A DATA QUALITY METHODOLOGY FOR HETEROGENEOUS DATA

Batini Carlo[1], Barone Daniele[1], Cabitza Federico[1] and Grega Simone[2]

[1] Università degli Studi di Milano Bicocca, Milan, Italy
{barone,batini,cabitza}@disco.unimib.it
[2] Nextt Lab S.r.l.Via Benedetto Croce 19, 00142 Roma, Italy
simone.grega@nexttlab.it

ABSTRACT

We present a Heterogenous Data Quality Methodology (HDQM) for Data Quality (DQ) assessment and improvement that considers all types of data managed in an organization, namely structured data represented in databases, semistructured data usually represented in XML, and unstructured data represented in documents. We also define a meta-model in order to describe the relevant knowledge managed in the methodology. The different types of data are translated in a common conceptual representation. We consider two dimensions widely analyzed in the specialist literature and used in practice: Accuracy and Currency. The methodology provides stakeholders involved in DQ management with a complete set of phases for data quality assessment and improvement. A non trivial case study from the business domain is used to illustrate and validate the methodology.

KEYWORDS

Data quality, Methodology, structured data, semistructured data, unstructured data

1. INTRODUCTION

1.1 Background and Motivations

Information and communication technologies have extraordinarily increased the amount of information that is managed within organizational processes. In the last years, academic researchers and private consultants have shed light on the correlation between information and organizational processes that either consume or produce data. Consequently, the quality of these data assumes a crucial role [27]. In fact, low quality of data a) can be a proxy indicator of either low process quality or loose control of process performance, and b) can have an impact on the organizations' ability to fulfil the needs of their customers and create value efficiently and effectively. Since data are almost ubiquitous in modern organizations, assessing and improving their quality can be complex, mainly for two reasons: on the one hand, data are not given on their own in the organizational domain; rather, data are always given in a specific context. This context encompasses precise requirements about the inner quality of data as well as about how tasks are actually performed (practices) with respect to reference models of action (processes). This means that Data Quality (DQ) must be evaluated according to the expectations of users and consumers (cf. the concept of quality as fitness for use [35]). Accordingly, human actors must be involved in any DQ assessment program and the related consequences must be addressed coming from the involvement of practitioners who could have different competencies, responsibilities and diverging interests about data.

On the other hand, organizational data are increasingly distributed in heterogenous resources and represented with different formats, even if they refer to the same organizational entities, ranging from almost unstructured, e.g. in file systems, document repositories and on the web, to

DOI: 10.5121/ijdms.2011.3105

highly structured, e.g. in database management systems. In the literature, data sources are classified depending on the level of structure that characterizes them. According to a widely accepted yet quite informal convention, the literature distinguishes between data sources in terms of

1. *Structured* data, if their formal schema (i.e., formats, types, constraints, relationships) is defined and explicit.
2. *Semistructured* data, (sometimes also called "self-describing" data [2]) if data are something in between raw data and strictly typed data, i.e., when they have some structure, but it is not as rigidly structured as in databases [1]. They are usually represented with XML markup language.
3. *Unstructured* data, if they are but sequences of symbols (at least at a human reader), e.g., a string in natural language, where no specific structure, type domains and formal constraints are defined.

Differences in the format of data are necessarily reflected in the methods and techniques that organizations use in order to assess and improve the quality of their information resources. These differences lead to methodologies that support the selection and implementation of DQ improvement programs that are tailored to specific needs and domains. For instance, large databases are maintained by data cleansing and record matching techniques, whereas organization's documents are improved in quality by adopting more structured templates and more formal and unambiguous lexicons and by establishing internal auditing procedures.

1.2 Research Objectives

Nowadays, organizations must cope with data represented in different format; therefore, the main point that we address in this paper is that a comprehensive and general approach towards the DQ improvement of all these types of data is feasible and indeed necessary. Within a single organization, for instance, an important concept such as that of 'customer' can be mapped either into a corporate database, in electronic accounts or invoices, as well as in the address books of its agents. In order to give a contribution to the literature on data quality that focuses on such a comprehensive approach, we present the Heterogeneous Data Quality Methodology (HDQM) and the underpinning meta-model. Our aim is to address the topic of how to improve the quality of key single information sources; not only in light of the information sources involved in an organization, but above all, in light of the conceptual entities that they represent. In other words, the main idea underpinning HDQM is *to map the information resources used in an organization to a common conceptual representation and then to assess the quality of data considering such homogeneous conceptual representation*. As a result we achieve two goals: on one hand, we reach the flexibility and modularity that is needed when coping with users of different departments (e.g., a salesman and an IT developer) and levels (e.g., an IT project manager and the CEO). On the other hand, by assessing data quality on each organizational (information) resource and composing DQ values at the level of the common concepts represented we provide organizations with a wider selection of improvement strategies they can undertake to achieve their quality targets. For instance, provided that the overall quality of the information regarding customers must increase by, say, five percent (resulting in less costs and missed opportunities), HDQM encompasses methods to choose what resources referring to customers should be improved first or more intensively, and how the objective should be reached.

HDQM is an extension of the Comprehensive Data Quality (CDQ) methodology [7] towards semistructured and unstructured relational data sources. The scope of our contribution is within the DQ area and, therefore, does not cover issues that are addressed by the wide range of disciplines investigating unstructured textual data, such as natural language processing and information retrieval, or other unstructured sources (i.e., images and sounds). Covering also unstructured relational data means to consider information resources that are usually neglected in other approaches to data quality improvement but that bear great importance in modern organizations. In fact, raw tables in documents, item lists or files where aggregated sets of values are separated by some delimiter (e.g., a comma, a colon) are often the first stage in which relational data are managed by people, before their coding in either XML files (for electronic exchange) or databases (for querying and storing). In unstructured relational data, no information about data types and constraints is given, besides the mere content of the relation (or table) itself. For this reason, they are different from semistructured documents, in that they are associated with no well-defined schema. Focusing on relational data allows us to assume that the information resources considered within the HDQM are defined both in intensional terms, i.e., they cover and make different aspects of organizational concepts explicit, and in extensional terms, i.e., they associate different values to different aspects of the concept they represent. Future work will address how HDQM can be extended to other unstructured data types (like texts, images and sounds) by encompassing dimensions, metrics and improvement techniques tailored to specific type of data.

The paper is organized as follows. In Section 2, we present the related work. In Section 3, we introduce a simplified but real case study to which HDQM is applied and tested against results. In Section 4, we present the meta-model that encompasses the main concepts and dimensions that HDQM considers. Section 5 classifies and defines DQ dimensions and metrics for heterogeneous data. In Section 6, we outline HDQM, and highlight the new phases and activities that are introduced to cope with heterogeneous types of data. Sections 7, 8, and 9 describe in detail the three main phases of HDQM, namely, State reconstruction, Assessment, and Selection of the optimal improvement process. Finally, Section 10 draws the conclusions and outlines the future lines of work.

2. RELATED WORK

In this paper, we address several problems related to the quality of heterogeneous data, such as methodologies for DQ assessment and improvement, DQ dimensions and composing techniques. The large majority of research contributions in DQ methodologies focus on structured and semistructured data: see [8] for a comprehensive survey. In the following, we propose a comparison between HDQM and the DQ methodologies shown in Table I in order to emphasize our original contribution and peculiarity. In general, a DQ methodology provides a set of guidelines and techniques that, starting from input information that describe a given application context, define a rational process to assess and improve the quality of data. The main sequence of activities of a DQ methodology encompasses three phases: *State reconstruction*, *Assessment/Measurement* and *Improvement*. HDQM keeps this structure and enriches the steps defined in each phase or, in some cases, introduces distinctive ones.

In the State reconstruction phase, few methodologies consider different types of data managed by the organization. For example, AIMQ uses the generic term information, and performs qualitative evaluations using questions which apply to structured data, but may refer to any type of data, including unstructured ones. In regards to semistructured data, the DaQuinCIS methodology proposes a model that associates quality values with XML documents. The model, called Data and Data Quality (D2Q), is intended to be used in the context of data flows exchanged by different organizations in a cooperative information system. The quality values can be associated with various elements of the data model, ranging from individual data values to whole data sources. The HDQM proposes to consider all types of data in the State reconstruction phase by using a model that allows for the description of information depending on different levels of abstraction (see the HDQM meta-model in Section 4). Furthermore, HDQM drives the measurement and improvement of DQ dimensions that are associated with each of the different types of data. Another contribution of HDQM is related to the effectiveness and usability of techniques used to relate the data sources of an organization with processes, organizational units and dimensions. For example, TDQM uses suitable tools such as IP-Map [29] in order to diagram and analyze the process by which data are manufactured. Complex solutions such as IP-MAP cannot always be adopted due to their high costs and the practical unfeasibility of the step of thorough process modelling. For these reasons, HDQM adopts less formal, but more feasible solutions. For instance, the set of matrixes, within the HDQM meta-model, describe the main relationships among all data sources.

Table I: Methodologies considered in the comparison with HDQM

Methodology Acronym	**Extended name**	**Main reference**
TDQM	Total Data Quality Methodology	[33]
TIQM	Total Information Quality Management	[13]
AIMQ	A methodology for information quality assessment	[18]
CIHI	Canadian Institute for Health Information methodology	[19]
ISTAT	ISTAT methodology	[14]
COLDQ	Loshin Methodology (Cost-effect Of Low Data Quality)	[20]
DaQuinCIS	Data Quality in Cooperative Information Systems	[28]
CDQ	Comprehensive methodology for Data Quality management	[7]

In the Assessment/Measurement phase, a number of DQ methodologies (e.g. TIQM, AIMQ, ISTAT) manage a fixed set of DQ dimensions (and metrics) and their approach is strictly “hardwired” to these dimensions. A distinctive example is represented by the ISTAT methodology which focuses on how to guarantee the quality of data integrated from multiple databases of local Public Administrations. Besides providing a detailed measurement for a limited set of DQ dimensions, ISTAT suggests improvement procedures that are applicable only for these dimensions. HDQM addresses this problem by defining an approach that can be easily generalized to any dimension, despite the fact that here we present it with respect to Accuracy and Currency. The presented approach also tries to extend the assessment techniques proposed in the literature for structured and semistructured data to unstructured data. In particular, for structured and semistructured data, the quality is usually measured along DQ dimensions such as Accuracy, Completeness, Consistency and Currency, since they are context independent and associated with consolidated assessment algorithms (e.g., [23]). Instead, for unstructured data, the assessment techniques are less consolidated at a general level and they are only analyzed in particular domains such as Web Information Systems [10], design documentation [24] and the archival domain [17]. Another important issue within the Assessment/Measurement phase is the definition of techniques to compose DQ dimensions. The problem of defining a composition algebra for DQ dimensions has been considered in several papers (e.g., [23]). In particular, there are strong similarities between our approach and the one described in [30] that illustrates how to evaluate the Completeness DQ dimension, by extending a previous work on Timeliness and Accuracy [5] on the basis of the notion of managing information as a product -Information Product (IP) [34]. An IP is composed of both raw data elements and component data elements. Their evaluation approach encompasses the calculation of Completeness at any level of the IP by using aggregation operations. These are mathematical functions used to compose the Completeness value on the basis of the different levels of the IP. Our composing technique extends this approach by considering two weights in the composition function, which are measured by quantitative metrics and represent the relevance and the scope of the data.

In the Improvement phase, the DQ methodologies generally adopt two types of strategies, namely *Data-driven* and *Process-driven*. Data-driven strategies improve the quality of data by directly modifying the values of data. For example, an incorrect data value can be fixed with the data taken from a more accurate database. Process-driven strategies improve quality by redesigning the processes that create or modify data. For example, an acquisition process can be redesigned by including an activity that verifies the accuracy of data before storing them. TDQM offers the possibility of applying only the process-driven strategy by using the Information Manufacturing Analysis Matrix [5], which suggests when and how to improve data. It is worthwhile to note that a methodology exclusively adopting either a data-driven (as for DaQuinCIS) or a process-driven strategy may not be flexible for organizations that have DQ practices. The only methodologies that openly address this issue are the CDQ and the HDQM, which jointly select data-and process-driven techniques. The selection of the most suitable strategy and improvement technique is based on domain-dependent decision variables. Furthermore, HDQM discusses new techniques to improve the quality of unstructured data (see Section 9.2). Another HDQM contribution in the improvement phase is an extension of the specific techniques proposed only in TIQM, COLDQ and CDQ for cost-benefit analysis. For example, COLDQ focuses on the evaluation of the costs associated with poor data quality, as an argument for supporting the investment in a knowledge management program. Instead, HDQM proposes a more qualitative approach or cost-benefit analysis in order to define the data quality targets and to guide the selection of the most suitable improvement process (see Section 9.3).

3. THE RISTOBILL CASE STUDY

In this section we outline the main requirements and characteristics of the case study that we consider to illustrate HDQM. The core business of a private firm, the Ristobill Ltd., is to develop innovative systems for wireless handheld order entry systems. These systems are used by waiters to collect orders from patrons at their tables and communicate with the kitchen in real time through a wireless connection. As the majority of businesses, the main entities at Ristobill are those of Customer and Supplier. In this example, we will concentrate on the Customer entity and mention two suppliers of customer profiling information. Typical customers of Ristobill are restaurants, pubs, bistros, snack bars and their licensees. At Ristobill, three business units are directly involved with customers, although in very different ways and with different quality requirements on the Customer entity.

The *Marketing Department* (MD) and its network of commercial agents are supposed to either seek new customers or propose new solutions and upgrades to old ones. MD agents need to have very precise information on the profile of potential customers as this can be acquired from specific vendors and aggregated along several dimensions, like region, turnover, and cuisine. MD agents are also the first contact between Ristobill and the market, and provide the rest of the enterprise with valuable information on what customers need, are willing to pay for and have paid for. The *Technical Department* (TD) is supposed to monitor the well running of sold installations and provide both ordinary and extraordinary maintenance upon on it. In order to give apt and timely assistance, TD members must then rely on information about customers regarding which systems they purchased, at which level of quality of service and where they are exactly located. Lastly, the *Accounts Department* (AD) needs accurate and up-to-date administrative information for invoice drawing and accounting.

The Customer entity at Ristobill is represented by three main data sources depicted in Figure 1, that also shows the data flow and processing phases for them. They are:

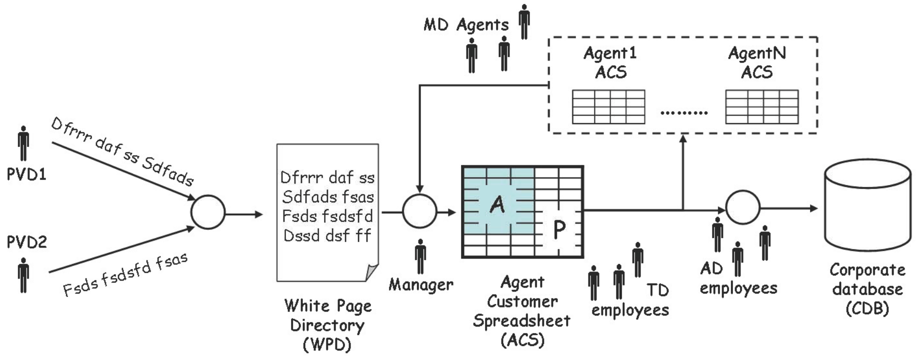


Figure 1: The data processing described in the case study

1. A huge *White Page Directory* (from now on, WPD) that is created starting from the data streams coming from two different profiling providers (PVD1 and PVD2 in Figure 1). The two providers focus on different types of information on potential customers and their output is assumed to be an accurate representation of the reality of interest. PVD1 was chosen for provision of complete administrative data (e.g., who is the owner of a facility and the facility's accurate name) as well as geo-localization data (like facility address, phone number and GPS coordinates). Conversely, PVD2 has been chosen for its highly valuable information about business typologies and market shares (e.g., market sectors, average billing and turnover). These two providers offer a complete service catalog, that Ristobill subscribed for the cheapest service, i.e. the provision of unstructured web pages. These two parallel data flows are then merged at Ristobill into one long web page which is updated every two weeks and rendered on the Ristobill intranet. The merging program assumes that different rows from the HTML streams are between < tr >< /tr > tags; this program accomplishes some syntactic normalization by resolving common abbreviations on addresses and conventions on phone numbers (like sq for square or slashes for dashes, respectively). After this normalization the merging program performs a first record linkage activity to identify information referring to the same customer, performed with simple distance functions and regular expression evaluation (for instance, fields are any string between two consecutive < td > tags). Elements (simple rows within the HTML stream) that are recognized as pertaining to the same object are just appended without further processing in one single WPD row. Since the customer directory comes from two different data sources with no particular constraint on data types and values, it is rendered into an almost unstructured source, where data are represented with possible redundant and overlapping data on trade names.

2. An *Agent-Customer Spreadsheet* file (ACS) that is partly derived from the WPD and contains some further fields to be filled in the field of work. Examples of such fields are i) a valuable "remark" field (in which agents jot down their personal experience on how to approach a certain customer), ii) a field to annotate the secondary address for the charge and discharge of goods and iii) a field for the typology of a stipulated contract. Each agent gets a copy of this source stored in laptop, and is supposed to use and compile it in daily job of social networking, customer first contact and business negotiations. In this process, the source is partitioned in two subsets: the portion of records pertaining to acquired customers (A in Figure 1) and the portion of potential customers (P in Figure 1). Every week an account manager merges the various spreadsheets coming from their agents and creates one global file where buyers are characterized with data that will be then used by the TD (e.g., serial number of the purchased product, type of contract and restaurant address) and the AD (e.g., identifiers such as fiscal codes and V.A.T. numbers, ownership data and company full

names). Since any field in the spreadsheet is strictly typed but with no semantic constraints, the ACS can be considered semistructured.

3. Every two weeks, a semiautomatic procedure parses the shared ACS spreadsheet and feeds its data into a *Corporate Data Base* (CDB). Some clerks are supposed to watch for duplicates, evident outliers and errors that are reported by colleagues from the TD and AD. Specific business rules check the semantic integrity between fields and foreign keys. While TD employees must usually use both the ACS and CDB to know exactly where to ensure service assistance, clerical workers from the AD exclusively rely on the CDB data to issue invoices and to keep Ristobill books updated.

4. THE HDQM META-MODEL

The HDQM meta-model shown in Figure 2 represents the main types of organizational knowledge managed within the steps of HDQM. In order to represent the concepts of the HDQM meta-model, we use the Entity Relationship Model [12]). The concepts are as follows.

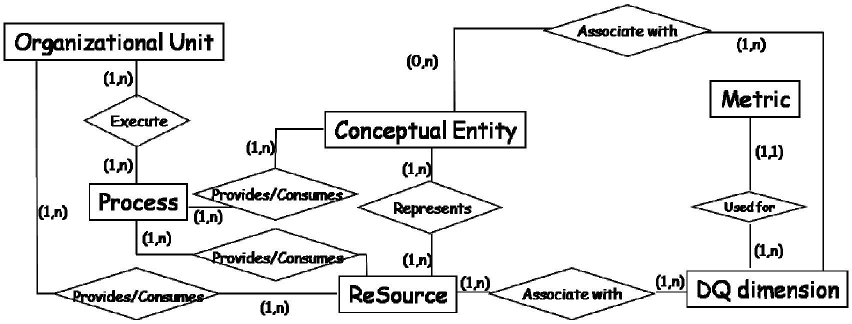


Figure 2: The HDQM meta-model

A *Process* is an articulated sequence of activities that are performed by and within the considered organization. Processes are considered since their efficiency and effectiveness are influenced by data quality. An *Organizational Unit* is a significant element of an organization that is involved in data production, use and processing; it is characterized by an internal structure and a set of internal rules. A *ReSource* (RS) is any information source that an organization can either use or get access to in order to represent some aspects of the reality of interest. The term ReSource comes from the fact that RSs can be seen as either business assets – i.e., sources of supply by which to get information (i.e., a resource) – or also as the origin itself of these data (i.e., a source). Within a business organization, RSs are managed during processes of either information production or information consumption. Typical examples of RSs are databases, data flows, either electronic or paper-based documents. Data represented in RSs can be either structured, semi-or unstructured.

In our case study, there are three RSs: i) the White Page Directory (WPD), ii) the Agent-Customer Spreadsheet (ACS) file, and iii) the Corporate Data Base (CDB). These are, respectively, an unstructured, semistructured and structured data RS.

A *Conceptual Entity* (CE) is any concept that refers to a single phenomenon of the reality of interest and that is possible to abstract from the RSs used within an organization: e.g., customer, supplier, facility, commodity. A CE refers to a concept that is independent of the specific way a single RS represents it, as well as of the RS's physical medium and format.

In our case study, the White Page Directory (WPD) refers to a single CE, that of Customer; The Agent-Customer Spreadsheet (ACS) refers to the Potential CustomerCE and to the Acquired CustomerCE; the tables of the Corporate Data Base (CDB) refer to different CE (e.g., Customer, Solution, etc. . .).

The HDQM meta-model proposes to explicitly distinguish between RSs and CEs because both the dimensions pertaining to the quality of RSs and to the quality of CEs must be addressed, each with the most suitable techniques for the dimension under consideration. Moreover, our point is that the tight tie between RSs and the corresponding CEs (represented therein) must be considered to enable and support a comprehensive approach to data quality assessment and improvement.

> *In our case study, Customer is an example of CE. A customer table within a CDB can be the most convenient and reliable means of storing confidential information that must be queried along several and often quite complex dimensions; conversely, ACS files can be convenient means of quickly and easily transmitting customer directories to multiple travelling salesmen asynchronously. The WPD can be used by the agents of the MD to quickly get access to customer's information.*

In the HDQM meta-model, one CE can be referred by multiple RSs within an organization, and vice versa. Moreover, each CE is also associated with the processes in which it is involved to the organizational units it belongs to and to the RSs that refer to it.

5. DQ DIMENSIONS FOR HETEROGENEOUS DATA

DQ dimensions are quality properties that characterize a certain resource and are measured by applying a suitable metric to the RSs. We follow the definition given within the ISO standard 9126-1 and conceive of a metric as both a measurement procedure and a proper unit of measure, i.e., the domain of values returned by the measurement procedure. In what follows, our aim is to describe the considered dimensions and related metrics that can be associated to RSs and the corresponding CEs. Since a single CE can be related to several RSs, we also need to introduce the concept of composing function. This function, given a CE, allows us to compute the value of a given quality dimension for the CE at hand from the values of the dimension regarding the related RSs.

We consider two dimensions that have been widely analyzed in the specialist literature and used in practice: *Accuracy* and *Currency*. Accuracy is defined as the closeness between a value v and another value v. of the domain D, which is considered as the correct representation of the real-life phenomenon that v aims to represent [9]. In the following, we will discuss what in the literature is called Syntactic accuracy, since this is the type of Accuracy mainly proposed in the assessment of the DQ methodologies. Syntactic accuracy is measured by means of comparison functions that evaluate the distance between v and the elements belonging to D. For structured data we can calculate the Syntactic accuracy of a tuple t of a relation using the metric (1) based on [26], where r_i is the i-th value of the tuple t, |t| is the number of attributes of the considered tuple and acc(r_i, D(r_i)) is defined as (2).

$$(1)\ Acc(t) = \frac{\sum_{i=1}^{|t|} acc(r_i, D(r_i))}{|t|} \qquad (2)\ acc(r_i, D(r_i)) = \begin{Bmatrix} 1 & if\ r_i \in D(r_i) \\ 1 - NED(r_i, D(r_i)) & otherwise \end{Bmatrix}$$

The acc(·,·) returns a value equal to 1 if there is an exact matching between the value r_i and its closest value in D(r_i), that is, r_i is syntactically accurate. Otherwise the function returns the normalized edit distance [11] which takes into account the minimum number of character insertions, deletions, and replacements required to convert a value r_i to a value in D(r_i). In particular, NED is a metric valued in [0, 1] where the value 0 means an exact match. For semistructured data, the metrics for Accuracy are more complex since no complete information about the structure and semantics of data is usually available; thus, we have to preliminarily create an association between the values contained in the semistructured data sources and the related domains. To do this, the XML schema can be extracted (as described in [15]) in order to identify (i) the different elements in the XML document (e.g. the ‘Customer’ element), (ii) the different properties of an element (e.g. the ‘Name’ property for the ‘Customer’ element) and (iii) the associated domain for each property (e.g. the Name domain for the ‘Name’ property). For each property p of an element, it is then possible to identify its associated value ri, where each ri pertains to a specific domain D(ri). Finally, applying the same formula (1) given for the structured data, we calculate the Accuracy of the element by using ri as a value speci.ed in a property and |p| as the number of properties. For unstructured data, we adopt a metric that uses the structured techniques mentioned above, once single objects have been extracted and classified by means of information retrieval techniques. To our aims, we propose an approach to perform the object identification and classification that is based on the standard information extraction task presented in [22]. Information extraction is a process employed to fill in the fields of a database when only unstructured or loosely formatted text is available. The process encompasses i) the object identification finds the starting and ending boundaries of the text segments that are traceable back to a particular CE; ii) the object classification determines which domain D has to be associated to each object. For instance, “John Smith” could be a text snippet associated with the Name domain.

Once structuring techniques, as the above mentioned, have been applied, it is possible to use the same formula (1) given for the structured and semistructured data to calculate the Accuracy of the object o, where r_i is a segment specified for the object o and |o| is the number of these segments. For the three types of data, it is now possible to apply the following general formula to calculate the Accuracy of an RS, where the accuracy(X_i) function is applied on each tuple t_i for structured data, element e_i for semistructured data and object o_i for unstructured data and where n is the number of tuples, elements or objects identified in the RS.

$$(3)\ Accuracy(RS) = \frac{\sum_{i=1}^{n} Accuracy(X_i)}{n}$$

The second DQ dimension that we consider is Normalized currency. We define Normalized currency as the ratio between Actual and Optimal currency. Currency is usually defined (e.g., [9]) as the “temporal difference between the date in which data are used and the date in which data have been updated”. Therefore, Normalized currency is the ratio between the minimum timespan that data have become old (Optimal currency) and the Actual currency of these data. More specifically, Normalized currency concerns how promptly data are updated with respect to how promptly they should be with respect to users needs and the main domain constraints (e.g., service costs, providers’ availability). Here we consider this dimension instead of the related Currency because we use percentages instead of time intervals as in the case of Currency. Percentages indicate how much an indicator deviates from the ideal standard; in doing so, they result in quality measures that can be easily compared with each other and make related metrics of different DQ dimensions homogeneous throughout the next phases of the methodology. The metric of Normalized currency is defined as:

$$(4)\ Cur(RS) = \frac{OptimalCurrency}{ActualCurrency} = \frac{OptimalCurrency}{Age(RS) + (DeliveryTime(RS) - InputTime(RS))}$$

Actual currency is measured as the sum of how old data are when received from a data provider (Age), plus a second term that measures how long data have been in the organization before its actual use (DeliveryTime - InputTime) [5]. In our application domain, we assume that i) Age > 0 (e.g., addresses cannot be updated instantaneously by data providers when they change in real life), and ii) OptimalCurrency = Age (since there is no way an organization could have data more up-to-date than its providers).

In our case study, the Age of the WPD is measured by considering the difference between the updating date of the data streams coming from the two providers (PVD1 and PVD2) and the InputTime when such data are received by the merging program. The DeliveryTime is when the merging data are updated on the WPD. In regards to the ACS, the Age is the difference between the time in which the MD agents compile their spreadsheets and the time in which the account manager receives them. The Delivery Time is when the spreadsheets are merged creating the ACS. Finally, the Age of the CDB is the Currency measured for the ACS, the Input Time is when the AD employees parse the ACS and the Delivery Time is when they feeds the CDB.

In order to evaluate the DQ dimensions at the level of CEs, it is necessary to define an algebra to compose DQ dimensions values. In our approach, we consider the following aspects that are relevant for the DQ composition: *relevance* and *scope* of each RS. The relevance expresses the importance and usefulness of a single RS. The relevance values rel(RS) can be evaluated by means of domain experts panels , or else by considering the number of critical business processes that use the RS. In our study, we focus on the second approach that is characterized by the metric (5) where P(RS,i) is a boolean function which returns 1 if the i-th process uses the RS, 0 otherwise and |proc| represents the total number of processes of the organization.

$$(5) \quad rel(RS) = \frac{\sum_{i=1}^{|proc|} P(RS, j)}{|\, proc \,|}$$

$$(6) \quad dq(CE)_{rel} = \frac{\sum_{i=1}^{n} rel(RS_i) \cdot dq(RS_i)}{\sum_{i=1}^{n} rel(RS_i)}$$

Such values are considered as weights of a weighted arithmetic mean of the DQ values measured over different RSs. Now, we are able to define the dq value for a single CE according to relevance by means of the composing function (6) where dq(RS_i) is the value of the DQ dimension (Accuracy or Normalized currency) measured for each RS.

The scope values scp are calculated by Formula (7), where the numerator is the number of instances of the RS n and the denominator is the number of instances of the union set between the RSs. The scp expresses the extent an RS covers the total number of the instances (either tuples, elements or objects according to the type of data) represented in all the RSs associated to the same CE.

$$(7) \quad scp(RS) = \frac{|\, RS \,|}{|\bigcup_{i=1}^{n} RS_i \,|}$$

$$(8) \quad dq(CE)_{scp} = \frac{\sum_{i=1}^{n} scp(RS_i) \cdot dq(RS_i)}{\sum_{i=1}^{n} scp(RS_i)}$$

The scp values are considered as weights of a weighted arithmetic mean of the DQ values measured over different RSs. Now, we are able to define the dq value for a single CE according to scope by means of the composing function (8). Finally, we are able to define the general DQ value associated to the single CE represented in the above mentioned RSs, by means of the following composing function:

$$(9) \quad dq(CE) = \frac{dq(CE)_{rel} + dq(CE)_{scp}}{2}$$

In the Ristobill case study, let us consider the following data: (i) the 100 instances of the ACS are a subset of the 300 instances contained in the WPD; (ii) Accuracy scores an 80% for the ACS and a 70% for the WPD; (iii) relevance values are 0,8 for the ACS and 0,9 for the WPD. At the beginning we apply the formula (6) to calculate the DQ value of the CE on the basis of the relevance values and we obtain 0,75 (i.e., 75%). Now, we apply the formula (8) considering that the instances of the union set are 300 because the ACS is a subset of the WPD. We obtain 0,73 *(i.e., 73%) as the DQ value of the CE. Finally, applying formula (9), we obtain 74% as final Accuracy value of the CE.*

6. THE HDQM AT A GLANCE

HDQM aims to provide indications on the optimal DQ improvement program that an organization should undertake with respect to its peculiar needs and constraints. HDQM consists of three main phases and each of them is composed of a number of steps. In particular, the main phases are:

1. State reconstruction, which aims to reconstruct all the relevant knowledge regarding the organizational units, processes, resources and conceptual entities involved in the organization.
2. Assessment, which aims to obtain a quantitative evaluation of data quality problems. DQ dimensions are measured in order to assess the current level of data quality and to set the new DQ targets that must be reached at the end of the DQ improvement program.
3. Improvement, where apt improvement activities are selected by evaluating their effects in terms of DQ dimensions/cost ratio.

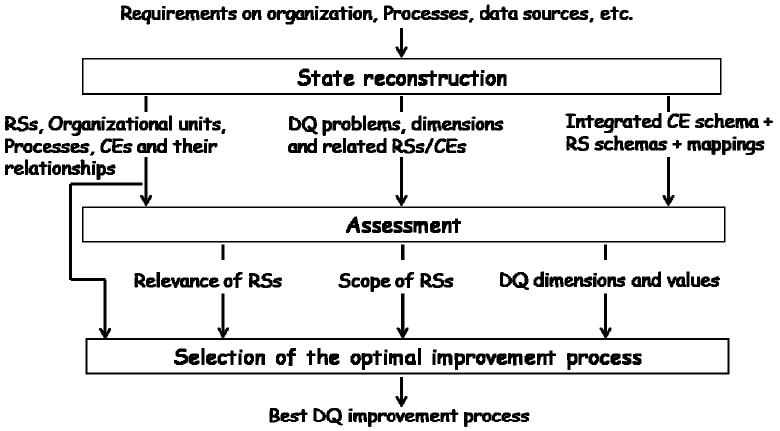

Requirements on organization, Processes, data sources, etc.
State reconstruction
RSs, Organizational units, Processes, CEs and their relationships
DQ problems, dimensions and related RSs/CEs
Integrated CE schema + RS schemas + mappings
Assessment
Relevance of RSs
Scope of RSs
DQ dimensions and values
Selection of the optimal improvement process
Best DQ improvement process

Figure 3: A schema of the phases, inputs and outputs of HDQM

The three phases of the methodology and their inputs and outputs are shown in Figure 3. Figure 3 represents the deployment of the HDQM in a streamlined and sequential fashion and does not depict possible feedback for the sake of simplicity. Moreover, Figure 3 highlights the modularity of the methodology, a characteristic that makes it capable of being adapted depending on either the application domain or the overall knowledge that is actually available from a specific organization. For instance, the State reconstruction phase can be simplified in all those cases when either knowledge on the status of resources and data quality values is already available or when undergoing a comprehensive reconstruction of this knowledge would be too expensive. Likewise, the Assessment phase can be adapted in terms of dimensions and metrics in those contexts where more domain-specific quality dimensions and activities must be considered, e.g., in the geographical domain where a spatial Accuracy dimension could be considered. In the next sections, we describe in detail each phase of the HDQM and exemplify its application to the case study outlined in Section 3.

7. STATE RECONSTRUCTION

State reconstruction is a complex phase that has been described to great extent in [9] and [11]. In this paper suffices to say that it encompasses a preliminary task of problem identification and three tasks of reconstruction regarding a) ReSources b) Conceptual Entities and c) Processes. The problem identification task aims to identify the most relevant data quality problems as they are perceived by all the actors involved in the business processes. This means to focus on the most important data only, the so called master data [20] and on those data that are involved in some organizational shortcoming. The subjective perception given by internal and external actors is quantified by means of focused interviews and survey questionnaires.

In the case study, the results of the interviews highlight the following data quality problems: (i) The information about the customers is not frequently updated. A frequent and significant case reported in the interview sessions with the MD agents regards the fact that a restaurant had moved before the phone contact and therefore all its identifying data became wrong. This problem is related to the Normalized currency dimension, and is associated to the CustomerCE; (ii) The information about customers exhibits several mistakes. For instance, a restaurant has two addresses, while different customers are associated with the same restaurant. This problem is related to the Accuracy dimension, and it is associated to the CustomerCE.

Once the RSs involved in data quality problems have been identified, it is possible to model the CEs represented in each RS and their relationships in terms of a conceptual schema. To perform this task, two activities are needed: reverse engineering and schema integration. Reverse engineering has the goal of translating the intensional part of each RS into the corresponding conceptual schema. We apply reverse engineering techniques according to the specific type of data of the RS at hand. In literature, several techniques to extract a conceptual schema from a relational database [36] and from XML documents or forms [25] are described. In regard to unstructured data, the extraction of a conceptual schema is a complex task. To perform this task we apply the same two tasks of object identification and classification that we presented in Section 5 by considering the referred CE with respect to the associated domain. Then, we add the task of CE-CE relationship extraction, which aims to determine the relationships between the extracted CEs. For instance, the Customer CE has a relationship, named "has", with the Business info CE. The mapping between each RS and the corresponding CE schema S is denoted as mapping(RS,S).

By applying the reverse engineering activity to the case study we detected three schemas, specific for each RS, that then we had to merge in the schema integration activity. In this activity, CE schemas of the selected RSs are analyzed for conflict resolution, e.g., for the resolution of synonymies, homonymies and type conflicts. Several integration methodologies are available to this aim, see [31] for a comprehensive survey and discussion. We call the schema obtained in this step integrated schema. For each RS we have to produce the mapping(RS,S), where S is a sub-schema of the integrated schema. The knowledge on the RSs extends the typical approach defined in data integration and encompasses: 1) the set of RSs, each with its relevant DQ dimensions; 2) the integrated schema, with relevant DQ dimensions for each CE; 3) the mapping (RS,S) for each RS. This knowledge will be used in the Assessment and Improvement phases to allow for the measurement of the DQ dimensions and the selection of the best improvement process.

The result of the schema integration activity in the case study is shown in Figure 4, where schemas associated to RSs are highlighted.

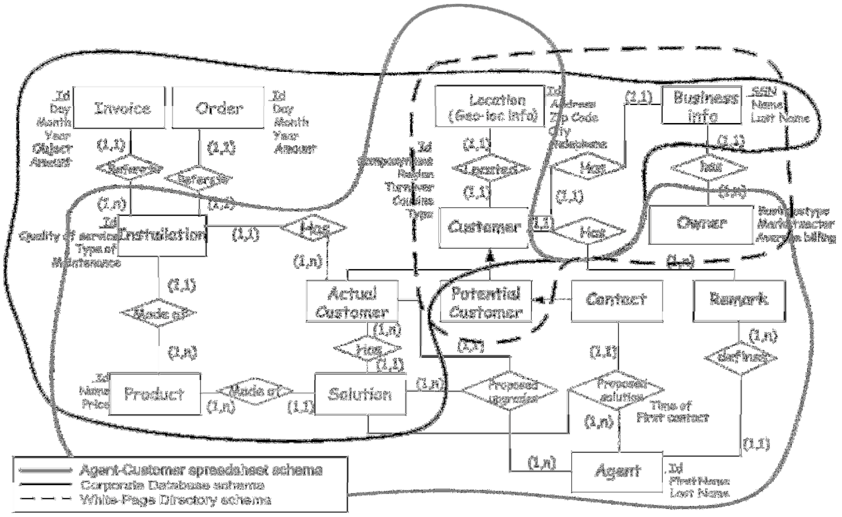


Figure 4: The Integrated schema of the case study

8. ASSESSMENT

The Assessment phase is made up of two steps: the resources ranking, and the data quality measurement.

8.1 Resources Ranking

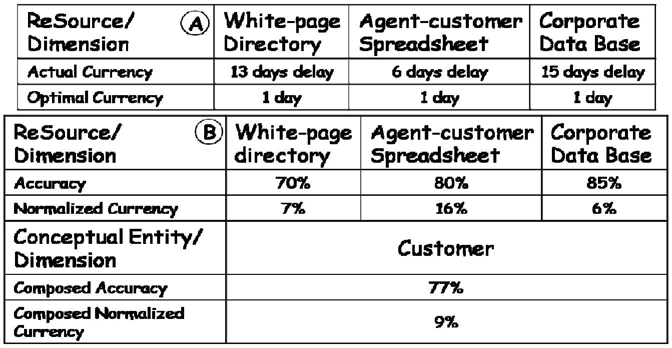
This step aims to measure the effective relevance and significance of the RSs that have been identified during the problem identification step. Starting from the integrated schema, the relevance weights are measured to (i) obtain a precise indication of which RSs and related DQ dimensions should be addressed in the Improvement phase; (ii) facilitate the evaluation of the risk/feasibility of DQ improvement programs; (iii) apply the composing functions as will be shown in the next section. To perform the first task, relevance weights are compared with a relevance threshold in order to identify a ranking of RSs. The relevance threshold is qualitative and can be defined in a number of ways: in most of the cases, analysts would fix it on the basis of their personal experience or by relying on the suggestions given by business experts and key users of the organization. If these indications are not in agreement, a more structured method can be employed, like the Delphi method, in order to decrease the range of answers and have the panel of experts converge towards a reliable value of threshold for the setting at hand [3]. The RS ranking is used to select which RSs should be object of the improvement program.

In the case study, a relevance weight is measured for each considered RS obtaining the following values: 0,9 WPD, 0,8 ACS and 0,6 CDB. The relevance values are then compared with the reference threshold. For instance, assuming the threshold is equal to 0,5, the performed analysis confirms the relevance of the WPD in the organization and the necessity to apply an improvement program on it. Since values for the other RSs are close to the WPD's value, we keep considering all the RSs in the next steps.

8.2 Data Quality Measurement

This step aims to obtain a quantitative evaluation of the quality issues identified in the problem identification step. For this, it is necessary to select the relevant DQ dimensions and related metrics. Then, metrics are applied to the RSs to provide a quantitative evaluation of the quality problems. In what follows, we describe how the DQ dimensions and their metrics defined in Section 5 are applied to the case study.

> *The integrated schema specifies that the CustomerCE is associated with all the critical RSs, namely WPD, ACS and CDB. Considering the Normalized currency dimension, we have to measure this DQ dimension for each RS and compose the obtained values to define the dimension for the Customer CE. To perform this activity, we apply metric (4) shown in Section 5. The Actual currency values calculated on the basis of the scenario described in Section 5 are shown in Figure 5.A. In regard to the maximum acceptable delay, the standard to which to tend posed at Ristobill is one day delay, i.e., the updating should be performed the very same day that data are modified by any provider. Then, we can calculate the Normalized currency value for the CustomerCE, applying the composing function (9) to the three RSs involved. This approach is repeated for the Accuracy dimension by applying the referred metric shown in Section 5. At the end of the measurement process, we may fill the matrix shown in Figure 5.B and then easily identify to which RSs poor quality is more attributable.*



ReSource/ Dimension Ⓐ	White-page Directory	Agent-customer Spreadsheet	Corporate Data Base
Actual Currency	13 days delay	6 days delay	15 days delay
Optimal Currency	1 day	1 day	1 day

ReSource/ Dimension Ⓑ	White-page directory	Agent-customer Spreadsheet	Corporate Data Base
Accuracy	70%	80%	85%
Normalized Currency	7%	16%	6%
Conceptual Entity/ Dimension	Customer		
Composed Accuracy	77%		
Composed Normalized Currency	9%		

Figure 5: A) Currency values used to calculate the Normalized Currency values; B) DQ values measured on RSs and composed for the CE

9. IMPROVEMENT

The Improvement phase encompasses three steps: (i) the DQ requirement definition, (ii) the DQ improvement activity selection, and (iii) the choice and evaluation of the optimal improvement process.

9.1 Data Quality Requirement Definition

This step aims to set the target quality values to be reached through the improvement process. This activity is based on the present quality values dqi j, that are associated with the i-th CE or RS and the j-th quality dimension. Data quality targets are defined by performing a process-oriented analysis [9], as summarized in what follows.

The process-oriented analysis is based on the information collected in the State reconstruction phase, regarding the involved CEs, RSs, processes, organizational units and their mutual interrelations. In order to define feasible target quality values, we use the formula (10) that holds for a specific CE assuming the linearity of the correlation between process quality (performance) and data quality. In (10) pq_x is the current value of process quality for process x and dq_y is the current level of the composed data quality for the y-th CE. Then, we can state that the target quality of CE, dq*, is obtained by considering the wished performance of the considered process, pq, as in (11).

$$(10)\ \alpha_{xy} = \frac{pq_x}{dq_y} \qquad (11)\ dq_y^* = \frac{pq_x^*}{\alpha_{xy}}$$

For instance, let us consider a process whose performance indicator is the amount of interest accrual derived from sale receipts. The longer the delay by which invoices are sent to customers, the lower the yearly interest, and hence the performance of the process. Once the overall DQ (target) value has been fixed at CE level, improvement techniques can be applied to the single RSs representing it. To this aim, the target value at CE level is propagated at the RS level in order to define the target values for each component RS. The propagating operation takes the relevance and scope weights into account, which were defined in formula 5 and 7.

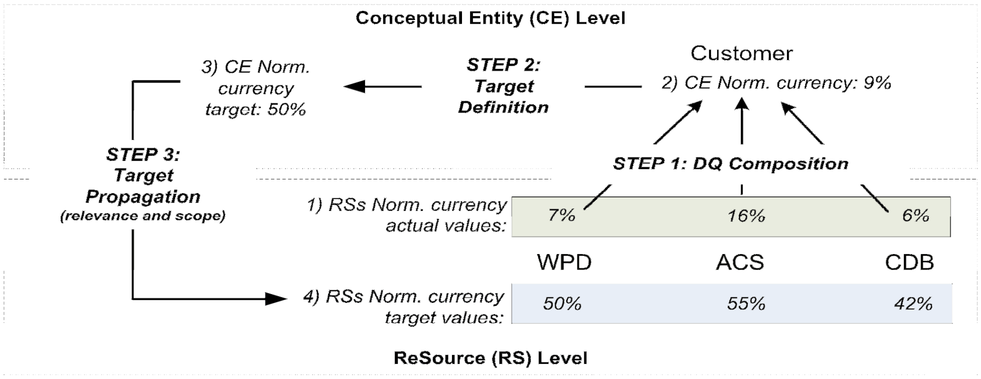
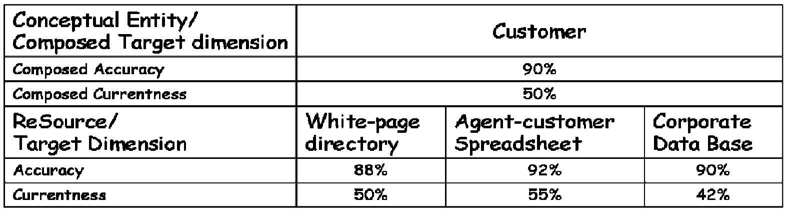

Conceptual Entity (CE) Level
Customer
3) CE Norm. currency target: 50%
STEP 2: Target Definition
2) CE Norm. currency: 9%
STEP 3: Target Propagation (relevance and scope)
STEP 1: DQ Composition
1) RSs Norm. currency actual values:
7%
16%
6%
WPD
ACS
CDB
4) RSs Norm. currency target values:
50%
55%
42%
ReSource (RS) Level

Figure 6: Definition of target values for the Normalized currency dimension

In regard to the normalized Currency and Accuracy dimensions, the obtained target values refer to the Customer CE only. As regards Normalized currency, a feasible delay is considered two days from data provision (i.e., 50% of Normalized currency). As regards Accuracy, a feasible level is 90%. At this stage, we choose to select a different level of abstraction and define a particular target value for each component RS. The definition of target values is influenced by the relevance and scope weights assigned in the Assessment phase. Figure 6 illustrates this approach with respect to the Normalized currency dimension. In step 1, we get the current value of Normalized currency at customer (CE) level by applying the formula (9); in this case, 9%. In step 2, we obtain the target values for Normalized currency at the CustomerCE level by applying the formula (11) to the current Normalized currency value; in this case, to obtain the correlated process performance, a huge improvement is needed, and the new Currency target is set to 50%. In step 3, we propagate the customer-related (CE level) value to the level of single RS taking the relevance and scope weights into account. In doing so, we obtain the target value for Normalized currency for each component RS. This problem has several integer solutions, but heuristics based on costs and pertinent values usually suffice to delimit the definition of the resource-related values (in this case, the three RSs need an DQ improvement up to a Currency of 50%, 55% and 42%). A possible solution is shown in Figure 6. Using the same approach to define the target value of the Accuracy dimension, we fill the matrix shown in Figure 7.



Conceptual Entity/ Composed Target dimension	Customer		
Composed Accuracy	90%		
Composed Currentness	50%		
ReSource/ Target Dimension	White-page directory	Agent-customer Spreadsheet	Corporate Data Base
Accuracy	88%	92%	90%
Currentness	50%	55%	42%

Figure 7: New Conceptual Entities and Target quality values matrix

9.2 Selection of the Data Quality Improvement Activity

This step selects the data-driven and process-driven activities that are candidate to be chosen for optimal improvement process. Although in the previous step we discussed a process-driven approach, in what follows we consider both process-and data-driven techniques, since they are optimal in different contexts and hence complementary. The selection of these activities depends not only on the target set of quality values given in the previous step, but it also considers the different types of RSs that were identified in the State reconstruction phase. Thus, a specialization of the data-driven and process-driven techniques defined in the literature (see [27] for a survey) is required. In what follows we will employ Source improvement (i.e., an existing data source is improved in order to also improve the quality of its data provision [4]), Record linkage (i.e., it allows for the identification of the same object in different data sources), Process control (i.e., existing processes are modified so that critical points are either controlled, verified or audited more effectively) and Process re-design (i.e., processes are redesigned from scratch in order to avoid the intrinsic causes of bad quality and introduce new sub-processes that produce data of better quality [32]) activities. At the end of this step, HDQM produces a ReSource / Improvement-Activity matrix, where a cross is marked for all relevant RSs for which the improvement activity is scheduled.

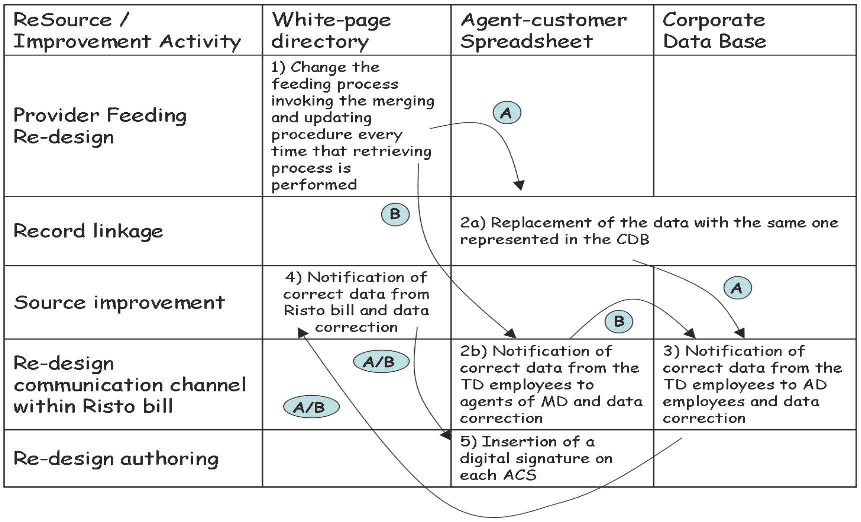
Applying the activities presented above to the case study, we obtain:

- *In regards to source improvement activities, there are two cases. In order to improve the Accuracy of the ACS, it is possible to leverage on the fact that TD employees have to accomplish on-site installation and maintenance interventions for each acquired customer. To do so they must have obtained the correct phone number to contact storekeepers and schedule these interventions. Moreover, they can verify directly at the location the right address of the customer and get other accurate information by contacting customers personally. A first source improvement activity hence regards the establishment of an official channel within Ristobill. Through this channel, the TD manager sends to the MD manager and to the AD employees an amended version of the ACS. This is performed once a month, after that the TD manager has collected the error reports from her service engineers. Another source improvement activity can be performed if a similar channel is established between Ristobill and at least one of the two providers. Thus, amended data are communicated to the original source of customer data, so that the Accuracy of the WPD can also be improved. Moreover, the risk of overwriting correct values into the spreadsheet with inaccurate and out-of-dated data from the WPD is also significantly reduced when providers feed Ristobill with new data.*
- *Record linkage can be applied to improve the Accuracy of the ACS. In particular, we can use the CDB to replace the data referred to the same CustomerCE represented in the ACS. For instance, a tuple that in the CDB represents the customer "John Smith" can be used to amend the same data represented in the ACS.*
- *In order to produce more reliable ACSs, a process control activity can be accomplished on the processes that compile and update the ACS. The MD manager could enforce a policy within his department by which agents have to digitally sign their own copy of the spreadsheet before submitting it for the merging into the corporate one. In so doing, single agents are motivated to double-check the Accuracy of data they are responsible for.*
- *Process re-design can be performed to improve the Normalized currency on the WPD. In particular, it is necessary to focus on the feeding process by which the RS is made up-to-dated. This process is independent of the time in which the two providers, PVD1 and PVD2, perform the updating of the customer-related data. This implies delays in the data arriving from the two providers to Ristobill. A possible solution is the retrieval and publication of new data on the WPD at the time when the two providers change their data. This solution is applied by invoking the merging and updating process every time that the retrieving process is performed.*

9.3 Selection and Evaluation of Improvement Processes

In this step, crossings in the ReSource/Activity matrix are linked with each other by identifying candidate improvement processes. A property that a candidate improvement process should satisfy is Completeness, i.e. the inclusion of all RSs involved in the improvement program.

Moreover, the candidate improvement process must include all activities needed to improve the entire set of DQ dimensions measured in the Assessment phase.



ReSource / Improvement Activity	White-page directory	Agent-customer Spreadsheet	Corporate Data Base
Provider Feeding Re-design	1) Change the feeding process invoking the merging and updating procedure every time that retrieving process is performed		
Record linkage		2a) Replacement of the data with the same one represented in the CDB	
Source improvement	4) Notification of correct data from Risto bill and data correction		
Re-design communication channel within Risto bill		2b) Notification of correct data from the TD employees to agents of MD and data correction	3) Notification of correct data from the TD employees to AD employees and data correction
Re-design authoring		5) Insertion of a digital signature on each ACS	

Figure 8: Two possible improvement processes (see flows denoted as A and B)

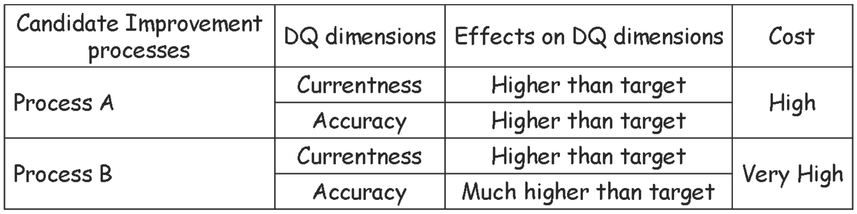
Two or three candidate improvement processes are usually sufficient to cover all relevant choices.

> *In the case study, two improvement processes (Process A and B in Figure 8) are considered. The numbers show the sequence of activity execution for both processes. The difference between the two processes is represented by the selection of the improvement activity related to the ACS for the Accuracy dimension. In process A, the record linkage activity is applied, while in process B, a re-design process of accurate data notification is applied alternatively.*

For each candidate improvement process an evaluation of costs and the achievement of target DQ dimensions values must be performed. In regard to costs, we may adopt one of the quantitative approaches to cost/benefit evaluation described in the literature, such as [20]. These approaches provide a classification of the different types of costs involved in the improvement process. Such costs include, e.g., the costs of personnel, the costs of equipment and the costs of licensees of software tools. Here, we propose a more qualitative approach in which costs are quantified in the categorical domain encompassing “very low, low, medium, high, very high” according to the designer’s experience.

In regards to DQ dimensions, the achievement of target values of dimensions for different improvement processes has to be checked at this stage by a qualitative heuristic that associates a quality improvement value in the categorical domain encompassing “below target, on target, higher than target, much higher than target”. The different candidate improvement processes are then compared with each other on the basis of the pairs of qualitative and cost-oriented values – i.e., <effects on DQ dimensions, costs> pairs – in order to select the best one with respect to the best effects on DQ dimensions/cost trade-off. For instance, a possible result of the evaluation can lead to select a process that ranked ‘much higher than target’ for quality and ‘high’ for costs while the cost is not compatible with the budget available for the organization. In this case, an alternative solution could be to selecting another process whose application allows for quality values ‘much higher than target’ or ‘higher than target’ with ‘low’ costs.

In the case study, we have applied the evaluation approach proposed above. The effects of the two candidate processes on the selected quality dimensions and their costs are represented in Figure 9. The process A results in a lower improvement of the Accuracy dimension than process B, but it allows the organization to curb spending by achieving the target value for the considered dimensions. We select process A as the best candidate improvement process.



Candidate Improvement processes	DQ dimensions	Effects on DQ dimensions	Cost
Process A	Currentness	Higher than target	High
	Accuracy	Higher than target	
Process B	Currentness	Higher than target	Very High
	Accuracy	Much higher than target	

Figure 9: A sample improvement process

10. CONCLUSION AND FUTURE WORK

This paper describes the Heterogeneous Data Quality Methodology for data quality assessment and improvement. It provides database and information system practitioners with guidelines to consider different types of data (structured, semistructured and unstructured) for the analysis of the quality of the information managed in an organization and the selection of effective solutions for data quality improvement. Yet finding an adequate trade off between simplicity and comprehensiveness is a difficult task that greatly depends on the application domain and single case at hand. HDQM is a high-level and general-domain methodology that we propose more as a frame according to which to articulate and apply known techniques and methods rather than as a toolbox, i.e., a collection of novel interventions by which to either assess or improve organizational data. Consequently, we are aware that implementing HDQM in complex cases could be very challenging. The quantitative assessment metrics we discussed for the operative application of HDQM can require a significant effort, whereas they require a tuple-by-tuple approach and do not leverage on a statistical (sample-based) approach. On the other hand, qualitative metrics rely on the availability and above all, on the competency of domain experts. They are called into question in order to either weigh the contribution of single resources to the overall quality/performance of the information system or to fix judicious thresholds in order to rank processes accordingly and find criteria to decide priorities. Moreover, we are aware of the fact that state reconstruction, as conceived within the HDQM, is a time consuming and resource demanding activity. This is true especially in the task of extracting the schema of relevant resources if they are not yet well known and documented. Yet, HDQM is highly modular in considering single resources and can be effectively applied to portions of the whole set of resources used in the organization, if a comprehensive approach were too costly with respect to the available budget. Last, HDQM shares with other rationalizing approaches of the organizational domain the main success factors to achieve bottom-line business results: above all, a great deal of motivation from all the key stakeholders involved, the concrete sponsorship of top management and the strong commitment of the middle management [16].

In regards to the improvement phase, we are addressing a stronger quantification of the methods proposed for the cost evaluation. We also are considering inter-dependencies between DQ dimensions in the calculation of target values for each data quality dimension. To this aim, we will adopt and extend the formal and data-driven approach we proposed in [6]. This analytical framework provides the main models of dependencies and analytic formulas, based on the entropy of Shannon, to measure the correlations between dimensions. These correlations would allow us in Formula 11 to consider dependencies among data quality dimensions besides their independently judged values. Considering these dependencies would lead to several potential advantages in the achievement of target values: (i) choosing the most effective improvement activity on involved dimensions, (ii) unfolding progressively the effects from independent to dependent dimensions, and (iii) minimizing the cost of the improvement process, since redundant and overlapping activities are avoided. In this way, for each DQ value, we can achieve target quality values while saving resources in the improvement process.

Finally, in regards to the application of the methodology, we will extend its deployment to meaningful large scale domains, such as the public and the financial sector. To this aim, one current direction of research regards the application of HDQM to the Basel II norms concerning measurement and mitigation of various types of risks in financial and credit institutes. Among them, the operational risk is strictly related to the quality of information used in bank transactions and bank loan authorizations. Moreover, a tool that supports HDQM is under construction; some of the functionalities of the tool, for the state reconstruction phase and for the data quality measurement of the Accuracy dimension, are currently operational [21].

REFERENCES

[1] Abiteboul, S. (1997). “Querying semi-structured data”. In *Proceedings of the 6th International Conference on Database Theory*, Delphi, Greece.

[2] Abiteboul, S., Buneman, P. and Suciu, D. (2000). *Data on the Web: from relations to semistructured data and XML.* Morgan Kaufmann Publishers Inc.

[3] Anderson, J. C., Rungtusanatham, M. and Schroeder, R. G. (1994). “A theory of quality management underlying the deming management method”. *The Academy of Management Review,* Vol. 19, No. 3, pp472-509.

[4] Avenali, A., Bertolazzi, P., Batini, C. and Missier, P. (2008). “Brokering infrastructure for minimum cost data procurement based on quality -quantity models”. *Decision Support Systems,* Vol. 45, No.1, pp95-109.

[5] Ballou, D., Wang, R., Pazer, H. and Tayi, G. K. (1998). “Modeling information manufacturing systems to determine information product quality”. *Management Science,* Vol. 44, No. 4, pp462-484.

[6] Barone, D., Batini, C. and De Amicis, F. (2006). “An analytical framework to analyze dependencies among data quality dimensions”. In *Proceedings of the 11th International Conference on Information Quality*.

[7] Batini, C., Cabitza, F., Cappiello, C. and Francalanci, C. (2008). “A comprehensive data quality methodology for web and structured data”. *International Journal of Innovative Computing and Applications* Vol. 1, No.3, pp205-218.

[8] Batini, C., Cappiello, C., Francalanci, C. and Maurino, A. (2009). “Methodologies for data quality assessment and improvement”. *ACM Computing Survey*, Vol. 41, No. 3, pp1-52.

[9] Batini, C. and Scannapieco, M. (2006). *Data Quality: Concepts, Methods, and Techniques.* Springer Verlag.

[10] Cappiello, C., Francalanci, C., Pernici, B. and Plebani, P. (2003). “Data quality assurance in cooperative information systems: a multi-dimension certificate”. In *Proceedings of the International Workshop on Data Quality in Cooperative Information Systems*, Siena, Italy.

[11] Elmagarmid, A. K., Ipeirotis, P. G. and Verykios, V. S. (2007). “Duplicate record detection: A survey”. *IEEE Transactions on Knowledge and Data Engineering,* Vol.19, No. 1, pp1-16.

[12] Elmasri, R. and Navathe, S. B. (1994). *Fundamentals of Database Systems, 2nd Ed.* Benjamin-Cummings.

[13] English, L. P. (1999). *Improving data warehouse and business information quality: methods for reducing costs and increasing profits*. John Wiley & Sons Inc.

[14] Falorsi, P. D., Pallara, S., Pavone, A., Alessandroni, A., Massella, E. and Scannapieco, M. (2003). “Improving the quality of toponymic data in the italian public administration”. In *Prooceedings of the International Workshop on Data Quality in Cooperative Information Systems*.

[15] Hegewald, J., Naumann, F. and Weis, M. (2006). “XStruct: Efficient schema extraction from multiple and large xml documents”. In *Prooceedings of 22th International Conference on Data Engineering Workshops*.

[16] Hengst, M. D. and Vreede, G. D. (2004). “Collaborative business engineering: A decade of lessons from the field”. *Journal of Management Information Systems*, Vol.20, No.4, pp85-114.

[17] Krawczyk, H. and Wiszniewski, B. (2003). “Visual gqm approach to quality-driven development of electronic documents”. In *Proceedings of the 2nd International Workshop on Web Document Analysis*.

[18] Lee, Y. W., Strong, D. M., Kahn, B. K. and Wang, R. Y. (2002). “Aimq: A methodology for information quality assessment”. *Information & Management*, Vol.40, No.2, pp133-146.

[19] Long, J. and Seko, C. (2002). “A new method for database data quality evaluation at the canadian institute for health information (CIHI)”. In *Proceedings of 7th International Conference on Information Quality*.

[20] Loshin, D. (2008). *Master Data Management*. Morgan Kaufmann.

[21] Maurino, A., Batini, C., Barone, D., Mastrella, M. and Ruffini, C. (2007). “A framework and a methodology for data quality assessment and monitoring”. In *Proceedings of the 12th International Conference on Information Quality*. Boston, MA, USA.

[22] McCallum, A. (2005). “Information extraction: distilling structured data from unstructured text”. *ACM Queue*, Vol.3, No.9, pp48-57.

[23] Naumann, F., Freytag, J. C. and Leser, U. (2004). “Completeness of integrated information sources”. *Information Systems*, Vol.29, No.7, pp583-615.

[24] Pareto, L. and Boquist, U. (2006). “A quality model for design documentation in model-centric projects”. In *Proceedings of the 3rd international workshop on Software quality assurance*. New York, NY, USA.

[25] Penna, G. D., Marco, A. D., Intrigila, B., Melatti, I. and Pierantonio, A. (2006). “Interoperability mapping from XML schemas to ER diagrams”. *Data Knowledge Engineering*, Vol.59, No.1, pp166-188.

[26] Pipino, L., Lee, Y. W. and Wang, R. Y. (2002). “Data quality assessment”. *Communications of ACM*, Vol.45, No.4, pp211-218.

[27] Redman, T.C. (1997). *Data Quality for the Information Age*. Artech House, Inc.

[28] Scannapieco, M., Virgillito, A., Marchetti, C., Mecella, M. and Baldoni, R. (2004). “The daquincis architecture: a platform for exchanging and improving data quality in cooperative information systems”. *Information Systems*, Vol.29, No.7, pp551-582.

[29] Shankaranarayan, G., Wang, R. Y. and Ziad, M. (2000). “Modeling the manufacture of an information product with IP-MAP”. In *Proceedings of the 5th International Conference on Information Quality*. Massachusetts Institute of Technology, USA.

[30] Shankaranarayanan, G. and Cai, Y. (2006). “Supporting data quality management in decision-making”. *Decision Support Systems*, Vol.42, No.1, pp302-317.

[31] Shvaiko, P. and Euzenat, J. (2005). “A survey of schema-based matching approaches”. *Journal on Data Semantics IV*, pp146-171.

[32] Stoica, M., Chawat, N. and Shin, N. (2004). “An investigation of the methodologies of business process reengineering”. *Information Systems Education Journal*, Vol.2, pp1-11.

[33] Wang, R. Y. (1998). “A product perspective on total data quality management”. *Communications of ACM*, Vol.41, No.2, pp58-65.

[34] Wang, R. Y., Lee, Y. W., Pipino, L. and Strong, D. M. (1998). “Manage your information as a product”. *Sloan Management Review*, Vol.39, No.4, pp95-105.

[35] Wang, R. Y. and Strong, D. M. (1996). “Beyond accuracy: what data quality means to data consumers”. *Journal of Management Information Systems*, Vol.12, No.4, pp5-34.

[36] Yeh, D. and Li, Y. (2005). “Extracting entity relationship diagram from a table-based legacy database”. In *Proceedings of the 9th European Conference on Software Maintenance and Reengineering*, Manchester, UK.