Université de Versailles Saint-Quentin-en-Yvelines (France) and
Universidad de la República (Uruguay)

Data Quality Evaluation in Data Integration Systems

by

Verónika PERALTA

PhD Thesis

for obtaining the degree of **Doctor of Philosophy**
in **Computer Science**

Versailles, November 17th, 2006

Committee

Jacky AKOKA	Professor, CNAM, Paris (reviewer)
Catherine BERRUT	Professor, IMAG, Grenoble
Mokrane BOUZEGHOUB	Professor, Université de Versailles (advisor)
Sylvie CALABRETTO	Assistant Professor, INSA, Lyon (reviewer)
José COCH	Director of Services and Support, LINGWAY, Paris
Nicole LEVY	Professor, Université de Versailles (chair)
Raúl RUGGIA	Professor, Universidad de la República, Uruguay (advisor)

Acknowledges

I extend my sincere gratitude and appreciation to many people who made this PhD thesis possible. The first persons I would like to thank are my supervisors Mokrane Bouzeghoub and Raúl Ruggia, who supported me and placed their trust in me. I owe you lots of gratitude for having taught me so much during these years. *Merci Mokrane ! ¡Gracias Raúl!*

I would like to thank Professors Jacky Akoka, Catherine Berrut, Sylvie Calabretto, José Coch and Nicole Levy for having accepted to be members of my PhD committee. You monitored my work and took effort in reading and providing me with valuable comments.

I also owe a lot of thanks to all my colleges of the InCo Laboratory at the University of the Republic of Uruguay, who introduced me to the fabulous world of research during my master, encouraged me to start PhD and gave me great advices all along my career. Special thanks to Adriana Marotta, with whom I shared this quest from the beginning. I really enjoy working with you and I wish you a lot of success for your PhD dissertation. Many thanks to Regina Motz for sharing so many adventures with me. Many thanks to Martín Barrere, Fernando Carpani, Lorena Etcheverry, Pablo Gatto, Joaquín Goyoaga, Alejandro Gutiérrez, Jacqueline Guzmán, Ignacio Larrañaga, Federico Piedrabuena, Lydia Silva, Salvador Tercia and Diego Vallespir for being so nice friends and giving me the feeling of being at home at work. I would also like to thank Raquel Sosa, Enrique Delfino, Pablo Alzuri and María Freira for being fantastic friends and working late with me so many times. Very special thanks to Alvaro Illarze for his unconditional friendship and for the numerous times he helped me with my English. *¡Muchas gracias a todos!*

I would also like to thank all my colleagues of the PRiSM Laboratory at the University of Versailles St-Quentin-en-Yvelines, who received me and gave me the feeling of being in my country. Special thanks to Xiaohui Xue and Tao Wan, my Chinese sisters, who started and finished PhD with me. It was very nice to share so many hours with you during my studies; I hope we find new challenges to try together. I would also like to thank Dimitre Kostadknov and Juan Carlos Corrales. It was nice to work with you in a friendly environment; good luck for your PhD dissertations! Many thanks to Zoubida Kedad, who always was there for giving her sincere advice. Many thanks to Daniela Grigori, Stéphane Lopes and Assia Soukane for bringing a cheerful atmosphere at work. I would also like to thank Armando Borrero, Michel Cavaillé, François Dang-Ngoc, Tuyet-Trâm Dang-Ngoc, Sébastien Donadio, Clément Jamard, Huaizhong Kou, Octavio Ramírez, Romain Ravaux, Nicolas Travers, Daniel Vila Monteiro and Lilan Wu for being fantastic colleagues and friends and for the numerous times they helped me with my French. Special thanks to David Nott for his invaluable encouragement, patience and sense of humor. *Merci à vous tous !*

I must thank Lucyla Alonso, Annick Baffert, Laura Bermúdez, Joseline Cortazo, Chantal Ducoin, Isabelle Kretz, Catherine Le Quere, Isabelle Moudenner, Martiniano Olivera, Lucía Píriz and Mabel Seroubian, who helped me many times on administrative issues. I would also like to thank Julián Adib, Alejandro Blanco, Juan Diego Ferré, Jean-Louis Jammier, Marcelo Rodríguez, Alexander Sklar, Jorge Sotuyo and Felipe Zipitría for their useful support in networking issues.

I also thanks some people I met during my studies: Stéphane Boll, Daniel Calegari, Cosmin Cremarenco, Frédéric Dang-Ngoc, Mathieu Decore, Nicolas Dieu, Sophie Giraud, Aurelian Lavric, Cristian Saita, Christophe Salperwyck and Fei Sha for their kind friendship. Special thanks to Abdelkrim Lahlou, who taught me how to live in France and always offered me his friendly support. Many thanks to Noel Vidart and Gustavo Betarte for being my adoptive parents in France. *Merci pour votre amitié !*

I also owe a lot of thanks to my friends Eduardo Assandri, Santiago Fraschini, Luis Lena, Mario Maneyro, Julio Russo and Karina Talasimov, who accompanied me since my first day at University. Thank you for your encouragement and friendship. Many thanks to Beatriz Castro, Viviana García, Diego Olave, Marcos Orfila, Andrea Pereira, Mariela Questa, Leonardo Renard, Andrea Rodríguez and Rossanna Wirth for being so fabulous friends and for encouraging me, even by mail, all along this adventure. *¡Gracias por ser tan buenos amigos!*

Finally, I owe a lot of thanks to my family, who always supported me along my life: my parents, my sister Ana Laura and his husband Diego, my little sister Camila, my aunts Elbia, Esther, Maruja and Virmar, my cousins Andreia, Gustavo, Joaquín, Jorge, Juan Manuel, Pilar, Silvia and Valeria. *¡Los quiero muchísimo!*

Abstract

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. With the development of Data Integration Systems (DIS), information quality is becoming a *first class* property which is more and more required by end-users.

This thesis deals with data quality evaluation in DIS. Specifically, we address the problems of evaluating the quality of the data conveyed to users in response to their queries and verifying if users' quality expectations can be achieved. We also analyze how quality measures can be used for improving the DIS and enforcing data quality. Our approach consists in studying one quality factor at a time, analyzing its impact within a DIS, proposing techniques for its evaluation and proposing 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 analyze the different definitions and metrics proposed for data freshness and data accuracy and we abstract the properties of the DIS that impact on their evaluation. We summarize the analysis of each factor with a taxonomy, which allows comparing existent works and highlighting open problems.

We propose a quality evaluation framework that models the different elements involved in data quality evaluation, namely: data sources, user queries, DIS processes, DIS properties, quality measures and quality evaluation algorithms. In particular, DIS processes are modeled as workflow processes in which the workflow activities perform the different tasks that extract, integrate and convey data to end-users. Our reasoning support for quality evaluation is a direct 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. We develop quality evaluation algorithms that take as input source data quality values and DIS property values and combine such values obtaining a value for the data conveyed by the DIS. They are based on the graph representation and combine property values while traversing the graph. Evaluation algorithms can be instantiated for taking into account the properties that influence data quality in a particular application. The idea behind the framework is to define a flexible context which allows specializing evaluation algorithms for specific application scenarios.

The quality values obtained during data quality evaluation are compared to those expected by users. If quality expectations are not satisfied, several improvement actions can be taken. We suggest some elementary improvement actions that can be composed to improve data quality in concrete DISs. For enforcing data freshness, we propose the analysis of the DIS at different abstraction levels in order to identify its critical points (the portions of the DIS that cause the non-achievement of freshness expectations) and to target the study of improvement actions for these points. For enforcing data accuracy, we propose the partitioning of query result in areas (some attributes, 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 areas (e.g. displaying first the most accurate areas). Our approach differentiates from existing source selection proposals because we allow the selection of the areas having the best accuracy instead of only selecting whole relations.

The main contributions of this thesis are: (i) a detailed analysis of data freshness and data accuracy quality factors; (ii) the proposal of techniques and algorithms for the evaluation and enforcement of data freshness and data accuracy; and (iii) a prototype of a quality evaluation tool oriented to be used in practical contexts of DIS management.

Résumé

Les besoins d'accéder, de façon uniforme, à des sources de données multiples, sont chaque jour plus forts, particulièrement, dans les systèmes décisionnels qui ont besoin d'une analyse compréhensive des données. Avec le développement des Systèmes d'Intégration de Données (SID), la qualité de l'information est devenue une propriété de premier niveau de plus en plus exigée par les utilisateurs.

Cette thèse porte sur la qualité des données dans les SID. Nous nous intéressons, plus précisément, aux problèmes de l'évaluation de la qualité des données délivrées aux utilisateurs en réponse à leurs requêtes et de la satisfaction des exigences des utilisateurs en terme de qualité. Nous analysons également l'utilisation de mesures de qualité pour l'amélioration de la conception du SID et de la qualité des données. Notre approche consiste à étudier un facteur de qualité à la fois, en analysant sa relation avec le SID, en proposant des techniques pour son évaluation et en proposant des actions pour son amélioration. Parmi les facteurs de qualité qui ont été proposés, cette thèse analyse deux facteurs de qualité : *la fraîcheur* et *l'exactitude* des données.

Nous analysons les différentes définitions et mesures qui ont été proposées pour la fraîcheur et l'exactitude des données et nous faisons émerger les propriétés du SID qui ont un impact important sur leur évaluation. Nous résumons l'analyse de chaque facteur par le biais d'une taxonomie, qui sert à comparer les travaux existants et à faire ressortir les problèmes ouverts.

Nous proposons un canevas qui modélise les différents éléments liés à l'évaluation de la qualité tels que les sources de données, les requêtes utilisateur, les processus d'intégration du SID, les propriétés du SID, les mesures de qualité et les algorithmes d'évaluation de la qualité. En particulier, nous modélisons les processus d'intégration du SID comme des processus de workflow, dans lesquels les activités réalisent les tâches qui extraient, intègrent et envoient des données aux utilisateurs. Notre support de raisonnement pour l'évaluation de la qualité est un graphe acyclique dirigé, appelé graphe de qualité, qui a la même structure du SID et contient, comme étiquettes, les propriétés du SID qui sont relevants pour l'évaluation de la qualité. Nous développons des algorithmes d'évaluation qui prennent en entrée les valeurs de qualité des données sources et les propriétés du SID, et, combinent ces valeurs pour qualifier les données délivrées par le SID. Ils se basent sur la représentation en forme de graphe et combinent les valeurs des propriétés en traversant le graphe. Les algorithmes d'évaluation peuvent être spécialisés pour tenir compte des propriétés qui influent la qualité dans une application concrète. L'idée derrière le canevas est de définir un contexte flexible qui permet la spécialisation des algorithmes d'évaluation à des scénarios d'application spécifiques.

Les valeurs de qualité obtenues pendant l'évaluation sont comparées à celles attendues par les utilisateurs. Des actions d'amélioration peuvent se réaliser si les exigences de qualité ne sont pas satisfaites. Nous suggérons des actions d'amélioration élémentaires qui peuvent être composées pour améliorer la qualité dans un SID concret. Notre approche pour améliorer la fraîcheur des données consiste à l'analyse du SID à différents niveaux d'abstraction, de façon à identifier ses points critiques et cibler l'application d'actions d'amélioration sur ces points-là. Notre approche pour améliorer l'exactitude des données consiste à partitionner les résultats des requêtes en portions (certains attributs, certaines tuples) ayant une exactitude homogène. Cela permet aux applications utilisateur de visualiser seulement les données les plus exactes, de filtrer les données ne satisfaisant pas les exigences d'exactitude ou de visualiser les données par tranche selon leur exactitude. Comparée aux approches existantes de sélection de sources, notre proposition permet de sélectionner les portions les plus exactes au lieu de filtrer des sources entières.

Les contributions principales de cette thèse sont : (1) une analyse détaillée des facteurs de qualité fraîcheur et exactitude ; (2) la proposition de techniques et algorithmes pour l'évaluation et l'amélioration de la fraîcheur et l'exactitude des données ; et (3) un prototype d'évaluation de la qualité utilisable dans la conception de SID.

Resumen

La necesidad de acceder de manera uniforme a múltiples fuentes de datos es cada día más fuerte y generalizada, especialmente en el contexto de aplicaciones para toma de decisiones, las cuales necesitan de un análisis comprensivo y exploratorio de los datos. Con el desarrollo de Sistemas de Integración de Datos (SID), la calidad de la información se ha transformado en una propiedad de primer nivel, cada vez más requerida por los usuarios.

Esta tesis trata sobre la evaluación de la calidad de los datos en los SID. En particular, se abordan los problemas de la evaluación de la calidad de los datos que responden a consultas de usuarios y la satisfacción de las exigencias de dichos usuarios en términos de calidad. Se analiza también la utilización de medidas de calidad para mejorar el diseño del SID e incrementar la calidad de los datos. Nuestro enfoque consiste en estudiar un factor de calidad a la vez, analizando su impacto en el SID, proponiendo técnicas para su evaluación y proponiendo acciones para su mejora. Entre los factores de calidad que se han propuesto en la literatura, esta tesis analiza dos de los más usados: *la frescura* y *la exactitud* de los datos.

Analizamos las diferentes definiciones y métricas que se han propuesto para la frescura y la exactitud de los datos y abstraemos las propiedades del SID que juegan un rol importante en su evaluación. El análisis de cada factor se resume en una taxonomía, la cual permite comparar los trabajos existentes y resaltar los problemas abiertos.

Proponemos un marco de trabajo que modela los diferentes elementos relacionados a la evaluación de la calidad: fuentes de datos, consultas de usuarios, procesos de integración del SID, propiedades del SID, medidas de calidad y algoritmos de evaluación de la calidad. En particular, los procesos de integración del SID se modelan como flujos de trabajo, cuyas actividades realizan las tareas de extracción, integración y entrega de los datos a los usuarios. Nuestro soporte de razonamiento para la evaluación de la calidad es un grafo acíclico dirigido, llamado grafo de calidad, que tiene la misma estructura del SID y está etiquetado con las propiedades del SID que son relevantes para la evaluación de la calidad. Los algoritmos de evaluación de la calidad toman como entrada los valores de calidad de los datos fuentes y las propiedades del SID y combinan dichos valores obteniendo una medida de la calidad de los datos retornados por el SID. Los algoritmos se basan en la representación de grafo y combinan los valores de las propiedades mientras recorren el grafo. Los mismos pueden instanciarse para tener en cuenta las propiedades que influencian la calidad en una aplicación concreta. La idea detrás del marco de trabajo es definir un contexto flexible que permita la especialización de los algoritmos para escenarios de aplicación específicos.

Los valores de calidad obtenidos durante la evaluación se comparan con los valores exigidos por los usuarios. Si las exigencias de calidad no son satisfechas, se pueden realizar acciones de mejora al SID. Sugerimos un conjunto básico de acciones de mejora que pueden componerse para mejorar la calidad en SID concretos. Para mejorar la frescura de los datos proponemos analizar el SID a diferentes niveles de abstracción, de manera de identificar sus puntos críticos (las porciones del SID que causan la no satisfacción de las exigencias de frescura) y concentrar la aplicación de acciones de mejora sobre esos puntos. Para mejorar la exactitud de los datos proponemos partir los resultados de las consultas en áreas (algunos atributos, algunas tuplas) de exactitud homogénea. Esto permite que las aplicaciones de los usuarios desplieguen solamente los datos más exactos, filtren los datos que no satisfacen las exigencias de exactitud o desplieguen los datos en camadas según sus exactitudes. Nuestro enfoque se diferencia de los enfoques existentes de selección de fuentes porque podemos seleccionar áreas de buena exactitud en lugar de sólo seleccionar fuentes enteras.

Las principales contribuciones de esta tesis son: (i) un análisis detallado de los factores de calidad frescura y exactitud, (ii) la propuesta de técnicas y algoritmos de evaluación y mejora de la frescura y la exactitud de los datos, y (iii) un prototipo de evaluación de la calidad utilizable en contextos prácticos de diseño de SID.

Content