
Chapter 1. Introduction

This chapter introduces data quality evaluation in data integration systems, describes the addressed problems and presents an overview of our proposal for solving such problems.

1. Context

The technological advances of the last years allowed the development of information systems of wide scope, which offer access to large volumes of information distributed in multiple heterogeneous data sources. Although these systems have been proposed and used for more than a decade, they became more important in the last years, both at academic and industrial level. The increasing interest in this type of systems is mainly due to the proliferation of data available in remote sites, frequently stored in diverse platforms and with diverse formats. In particular, the World Wide Web has become a major source of information about all areas of interest, which can be used by information systems as any data supplier.

The needs of accessing in a uniform way to information available in multiple data sources are increasingly higher and generalized, particularly in the context of decision making applications which need a comprehensive analysis and exploration of data. The *Data Integration Systems* (DISs) appeared in response to these needs. A DIS is an information system that integrates data of different independent data sources and provides the vision of a unique database to users. Users pose queries via an access interface and the DIS answers their queries with information obtained and synthesized from source data. Information integration is the problem of combining the data residing at different sources and providing the user with a unified schema of these data, called global schema [Calvanese+2001], which represents the potential requirements of users. The global schema is therefore a reconciled view of the information, which can be queried by the user. It can be thought as a set of relations, potentially virtual (in the sense that their extensions may not be actually stored anywhere). A data integration system liberates the user from having to locate the sources relevant to a query, interact with each source in isolation, and manually combine the data from different sources.

Examples of DISs are *Mediation Systems* [Wiederhold 1992], which provide access to data extracted from several sources and integrated in a transparent way in response to user queries. *Data Warehousing Systems* [Inmon 1996], also extract, transform and integrate data from various, possibly heterogeneous, sources, aggregate and materialize information from this data and make it available for strategic analysis to the decision makers. Other examples of DISs are *Web Portals*, which provide access to subject-oriented information acquired and synthesized from Web sources, generally caching important amounts of data [Bright+2002].

In a context of DISs providing access to large amounts of data from alternative sources and conveying alternative query answers to users, information quality is becoming a *first class* property increasingly required by end-users. As the potentially retrieved data grows, users are more concerned about data quality [Wang+1996] [Gertz+2004] [Ballou+1998]. Some surveys and empirical studies have showed the importance of data quality for end users, in particular, when dealing with heterogeneous data coming from distributed autonomous sources [Wang+1996] [Mannino+2004] [Shin 2003].

Assuring the quality of the data conveyed to users is an important problem, which is closely related to the success of information systems. Numerous works in the areas of Information Systems and Software Engineering deal with quality control and quality assurance [Ballou+1998] [Pipino+2002] [Bobrowski+1998]. In the case of DISs, the problem is particularly complex due to the integration of data coming from multiple sources and possibly having different quality. Because of the great number and high diversity of data sources as well as their autonomy, it is important to have a fine knowledge of their quality and to take it into account during DIS design. In addition, information quality problems have been reported as critical in several scientific and social areas such as Environment [Jankowka 2000] [USEPA 2004], Genetics [Müller+2003a], Economy [Mazzi+2005] and Informatics in the Web [Gertz+2004]. Solving data quality problems opens a door to consider data production as any other item production.

2. Motivations and problems

Data quality evaluation in DISs consists in calculating the quality of the data conveyed to the users. Generally, data quality is best described or characterized via multiple attributes or factors, which describe certain properties about the data delivered to users (e.g. freshness, accuracy, completeness) and about the processes that manipulate this data (e.g. response time, reliability, security). Consequently, data quality evaluation consists in calculating several quality attributes, each one describing a specific quality aspect of data.

The quality of the information conveyed to users depends on the internal quality of source data (coherence, completeness, freshness, etc.) and on the characteristics of the processes that extract and integrate such data (policies, constraints, costs, delays, implementation features, etc.).

For example, let us consider a user asking for the most popular children videos and demanding certain levels of completeness, accuracy, freshness and confidence for query result. The quality of the delivered data depends on the intrinsic data sources and on the way of integrating data (e.g. the execution cost of the integration process may influence data freshness). As there may be several sources providing the same type of data, several extraction and integration processes can be conceived for conveying data to users. For example, different processes can provide children popular videos, e.g.: (P1) returning data from Disney®, (P2) returning data from Amazon®, (P3) integrating data from Amazon® and Block-Buster®, and (P4) conveying data from Film-Critiquer®, as illustrated in Figure 1.1.

Two problems arise: (i) deciding which sources to query and (ii) deciding how to merge the obtained data. These decisions can be taken considering the quality of data delivered by each process. In addition, the perception on data quality depends on users expectations, for example, some users do not care if the video list is quite incomplete if they have a rapid and accurate response; conversely, other users may want to examine all videos, even if they must wait additional time.

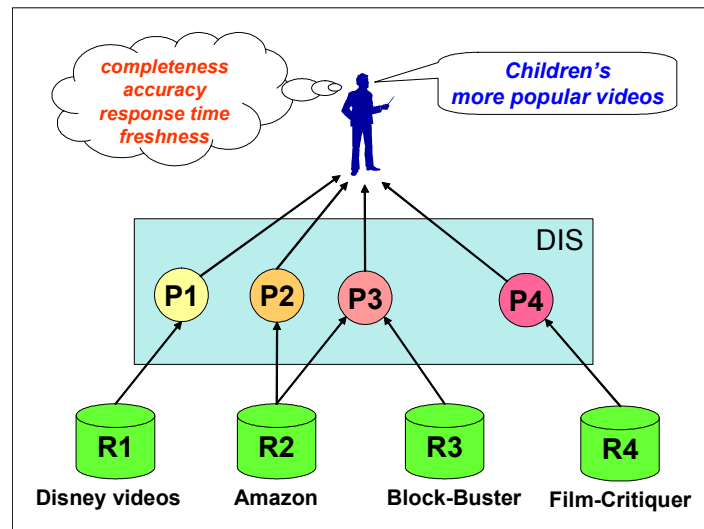


Figure 1.1 – Alternative responses to a DIS query

Consequently, in order to evaluate the quality of data conveyed to users we should study the quality of source data and the characteristics of the DIS integration processes. We should also analyze how to combine all these elements in order to aggregate a value that qualifies the delivered data.

In summary, data quality evaluation in DISs involves a certain number of technical issues:

- *Analysis of quality factors and metrics*: This concerns the analysis of quality factors, the definition of appropriate metrics for measuring them and the study of the DIS properties that have impact in their evaluation (for example, the execution delay of the system may impact data freshness).
- *Definition of user quality expectations*: This concerns the study of the quality factors that are more appropriate for representing user quality needs and the specification of user quality expectations according to such quality factors and metrics.

- *Assessment of source data quality*: This concerns the specification and implementation of measurement procedures for measuring the quality of source data (or a sample of source data) according to the analyzed quality factors and metrics.
- *Assessment of DIS property values*: This concerns the specification and implementation of measurement processes for taking measures or estimating DIS property values (for example, routines for measuring the average response time of the DIS).
- *Quality auditing*: Having measures of source data quality and DIS properties, such measures should be aggregated into a value that qualifies the delivered data. This concerns the study of the influence of those properties in specific quality factors and the specification of aggregation functions. The confrontation of the calculated quality values with those expected by users allows deciding if user quality expectations can be achieved and in which degree.
- *Quality improvement*: This concerns the specification of improvement actions for enforcing data quality. There may be a wide range of improvement actions involving different aspects of DIS design, varying from simple source selection (for example, if two sources provide the same type of data, the DIS can convey data from the source having the best quality) to changes in DIS design (for example, if a source is accessible only during certain periods, the DIS can materialize some data in order to enforce data availability).
- *Quality-driven DIS design*: This concerns the design of DISs taking into account quality guidelines, which may have the form of quality constraints or improvement actions.

Many quality factors have been proposed for modeling data quality, some of them defined in different manners. Several works propose classifications of quality factors according to semantic criteria [Wang+1996] [Naumann+2000], process-oriented criteria [Naumann+1999] [Weikum 1999] or goal-oriented criteria [Jarke+1997]. Other works propose formal frameworks for describing quality factors and deal with the management and storage of the appropriate metadata [Strong+1997] [Jarke+1997] [Jeusfeld+1998] [Helfert+2002] [Gertz 1998] [Missier+2001]. However, there is no consensus in the definition of quality factors. Each application domain has its specific vision of data quality as well as a battery of (generally ad hoc) solutions to solve quality problems [Berti-Equille 2004]. Furthermore, even if quality factors are frequently treated as being independent, there exist lots of relationships among them. The great number of quality factors and their inter-relationships cause quality evaluation to be a complex problem of many variables. Improving the quality of a system corresponds to optimizing a problem of N variables, which is of high complexity if done in a general context. As a consequence, it is difficult to consider many quality factors at a time. In order to study data quality in depth, we believe that it is necessary to study separately each quality factor as well as the properties of the environment that impact it. Afterwards, we can consider interaction between quality factors.

In [Wang+1996], the various quality attributes are analyzed from the user perspective. However, the definition of accurate profiles that allow users to understand the different quality metrics and to express their expectations is an open problem.

Regarding the assessment of source data quality, several works study and classify assessment techniques, types of metrics, units and aggregation functions [Pipino+2002] [Naumann+2000] [Ballou+1998]. For some quality factors, there are also detailed techniques for specific DISs scenarios, for example, the assessment of data accuracy in customers address attributes [Laboisie 2005]. Analogously, existing works treating the estimation of specific DIS properties (e.g. costs) can be reused for quality evaluation, for example, the estimation of data change frequency in caching systems [Cho+2003].

Despite the existence of several proposals for assessment of source data quality and assessment of DIS property values, putting such capabilities to use in DISs still requires modeling effort. There is a proposal for combining quality values of source data using simple arithmetic operations (minimum, maximum, average, sum and product) [Naumann+1999]. The need of an algebra for combining source quality values was also highlighted in [Gertz+2004]. Other works analyze the combination of source quality values for some specific quality factors, as accuracy and completeness, for example [Braumandl 2003] [Motro+1998]. But none of these works takes into account DIS property values.

The few proposals addressing quality-driven design deal with source selection, i.e. when several sources provide the same type of data, selecting the one whose data has the best quality or building a ranking of sources according to its quality [Mihaila+2000] [Naumann+1998] [Nie+2002] [Zhu+2002]. In [Naumann+1999], several query plans, accessing to different sources are compared, selecting the one having the best quality. Quality values are also taken into account in [Braumandl 2003] for selecting sources and servers for executing a user

query and monitoring the execution. An alternative approach consists in improving data quality by correcting data errors (e.g. typing errors) or reengineering DIS processes in order to avoid introducing errors (e.g. loose precision during calculations). Several data cleaning tools have been proposed, each one proposing a wide range of correction functions [Galhardas+2000] [Sattler+2000] [Raman+2001] [Vassiliadis+2001] [Lee+2000]. In [Ballou+1998], authors present some guidelines for DIS reengineering in order to improve the relation quality/cost of the information. Other works propose solutions for specific DIS scenarios, e.g. [Bright+2002] [Gancarski+2003]. But despite the existence of specific proposals, the land of quality improvement and quality-driven design is almost unexplored.

In this thesis we address the problems of evaluating the quality of the data delivered to users in response to their queries and deciding if users' quality expectations can be achieved. We also discuss how quality measures can be used for improving the DIS and enforcing data quality.

3. Our proposition

In order to introduce our proposal, we describe the technical issues addressed in this thesis and we present our contributions.

3.1. Technical issues addressed in this thesis

This thesis focuses on three main technical issues: (i) *analysis of quality factors and metrics*, (ii) *quality auditing*, and (iii) *quality improvement*.

Our approach consists in studying one quality factor at a time, analyzing its relationship with the DIS, proposing techniques for its evaluation and analyzing improvement actions for its enforcement. Among the quality factors that have been proposed, this thesis analyzes two main ones: *data freshness* and *data accuracy*. We summarize the analysis of each factor with a taxonomy, which allows comparing existent works and highlighting open problems. These taxonomies will serve to study the properties that impact the evaluation of each quality factor.

Concerning quality auditing, we propose the development of quality evaluation algorithms that take as input source data quality values and DIS property values generating as output a value for the data conveyed to the users. To this end, we model the different elements involved in data quality evaluation in a quality evaluation framework. Among these involved elements there are: data sources, user queries, DIS processes, DIS properties, quality measures and quality evaluation algorithms. In particular, we model DIS processes as workflow processes in which the workflow activities perform the different tasks that extract, integrate and deliver data to end-users. Quality evaluation algorithms are based on the workflow graph representation and consequently, the freshness evaluation problem turns into value aggregation and propagation through this graph. The idea behind the framework is to define a flexible environment, which allows specializing evaluation algorithms in order to take into account the characteristics of specific application scenarios.

Concerning quality improvement, we propose different kind of improvement actions to enforce data freshness and data accuracy when user expectations are not satisfied. Such actions are building blocks that can be composed to improve data quality in concrete DISs. For data freshness, we propose an enforcement approach that supports the analysis of the DIS at different abstraction levels in order to identify critical points (the portions of the DIS that are the bottlenecks for achieving freshness expectations) and to target the study of improvement actions for these critical points. The graph representation of the DIS allows the rapid visualization of critical points. For data accuracy, we propose the partitioning of query results in areas (some attributes of some tuples) having homogeneous accuracy. This allows user applications to retrieve only the most accurate data, to filter data not satisfying an accuracy threshold or to incrementally convey data (e.g. displaying first the most accurate areas). This represents an improvement to source selection proposals because accurate areas of several source relations can be combined while discarding inaccurate areas (instead of discarding whole sources).

Among the technical issues presented as motivation, we analyze freshness and accuracy factors, and we propose some solutions for quality auditing and quality enforcing problems. The proposed quality evaluation algorithms take as input the DIS processes and a set of values qualifying source data, DIS properties and user expectations. However, we do not deal with the acquisition of such values, which is carried out by quality assessment, property assessment and profile management techniques. Analogously, we propose basic improvement actions for enforcing freshness and accuracy in DISs, but we do not deal with the development of design (or reengineering) methodologies for adapting and combining the improvement actions in concrete DISs.

3.2. Main contributions

The main contributions of this thesis are:

- ❑ ***A detailed analysis of data freshness and data accuracy quality factors.*** The main result of this analysis is a survey on data freshness and data accuracy definitions and metrics used in the literature, which intends to clarify the meanings of such quality properties. We also elaborated a taxonomy of elements that influence quality evaluation. Additionally, this analysis highlights major research problems which still remain to be solved. As far as we know, such deep analysis has not yet been done for these quality factors.
- ❑ ***The proposal of techniques and algorithms for the evaluation and enforcement of data freshness and data accuracy.*** Our contribution consists in the specification of evaluation algorithms and improvement policies for data freshness and data accuracy. The definition of a homogeneous framework to manipulate quality factors constitutes a basis to the identification of the DIS properties that impact freshness and accuracy evaluation. It also allows an easy development of evaluation algorithms that consider such properties.
- ❑ ***A prototype of tool intended to be used in practical contexts of DIS management.*** The main results concerning the implementation of the proposed framework are the specification and prototyping of a quality evaluation tool that manages the framework. The framework components are specified in an abstract model, which supports the dynamic incorporation of new components to the tool, especially the inclusion of new quality factors and their evaluation algorithms. This brings support for the extensibility of the tool regarding the evaluation of other quality factors. The operational-style of the proposal, in particular the graph representation of DIS processes and the specification of quality evaluation algorithms as graph propagation methods facilitate their reuse for a wide range of DIS applications.

4. Outline of the thesis

The remaining of this thesis is organized in five chapters:

Chapter 2 presents the state of the art on the evaluation of data freshness and data accuracy quality factors. It provides a survey of freshness and accuracy definitions and their various underlying metrics proposed in the literature. We explore the dimensions that influence their evaluation, which are organized in taxonomies. Guided by the taxonomies, we analyze and classify the relevant work proposed for dealing with freshness and accuracy in different kinds of DISs. Both analysis, for data freshness and for data accuracy, show that existing work focus on specific types of DISs, specific characteristics (e.g. materialized data, some types of errors) and specific metrics, but other configurations remain untreated.

We identify several open research problems: specification of user expectations, acquisition of source data quality, formulation of cost models, data quality auditing and quality-driven engineering. We conclude by positioning our work with respect to these research problems; concretely, we focus on data quality auditing and quality-driven engineering issues.

Chapter 3 describes our proposal for data freshness evaluation and enforcement. We address two main problems: (i) evaluating the freshness of the data delivered to users in response to their queries, and (ii) enforcing data freshness when users' expectations cannot be achieved.

We propose a quality evaluation framework that models the different elements involved in data freshness evaluation, namely: data sources, user queries, DIS processes, DIS properties, quality measures and quality evaluation algorithms. In this framework, a DIS is modeled as a directed acyclic graph, called quality graph, which has the same workflow structure than the DIS and contains (as labels) the DIS properties that are relevant for quality evaluation. Quality evaluation is performed through evaluation algorithms that compute data quality traversing the quality graph and propagating property values. We discuss two kinds of propagations: (i) propagation of actual values, from sources to user interfaces, in order to calculate the freshness of delivered data, and (ii) propagation of expected values, from user interfaces to sources, in order to determine quality constraints for source providers. Therefore, we propose two propagation algorithms for data freshness. These algorithms take into account the DIS properties that impact data freshness, namely, the processing cost of DIS activities and the delays among them. The algorithms can be instantiated for different application scenarios by analyzing the properties that influence the processing costs and delays in those scenarios.

In addition, the framework proposes facilities to perform data freshness enforcement. If users' freshness expectations are not achieved, we may improve DIS design in order to enforce freshness or we may negotiate with source data providers or with users in order to relax constraints. The proposed enforcement approach allows analyzing the DIS at different abstraction levels in order to identify the portions that cause the non-achievement of freshness expectations. We suggest some elementary improvement actions, which can be used as building-blocks for specifying macro improvement actions adapted to specific DISs. As an application, we study the development of an improvement strategy for a concrete application scenario. The strategy consists in analyzing (and eventually changing) the execution frequency of DIS processes in order to satisfy user freshness expectations.

Chapter 4 describes our proposal for data accuracy evaluation and enforcement. We consider a relational context where user queries consist in selections, projections and joins over a set of source relations. We address two main problems: (i) evaluating the accuracy of the data delivered to users in response to their queries, and (ii) enforcing data accuracy when users' expectations cannot be achieved. Several algorithms and techniques that are used in the proposal are detailed in this chapter, as a complement to the state of art presented in Chapter 2.

Our approach for accuracy evaluation consists in two main phases: (i) partitioning source relations in areas with homogeneous accuracy in order to represent their distribution of inaccuracies, and (ii) for each user query, partitioning query result based on the partition of source relations. Each area is a virtual relation (view) over a source relation. User queries are reformulated (rewritten) in terms of areas. Specifically, the result of a user query consists in the union of a set of sub-queries, each one extracting data from several areas. We evaluate the accuracy of each sub-query and aggregate an accuracy value for the query result. We reuse the quality evaluation framework proposed for data freshness. We present an accuracy evaluation algorithm that takes into account the partitions of source relations and propagates them to query result. We focus on a priori evaluation, i.e. estimating data accuracy before executing user queries. Our algorithm explicitly indicates the areas that have lower accuracy, which allows filtering data not satisfying accuracy expectations.

Thereupon, we deal with accuracy improvement. We propose to use the partition of query result in order to select the areas that have the best accuracy. Note that we do not select whole relations but the portions that have the best accuracy, which differentiates our approach from the existing source selection approaches. We also proposed some basic improvement actions, for filtering data with low accuracy in order to satisfy several types of accuracy expectations.

Chapter 5 presents our experimentation results. Firstly, we describe a prototype of a data quality evaluation tool, DQE, which implements the framework. This tool allows displaying and editing the framework components as well as executing quality evaluation algorithms.

The prototype is used to evaluate data freshness and data accuracy in several application scenarios, enabling to validate the approach. Specifically, we describe three applications: (i) an adaptive system for aiding in the generation of mediation queries, (ii) a web warehousing application retrieving movie information, and (iii) a data warehousing system managing information about students of a university. We briefly describe each application, we model it in DQE (quality graphs, properties, etc.) and we explain the evaluation of data freshness and data accuracy for them.

Afterwards, we describe some tests for evaluating performance and limitations of the tool. To this end, we generated some data sets (quality graphs adorned with property values) and we executed a quality evaluation algorithm over each graph. The test results allow affirming that the tool can be used for large applications (modeling hundreds of graphs with hundreds of nodes each).

Chapter 6 presents conclusions and research perspectives.