HDQ: a Meta-Model for the Quality Improvement of Heterogeneous Data

Daniele Barone, Federico Cabitza and Simone Grega
Dipartimento di Informatica Sistemistica e Comunicazione
Università di Milano-Bicocca
Via Bicocca degli Arcimboldi 8, 20126 Milano Italy
Email: {barone,cabitza,grega}@disco.unimib.it

Abstract—In this paper, we outline the HDQ meta-model and exemplify its application to a realistic organizational scenario within a methodology for assessing and improving corporate data quality. The main aim of the HDQ meta-model is to support the modelling of the information resources used within an organization towards the application of comprehensive but yet lean methodologies of data quality improvement. The comprehensive approach allowed by HDQ meta-model is aimed at addressing the scalable and incremental management of the quality of data represented in either structured, semi-structured or unstructured information sources.

I. Background and Motivations

Information and Communication Technologies have extraordinarily increased the amount of information that is processed by organizational processes. With the advent of large-scale networks, web information systems, and peer-to-peer information systems, groups of corporate organizations have been enabled to exchange data, interact and cooperate as never before and several organizations, whose core business is to provide information, have been founded and flourish. On the other hand, in the current scenario of rushing migration from hierarchical/monolithic to network based information systems, also data sources have increased dramatically: an increasing number of organizational processes get involved in complex information exchanges and rely on inputs that are obtained from external sources on which low or no control can be exerted since these sources are usually unknown a priori.

Consequently, modern organizations, both in the public and private sectors, are becoming more and more “information intensive”; the more complex these organizations get, the more complex and intertwined the information flows, which are processed in order to manage and maintain the organization competitive or able to make value for its stakeholders. In this scenario, the quality of data assumes a crucial role in that researchers have collected clear evidence that poor quality of data has a negative impact on customer satisfaction and, thus, competitiveness [1], [2], in terms of cost implications of data errors for both internal and external users.

Despite data and information are almost ubiquitous in modern organizations – or maybe just because of it – assessing and improving Data Quality (DQ) is a complex task. Indeed, data can reside in very different forms, ranging from almost unstructured in file systems, document repositories and on the web, to highly structured in relational Data Base Management Systems (DBMS).

In the literature, data sets are classified, depending on the medium used to represent them, according to wide-accepted yet quite informal conventions in terms of:

1) structured if their formal schema (i.e., formats, types, constraints, relationships) is defined and explicit.
2) semi structured (sometimes also called “self-describing” [3], [4]) 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 relational data [5].
3) unstructured, if data sets are but sequences of symbols, typically coded in a natural language, where no specific structure and definition domains are defined.

Our main point is that a comprehensive and general approach towards the improvement of all these types of data is feasible and indeed necessary since almost every modern organization has usually to cope with all these different formats at the same time. 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. Consequently, the requirements and demands that an organization exhibits about its customers at intensional level are necessarily intertwined to all its heterogeneous sources where information about customers is represented, in deep and not always trivial ways. In this paper, we present the Heterogeneous Data Quality (HDQ) meta-model with the aim to address the topic of how to improve the quality of key single information sources used in an organization not irrespective of the concepts they refer to, but rather in light of the concepts they represent. The HDQ meta-model is intended to encompass the main concepts and related quality dimensions that analysts must consider in an organizational context when they want to apply a data quality improvement methodology to that organization. Moreover, the instances of these concepts represent a significative repository of information for the considered organization. Currently, the HDQ meta-model is the underpinning model of the HDQ methodology – an extension of the Comprehensive Data Quality methodology presented in [6], [7] but more focused on semi-structured.
and unstructured data – in order to facilitate and make more scalable its phases. In what follows, we will show the main aspects of the HDQ meta-model (see Section III) and we will exemplify its application to a realistic case study focusing on a couple of significant quality dimensions (see Section IV). The case study is taken from the private sector and is depicted in Section II. Section V closes the paper and gives some indications on our current research agenda.

II. THE RISTOBILL CASE STUDY

In the following, we provide the requirements of our case study.

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. Typical customers of Ristobill are restaurants, pubs, bistros and snack bars or, better yet, their proprietary or licensees. At Ristobill, three business units are directly involved with customers, although in very different ways and with different quality requirements. 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 access to a sort of huge White Page Directory (from now on, simply WPD) that is created starting from the data streams coming from two different profiling providers, PVD1 and PVD2. The first one has been chosen for complete administrative data (e.g., who is the owner of a facility and the facility’s accurate name) as well as its precise geo-localization data (like facility addresses, phone number and GPS coordinates). Conversely, the second one 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 catalogue and Ristobill subscribed for the cheapest service, i.e., the provision of unstructured web pages. These two parallel data flows are then appended at Ristobill into one long web page which is updated every two weeks and rendered on the Ristobill intranet.

The second business unit is the Technical Department (TD) that is supposed to monitor the well running of sold installations and provide both ordinary and extraordinary maintenance upon it. In order to give an 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. They use an Agent-customer spreadsheet file (ACS) which is compiled by the MD agents and is partly derived from the white pages. 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., fiscal codes, ownership data and company full names). Since any field in the spreadsheet is strictly typed but with no semantic constraints, the ACS ReSource can be considered semi-structured.

Lastly, the Accounts Department (AD) needs accurate and up-to-date administrative information for invoice drawing and accounting. Every two weeks, a semiautomatic procedure parses the shared ACS spreadsheet and feeds its data into a corporate database table, i.e., the Corporate Data Base (CDB). While TD employees must usually use both the ACS and CDB to know where exactly to ensure the service assistance, clerical workers from the AD exclusively rely on the CDB data to issue invoices and keeping Ristobill books updated. Figure 1 shows the data flow and processing phases described above.

III. THE HDQ META-MODEL

The HDQ meta-model shown in Figure 2 represents the types of organizational knowledge and the mutual relationships that we consider to feasibly and synthetically express all the relevant information that is managed within an organization. The HDQ meta-model is a meta-model since it is intended as general enough to be used to build knowledge models of single organizations by simple specialization of its concepts. These concepts are:

- **Organizational Unit**: 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.

**ReSource (RS):** a ReSource is any information source that an organization can either use or get access to in order to hold a representation of 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). A ReSource can be either a physical, tangible object or a computational data structure (i.e., a virtual object) with which the user interacts in order to fulfill her information needs or accomplish her duties. Within a business organization, RSs are managed during activities of either information production or information consumption. These activities are usually parts of greater sets of activities within a business process. Typical examples of RSs are single databases, specific data flows, either electronic or paper-based documents and, more generally, any set of data of interest and the medium used to represent it. As such, every RS can be seen from either the extensional or the intensional point of view, i.e., from the point of view of either its content or its conceptual schema. The fact that any resource – or better yet, its conceptual schema – can refer to one or multiple abstract concepts, leads us to introduce the concept of Informative Object.

*In our case study there are three resources: i) the White Page Directory (WPD), ii) the Agent-customer spreadsheet file (ACS), and iii) the Corporate Database (CDB).*

**Informative Object (IO):** IOs represent the main conceptual entities that are possible to “extract” from the RSs used and managed within the organization. Accordingly, IOs can be seen as what RSs give access to, what they refer to and also what users really retrieve and retain when they exploit/consume/access a RS. In other words, an IO is a conceptual entity that is independent of the specific way a resource represents it as well as of the resource’s physical medium, format or structuredness. The HDQ meta-model proposes to explicitly distinguish between RSs and IOs right because both the contextual and situated dimension pertaining to the quality of resources and the conceptual and abstract dimension pertaining to the quality of informative objects must be addressed, each with the most suitable techniques for the dimension under consideration. Moreover, our point is that the tight tie between data sets and the corresponding information must be acknowledged to enable and support a comprehensive approach to data quality assessment, improvement and management – each to be accomplished at the most suitable level of description.

*In our case study, an example of IO is the concept of Customer. A customer table within a CDB can be the most convenient and reliable means to store within a single server confidential information that must be queried along several and often quite complex dimensions; conversely, ACS file can be a convenient means to quickly and easily transmit customer directories to multiple traveling salesmen asynchronously. Moreover, the WPD can be used to quickly access to customer’s information by the agents of the MD. Accordingly, each resource is conceived and developed to conveniently represent only some aspects of the Customer concept it refers to and the whole set of characteristics of a given IO that are important for the Ristobill organization can be gathered only by considering all the mediums where this information is represented and used.*

We now address with more detail the concepts of RS and IO. In the HDQ meta-model, the concept of RS can be characterized in terms of properties along three main dimensions: (i) the identification of the resource; (ii) its intrinsic description; (iii) the characterization of the relationships between the RS and its application environments, e.g., either the organization or the business processes in which it is used. Taking inspiration from the Dublin Core (DC) – a widely applied standard for describing networked resources – we consider properties such as name (cf., the DC title), and location (cf., the DC identifier and source) resource-identifying properties. Resource-characterizing properties are description, type (cf. the DC type and format) and a categorical property we call *structuredness* (read more below). The main properties that put a RS in relationships with its environment are ownership (telling which organizational unit owns and is responsible for the given RS, cf. the DC creator and rights elements), and use (telling whether the resource is either created, produced or consumed within a specific activity, i.e., it is either input, output or both of some process).

In regards to *structuredness*, we adopt a more formal way than just relying on macro-types of data sets (e.g., XML files for semi-structured) or anecdotal enumerations (e.g., spreadsheets are semi-structured, databases are structured) to assign to a RS one of the conventional labels mentioned in Section I, i.e., to categorize a resource as either unstructured, semi-, or structured. The structuredness of a resource (or part of it) is assessed on the basis of the extent an interpreter (either human or automatic) can distinguish *structuring meta-data* within the content of the RS. A resource is completely unstructured if no interpreter can recognize any metadata by which to identify the logical structure of the content. If, conversely, the interpreter recognizes, in this order, (a) data and domain types, (b) the existence of logical constraints, (c) semantic domains and (d) semantic constraints, the resources are associated to different levels of structuredness, from partly unstructured to highly structured, respectively.

In the HDQ meta-model, the concept of IO can be referred by multiple RSs within an organization but also the vice versa holds, since a single RS can refer to all the IOs of interest of an organization (e.g., its main database). Like the RSs that represent them, several properties can be considered to

---

characterize each IO, such as the obvious DC elements like name and description. Moreover, each IO is also associated to the processes in which it is involved, to the organizational units it belongs to and to the RSs from which it has been extracted. While RSs have heterogeneous representations, IOs are abstract descriptions of resources and of relationships among them. For this reason, IOs are represented using a conceptual model, the Entity Relationship model enriched with generalization hierarchies [8], since it is the most widely used conceptual model used in design methodologies of database applications.

**DQ dimension:** DQ dimensions characterize quality properties related to data. As such, they are usually not directly measurable: metrics express quantitative or qualitative measures of dimensions. Several dimensions and metrics have been proposed in the literature to characterize the quality of data managed in an information system [6]. Usually dimensions are only implicitly referred to the type of data. In the HDQ meta-model, dimensions and metrics are classified accordingly the type of data: structured, semi-structured and unstructured. The metrics of the DQ dimensions are adapted for the three types of data, and are progressively more complex to conceive and use from structured to unstructured data. Quality dimensions can refer either to the extension of data, i.e., to data values, or to their intension, i.e., to their schema. Despite the recognized importance of the quality of schema, the prevalent attention to the definitions of DQ dimensions has been devoted to data values, which, more extensively than schemas, are used in business and administrative processes. As a consequence, in this paper we especially deal with data dimensions.

In the HDQ Meta-model, we propose a new classification of the DQ dimensions according to the two different level of abstraction: RS and IO. More specifically, the DQ dimensions can be classified as:

- **ReSource dependent dimension:** it is associated to RS because it is dependent of the medium in which the data is re-presented (e.g. usability, volatility, portability). This kind of dimension can not be associated to IOs.
- **ReSource independent dimension:** it is associated, besides RSs, to IOs, since it is intrinsic of the conceptual meaning of data (e.g., accuracy, completeness and timeliness).

IV. AN EXAMPLE OF METHODOLOGICAL USE

The HDQ meta-model is mainly used in the assessment phase of the HDQ methodology. This phase aims at obtaining a quantitative evaluation of the quality issues identified previously in the State reconstruction phase [7] where the reconstruction of all the relevant relationships between organizational units, processes, resources and conceptual entities is performed. To this aim, it is necessary to select a subset of relevant DQ dimensions and related metrics. Then, these metrics are applied to the RSs in order to calculate their actual quality value.

The identification of DQ dimensions and metrics depend on the classification of them proposed in the HDQ meta-model. In particular, in this step it is necessary to classify each considered DQ dimension in order to establish which dimension is an instance of the RS-dependent concept or RS-independent concept in the HDQ meta-model. Thus, if a DQ dimension can be referred to an IO since it is either intrinsic or referable to a conceptual entity, or it concerns only the level of RS and directly relates to the type, medium and context of use of the RS at hand.

**In our case study, the following DQ dimensions are considered:**

- **Currency** concerns how promptly data are updated. As an example, if the residential address of a person is updated, i.e., it corresponds to the address where the person lives, then the currency is high. It is classified as a ReSource independent dimension.
- **Accuracy** is defined as the closeness between a value $v$ and a value $v'$, considered as the correct representation of the real-life phenomenon that $v$ aims to represent. As an example if the name of a person is John, the value $v' = \text{John}$ is correct, while the value $v = \text{Jhn}$ is incorrect. It is classified as a ReSource independent dimension.
- **Relevance** expresses the importance and the usefulness of a data in the business processes. It is classified as a ReSource independent dimension.
- **Originality:** expresses the extent a resource has either unique or identifiable author and origin. It is classified as a ReSource dependent dimension.

The figure 3 highlights the RSs and Customer IO with the associated DQ dimensions considered in the case study.

Since single IOs can refer to several different RSs, we need a composing function that, given an Informative Object
the value is estimated optimistically as the highest score of currency of the RSs composing the function:

\[ \text{Currency} = \text{Age} + (\text{DeliveryTime} - \text{InputTime}) \]  (1)

Another application of the HDQ meta-model in the HDQ methodology regards the choice of the optimal improvement process. In particular, it is used at the time when the target quality values to reach through the improvement process are set [7]. This activity is based on actual quality values, \( DQ_{ij} \), associated with the \( i \)-th Informative Object or ReSource and the \( j \)-th quality dimension. Data quality targets are defined by performing a process-oriented analysis [6], as summarized in the following.

The process-oriented analysis is grounded on the information expressed by means of the HDQ meta-model, and concerns the IOs, RSs, processes, organizational units involved and their mutual interrelations. In order to define feasible target quality values, it is necessary to fix the expected \( y \)-th process performance index \( \text{index}_{xy} \) along each \( x \)-th process \( p_x \). By considering the Informative Object \( IO_i \) involved in the process, the related data quality dimensions \( DQ_{ij} \) to improve can be identified and target values \( DQ_{proc}^*_{ij} \) can be determined. In order to relate process performance indexes with data quality measures, linear correlation indexes \( \alpha_{xyij} \) are used:

\[ DQ_{proc}^*_{ij} = \alpha_{xyij} \cdot \text{index}_{xy} \]  (2)

This analysis is driven by the characterization of the DQ dimensions proposed in the HDQ meta-model. Indeed, an actual quality value \( DQ_{ij} \) can be either associated to an IO – if it regards an RS-independent dimension – or to a single RS, if it regards an RS-dependent dimension (see Section III). In the former case, the value results from composing the data quality dimensions measured for each RS. Therefore, the obtained value of \( DQ_{proc}^*_{ij} \) represents an expected target independent of the single RS. At this stage, two options are possible: on the one hand, it is possible to keep the IO level of abstraction and then choose to apply improvement techniques that are oriented to the achievement of the composed target value only. Alternatively, it is possible to propagate the target values in order to define a target value for each RS. The propagation (see point 3 in Fig. 5) is driven by the Relevance DQ dimension measured in the assessment phase of the HDQ methodology\(^2\). In particular, it is necessary to specialize the formula to calculate the target values \( DQ_{proc}^*_{ij} \) by considering the following constraints:

1) the weights that correlate the composition of the RSs into the abstracted IO must be considered in the formula IV as follows:

\[ DQ_{proc}^*_{ij} = \alpha_{xyij} \cdot \text{index}_{xy} \cdot \text{weight}_{ij} \]  (3)

\(^2\)This dimension is evaluated by means of both qualitative (e.g., likert-scale surveys) and quantitative techniques (e.g., number of queries or hits) that are besides the paper’s scope.
2) when propagating the obtained target values for the single RSs, the target value of the abstracted IO must be equal to the initial one \( \text{DQ}_{\text{proc}} \) by using the same composing function considered in the assessment phase. In regards to the Currency dimension, the obtained target value refers to the Customer IO. At this stage, we choose to select a different level of abstraction and define a particular target values for each component RSs. We are able to identify the component RSs using the HDQ meta-model and hence define the associated target values corresponding to each of them. Figure 5 illustrates this example in a step-like fashion from step 1 (composition), through step 2 (correlation by formula IV), to step 3 (propagation).

![Figure 5: Definition of target values for the currency DQ.](image)

The weights shown in Figure 5 represent the Relevance DQ dimension for each RSs. To obtain the target values, we apply the formula \( 3 \) considering the relevance weight (first constraint). For example, we considerer 0.5 as relevance weight to calculate the target value for the white page directory. Then, by applying the composing function depicted for currency in Fig. 4 (i.e., the max between the currency values of the single RSs), we obtain the target value associated to the IO (second constraint).

V. CONCLUSIONS

In this paper, we outline the HDQ meta-model and exemplify its application to a realistic organizational scenario within a methodology for assessing and improving corporate data quality. This model is aimed at the modelling of the information assets of an organization in heterogeneous data contexts. The main idea underpinning the HDQ meta-model is to decouple the quality of organizational information and the quality of the data sets where this information is represented. This decoupling is aimed at enabling the correlation between the information qualities of the main concepts used in an organization and the use and management of its main information resources. This meta-information can be used in the process of identifying which resources should be comprehensively assessed and which improved in a quality improvement process. We are currently testing the feasibility of the HDQ model in addressing the typical concerns of private organizations in terms of data quality targets and interventions. To this aim, we are deploying the model within a comprehensive methodology to manage the data quality of structured, semi-structure and unstructured information sources. Future work encompasses the extension of the HDQ meta-model to resources representing information in other format than mere text, like pictures and audio samples.

ACKNOWLEDGMENT

The authors would like to thank Carlo Batini for the invaluable help and insightful comments he provided throughout the whole process of development and application of the HDQ model.

The work presented in this paper has been partially supported by the European IST project n. 27347 SEEMP - Single European Employment Market-Place, the FIRB project NeP4B - Networked Peers for Business and the FIRB project eG4M - eGovernment for Mediterranean countries.

REFERENCES