UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

DIPARTIMENTO DI INFORMATICA

PhD Program in Computer Science and Mathematics
XXXIII cycle

Research Project

PhD Student: Bibiano Rivas Garcia

Supervisor: Prof. Dr. Maria Teresa Baldassarre

Coordinator: Prof. Maria F. Costabile

PhD student signature _____

Supervisor signature _____

1. Research title:

A framework for the evaluation and improvement of Data Quality from an ISO/IEC 25000 perspective.

2. Research area:

Software & Data Quality

3. Research motivation and objectives

Companies need data for their organizations [1], [2]; furthermore it is an objective fact, that data is one of their most important assets [3]. In fact, data tends to be managed as such [4] because most companies see data as an indispensable pillar to face up with an increasingly competitive market [5].

Some organizations think that the greater the amount of data, the greater the advantages they will have to exploit them. For several years, the easiness and low cost with which data has been captured and stored thanks to the improvement and dissemination of data entry technologies and the large amount of data that can be obtained through the Internet; as well as the fact that the progressive cheaper costs of physical media favors the accumulation of large amounts of data. In addition, with the increasing implementation of Big Data solutions in companies, not only are these data quantities sought, but also quickly incorporated into the organization from different data sources - in heterogeneous cases - in order to analyze them in accordance with the organizational strategies and retrieve performance from them.

Unfortunately, these data is rarely properly governed and/or managed and due to daily operations, the information systems grow in a disorganized and unplanned manner. Furthermore, in many cases, there is no an adequate data architecture in the company. That is why it is often possible to find large amounts of "expired" historical data, which can no longer be used to carry out any process or obtain any type of other relevant information. This fact obliges organizations to face a serious problem of data pollution after a while: too much data is available, most probably useless and not analyzable. Fortunately, in recent years, organizations are beginning to realize that a situation of data with insufficient quality is unacceptable since their performance depends heavily on data ([6] and [7]). Indeed the decisions that can be made will not be better than the data they are based on [8], and therefore such decisions may generate or lead to errors that will negatively impact on the overall efficiency of the organization and information systems. Inadequate levels of data quality will mean: unused data, barriers to its accessibility or difficulty in its use in the tactical and strategic processes of product positioning [9]. These drawbacks can be the source of serious organizational problems both from the economic ([10], [6], [11],[12]), technical ([13],[14]), and social ([15],[16], [6],[17]) perspective. Furthermore, it is worth remembering the current laws on data protection [18].

Taking into account that a large number of organizations do not yet have the means, nor the tools, nor a description of the adequate processes to achieve a high level of data quality ([19], [20], [21]), and that they tend to neglect this area ([22], [23]), it is necessary to introduce organizational changes and to raise awareness among employees so that the level of information quality can be improved ([6], [13]). That is why organizations are implementing actions aimed at taking care of the quality of the data, as it is concluded in the report presented in [20], in which it is revealed that more than three quarters (77%) of the companies surveyed intend to increase their investments in projects to improve data quality as well as data-centric and data-driven projects.

In this thesis, data quality aspects will be addressed in software environments. Our objective is to quantify the quality characteristics of the different datasets used in those projects to determine if the data fit the specific use for which they were designed.
As far as we know, a data quality model that can be used as a reference to manage the data quality in contexts with large amounts of data is still not mature. Nor, of course, do the assessment and certification mechanisms of this data quality model exist. Considering that many experts [24] believe that Big Data will be a significant business volume for companies, and that data is at the core of such initiatives, we believe that it is very important to evaluate and certify the quality of data in similar contexts as it will most likely be a notable business opportunity that currently deserves to be investigated.

The starting hypothesis is contemplated below:
It is possible to develop a Data Quality Evaluation Environment that allows organizations to ascertain the quality levels of their data in different Software Environments?

In accordance to this hypothesis, the following research questions have been identified:

R.Q.1. Which are the data quality characteristics that should be included into a suitable Data Quality Model and how are they related with big data V's?

R.Q.2. Which are the measurement methods and functional activities that should be conducted during any Data Quality evaluation?

4. State of the art

Data Quality concept has a plethora of definitions, but the most accepted one is "fitness for use" [25]: data should meet needs and adequate characteristics in order to be used in the tasks that require them.

It is said that the data have quality if they serve the purpose (task) for which they intend to be used. This definition ([9], [26], [27]) has two important implications: (1) data quality is a multidimensional concept, which means that it is necessary to evaluate the quality of data, using several criteria called data quality dimensions [9], and (2) data quality is a subjective concept, which implies that different users may have a different perception of the level of quality of a dataset, and even that the same user may have different perceptions of the quality of the same dataset in different contexts ([2], [28]).

It is a proven fact that using data with inadequate levels of quality motivates the appearance of problems that will end up negatively affecting the tasks performed by the workers, and consequently the performance of the organizations ([29], [6], [30]), with negative consequences ([31], [32] , [13], [33],[6], [10], [11], [34]), both technical [13], and social ([15], [16], [6], [17]). This awareness for data quality is becoming so important for companies that some initiatives have launched to remedy situations of inadequate levels of data quality. To help in this concern, several consultancy organizations have conducted surveys to try to estimate the negative impact of using data with inadequate levels of quality.
Data quality evaluation requires identifying a set of data quality characteristics or dimensions. Each one of these dimensions represents the user criteria that is used to evaluate data quality, the election of these criteria respond to the user's data quality requirements [35].
ISO 25000 allows to study the quality of a Software Product, specifically, [36] provides a quality model for this type of products. In addition, in [37] the main characteristics that a Software Product should implement to treat data with an adequate level of quality are collected and outlined. These

characteristics represent what a software must do to meet certain Data Quality (DQ) requirements. These characteristics are the result of implementing certain DQ Software Requirements (RSDQ) [38] for each of the functionalities required for a Software Product in a specific context of use. The ability of the Software Product to manage data with adequate levels of quality would be given by the implementation of these DQ Software Requirements.
International standards are a good starting point since they are developed by experts aggregating the knowledge of the field of Data Quality. Additionally, and fundamentally, international standards are usually defined in a holistic manner to cover every possible business domain. Unfortunately,
ISO/TS 8000-8 only provides theoretical metrics and ISO/IEC 25012 is just a Data Quality Model. ISO/IEC 25024 [2] is not sufficient either, since it only provides basic metrics for the Data Quality Model of ISO/IEC 25012. Moreover, ISO/IEC 25040 provides a basis for creating a Data Quality Assessment Process; ISO/IEC 25012 provides as basis for selecting the characteristics that take
part of the Data Quality Model and ISO/IEC 25024, as basis for creating the measures to quantify the Data Quality characteristics. As one can imagine, there is still a lot of research to be done in this context.

5. Problem approach

The first phase to approach the problem is a deep study of the state of art and the literature in the field of Data Quality followed by a study of tools, techniques, technologies and standards in the field. For this reason the first part of the research thesis will by a systematic literature review focusing on pointing out state of art, state of practice and current research gaps.
The project aims to define a data quality model based on ISO/IEC 25000 applicable in various software contexts that are called to handle large amounts of data, among others, Big Data environments. This requires a detailed study of the different characteristics or dimensions of data quality and its influence in the different dimensions involved.
Once the relationships between the dimensions of Data Quality and big data have been identified, this relationship will be validated by implementing data quality measurement algorithms (data quality metrics) referred to various possible data sets.
A variant of "Technical Action Research" (TAR), has recently been proposed in [39] and [40] as a combination between design science and TAR, that starting from an artefact identifies the best way to validate it. Therefore, it is a type of research, "artifact-directed" as opposed to traditional ("problem-driven"). In TAR, an artifact is created, and it begins by carrying out a proof of concept, then testing it through small ("toy") problems under ideal circumstances, and then scaling conditions for solving more realistic problems, until it can be tested in organizations to solve specific problems.
To validate the thesis, a series of case studies will be carried out:
The first case study will analyze how complete the Data Quality Model is in terms of the selected Data Quality Characteristics in a specific software environment.
The second case study will address the process of the Evaluation Process with the Data Quality Model and the Data Quality Metrics developed.
The last case study will validate the data quality environment which covers the data quality model definition, the data quality metrics, the validation of the results and technological environment.

6. Expected results

This research aims to assess the extent to which different characteristics determine the data quality model for software systems and how we can evaluate the data quality of a system. Data will be collected and analyzed through the three case studies mentioned in the previous section. In particular, we are interested in investigating data quality metrics for different dimensions like volume, variety and velocity (3Vs) of data.

The expected results are the following:

- A framework for data quality evaluation
- A data quality metamodel to represent the results
- Data Quality metrics for different characteristics from an ISO 25000 perspective
- A data quality evaluation methodology

The main research domain of this proposal is Data Quality which is beginning to have importance due to the increase of related topics as Big Data, IoT, data science, etc. Many banks and insurance companies are starting to develop their own data quality projects in order to enhance the quality of the services to their clients to meet with national and European regulations. This proposal aims to identify and develop a solution that is appropriate for various environments that handle large amounts of data, and at the same time respects quality standards such the ISO/IEC family.

7. Phases of the project

The project is divided into 3 academic years of the doctorate course:

- **First Year: preliminary study of literature and the state of the art**
  - Activity 1.1: study of literature and the state of the art.
  - Activity 1.2: deepening the problems related to data quality measurement and evaluation techniques
  - Activity 1.3: research and study of methods for measurement and data quality
  - Activity 1.4: publication of results obtained in international journals and conferences
  - Activity 1.5: participation in international schools, conferences and seminars on topics related to the activity and planned objectives.
- **Second Year: synthesis, development and implementation of methods**
  - Activity 2.1: training period in a foreign university and comparison with the activity carried out in other research groups with similar objectives
  - Activity 2.2: synthesis, design and implementation of methods for the measurement and evaluation of data quality in environments that manage large amounts of data such as Big Data, IoT, , etc.
  - Activity 2.3: evaluation of the implemented methods, comparison with existing approaches in publication of the results obtained in journals and / or international conferences
  - Activity 2.4: planning of case studies
- **Third Year: practical applications and writing of the doctoral thesis dissertation:**
  - Activity 3.1: execution of explorative case studies
  - Activity 3.2: refinement of methods and implementation of specific characteristics for selected application domains by means of the case studies
  - Activity 3.3: analysis of the experimental results obtained in the domains of the selected application.
  - Activity 3.4: Writing the doctoral thesis dissertation.

	1st YEAR				2nd YEAR				3rd YEAR			
Activity	I TRIM	II TRIM	III TRIM	IV TRIM	I TRIM	II TRIM	III TRIM	IV TRIM	I TRIM	II TRIM	III TRIM	IV TRIM
1.1												
1.2												
1.3												
1.4												
1.5												
2.1												
2.2												
2.3												
2.4												
3.1												
3.2												
3.3												
3.4												

Figure 1: Gantt Diagram of the project

8. Result evaluation

Results of the research will be evaluated through in vivo case studies carried out in collaboration with local and international companies. Moreover, we will carry out two different types of approaches in planning and executing case studies:

- Analytical Approach: The case study is examined in order to attempt and understand what has happened and why. It is not necessary to identify problems or suggest solutions.
- Problem-Oriented Method: The case study is analyzed to identify the major existing problems and to suggest solutions to these problems.

All case studies will be defined according to the experimental guidelines outlined by Wohlin et al. [41]

9. Possible reference persons external to the department

A stay abroad will be made during the second year of doctorate in Spain in the Alarcos Research Group in the University of Castilla-La Mancha (Spain), under the supervision of:

- Prof. Mario Piattini Velthuis - Full Professor at University of Castilla-La Mancha (UCLM), Spain
- Dr. Ismael Caballero Muñoz-Reja - Associate Professor at the University of Castilla-La Mancha (UCLM), Spain

Moreover, the collaboration will be made formal through a cotutele agreement between the two universities, based on current existing Memorandum of Understanding between Spain and Italy.

During the period of the doctoral school we want to establish collaborative relationships with researchers from Italian and other foreign institutes or universities. The main contacts and external references are:

- ISO TC184/SC4/WG13 for the development of ISO 8000 series
- ISO TC184/SC4/WG23 for the development of ISO 8000 series
- DQTeam S.L a Spin-off from UCLM about Data Quality Assurance
- Ser&P a Spin-off from Bari University about Software Quality ISO 25000

- AENOR leading entity in certification of management systems, products and services, and responsible for the development and dissemination of UNE standards.

Finally, there are some important conferences and summers schools I plan to attend:

- International Conference on Information Quality (ICIQ)
- The Empirical Software Engineering International Week, ESEIW 2018 which comprises the International Doctoral School on Empirical Software Engineering (IDoESE) and ACM/IEEE Conference on Empirical Software Engineering and Measurement (ESEM)
- The International Conference on Software Maintenance and Evolution (ICSME 2018)
- International Summer School on Software Engineering 2018 (Salerno Italy and Oulu Finland)

10. References

[1] T. C. Redman, *Data Quality for the Information Age*, 1st ed. Norwood, MA, USA: Artech House, Inc., 1997.

[2] R. Y. Wang, *A Product Perspective on Total Data Quality Management*, vol. 41. 1998.

[3] Dama International, *The DAMA Guide to the Data Management Body of Knowledge: (DAMA-DMBOK Guide)*. Technics Publications, 2010.

[4] P. Woodall, A. K. Parlikad, y L. Lebrun, «Approaches to Information Quality Management: State of the Practice of UK Asset-Intensive Organisations», en *Asset Condition, Information Systems and Decision Models*, Springer, London, 2012, pp. 1-18.

[5] K.-T. Huang, Y. W. Lee, y R. Y. Wang, *Quality Information and Knowledge*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1999.

[6] T. C. Redman, *Data Driven: Profiting from Your Most Important Business Asset*. Boston, Mass: Harvard Business School Press, 2008.

[7] A. Rodríguez, A. Caro, C. Cappiello, y I. Caballero, «A BPMN Extension for Including Data Quality Requirements in Business Process Modeling», en *Business Process Model and Notation*, 2012, pp. 116-125.

[8] T. C. Redman, *Data Quality for the Information Age*. Artech House, 1996.

[9] D. M. Strong, Y. W. Lee, y R. Y. Wang, «10 potholes in the road to information quality», *Computer*, vol. 30, n.º 8, pp. 38-46, ago. 1997.

[10] D. McGilvray, «Ten Steps to Quality Data and Trusted Information™», 29-ago-2016. [En línea]. Disponible en: http://mitiq.mit.edu/IQIS/Documents/CDOIQS_200977/Papers/01_02_T1D.pdf.

[11] C. Cappiello, F. Daniel, M. Matera, y C. Pautasso, «Information Quality in Mashups», *IEEE Internet Comput.*, vol. 14, n.º 4, pp. 14-22, jul. 2010.

[12] J. Barua, *Enterprise architecture led data quality strategy*. 2011.

[13] S. Sarsfield, *The Data Governance Imperative*. Ely: IT Governance Publishing, 2009.

[14] D. Loshin, *Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2013.

[15] K. C. Laudon, «Data Quality and Due Process in Large Interorganizational Record Systems», *Commun ACM*, vol. 29, n.º 1, pp. 4–11, ene. 1986.

[16] M. J. Eppler y M. Helfert, «A Classification and Analysis of Data Quality Costs», en *Proceedings of the 9th MIT Information Quality Conference*, 2004, p. 1.

[17] K. Mehmood, S. Si-Said Cherfi, y I. Comyn-Wattiau, «Data Quality Through Model Quality: A Quality Model for Measuring and Improving the Understandability of Conceptual Models», en *Proceedings of the First International Workshop on Model Driven Service Engineering and Data Quality and Security*, New York, NY, USA, 2009, pp. 29–32.

[18] Á. G. Vieites, «Principales aspectos del reglamento general de protección de datos (GDPR) de la Unión Europea», *Contact Cent. Call Cent. IP Solut.*, n.º 83, pp. 54–57, 2016.

[19] M. Ge y M. Helfert, «Cost and Value Management for Data Quality», en *Handbook of Data Quality*, Springer, Berlin, Heidelberg, 2013, pp. 75-92.

[20] C. Guerra-García, I. Caballero, y M. P. Velthius, «A Survey on How to Manage Specific Data Quality Requirements during Information System Development», en *Evaluation of Novel Approaches to Software Engineering*, 2010, pp. 16-30.
[21] T. C. Redman, «Data Quality Management Past, Present, and Future: Towards a Management System for Data», en *Handbook of Data Quality*, Springer, Berlin, Heidelberg, 2012, pp. 15-40.
[22] H. Hinrichs, *CLIQ - Intelligent Data Quality Management*. 2012.
[23] H. Markus y von M. Eitel, «A Strategy for Managing Data Quality in Data Warehouse Systems», presentado en Sixth International Conference on Information Quality (ICIQ), 2001.
[24] D. Laney, «Gartner Predicts Three Big Data Trends for Business Intelligence.», 2015.
[25] J. M. Juran, *Juran on Leadership For Quality*. Simon and Schuster, 2003.
[26] C. Batini, C. Cappiello, C. Francalanci, y A. Maurino, «Methodologies for Data Quality Assessment and Improvement», *ACM Comput Surv*, vol. 41, n.º 3, p. 16:1–16:52, jul. 2009.
[27] C. Batini y M. Scannapieco, *Data Quality Concepts, Methodologies and Techniques. 2006*. Springer-Verlag.
[28] M. Á. Moraga, M. Piattini, C. Guerra-García, y R. Pérez, «Developing Data Quality Aware Applications», en *2009 9th International Conference on Quality Software (QSIC 2009)*, Los Alamitos, CA, USA, 2009, pp. 458-464.
[29] A. Caro, C. Calero, I. Caballero, y M. Piattini, «A proposal for a set of attributes relevant for Web portal data quality», *Softw. Qual. J.*, vol. 16, n.º 4, pp. 513-542, dic. 2008.
[30] D. Loshin, *The Practitioner's Guide to Data Quality Improvement*. Burlington, MA: Morgan Kaufmann, 2010.
[31] M. Levis, M. Helfert, y M. Brady, «Information Quality Management: Review of an Evolving Research Area», 01-ene-2007.
[32] D. Loshin, *Master Data Management*. Morgan Kaufmann, 2010.
[33] D. O. Brien, «A Website Dedicated to Information/Data Quality Disasters from Around the World», *IQTrainwrecks.info*, 2016. .
[34] C. Cappiello, A. Caro, A. Rodriguez, y I. Caballero, «An Approach To Design Business Processes Addressing Data Quality Issues», *ECIS 2013 Complet. Res.*, jul. 2013.
[35] I. Caballero, I. Bermejo, M. T. G. López, R. M. Gasca, y M. Piattini, «I8K: AN IMPLEMENTATION OF ISO 8000-1X0 », presentado en 17th International Conference on Information Quality (ICIQ), 2013.
[36] «ISO/IEC 25010:2011 - Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- System and software quality models», 2011. [En línea]. Disponible en: https://www.iso.org/standard/35733.html. [Accedido: 07-dic-2017].
[37] «ISO/IEC 25012:2008 - Software engineering -- Software product Quality Requirements and Evaluation (SQuaRE) -- Data quality model», 30-mar-2016. [En línea]. Disponible en: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=35736. [Accedido: 30-mar-2016].
[38] C. Guerra-García, I. Caballero, y M. Piattini, «Capturing data quality requirements for web applications by means of DQ_WebRE», *Inf. Syst. Front.*, vol. 15, n.º 3, pp. 433-445, jul. 2013.
[39] R. Wieringa y A. Moralı, «Technical Action Research as a Validation Method in Information Systems Design Science», en *Design Science Research in Information Systems. Advances in Theory and Practice*, 2012, pp. 220-238.
[40] M. G. Bocco y M. G. P. Velthuis, *Métodos de investigación en ingeniería del software*. Bogotá, Colombia: Ra-Ma S.a. Editorial Y Publicaciones, 2014.
[41] C. Wohlin, P. Runeson, M. Höst, M. Ohlsson, B. Regnell, y A. Wesslén, *Experimentation in Software Engineering*. Springer, 2012.