Towards an Ontology of Quality Information Systems

Houda Zaidi  Faouzi Boufâres  Naoufel Kraiem  Yann Pollet
Laboratoire CEDRIC  Laboratoire LIPN  Laboratoire RIADI, Université de Manouba, Université Paris 13  Villetaneuse, Villetaneuse, France  Tunis, Tunisie
CNAM, Paris, France  Université Paris 13  Tunis, Tunisie
Laboratoire RIADI, Tunis, Tunisie
Université de Manouba, naoufel.kraiem@ensi.rnu.tn yann.pollet@cnam.fr
boufares@lipn.univ-paris13.fr
Tunis, Tunisie
zaidi.houda_h@yahoo.fr

Abstract: The quality of information systems is a multidimensional concept. It covers several aspects including data quality. Anomalies in the data sources such as duplicates, similar and inconsistencies may be due to the poverty of their semantic descriptions. The use of ontologies to improve the semantic and the interoperability between systems should contribute to the improvement of quality.

Keywords: Quality of information Systems, Data Quality, Data Warehousing, Data-integration of heterogeneous data sources, Semantic, Ontology of Information Systems, ETL (Extract-Transform-Load) tools.

I. INTRODUCTION

The quality of information systems (IS) is a multidimensional concept. It covers several aspects. It concerns the software quality and data quality. It includes more and more the quality of conceptual models, processes and ontologies [1].

In the literature there are many proposals visants to define a global view on the quality of IS. Hoxmeier [2] described the quality of the database (DB) through four aspects: processes, data, models and behavior. He produced a first list of Unified dimensions (Figure 1 is given in the appendix). Stylianou et al. [3] proposed a definition of IS quality in a six themes: quality of data that entering in the IS, the quality of information from the IS, software quality, quality of infrastructure (hardware, networks, basic software), the quality of services which are provided by the IS and finally the administrative quality (management of the IS function) of the IS direction.

In this context, we propose an ontology that unites the different approaches of quality including the components of an IS (Figure 2 is given in the appendix).

Our future work will focus on the aspect of data quality. The control of data quality is a critical issue especially for large organizations which need to exchange information between systems. It is more difficult to manage in the data-integration system due to data provenance from multiple sources. These sources are distributed, autonomous and heterogeneous and they have different schema and levels of quality.

There are many problems due to the heterogeneity of IS such as semantic and structural aspects of the information. Problems of quality of data stored in databases and data warehouses spread to all types of data (structured and unstructured). In particular the data errors i.e. duplicates, inconsistencies, missing values, incomplete, uncertain, obsolete, aberrant or unreliable. The consequences of non-quality data on decision-making and financial cost are considerable.

With the proliferation of information sources and the increase of volumes of data potentially available, data quality, more broadly, the information quality has taken a leading role within organization and in the last ten years in academia. Improve the quality of data in heterogeneous information systems returns to correct errors in each of DB and data warehouses (ED), to consider the structure and semantics of data in the process of integration, and to model the different interaction and interoperability between databases.
This paper is organized as follows: Section 2 provides a state of the art of work addressing the problematic of quality. In Section 3, we discuss the concepts of IS quality. An overview of our future work is given at the end.

II. STATE OF THE ART

A. Software Quality

In computer science, particularly in software engineering, quality is a global appreciation of software, it is based on many indicators. An overall appreciation of quality takes into account external and internal factors: External factor are measured by the dynamic properties of the code when executed (such as response time). Internal factors which are measured by the static properties of the code, typically by inspection (such as path length).

Several approaches of quality have been proposed in the literature. The approach of Bohem and al. [4] evaluates the quality of software by combining quality attributes of quality factors leading to quality measures. They identified eight criteria i.e understandability, conciseness, completeness, testability, structure, maintainability, reliability and efficiency. Research in software quality has also produced several standards such as ISO 9126.

B. Data Quality

Different research communities in DB, Statistics, and process management proposed several approaches to assess each dimension by various metrics, and control aspects of data quality ([5], [6], [7]).

Most work which addressing the problematic of data quality can be classified into four major types of approaches:

1) **Preventive approaches** focused on the IS engineering and process control with techniques for assessing the quality of conceptual models, the quality of software development and the processes used in the treatments of data.

The approach of Naubourg et al. [8] ensures the initial data quality at the moment of their importation in the context of clinical proteomics.

In their work, Akoka et al. [9] have presented a framework for assessing the quality of multi-sources IS in QUADRIS project (Quality of Data and Multi-Source Information Systems). This framework has identified a meta-model to examine the inter-dependencies between the dimensions of quality of a conceptual data model and those of data quality instantiating this model. This meta-model focuses on describing the different dimensions of quality. Each dimension can be declined into several factors. Each factor can be associated with a different set of metrics. A metric can correspond different measurement methods. This proposition has been validated in the three following fields: biomedical, commercial and geographical.

2) **Diagnostic approaches** focused on statistical, analysis and data mining methods used for detection anomalies in the data.


This approach indicates the advantage of the use of association rules as an automatic tool for audit of the actual data through the analysis of some files generated by the billing and order management system.

3) **Corrective approaches** focus on techniques for cleaning and consolidating data (imputation of missing values, removing duplicates) and using languages of manipulation of extensive data and tools for extraction and data transformation (ETL - Extraction-Transformation-Loading).

Boufares and Ben Salem [12] described an ongoing research project dedicated to the evaluation and improvement of data quality in enterprise information systems. They presented different strategies to handle with conflicts. They focused on the constraints at any level of integration: the domain constraints global consistency, the cardinality constraints and functional dependencies. They proposed algorithms for the elimination of similar data and data enrichment.

4) **Active or adaptive approaches** commonly applied in the mediation or integration of data. They focus on the adaptation of treatments (requests or cleaning the data) so that they include the execution in real-time and the checking constraint on data quality.

The approach of Peralta and al. [13] uses a set of dimensions quality (freshness and accuracy). They defined a Framework and algorithms based on the quality factors for data-integration systems.

C. Conceptual Modeling Quality

Nowadays, the software engineering is moving towards the model-driven engineering (MDE), in which a system is seen not as a series of lines of code, but as a set of abstract models describing each particular view (a concern) on the system. Then, these models determine all the rest of the development process and their quality impact heavily and directly the quality of the final system. The first approach of models quality is proposed by Batini and al. [14]. The criteria proposed are completeness, accuracy, expressiveness, readability, minimality, self-explanation, extensibility and normality. Cherfi et al. [15] proposed a framework for analyzing the quality of conceptual models in three dimensions: the specification includes all the items assessing the clarity, richness and correction of the model, the
dimension of the use is determined by the ease of understanding and completeness of the model. Finally the dimension of the implementation is defined by the ability of the model to be implemented and to facilitate the maintenance of resultant IS.

Vanbelle [16] proposed a Framework for evaluation and comparison of business models and an empirical test to validate it. He presented two types of classification criteria for quality assessment models. The first classification is composed of three axes of analysis namely syntactic, semantic and pragmatic and the second is composed of two measures: absolute measure (intrinsic quality, technical factors) and relative measure (extrinsic quality, commercial factors). In their work Genero et al. [17] have defined the factors that impact the maintainability of conceptual models and improve the modeling process. They defined a set of metrics to evaluate Entity-Relationship (ER) models. They defined a set of metrics in order to measure the structural properties of ER diagram. These metrics can be used as indicators of understandability which represent a key factor to determining maintainability.

III. INFORMATION SYSTEMS QUALITY

The components of an information system contribute to the success of the activities of an organization. Their quality is a very important issue. The cost of non-quality can be very high, for example, take a decision based on bad information can harm the organization or its customers and partners. Hence, the need for a global overview of IS quality, to capitalize knowledge around quality in various areas related to IS. The use of quality ontology which centered information system represents a very promising approach to capitalize the knowledge around the quality and to interoperate between the quality factors and metrics in the various fields.

We propose a quality ontology ([18]) which unites the different approaches incorporating quality components of an IS (Figure 2 is given in the appendix). We chose to structure the knowledge about the quality in the form of ontology; this concept has many advantages [19]:

- To provide a way to reuse the knowledge.
- To facilitate the understanding of knowledge related to a domain.
- To separate between the fundamental knowledge part and the part related to its implementation.
- To provide a means to analyze the structure of knowledge.
- Helping to achieve consensus around this knowledge.
- To facilitate the sharing and dissemination of knowledge.
- To facilitate the exploitation of knowledge through programs and tools.

First, we have instantiated the model of the ontology for the following areas: software quality, data quality, conceptual modeling quality, quality of ontologies. Then, we defined a set of inference rules in order to determine correspondences between concepts which are semantically close. Finally we have proposed correspondences between the concepts of quality from different areas (ontology alignment).

This proposition can group various work on the quality and move towards the federation and mutual enrichment of this work.

IV. CONCLUSION AND PERSPECTIVES

Our goal is to present an approach to guide the detection of anomalies in databases and data warehouses in order to provide a method of correction (cleaning process). It consists to address the following issues:

- To detect errors (missing values, incorrect values, inconsistent values) from a single source or multiple heterogeneous sources (based on different schemas or based on different meta-models).
- To detect duplicates and data integration. Indeed, it consist to locate the entities that represent the same real world, i.e to identify close or equal values for simple (string, numeric, date, binary) and complex data types (nested tables, user-defined types, image, text).
- To guide the user in the corrections through business rules.
- To mask the semantic heterogeneity between the IS in order to exchange, combine and merge the correct information from different sources. We propose an approach based on ontologies. This approach facilitates the communication of IS via common knowledge and it allows explaining and clarifying the meaning of data in order to be interpreted correctly for different systems. This knowledge can be captured through formal ontologies. We will use ontologies to formally describe the semantics of the concepts used in data sources, to describe each source by its own ontology, and to link all sources ontologies (local ontology) to a global ontology which sharing a common vocabulary.

Our ultimate goal is to participate in the development of new tools for data integration ETL that do not require user knowledge of the structure and semantics of the data handled from the sources. The data collected in the data warehouse should have more sense.
REFERENCES


Figure 1. Dimension of Database Quality [2]
Figure 1. Structure of IS Quality Ontology [18]