality criteria are defined. These criteria are aggregated by means of

algorithms of composition into 24 quality characteristics at the second hierarchical level, and finally, these are aggregated into five DQ dimensions at the third level. Finally, the five dimensions are aggregated into one overall database evaluation at the fourth level.
IQM methodology (*Information Quality Measurement* [5]): **IQM** conceived the provision of a quality framework adapted to the Web data. Among its entries, besides of the quality criteria, it has the tools and techniques used to measure the DQ. The result of evaluation is the most important outputs, which is a valuable guide for selecting and customization of the tools used by web administrators for creating, managing websites. IQM describes the following main phases: assessment planning, assessment configuration, Measurement and follow-up activities, where the most important processes are: the diagnosis of the data, the requirements analysis and evaluation of the DQ.
AMEQ methodology (*Activity-based Measuring and Evaluating of Product information Quality* [19]): **AMEQ** provide a rigorous basis for Product Information Quality assessment and improvement in compliance with organizational goals. The methodology is specific for the evaluation of DQ in manufacturing companies, where product information represents the main component of operational databases. In manufacturing companies, the association between product information and production processes is straightforward and relatively standard across companies [1]. AMEQ has five phases. The first one assesses the cultural preparation of the organization. The second one focuses on all information related to the product by process modeling and identification of critical areas. One of the outputs of this phase is a model of measurement techniques. The third phase focuses on the implementation of all activities and techniques for the measurement and evaluation. During the fourth phase the causes of DQ problems that have been detected after diagnosis of the dimensions will be investigated. The last one is responsible for monitoring and improving the quality of product information, through mechanisms of accountability of the processes and data.

After studying the characteristics of these audit methodologies, we consider that they are very useful, depending on its features and goals. However, according with some aspects like: the focus on the business processes of the organization, the definition of roles and responsibilities, the use of artifacts for document the process and the inclusion of data profiling techniques for the DQ evaluation; we conclude that, except AMEQ that utilizes the organizational processes in its process modeling, the rest of the methodologies are not based on business processes. They do not use roles and responsibilities in its phases and activities, they do not include data profiling techniques for the DQ evaluation and only CIHI has a well-defined documentation process.

Data profiling models, techniques and tools

Several data profiling methods and techniques also contribute to the necessary assessment for the DQ control, where the fundamental approach is performed on the data collections. The DQ dimensions more widely used to assess DQ are: correctness, completeness and accuracy. One of the models available today is [4], which consists of one or more inputs of data and metadata, the application of research techniques, and as outputs, corrected metadata and information related with data, as shown in Figure 1.

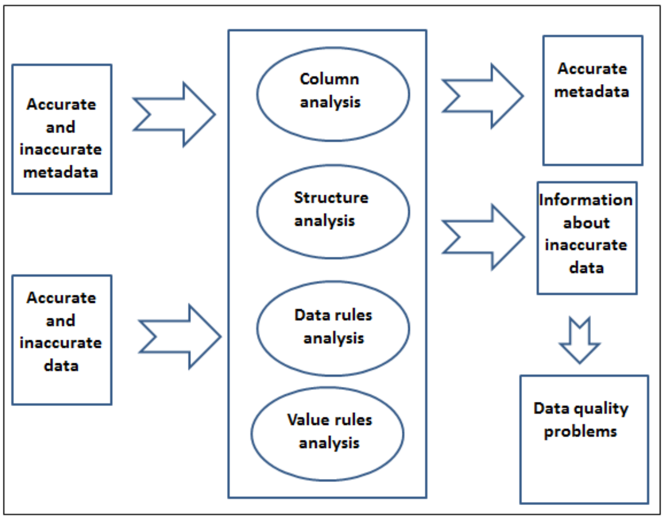


Fig. 1. Data profiling model of [4]

Oracle Corporation, a company that has developed a system for profile data, is oriented to the thorough investigation and close monitoring of its quality [15]. With a tool named *Oracle Data Profiling*, the user has the possibility to discover and infer rules based on data, and monitor their quality over time. As shown in Figure 2, the inputs and outputs are well defined, where data and metadata that were profiled, can be profiled again.

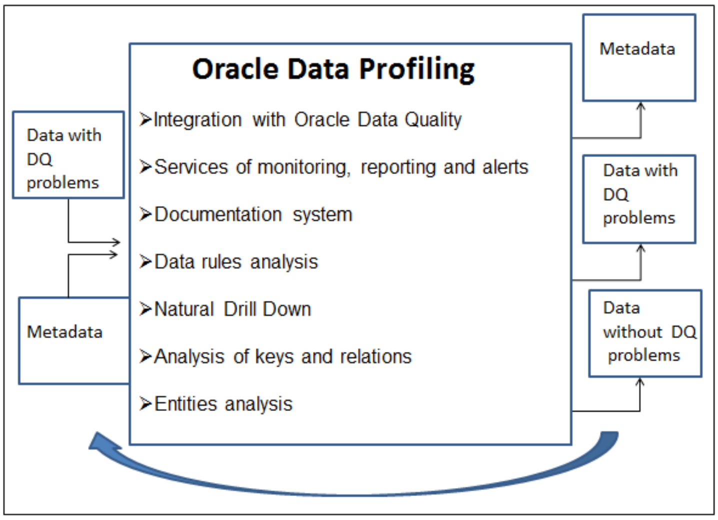


Fig. 2. Data profiling techniques and process of [15]

Microsoft offers a tool named Data Quality Services 2008 [11], with techniques and mechanism of data profiling, such as: candidates keys profiling, column profiling, data profiling using patterns, functional dependences profiling, and inclusion values profiling [11].

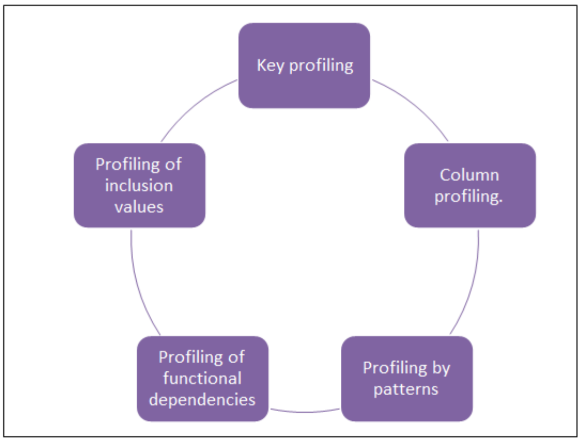

Key profiling
Column profiling.
Profiling by patterns
Profiling of functional dependencies
Profiling of inclusion values

Fig. 3. Data profiling model of [11]

The Embarcadero Company is noted for its software design: *ER/Studio*. Through its *CA ERwin Data Profiler* tool, the user can combine the analysis and the data modeling in a practical way. In its own model highlights four key activities: analysis column, integration with data models, the discovery of keys and Extended Analysis of attributes [6].

Informatica Corporation [27], with its tool named *Informatica PowerCenter*, an enterprise platform that offers access, research, data profiling and data integration from any data source, and any format. It is a very important tool for data profiling and diagnoses the DQ. As shown in Figure 4, *Informatica PowerCenter* has five subsystems: Access, Discovery, Cleaning, Integration and Delivery [23].

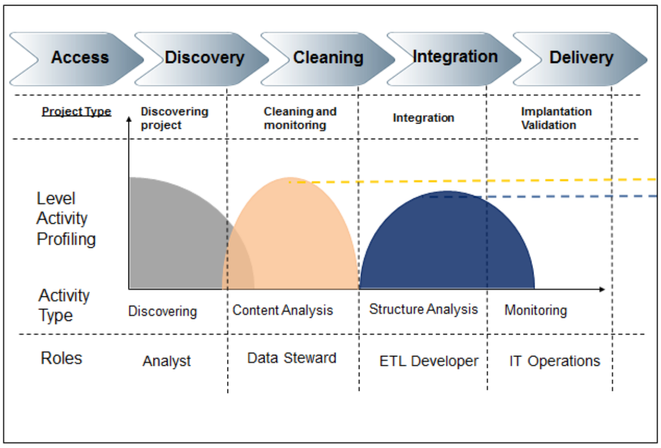


Fig. 4. Subsystems, levels, roles, activities, and techniques of data profiling of *Informatica Power Center*

CALYDAT: A METHODOLOGY FOR DATA QUALITY CONTROL, ANALYSIS AND EVALUATION BASED ON DATA PROFILING TECHNIQUES

The contribution of this paper is a methodology to control, analyze and evaluating of DQ through the use of data profiling techniques and diagnosis of the DQ attributes. It consists of three phases, each of which contains processes, activities, artifacts, people in charge and tools. It is named as: *Methodology for the **C**ontrol, **A**na**LY**sis an**D** Evalu**AT**ion of Data Quality based on Data Profiling Techniques* (CALYDAT).

Scope of CALYDAT

CALYDAT guides to the establishment of the control of DQ, based on the diagnosis of DQ dimensions and processes related with data profiling of relational databases, where activities, techniques and mechanisms are involved, as a guide for its implementation.

Unlike the others audit methodologies; CALYDAT is based on business processes of the organization, besides it defines roles and responsibilities for a better organization of the execution of its phases and activities. CALYDAT also proposes well-defined artifacts that help for documenting the implementation of the methodology, and it includes an added value: the using of the results of data profiling techniques for the DQ evaluation.

Fundamentals of CALYDAT

CALYDAT is based on the following pillars:

- *It is focused on the DQ control of each organizational business process*: Its main objective is to implement a DQ monitoring system of the organizational process analyzed, and if a DQ problem raises, enable the possibility of detecting when occurred, the area, database or information system where the problems happened, and who are the people in charge.
- *Implication of roles that manage data*: It is based on committing all the roles involved in data management, which are in charge for monitoring or controlling the data quality.
- *Iterative and incremental*: Once a CALYDAT development cycle is completed, it should be executed again so that each iteration will cover each of the organizational processes involved in access, control and management of data in the organization.

Representation of CALYDAT

CALYDAT is based on three phases: Analysis, Evaluation y Transition, as shown in the figure 5:

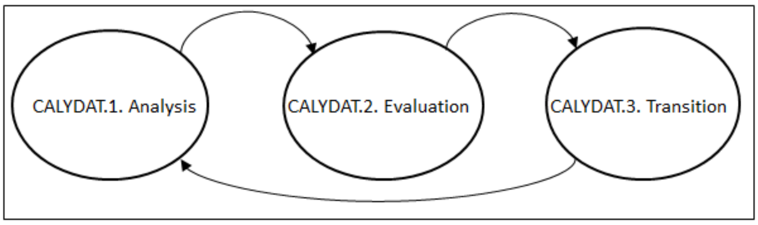


Fig. 5: Phases representation of CALYDAT.

The following subsections provide details of each one of the phases of the methodology:

CALYDAT.1. Analysis

At this phase, the current status of a particular organizational process is studied, this implies to take into account the types of existing users, the data types, database administrators, etc., for preparing the infrastructure for the application of data profiling techniques and for the survey of diagnostic of DQ dimensions. In new iterations new organizational processes will be diagnosed. Table 2 shows their characteristics:

Input Products	Description of the organizational process
Output Products	Identified information, selected dimensions for the DQ evaluation.
Activities	1.1. Diagnosis, 1.2. Election of Requirements .
Methods, techniques and tools	Expert judgment, brainstorming, artifact for the Diagnosis of organizational process (please see artifacts presented in Appendix A)
Roles	Business analyst, DQ analyst

Tab. 2: Phase of analysis.

In this phase, the activities are:

CALYDAT.1.1. Diagnosis

It must be executed to get a first assessment of the current status of the selected organizational processes. For doing so, it is necessary to take into account the databases used by the selected organizational processes, the user types and the existing roles, the types and formats of data handled by the organization and database administrators. As input of this activity, aspects related with the diagnostic process should be provided, and as output, the identified information related with the organization are to be generated.

CALYDAT.1.2. Election of the Requirements

The DQ dimensions selected will be involved in the entire cycle of execution of the methodology for each one of the selected organizational processes. As input of this activity, some aspects of the diagnostic process must be provided, and as output, the list with the selected DQ dimensions should be generated.

CALYDAT.2. Evaluation

In this phase, an evaluation of the level of DQ of a relational database should be performed. This implies the use of some techniques like structure profiling, relational profiling, data rules profiling and the implementation of surveys for the diagnosis of DQ dimensions Table 3 shows their characteristics:

Input Products	Result of the diagnosis of organizational process, data source, metadata source
Output Products	Data profiled, metadata profiled, result of the survey for the diagnostic of DQ dimensions
Activities	2.1. Structure profiling, 2.2. Relational profiling, 2.3. Data rule profiling, 2.4. Conductions of a Survey for the diagnostic of DQ dimensions
Methods, techniques and tools	Profiling of table structures, and its functional dependences, data rules profiling, questionnaire of the survey for the diagnostic of DQ dimensions (see Appendix B), data profiling tools
Roles	Business analyst, DQ analyst, database administrator, database designer

Table 3: Phase of evaluation.

To achieve the goals of this phase, the team should execute the following activities:

CALYDAT 2.1. Structure profiling

It consists of thoroughly investigate each one of the columns and rows of tables in the source systems, applying a set of techniques to calculate statistical information and metadata. The most significant DQ dimensions are completeness, accuracy and precision. As input of this activity, services of data access, profiled and not-profiled data and metadata should be provided, and as output, artifacts, data profiled and metadata profiled are to be generated.

Property profiling: It refers to applying profiling techniques to determine table properties, such as number and percent of null values, unique, duplicates, blanks, data types, minimum and maximum size of characters, maximum and minimum values and domains, among others.

- ***Regular expressions profiling*****:** It refers to applying pre-defined regular expressions to identify matches with the values of the attributes. You can define new expressions or use existing ones.
- ***Language profiling:*** Getting profiles of natural language terms and language elements stored as data, is very complex during the data profiling process. In this case, the domain plays an impor-

tant role in defining the dominant values in a column, and defining which values are written or spoken like a specific values (Matching Writing and Matching Sound respectively). For this, regular expressions or SQL statements and stored procedures or functions can be used. A repository of terms can also be used to store the letters or vowels of the alphabet and verify matches with the values of the columns.

CALYDAT 2.2. Relational profiling

The main aim of this activity is to determine possible relationships and functional dependencies between tables or business objects, and discovering primary and foreign keys. With this activity it is possible to evaluate the degree of consistency. According to [1], the DQ dimension consistency refers to the violation of semantic rules defined on a data or a particular data set. In this case it will be profiled the violations of integrity constraints, specifically inter-relational constraints. As input, the services of data access, data and metadata profiled and unprofiled are to be provided; and as output, artifacts, primary keys, foreign keys, relationships between entities and the relational matrix should be generated.

For this case, several rules, SQL statements and data mining techniques can be applied [24], for example association rules [24], to find dependency percentages of some attributes related to others, and thus find possible foreign keys.

Business analyst, DQ analyst, database designer can use the following techniques and tools to achieve their objectives:

- *Analysis of primary key*: It is used to determine those values in the attributes that are unique and are candidates for primary keys.
- *Analysis of foreign key*: It is used to determine those attributes that have been detected from the rules of inclusion and the relational matrix. The associations identified can be used to predict behavior, and to reveal correlations and occurrences of events [24]. To evaluate the rules, the support is used. As shown below, Equation 1.1 indicates the number of cases covered by the rule, and confidence; Equation 1.2 indicates the number of values of one item that belongs to another item; and Equation 1.3, referred to the confidence, it indicates the number of cases correctly predicted by the rule. Confidence is expressed as the ratio between the number of cases in which the rule is met and the number of cases in which it applies, because the premises are satisfied.

  If we consider the following item I = {A, B, C, D, E} where A, B, C, D, E are attributes of a particular database:

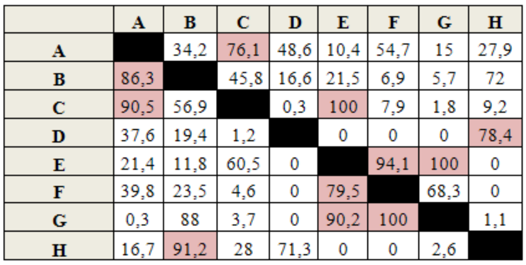
  $$\text{Support }(A) = P(A) \qquad \text{Equation. 1.1}$$

  $$\text{Support }(A \Subset B) = P(A \Subset B) \qquad \text{Equation. 1.2}$$

  $$\text{Confidence }(A \Subset B) = P(B \mid A) = \frac{P(A \Subset B)}{P(A)} \qquad \text{Equation. 1.3}$$

  Where $P(A)$ is the total value of the attribute A and $P(A \Subset B)$, the number of values of attribute A that belongs to attribute B, where B can be repeated. Confidence is the ratio between both. This will determine the confidence of each of the attributes related with the rest, identifying the higher value, which constitute potential foreign keys.
- *Relational Matrix specification*: Technique that uses a two dimensional array to detect high levels of confidence from the result of applying the rules of inclusion, and thus identify possible relationships between attributes. The intersection of two attributes corresponds to a functional dependency between them, being represented by a box with a percentage value. This value is the confidence that exists between the two attributes. An example of a relational matrix is shown in Table 4, where the highlighted values represent the largest confidences (86.3, 90.5, 91.2, 76.1, 100, 79.5, 90.2, 94.1, 100, 100 and 78.4), and therefore, possible relationships between the attributes (A, B), (A, C), (B, H), (C, A), (E, C), (E, F), (E, G), (F, E), (F, G), (G, E) and (H, D), respectively.

As tools, any of the data profiling tools presented in section 2.2 could be proposed.



	A	B	C	D	E	F	G	H
A	■	34,2	76,1	48,6	10,4	54,7	15	27,9
B	86,3	■	45,8	16,6	21,5	6,9	5,7	72
C	90,5	56,9	■	0,3	100	7,9	1,8	9,2
D	37,6	19,4	1,2	■	0	0	0	78,4
E	21,4	11,8	60,5	0	■	94,1	100	0
F	39,8	23,5	4,6	0	79,5	■	68,3	0
G	0,3	88	3,7	0	90,2	100	■	1,1
H	16,7	91,2	28	71,3	0	0	2,6	■

Table 4: Example of relational matrix.

CALYDAT 2.3. Data rules profiling.

Activity aimed to the researching, discovering, verification and validation of data rules. It helps to specify the degree of conformity, which determines whether the data has attributes that adhere to standards, conventions or regulations and similar rules relating to DQ in a specific context of use [11]. As input of this activity, services of data access, data and metadata profiled and unprofiled, and as output, artifacts and data rules.

Business analyst and DQ analyst could use the following techniques and tools to get the specified results:

- *Analysis of default data rules:* Based on existing data rules in information systems or in databases of the organization, are checked to see if the results match with what is expected of them.
- *Discovery of data rules*: These rules are conditions that may involve one or more columns. They generally use conditionals like (if, then, <,>, =).

As tools we propose the data profiling tools that implement data rules profiling.

CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ evaluation.

The survey is a system for collecting information to describe, compare and explain knowledge, attitudes and behavior [12]. In this process a qualitative diagnostic of DQ is performed from the application of the survey. As input of this activity, the result of organizational analysis is to be provided, and as output, the result of the survey of DQ dimensions should be produced. With the aim of improving the diagnosis and evaluation of DQ, there are five types of users to whom the survey is proposed. Table 5 shows some examples of types of users:

User types	**Examples of user types**
Data user	Database administrators, database developers, ETL specialists, etc.
Requirement user	Requirement analysts, requirement specialists, etc.
Technology user	Network administrators, server administrators, IT specialists , etc.
Business user	Business analysts, executives, leaders, managers, customers, area and department directors, final user, etc.
Interface user	Web programmers, designers, ads and marketing specialists, etc.

Table 5: User types and examples of user types.

Business analyst, DQ analyst and database administrator can use the following techniques and tools to achieve their objectives:

- *Conducting of survey for the diagnostic of DQ dimensions*: This method is based on the DQ characteristics provided by the ISO/IEC 25012. It should be conducted periodically to the members of the organization that interact and manage data involved in the organizational process, with questions related to each of these DQ dimensions, so as to provide a qualitative and quantitative value of the level of quality of the data used within the organization. The Details on the survey can be seen in Appendix B. As proposed tools, the data profiling ones and the questionnaire can be suggested.

3.1.1 CALYDAT.3. Transition

In this phase the organizational process analyzed is monitored, continuing to the analysis. It should be reported the status of the DQ, to all roles and members involved in the organizational business process. It implements activities related to the process of monitoring and alerting DQ. Table 6 shows their characteristics:

Input Products	Result of the survey for the diagnostic of DQ dimensions.
Output Products	Artifacts, notifications and alerts.
Activities	3.1. Monitoring and control
Methods, techniques and tools	Notification, and alerts of DQ
Roles	DQ analyst

Table 6: Phase of transition.

CALYDAT 3.1. Monitoring and control

The goal of this activity is to notify and alert events related with the detection of poor DQ in any of the selected business processes of the organization. The people in charge should ensure the beginning for repeating the phase of analysis in a new organizational process. As input, the result of the survey for diagnostic the DQ dimensions should be entered, and as output, the specification of the artifacts, notifications and alerts should be generated.

DQ analyst can uses the following technique and tools to achieve their objectives:

- *Execution of the monitoring and alert:* CALYDAT proposes the implementation of a reporting solution, for the notification of the DQ dimensions assessment to members and roles related with the organizational process, about the current diagnostic of DQ in that process.

RESULTS

In order to test the applicability of CALYDAT in a real environment, we used the methodology in an organization with well-defined business processes. Concretely, CALYDAT was applied to one business process named Control of mobile devices, where its main objective is to manage each mobile device in the agricultural fields where these mobiles work, its exact location, if they are stopped or moving, the fuel consumed, the kilometers traveled, etc. Obtained results are to be presented in this section.

We decided to apply CALYDAT to this scenario because of its own characteristics, for example, the organization has well-defined business processes, its data are stored in relational databases, it is possible to apply data profiling techniques for evaluating the DQ, also because it is a well-defined and complex organizational process where it is recommendable the use of artifacts that guide and document the application of CALYDAT. This experience involved the execution of techniques and activities of CALYDAT and the application of the survey (see the section CALYDAT 2.4. Survey) for the diagnostics of DQ dimensions.

Firstly, the members that will play the role of DQ analysts were identified; they would be the people in charge for the application of CALYDAT. DQ analysts and managers planned jointly the execution of the phases of the methodology, ensuring the availability of resources for the corresponding iterations in a periodical application of CALYDAT to other business processes of the organization.

Let’s explain the application of each phase of CALYDAT to the business process Control of mobile devices. For carrying out the phase of **CALYDAT.1. Analysis,** the activity of Diagnostic was performed, where the concepts and features of the business process were identified, using the artifact Diagnosis of organizational process (see Appendix A).

Activities	Selected DQ dimensions
CALYDAT 2.1. Structure profiling	Accuracy
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	
CALYDAT 2.1. Structure profiling	Completeness
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	
CALYDAT 2.2. Relational profiling	Consistency
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Credibility
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Currentness
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Accessibility
CALYDAT 2.3. Data rules profiling	Compliance
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Confidentiality
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Efficiency
CALYDAT 2.1. Structure profiling	Precision
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Traceability
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Understandability
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Availability
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Portability
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Recoverability

Table 7: Selected dimensions for the application of CALYDAT.

In addition, the dimensions that will be involved in the DQ evaluation were selected, as shown in the Table 7, where the *Activities* column refer to the activities of the phase of CALYDAT.2. Evaluation, where the DQ dimensions selected will be evaluated, and the *Selected DQ dimensions* column refers to the DQ dimensions that will be evaluated in each activities:

During the phase of **CALYDAT.2. Evaluation**, taking into account that the business process of the control of mobile devices and its data sources are stored in a relational database, we performed the activity of profiling structure, where highlighted the attributes **operation_date**, **mobile_state**, **year**, and **crop_cycle** (as shown in Table 8), which presented DQ problems, particularly with the completeness dimension. For example the attribute **date_operation** presented 17.5% of null values, the attribute **mobile_state**, 6.5% of null values and the attribute **year**, a minimum value of 1278.

Attribute	Entity	Description of the attribute in the data model
operation_date	operation	Attribute that stores the date of the operation performed by the mobile device
mobile_state	mobile	Attribute that stores the state of the mobile device in the mobile entity
Year	operation	Attribute that stores the year that the operation was performed
crop_cycle	crop	Attribute that stores the number of cycles of an agricultural crop determined
state_code	mobile	Attribute that stores the state code in the mobile entity.
state_id	state	Attribute that stores the identifier of the status of the mobile in the state entity
device_id	device	Attribute that stores the identifier of the tracking device

Table 8: Attributes susceptible to receive a data profiling analysis of the control of mobile devices.

After checking the degree of completeness, checking if all values for each row are complete or not, the result obtained is shown in Table 9, by each of the entities in the database (see figure 6 and Table 10):

Business entities	Percentage of the evaluation result of completeness
mobile	23 rows with incomplete values, 312 rows in total: 92,62 %
crop	7 rows with incomplete values, 215 rows in total: 96,74 %
device	49 rows with incomplete values, 378 rows in total: 87,04%
state	4 rows with incomplete values, 37 rows in total: 89,19 %
operation	18 rows with incomplete values, 193 rows in total: 90,67 %

Table 9: Verification of the degree of completeness

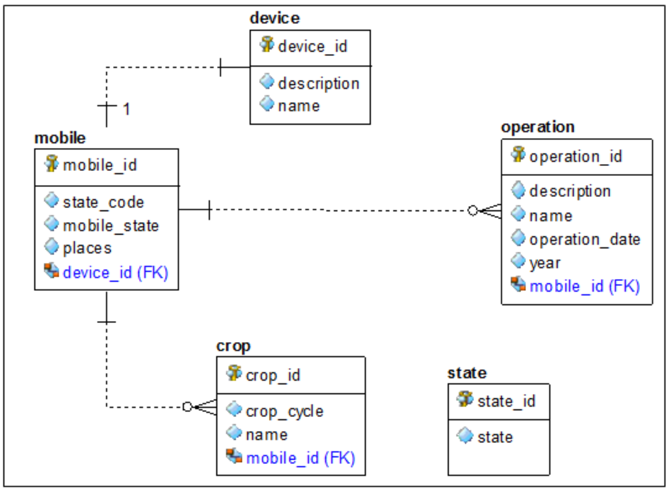


Fig. 6. Entity-Relation model of the control of mobile devices.

Entity	Description of the entity in the data model
Mobile	Entity that stores mobile devices: such as tractors, trucks, jeeps, etc.).
Crop	Entity that stores agricultural crops where worked the mobile devices
Device	Entity that stores the tracking devices carried by mobiles
State	Entity that stores the state of mobile devices
Operation	Entity that stores the operations of mobile devices

Table 10: Profiled entities of the control of mobile devices.

After the phase of CALYDAT.1. Analysis, specifically in the activity of diagnosis, it was noted that entities should not have rows with incomplete values, because in other processes it performed percentage calculations using these values, so that at least one row with incomplete values in some of these entities, represents a negative impact to the DQ dimension of completeness.

In the activity of CALYDAT 2.2. Relational profiling, after the processing, two attributes were detected as potentially relatable (state_code and state_id, see Table 8) between two unrelated entities (mobile and state, see Figure 6 and Table 10) with a 98.4% of confidence (see the section CALYDAT 2.2. Relational profiling, specifically the technique Analysis of foreign key). This helped to discover a violation of referential integrity, specifically in the dimension consistency.

Based on predefined business rules and in order to diagnose the DQ dimension of compliance of the organizational process analyzed, it was found that some rules were not complied with, such as the attribute values crop_cycle which must be in the range between 1 and 100, and were found values such as 134, 121, 106 and 189. Also that the attribute device_id must be unique, and was found the value 008 repeated twice, and the value 014 repeated three times. In crop_cycle, attribute with data type varchar, were found 7 strings, and according to business rules, should store only numeric values.

During the activity of CALYDAT 2.4. Survey, it was applied the questionnaire (see Appendix B). Candidates to participate in the survey were chosen from members who work directly with the analyzed business process. The members are the database administrators and workers of the technology department, related to the business process of control of mobile devices. In total there were 12 members: three (3) database administrators, four (4) network administrators, two (2) server administrators, one (1) security specialist, the manager and the vice-manager of technology. The result of the survey is shown in Table 11. The average column corresponds to the average values for each dimension of all applied surveys, and it is a value ranged between 0 and 5.

According with the context where data are used, in this case data are used for storing and managing information related with the exactly location of mobile devices, so the values of Table 11 become relevant. The highest percentage values, corresponds to DQ dimensions which quality is adequate. Conversely, the lower percentages are the DQ dimensions with data quality problems. As result, the critical dimensions that need an urgent attention are: compliance, precision and recoverability.

Dimensions	***Average***	**%**
Accuracy	4,12	82,4
Completeness	4,37	87,4
Consistency	4,10	82
Credibility	3,79	75,8
Currentness	4,05	81
Accessibility	4,17	83,4
Compliance	2,21	44,2
Confidentiality	3,78	75,6
Efficiency	4,19	83,8
Precision	2,92	58,4
Traceability	3,75	75
Understandability	4,56	91,2
Availability	3,98	79,6
Portability	4,46	89,2
Recoverability	2,73	54,6

Table 11: Results of the survey for the diagnostic of DQ dimensions of the process analyzed.

During the execution of the phase of **CALYDAT.3. Transition**, we proposed the creation of a web site in the organization, with the corresponding levels of access, and based on the types of users. The notification of the diagnosis of DQ dimensions should be weekly. It was advised to the managers that they should select the data profiling tool, according to their needs and possibilities, and repeat the survey frequently, including others business process of the organization.

CONCLUSIONS

This main contribution of this paper is CALYDAT, a methodology for the analysis, control and evaluation of DQ, through data profiling techniques and the application of surveys for the diagnostic of DQ dimensions, to various types of users. Its application in a real environment was satisfactory and provided the expected results, giving to the managers and members involved in the organizational process of control of mobile devices, a quantitative and qualitative evaluation of DQ. The type of user plays a fundamental role, which offers the possibility to detect more effectively the DQ problems, based on the role to which is directed the survey. As research methods during the process of developing the methodology, we used theoretical and empirical methods [28], including the method of survey. We empirically obtained the DQ dimensions used in CALYDAT, the types of users to which the questionnaire should be applied, the roles and responsibilities defined and the output products of the analysis phase. For the success of CALYDAT, it was necessary to consider the systemic method as a combined and integrated system of all phases and activities, with an iterative and incremental approach. Finally, the survey plays a key role for the evaluation of the DQ in CALYDAT.

In the future, we intend to develop a tool that supports the application of CALYDAT. This tool will have functionalities that allow execute data profiling techniques, and mechanisms for diagnosis the DQ dimensions: Accuracy, Completeness and Precision.

BIBLIOGRAPHY

[1] Batini C., Cappiello C., Francalanci C. and Maurino A., *Methodologies for Data Quality Assessment and Improvement*, ACM Computing Surveys, Vol. 41, No. 3, Article 16, 2009.

[2] Catarci, T., and Scannapieco, M. 2002. *Data quality under the computer science perspective*. Archivi Computer 2.

[3] Chengalur-Smith, I. N., Ballou, D. P., and Pazer, H. L. 1999. *The impact of data quality information on decision making: An exploratory analysis*. IEEE Trans. Knowl. Data Eng. 11, 6, 853–864.

[4] E. Olson, Jack. *Data Quality—The Accuracy Dimension*. San Francisco, Elsevier Science, 2003.

[5] Eppler, M. and M¨Unzenmaier, P. 2002. *Measuring information quality in the Web context: A survey of state-of-the-art instruments and an application methodology*. In Proceedings of the 7th International Conference on Information Systems (ICIQ).

[6] Erwin Studio, CA ERwin Data Profiler. 2009 [Accessed in 2009 October]; Available from: http://www.ca.com/us/products/Product.aspx?ID=8235

[7] Ferdinandi, P.L. *Data warehouse advice for managers*, New York: AMACOM American Management Association., 1999.

[8] Gomes, J.F., Maria José Trigueiros, *A Data Quality Metamodel Extension to CWM*, in *4th Asia-Pacific Conference on Conceptual Modelling (APCCM 2007)*. 2007: Australian.

[9] ISO-25012, *ISO/IEC 25012: Software engineering - Software Product Quality Requirements and Evaluation (SQuaRE)* - Data quality model. 2008.

[10] Jarke, M., Lenzerini, M., Vassiliou, Y., and Vassiliadis, P., Eds. 1995. *Fundamentals of Data Warehouses*. Springer Verlag.

[11] Wong, A., Sutcliffe D., (2009). *Data Quality Services 2008.* 2008 [Accessed in 2009 Septem-

ber]. Available from: http://msdn.microsoft.com/en-us/library/dd129900%28v=sql.100%29.aspx

[12] Lawrence, S., Kitchenham B., "Principles of Survey Research. Part 1: Turning Lemons into Lemonade" *Software Engineering Notes*, 2001. vol 26 no 6 pp. 16

[13] Lee, Y.W., Strong, D. M., Kahn, B. K., and Wang, R. Y. 2002. *AIMQ: A methodology for information quality assessment. Inform*. Manage. 40, 2, 133–460.

[14] Naumann, F. 2002. *Quality-driven query answering for integrated information systems*. Lecture Notes in Computer Science, vol. 2261.

[15] Oracle Corporation, Oracle Data Profiling. 2007 [Accessed in 2012 June]; Available from: http://www.oracle.com/technology/products/oracle-data-quality/pdf/oracledp_datasheet.pdf

[16] Microsoft Corporation, Microsoft Developer Network, MSDN [Accessed in 2012 July]; Available from: http://msdn.microsoft.com/en-us/library/ff877917.aspx

[17] Redman, T. 1996. *Data Quality for the Information Age*. Artech House.

[18] Rhind, Graham. *Poor quality Data. The pandemic problem that needs addressing*. 2007.

[19] Su, Y. and Jin, Z. 2004. *A methodology for information quality assessment in the designing and manufacturing processes of mechanical products*. In Proceedings of the 9th International Conference on Information Quality (ICIQ). 447–465.

[20] TDWI- The Data Warehouse Institute. 2009 [Accessed in 2009 October]; Available from: http://tdwi.org/research/2009/09/mr-who-ensures-clean-consistent-data.aspx?sc_lang=en

[21] Wand, Y. and Wang, R. 1996. *Anchoring data quality dimensions in ontological foundations*. Comm. ACM 39, 11.

[22] Wang, R. and Strong, D. 1996. *Beyond accuracy: What data quality means to data consumers*. J. Manage. Inform. Syst. 12, 4.

[23] Mínguez, A. *Fundamentos de Calidad de datos*. [Accessed in 2012, March]. Available from: http://seminarisempresa.fib.upc.edu/aulesempresa/2009/programes/POWERDATA.html

[24] R. Agrawal and R. Srikant. (June 1994). *Fast algorithms for mining association rules in large databases*. In Research Report RJ 9839, IBM Almaden Research Center, San Jose, CA.

[25] Bovee, M., Srivastava, R., and Mak, B. September 2001. *A conceptual framework and belief-function approach to assessing overall information quality. In Proceedings of the 6th International Conference on Information Quality*.

[26] Long, J. and Seko, C. April 2005. *A cyclic-hierarchical method for database data-quality evaluation and improvement. In Advances in Management Information Systems-Information Quality Monograph (AMISIQ) Monograph*, R. Wang, E. Pierce, S. Madnick, and Fisher C.W.

[27] Free informatica tutorials; [Accessed in May, 2012]. Available from: http://free-informatica-tutorials.blogspot.com/

[28] Hernández, R. and Coello, S. 2002. *El paradigma cuantitativo de la investigación científica*. EDUNIV. La Habana, Cuba. 82-96.

APPENDICES

Appendix A

Diagnosis of organizational process
Deliverable
<Organization name>
<Organizational process name>
<Version>
Version control

Date	Version	Description	Author
dd/mm/yy>	<x.x>	<Details>	<Name>

Introduction

Purpose

[Define the main objective for the evaluation and diagnosis of organizational process.]

Scope

[It specifies which business processes and DQ dimensions shall apply. In this case the artifact for the diagnosis of organizational process will integrate with the survey for the Diagnosis of the DQ dimensions.]

References

[List of referenced documents]

Code	Title
[1]	Document 1
[2]	Document 2

Glossary

[In the glossary specifies a group of basic terms that are managed for the diagnosis of organizational process.]

Description of the diagnosis application

[It describes the implementation strategy for the diagnosis of the organizational process.]

Summary of the diagnosis in the business process:

[Summary of the results of the diagnosis of organizational process.]

Analysis of significant results:

[Analysis of the most relevant results obtained in the diagnosis and a summary of the main factors to consider.]

Conclusions

[Conclusions of the diagnosis of the organizational process.]

Appendix B

Survey for the diagnostic of DQ dimensions

Deliverable

<Organization name>

<Organizational process name>

Introduction

This survey is defined for the investigation and the correct diagnosis of the DQ dimensions, where the participation of roles and members who interact and use the data is very important. Below are a number of aspects which should be marked with an X the value in the scale that is considered appropriate to characterize the current state of data quality dimensions. In case of indecision or ignorance in any aspect, please do not make any X in the corresponding aspect. The collection is a term used in the survey for referring to data or data set that will be analyzed by each of the dimensions.

General aspects:

Line/Area/Group where it belongs: ____________________________________

Role played: ____

Alternatives to respond to an aspect are listed below:

Nomenclature	A	B	C	D	E
Qualitative equivalence	Yes, quite	Yes, but not enough	Little	Very little, almost none	No, none

Survey: Diagnostic of DQ dimensions

Taking into account the following initial requirements:
The business area to diagnose: ___________________
The business concept to diagnose: _______________
The source, data source or agent in charge (person or system) to enter data: ________________
The data or the data set that must be evaluated:
In the range of time: From: __________(day/month/year)   To: __________(day/month/year)

		A	B	C	D	E
Accuracy	Does the collection have the value and the actual characteristics expected?					
Completeness	Is the collection completed and has all the expected values?					
Consistency	Is the collection free of inconsistencies, contradictions in relation to other data?					
Credibility	Does the collection have adequate credibility and reliability?					
Currentness	Do you think the collection is updated with respect to the specified time range or with respect to the current time?					
Accessibility	Can be the collection properly managed through its access?					
Compliance	Does the collection comply with business rules or restrictions?					
Confidentiality	Does the collection have the appropriate confidentiality and security?					
Efficiency	Does the collection have the expected levels of efficiency and performance?					
Precision	Does the collection have the adequate accuracy and precision?					
Traceability	Is the access to the collection being audited by traces or tracks?					
Understandability	Is the collection understandable and interpretable by users?					
Availability	Can the collection be properly retrieved by authorized users or applications?					
Portability	Will maintain the collection its quality if is moved from one system to another?					
Recoverability	Will maintain the collection its quality despite occurrences of failures?					