Chapter 6. Conclusions and Perspectives

This chapter presents our conclusions, highlighting contributions and discussing research perspectives. We also discuss the use of our proposal in some current research projects.

1. Summary and contributions

Data quality evaluation and assurance have been recognized as fundamental issues in Distributed Information Systems. Numerous works in the areas of Information System Design and Software Engineering deal with quality control and quality assurance. In a context of Data Integration Systems (DIS) providing access to large amounts of data extracted from alternative sources and conveying alternative query answers to users, information quality is becoming a *first class* property increasingly required by end-users. 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. In addition, information quality problems have been reported as critical in several scientific and social areas such as Environment, Biology, Genetics, Commercial, Economy and Informatics in the Web.

In spite of the large number of research works dealing with data quality, several problems still remain to be solved. An analysis of the state of the art in data quality evaluation revealed that some technical issues have not been sufficiently treated. In particular, we highlight: the analysis of quality factors and metrics, the definition of user quality expectations, the assessment of source data quality and DIS property values, the evaluation of data quality, the enforcement of data quality, and the design of a DIS driven by data quality expectations.

This thesis deals with some of these topics and proposes a framework for data quality evaluation in data integration systems. We analyze two major quality factors: data freshness and data accuracy. We presented a taxonomy of freshness measurement techniques based upon the nature of data, the architectural techniques and the synchronization policies of the underlying DIS. We also proposed a taxonomy of accuracy measurement techniques based upon the granularity of measurement, the types of errors, the data types and the architectural techniques of the underlying DIS. Both taxonomies allow identifying the DIS properties that impact freshness and accuracy evaluation.

Such DIS properties are taken into account by quality evaluation algorithms in order to estimate the quality of the data conveyed to users in response to their queries. Evaluation algorithms take as input the DIS processes and a set of values qualifying source data, DIS properties and user expectations and combine these values generating as output a value that qualifies 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, data targets, DIS processes, DIS properties, quality measures and quality evaluation algorithms. In particular, we model DIS processes as direct acyclic graphs, called quality graphs, which have the same workflow structure than the DIS and are labeled with the DIS properties that are relevant for quality evaluation. Quality evaluation algorithms are based on the graph representation and consequently, the quality evaluation problem turns into value aggregation and propagation through this graph. We propose two types of algorithms: (i) for propagating quality actual values (from sources to targets) and (ii) for propagating quality expected values (from targets to sources). The former are used for estimating the quality of delivered data and the latter are used for determining quality constraints for source providers. The framework provides a flexible environment, which allows specializing evaluation algorithms in order to take into account the characteristics of specific application scenarios.

The proposed freshness evaluation algorithms take into account source data freshness, processing costs of DIS activities and inter-process delays among them. They can be instantiated for different application scenarios by analyzing the properties that influence source data freshness, processing costs and inter-process delays in those scenarios. We also propose different kinds of improvement actions to enforce data freshness when user expectations are not satisfied. Such actions are building blocks that can be composed to improve data freshness in concrete DISs. Our enforcement approach supports the analysis of a DIS at different abstraction levels in

order to identify critical points of the DIS and to target the study of improvement actions for these critical points. The graph representation of the DIS allows the rapid visualization of such critical points.

The proposed accuracy evaluation algorithm takes into account the distribution of inaccuracies in source relations. Specifically, source relations are partitioned in areas and sub-areas having homogeneous accuracy. User queries are rewritten in terms of areas and accuracy values associated to those areas are propagated through query rewritings. As a result, we obtain a partition of query results according to data accuracy. An accuracy value is also aggregated for the whole query result. We also propose some improvement actions for enforcing data accuracy. The proposed actions consist in filtering areas and sub-areas not satisfying accuracy expectations and in incrementally conveying data according to their accuracy (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 source relations).

Finally, we developed a prototype of a quality evaluation tool, called DQE, which implements the proposed quality evaluation framework. The tool allows displaying and editing the framework components as well as executing quality evaluation algorithms. The prototype was used for evaluating data freshness and data accuracy in three applications: (i) an Adaptive Mediation application delivering data about scientific publications, (ii) a Web Warehousing application retrieving movie information, and (iii) a Data Warehousing application managing information about students of a university. This experimentation allowed the validation of the approach in real applications and raised the attention on the practical difficulties of modeling DISs as quality graphs, setting property values and instantiating evaluation algorithms. In addition, we described some tests for evaluating performance and limitations of the tool which proved that it can be used for large applications, modeling hundreds of graphs with hundreds of nodes each.

As summary, 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 their 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. We proposed some evaluation algorithms that consider such properties and some improvement actions that intend to achieve quality expectations.
- ***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.

In next section we discuss some research perspectives.

2. Perspectives

In this section we discuss some improvements that may be done to our proposal in order to complete the analysis of data freshness and data accuracy and provide extensibility to the quality evaluation framework. We aim to treat these issues as future work. In addition, the research topics analyzed in this thesis suggest further research perspectives. Next sub-sections discuss both, near future work and research perspectives and discuss the use of our proposal in some current research projects.

2.1. Near future work

In near future, we aim to improve some features of the quality evaluation tool and perform additional performance tests. In addition, we aim to analyze the relationship among data freshness and data accuracy quality factors. We describe these perspectives:

Improvement of the DQE tool

In Chapter 5, we described a prototype of the DQE tool, which implements the quality evaluation framework and allows the execution of quality evaluation algorithms over quality graphs. The functionalities of the tool were described in Sub-section 2.1 of Chapter 5. However, further functionalities are only partially provided or are scheduled for future versions. Concretely, we want to extend DQE with the following functionalities:

- Navigation in a hierarchy of quality graphs: Currently, the tool allows visualizing several quality graphs, which may represent the DIS at different abstraction levels. However, the methods for navigating in the hierarchy of graphs (level-up and level-down, as well as zoom+, zoom–, focus+ and focus–) are not yet implemented. Hence, the tool brings limited support to the top-down analysis approach presented in Chapter 3.
- Graphical representation of partitions: Partitions are treated as property labels (as explained in Sub-section 4.4 of Chapter 4) so they can be edited and displayed as formatted text. However, we should provide a graphical interface for visualizing partitions, e.g. coloring sub-areas. In addition, areas not satisfying accuracy expectations can be graphically highlighted in order to quickly visualize critical points of quality graphs (this is analogous to the coloring of critical paths for data freshness).
- Implementation of a library of improvement actions: By the moment, only elementary improvement actions (described in Sub-section 4.4 of Chapter 3) can be applied. We aim to provide a collection of macro actions as well as the mechanisms for developing new actions by composing elementary ones.

Concerning scalability, we aim to modify the persistency mechanism of the tool in order to support quality evaluation in larger applications. Test results presented in Chapter 5 showed that the tool can be used for large applications (modeling hundreds of graphs with hundreds of nodes each). However, the application does not scale to thousands of huge graphs. Scalability must be provided replacing the memory storage by the access to secondary storage. Remember that the second version of DQE (used in the tests) loads all quality graphs in memory. In the third version, we experimented a new loading mechanism, in which quality graphs are loaded in memory only when they are used (either for editing the graph or for executing an evaluation algorithm over it) and memory is liberated thereafter. As expected, this new mechanism is less performing, which is unnecessary paid in small applications. For that reason, we aim to improve the implementation of the persistency layer in order to cache, in main memory, a certain number of quality graphs (possibly all, in the case of small applications). Such cache should be reactive and proactive, i.e. it should contain the quality graphs that have been recently used as well as load the quality graphs that will be used soon (e.g. when executing an algorithm in batch mode we know which are the following graphs that will be accessed).

Test of performance and precision of the accuracy evaluation algorithm

In Chapter 5 we presented the results of some tests performed over DQE in order to determine the number of quality graphs that can be loaded simultaneously and the performance of executing the freshness evaluation algorithm over such graphs. We plan to perform similar tests for the accuracy evaluation algorithm.

Additionally, in Chapter 4 we observed that the precision of obtained accuracy values relies on the techniques used for measuring accuracy of source relations, partitioning them and estimating the selectivity of rewritings. We want to investigate which is the influence of such techniques, for example, which is the gain in precision if we use sophisticated statistical models instead of simple histograms for selectivity estimation. Our proposal for accuracy evaluation can be compared to the proposal of [Naumann+1999] in order to measure in which degree the partitioning of source relations allows obtaining a better approximation of the accuracy of query results. We should also compare the approach with the proposal of [Rakov 1998] which measures accuracy of the exact query result.

Exploring the relation between data freshness and data accuracy quality factors

In this thesis we studied, separately, freshness and accuracy quality factors, identifying the DIS properties that impact their evaluation. However, these quality factors are not independent, and hence, actions for improving one factor may have side-effects (improve or degrade) on the other. For example, some actions for improving data freshness (e.g. reducing processing costs of activities or refreshing materialized data more frequently) may also improve semantic correctness because they reduce the amount of obsolete data (data that no longer represent real-world because of changes in real-world). Conversely, corrective actions for enforcing data accuracy (e.g. data cleaning or format standardization) may consume significant time and degrade data freshness. In addition, in some application contexts cleaning processes cannot be executed on-line forcing data materialization, which also degrades data freshness.

A trade-off between accuracy and freshness is necessary. Such trade-off should take into account the relationship among freshness and accuracy factors. Table 6.1 sketches some correlations among them. However, as specific improvement actions can be defined for specific scenarios, specific relations may be analyzed for such scenarios. We made a preliminary analysis of correlations in some particular scenarios. We aim to formalize such analysis in near future and obtain quality improvement guidelines that take into account both factors.

Freshness factors / **Accuracy factors**	**Currency / Timeliness**
Semantic correctness	- Improving data freshness reduces expired data → improves data semantic correctness - Correcting errors takes time → degrades data freshness
Syntactic correctness	- Constraints (as format or belongness to a referential) are not time-related, so improving data freshness does not affect data syntactic correctness - Correcting errors takes time → degrades data freshness
Precision	- Improving data freshness does not affect data precision - Improving activities to avoid losing precision may take time → degrades data freshness

Table 6.1 – Correlations among freshness and accuracy factors

2.2. Other research perspectives

This thesis proposed some techniques and algorithms for quality evaluation and quality enforcement that we aim to extend in three lines: (i) the analysis of further quality factors and their inter-relationships, (ii) the development of specialized quality enforcement strategies, (iii) the application of the proposed framework to further application scenarios. We describe several research perspectives in these directions:

Generalization of the framework

In Chapter 3 we presented a quality evaluation framework, which allows modeling the DIS properties involved in data freshness evaluation and developing freshness evaluation algorithms. In Chapter 4 we extended the framework for evaluating data accuracy, representing further DIS properties and developing a new evaluation algorithm.

For accuracy evaluation we focused in a mediation scenario (relational model, JSP queries) and thus, we proposed evaluation and enforcement techniques for this scenario. As future work, we hope to extend the approach to consider other scenarios, especially those having activities for correcting errors (e.g. data cleaning and format standardization routines). The Data Warehouse application described in Chapter 5 can be taken as a case of study for proposing evaluation techniques and algorithms for this type of scenarios. The evaluation approach proposed in Chapter 4 is a first step towards the definition of general evaluation methods that might be instantiated to several types of application scenarios, as we proposed for data freshness.

On the other hand, the idea of partitioning relations can be reused for data freshness, i.e. we can partition source relations (and query results) in areas having homogeneous freshness. In this line, the analysis of the relationship

between freshness and accuracy involves an additional challenge: the partitioning of relations according to several quality factors.

Additionally, we think that the framework can be reused for evaluating other quality factors. To this end, we should analyze the DIS properties that impact those factors and we should implement the appropriate evaluation algorithms. Preliminary experiments performed with the Adaptive Mediation application (presented in Chapter 5) showed that such extension is feasible for certain quality factors, but of course, detailed surveys, as those presented in this thesis for data freshness and data accuracy, should be done for further quality factors. The abstract model of DQE, which supports the dynamic incorporation of new components to the tool (especially the inclusion of new quality factors and their evaluation algorithms) brings support for the extensibility of the tool regarding the evaluation of other quality factors.

Data quality in a context of information personalization

Personalization and quality of information are two major challenges for computer science industry. Relevance, intelligibility and adaptability of retrieved information are the key factors for success or rejection of many information retrieval systems. In this context, adaptability means that usual practices and preferences of users must be taken into account. Kostadinov's thesis analyzes the most important knowledge that should compose a user profile and proposes a generic profile model, which includes quality preferences [Kostadinov 2006]. The analysis of the quality factors that are the most relevant for end users and the definition of evaluation techniques adapted to their environments become important challenges for data personalization.

Another important issue is the comparison of actual quality values (aggregated from source quality values) with user quality expectations (expressed in user profiles). This comparison can be done at different moments: (i) during query rewriting, in order to reformulate queries taking into account user profiles, for example, to include further predicates for representing user preferences; (ii) during query execution, in order to consider the sources satisfying quality expectations; and (iii) during delivery of query results, in order to order results according to their quality. Some ideas presented in Chapter 4, specifically those concerning improvement actions, can be applied to this context. Analogous improvement actions should be defined for other quality factors.

Other aspects of user preferences (e.g. the relative importance of quality factors for users) should be modeled in user profiles and used in the personalization process. The filtering of data according to multi-criteria conditions should be also analyzed.

Integration with quality assessment techniques

Even the DIS and its relevant properties can be easily modeled in DQE, the assessment of source data quality and DIS property values may be a tedious task and may imply the development of specialized routines. In this thesis we do not dealt with the assessment of source data quality nor DIS property values, but we found that the analysis of assessment techniques can be an interesting line for future works.

Many enterprises are interested in the development of web-services or plug-ins in order to obtain statistics of DIS properties and invoke quality evaluation algorithms. In some cases (e.g. a telecommunications company having 900 servers) the automation of assessment techniques is essential. To achieve this, it is necessary to develop an extensible model for representing the relevant statistics, the logs that should store them and the processes for aggregating property values from logs. An example on the use of statistics and logs were presented in Sub-section 3.2 of Chapter 5 for Web Warehousing applications.

Development of specialized quality enforcement techniques

In this thesis we suggested basic improvement actions that are general enough to be applied to different types of DIS but their use for quality enforcement should be guided by some high-level strategies in order to be effective. In particular, the selection of the most appropriate actions for a given DIS may depend on addition criteria as DIS configuration, reengineering costs and user preferences. A wide range of improvement strategies (based on improvement actions) can be defined for specific DISs. Among these strategies, we are particularly interested in many ones that we partially explored during this thesis:

- Synchronization of activities: In Section 5 of Chapter 3 we analyzed the synchronization of activities in a concrete application scenario in order to reduce inter-process delays among activities and therefore improve data freshness. Similar analysis can be carried out for other application scenarios with different characteristics.
- Quality-driven data reconciliation. In the context of Web Warehousing applications (described in Sub-section 3.3 of Chapter 5), we experimented the incorporation of freshness values to disambiguate conflicts among data. The proposal of reconciliation policies that consider further quality factors becomes an interesting challenge.
- Quality-driven ordering and filtering of query answers: In Chapter 4 we proposed partitioning query results and filtering areas with low accuracy. Note that the most restrictive are users' accuracy expectations, the smallest is the result. Accuracy expectations should be balanced with completeness expectations for avoid filtering too much data and conveying a representative set of tuples to users. An alternative consists in ordering areas according to accuracy and incrementally deliver areas (the most accurate first), which allows users to dynamically decide the amount of data they want to analyze. The trade-off among data freshness, data accuracy and other quality factors should be also analyzed.
- Selective rewriting of user queries. Data filtering can be performed in early stages of query planning, precisely, during the rewriting of user queries in terms of partitions. In this context, sub-areas providing data with low accuracy can be excluded for query buckets and thus they will not be used in query rewritings. This strategy allows extracting the most accurate data of each source relation instead of discarding whole source relations. Analogously, the trade-off among accuracy, completeness and other quality factors should be analyzed.
- Quality-driven generation of mediation queries: In the Adaptive Mediation application described in Chapter 5, we described the use of data quality values for selecting the mediation query that best adapts to user quality expectations. Such selection is carried out by generating a set of candidate queries, evaluating their quality and comparing it with user profiles. Additionally, data quality can be used to improve query generation by eliminating intermediate results that cause the non-satisfaction of quality expectations. This filtering may considerable reduce the search space of the generation algorithms (proposed in Xue's thesis [Xue 2006]) and thus optimize generation performance. Furthermore, the generation of a limited number of queries also reduces the time spent in selecting the most appropriate one.

Although we have shown that our approach can be used for DIS design, maintenance and evolution, we don't treat these topics in this thesis. The thesis of Marotta [Marotta 2006], based in our framework, treats the problem of detecting changes in source data quality and propagating changes to the DIS. Its main goal is to propose techniques for maintaining as much as possible the satisfaction of users' quality requirements. On the one hand, she proposes a proactive strategy based on probabilistic techniques, which allow to model source quality behavior and to calculate the quality reliability of the system. On the other hand, she proposes a reactive strategy that must be applied for compensating source quality changes.

2.3. Towards quality-driven design of DIS

Figure 6.1 positions our proposal among some of the technical issues discussed as perspective. The proposed quality evaluation algorithms take as input the DIS processes and a set of values qualifying source data, DIS properties and user expectations. The acquisition of such values should be carried out by quality assessment, property assessment and profile management modules. Such modules should take into account the analysis of DIS and source properties that impact in quality evaluation. Thus, evaluation algorithms read property values from a metadata repository and combine them, obtaining quality values for the data conveyed to users. DIS processes adorned with quality values are the input for quality enforcement techniques, which propose improvement actions for DIS design. Quality-driven design methodologies should apply these improvement actions in the design or maintenance of DIS processes.

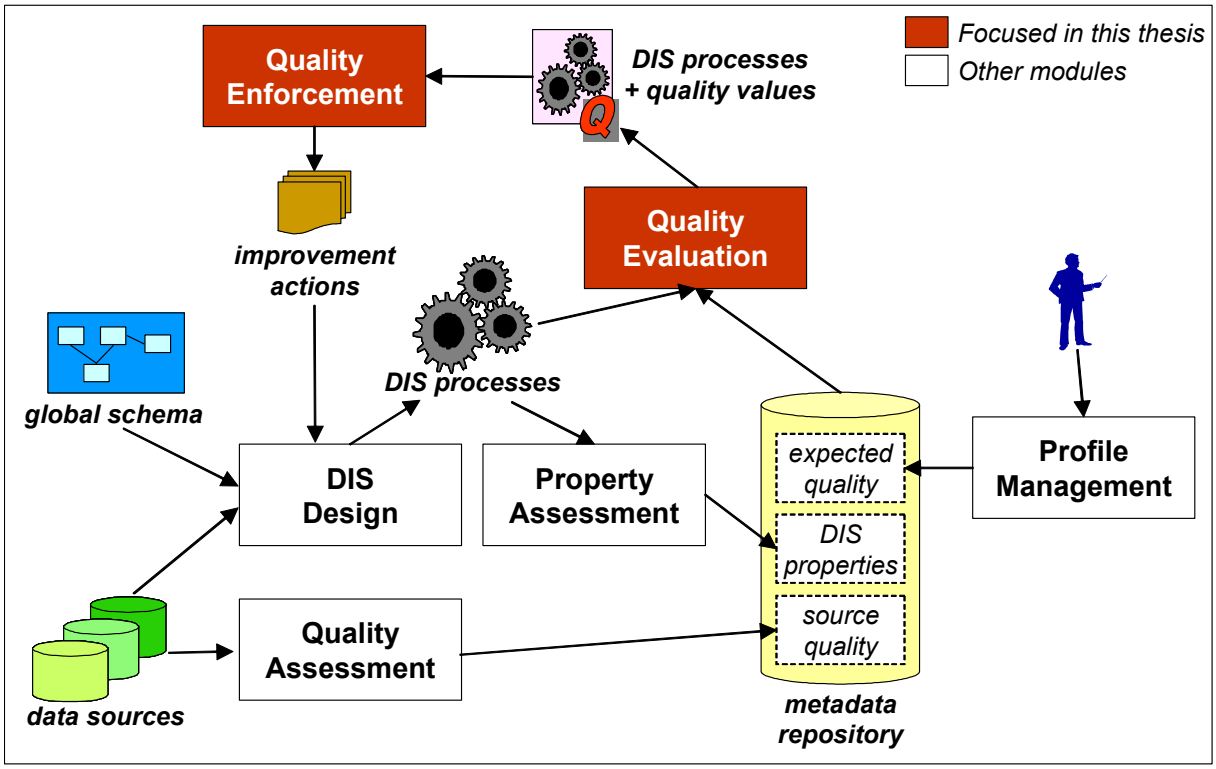

Figure 6.1 – Summary of the proposed solutions and perspectives

We are currently addressing these problems in some research projects: In the APMD* project, we will apply quality evaluation techniques and improvement actions to the context of data personalization. The main goal of the project is to carry out a cross investigation about personalization and quality of information. More precisely, its objectives are to propose formal models and robust algorithms, capable of capturing users' preferences which will be represented in user profiles and will be used for information filtering and adaptive display in a large scale environment. Concerning data quality, the project makes emphasis in the joint impact of profiles and information quality on the relevance of query results. A related ongoing thesis [Kostadinov 2006] proposes a taxonomy of the most important knowledge composing a user profile (including quality properties) and defines a generic profile model that can be instantiated and adapted for each specific application. In addition, a CSIC project† addresses quality evaluation and enforcement from a theoretical point of view, dealing with the analysis of several quality factors (including data freshness and data accuracy) and the proposition of techniques for quality-driven design and change management. An ongoing thesis [Marotta 2006] extends the quality evaluation framework with change management techniques. In the Quadris project‡ we analyze several quality factors and study their evaluation in several applications in the biomedical, commercial and geographical domains. The objective of the project is to solve the various data quality problems that appear when modeling DISs, when integrating and querying multi-source data and when evaluating data quality.

Finally, a cooperation agreement was signed with the Pasteur Institute§ in order to build a Laboratory Information Management System (LIMS). One of the technical aspects to be addressed is information quality management. There are two master thesis ongoing in this issues, one of them extending the quality evaluation framework with quality factors qualifying web services (e.g. avalaibility and performance), and the other one focusing in the construction of a decisional system for manipulating genetic data (including quality properties).

---

* Research project: "APMD - Accès Personnalisé à des Masses de Données" (Personalized access to massive data), ACI MASSES DE DONNEES - PROJET MD33/04-07, French Ministry of Research, France, 2004-2007. URL: http://apmd.prism.uvsq.fr/

† Research project: "Análisis de factores de calidad en sistemas de información multi-fuentes" (Analysis of quality factors in multi-source información systems), Instituto de Computación – Universidad de la República, financed by Comisión Sectorial de Investigación Científica (CSIC), Uruguay, 2005-2007. URL: http://www.fing.edu.uy/inco/grupos/csi/esp/Proyectos/Calidad/index.html

‡ Research project: "Quadris – Qualité des données et des systèmes d'information multi-sources" (Quality of data and multi-source information systems), ACI MASSES DE DONNESS, French Ministry of Research, France, 2006-2009. URL: http://www.irisa.fr/quadris/

§ Cooperation agreement between Institute Pasteur Montevideo and Instituto de Computación – Universidad de la República, Uruguay, 2006-2008.