A SYSTEMATIC LITERATURE REVIEW OF HOW TO INTRODUCE DATA QUALITY REQUIREMENTS INTO A SOFTWARE PRODUCT DEVELOPMENT

César Guerra-García, Ismael Caballero and Mario Piattini
ALARCOS Research Group, Department of Information Technologies and Systems(UCLM)
Paseo de la Universidad 4, Ciudad Real, Spain

Keywords: Data Quality, Requirements Specification, Systematic Review.

Abstract: In recent years many organizations have come to realize the importance of maintaining data with the most appropriate levels of data quality when using their information systems (IS). We therefore consider that it is essential to introduce and implement mechanisms into the organizational IS in order to ensure acceptable quality levels in its data. Only by means of these mechanisms, will users be able to trust in the data they are using for the task in hand. These mechanisms must be developed to satisfy their data quality requirements when using specific functionalities of the IS. From our point of view as software engineering researchers, both these data quality requirements and the remaining software requirements must be dealt with in an appropriate manner. Since the goal of our research is to establish means to develop those software mechanisms aimed at managing data quality in IS, we decided to begin by carrying out a survey on related methodological and technical issues to depict the current state of the field. We decided to use the systematic review technique to achieve this goal. This paper shows the principal results of the survey, along with the conclusions reached.

1 INTRODUCTION

Several authors have reported the problems caused by inadequate levels of data quality (DQ) in the use of IS (Caballero, Caro et al. 2008). These problems make a negative impact on an organization's performance, and additionally involve, among other things, certain types of damages, and an increasingly higher cost in economical terms (Laudon 1986; Wang, Storey et al. 1995; Eppler and Helfert 2004). Once organizations become aware of this situation, they are willing to eradicate these kinds of problems.

As a solution, Karel et al. propose that it is necessary to implement mechanisms by means of specific Data Quality Software (Karel, Moore et al. 2009). The Data Quality Software capabilities include data cleansing, standardization, matching, merging, enrichment and data profiling. However, these are "post-mortem" solutions and, although widely used, they are costly to buy and to implement. In addition, they are not focused on specific users' data quality requirements, which could embrace different data quality dimensions, such as those proposed by Strong et al. in (Strong, Lee et al. 1997) or those that appear in ISO/IEC 25012 (ISO-25012 2008), but solely on what are commonly called intrinsic requirements such as completeness, or accuracy. The latter are necessary but are not sufficient for a broader kind of related data quality problems, in which data quality requirements go beyond these intrinsic data quality dimensions.

We propose that these kinds of problems require preventive action (such as avoiding storing data which are not reliable or believable) rather than corrective actions (such as the use of data cleansing tools). We agree that corrective actions are necessary (Bertino, Dai et al. 2009), but they imply greater costs to organizations, e.g. the time and money necessary to execute cleansing processes on data that could already be used within a business process, in which unproductive costs might be incurred for nothing.

The goal of our research is, therefore, to discover how to introduce the appropriate mechanisms by implementing preventive actions at the point-of-entry of organizational IS software. These

Guerra-García C., Caballero I. and Piattini M. (2010).
A SYSTEMATIC LITERATURE REVIEW OF HOW TO INTRODUCE DATA QUALITY REQUIREMENTS INTO A SOFTWARE PRODUCT DEVELOPMENT.
In *Proceedings of the Fifth International Conference on Evaluation of Novel Approaches to Software Engineering*, pages 12-19

mechanisms might make it possible to attain a trade-off between preventive and "post-mortem" corrective action, thus minimizing investment or unnecessary costs.

We are conscious that preventive mechanisms must exist as part of the IS development, which requires bringing together specific software requirements with specific data quality requirements for the software being developed: it is important to delimit how software can be enhanced in order to satisfy data quality requirements. To achieve this goal, we must first identify existing proposals (both methodological and technological) that have dealt with some kind of solution for the introduction of DQ requirements management as part of IS software development.

To assist in this task, we have decided to use Systematic Review (SR) techniques in order to attain a strict view of the relevant literature. More precisely, we have decided to follow the formal Systematic Review protocol template proposed by Biolchini et al. in (Biolchini, Mian et al. 2007), since it is one of the most widely used techniques in the field of Software Engineering.

An SR focuses on integrating empirical research with the aim of creating generalizations. This integration challenge involves specific objectives which allow the researcher to critically analyze the data found, resolving conflicts in the material of the literature involved, and identifying aspects in order to plan future research areas.

The descriptions for the different phases of Biolchini et al.'s protocol are:

1. *Planning*, which primarily focuses on defining the research objectives, the selection of information sources and the definition of the inclusion and exclusion criteria of studies.
2. The *Execution* phase, which focuses on the selection and evaluation of the studies found, along with extracting information from the selected studies.
3. The *results analysis* phase, which is responsible for analyzing and presenting the results according to different criteria and perspectives that will facilitate their understanding and subsequent use.

This paper shows the results of the execution of an instance of this protocol for dealing with the *specification and modeling of DQ requirements*, and the consequent conclusions reached after analysing these results. We have structured the paper in order to present the main issues of each of the stages of the protocol in each section.

2 PLANNING THE SYSTEMATIC REVIEW

The expected outcome at the end of the SR is that of presentation the state-of-the-art of existing research proposals for the specification and modelling of DQ requirements, published in the resources available. Once the results have been obtained, the main beneficiaries of this work will be those people related to software development, such as: systems analysts, designers, programmers and project managers, in addition to, academics and researchers related to the data quality area, and other relative areas such as quality in Information Systems and Requirements Engineering.

2.1 Question Formularization

The research question which has motivated the SR is: *"are there any works that propose mechanisms (both methodological and technological) for the specification, representation and incorporation of data quality requirements during the process of developing a software product?"*

In order to seek a response to our research question, we elaborated a list with keywords, which could be used when querying the different search engines of the bibliographic resources in hand. These keywords are shown in Table 1.

Table 1: Keywords.

Data Quality Dimension	Data Quality Metadata
Data Quality Requirement	Data Quality Framework
Data Quality Metamodel	Data Quality Methodology
Data Quality Modeling	Data Quality Dimensions

2.2 Resource Selection

In this section, we show the set of criteria for the selection of resources, the search methods and the specification of the search strings (aforementioned keywords) used in the SR protocol. According to the recommendations of experts in the SR field, the searches should be conducted by using the search engines in the electronic databases of leading publishers. We have also added several specialized resources to the set of these databases: on the one hand, the conference proceedings of the "*International Conference on Information Quality*" (ICIQ, http://mitiq.mit.edu/) since it is the most important international event in the DQ area; and on the other hand the contents of the only two journals to deal with this area: *International Journal of Information Quality "IJIQ"* and *Journal of Data*

and Information Quality "JDIQ" (note that the search in both journals was carried out manually). All of these resources contain works of great importance and most of them offer search engines. The complete list of resources is presented in Table 2.

Table 2: List of resources.

ACM Digital Library
IEEE Computer Society
Information Quality at MIT (ICIQ)
Science Direct
Wiley InterScience
IJIQ
JDIQ

Upon considering the list of keywords mentioned above (see Table 1), and making a combination through the logical connectors "AND" and "OR" we decided to coin the following search string: *("Data Quality") AND ("requirements" OR "dimensions" OR "Metamodel" OR "Modeling" OR "Model" OR "metadata" OR "framework" OR "Methodology" OR "Modelling" OR "MDA" OR "representation" OR "accuracy" OR "completeness" OR "consistency" OR "credibility" OR "currentness" OR "accessibility" OR "compliance" OR "confidentiality" OR "efficiency" OR "precision" OR "traceability" OR "understandability" OR "availability" OR "portability" OR "recoverability")*. It is worth noting that it was decided incorporate in a particular way, each of the different DQ dimensions defined in ISO/IEC 25012, with the aim of broadening the range of search. Please note that the syntax of the string may differ according to the specific requirements of the different search engines of the available resources.

2.3 Studies Selection

Once the resources in which we intended to carry out the searches had been selected, we then defined the procedure for selecting the studies, which also included criteria for the inclusion and exclusion of the studies (works) found during the SR.

The procedure used to select studies was basically as follows: initially a researcher read only the title and the abstract of the set of papers found in each of the searches in order to select the most *relevant studies* from each set. After an initial coarse-grained filtering, the researcher analyzed the complete article, deciding which works he judged to be unsuitable since they did not make a particularly great contribution to the DQ requirement field. A list was then made of those studies that were considered to be very important (typically named *primary studies*). Once this list was considered to be complete, other researchers with a higher expertise in the field were encouraged to conduct an investigation of this list in order to verify that the studies really provided important knowledge with regard to that area.

The procedure for the selection of primary studies consists of an iterative and incremental process. It is said to be *iterative* because some of the main activities such as searching, reading and information extraction are carried out for each of the selected resources. In addition, if a search does not produce a minimal set of results, it is possible that the search string must be refined to obtain more accurate results. It is also said to be *incremental*, because when we perform the searches and extract information a set of potential studies grows from scratch, thus leading to the completion of the systematic review by obtaining a complete document. In our case, the first author of this paper acted as a researcher, and the remaining authors were those who had greater expertise in the area of DQ. The inclusion and exclusion criteria defined for this work are listed below.

Inclusion criteria:

- The articles should describe proposals or strategies for the specification and/or modelling of data quality requirements as a software specification.
- The articles must be written in English.
- There will be an analysis of the title, keywords and summary of each of the studies found.
- There is no restriction with regard to the date of publication.

Exclusion criteria:

- Articles that do not propose any methodology, strategy or model (or metamodel) with which to specify data quality requirements.

As a result of the application of these criteria, we were able to decide which studies found by the searchers could be considered as *primary studies*.

3 EXECUTION OF THE SELECTION

After executing the search procedure on the different resources, a total of 820 studies were found (once eliminated the duplicated studies). By applying the inclusion and exclusion criteria, 42 were considered to be important, while only 8 were eventually considered as *primary studies* by the experimental

researchers. Table 3 summarizes a report of our findings.

Table 3: Distribution of studies by resource.

	Studies			
Resources	*Search Date*	*Found*	*Relevant*	*Primary*
ACM Digital Library	Oct '09	164	6	4
IEEE Computer Society	Oct '09	169	9	1
ICIQ	Nov'09	44	7	2
JDIQ	Nov'09	12	0	0
IJIQ	Nov'09	34	0	0
Science Direct	Dec '09	100	16	1
Wiley InterScience	Dec '09	297	4	0
	Total	**820**	**42**	**8**

Of all the studies reviewed, only the following were considered as primary studies:

1. *Toward quality data: An attribute-based approach* (Wang, Reddy et al. 1995).
2. *Data Quality Requirements Analysis and Modeling* (Wang and Madnick 1993).
3. *A flexible and generic data quality metamodel* (Becker, McMullen et al. 2007).
4. *IP-UML: Towards a Methodology for Quality Improvement Based on the IP-MAP Framework* (Scannapieco, Pernici et al. 2002).
5. *A Product Perspective on Total Data Quality Management* (Wang 1998).
6. *DQRDFS: Towards a Semantic Web Enhanced with Data Quality* (Caballero, Verbo et al. 2008).
7. *Quality Views: Capturing and Exploiting the User Perspective on Data Quality* (Missier, Embury et al. 2006).
8. *A Data Quality Metamodel Extension to CWM (Gomes, Farinha et al. 2007)*.

Once the primary studies had been identified, the next step was to extract the relevant information for the systematic review from each one of them. A form was designed to better guide this process of extracting relevant information. All the information from the studies is displayed in Tables 5 to 12, located in the Appendix.

4 ANALYSIS OF OBTAINED RESULTS

Once the information had been extracted from all the primary studies, the main aim was to address the usability of the identified primary studies in accordance with our interest in discovering proposals dealing with methodological and technological issues. In this section, we show the results of the corresponding analysis. It is worth highlighting that the number of proposals is significantly poor in comparison to the degree of interest that data quality and information quality field has motivated in recent years. Our concern about this led us to ask various DQ researchers and practitioners from different countries and organizations why we had not found more works. Most of them agreed that since data quality is dealt with as a specific issue rather than an organizational issue, many organizations are not yet aware of the possible benefits of our research topic; moreover, most of them also agreed that the topic is quite relevant because the results could help organizations to improve their performance at a relative low cost.

Table 4 summarizes the information extracted from each proposal: the technology or data model used, the existence of a tool or prototype supporting it, the inclusion of an example or study case, and reports concerning whether the proposed results have already been tested in a real environment.

Table 4: Technology or Model used by the selected studies.

Studies	*Model*	*Tool*	*Example*	*Test*
(Wang and Madnick 1993)	Relational		Yes	
(Wang, Reddy et al. 1995)	Relational		Yes	
(Wang 1998)	Relational	Yes	Yes	
(Scannapieco, Pernici et al. 2002)	Object Oriented	Yes	Yes	
(Missier, Embury et al. 2006)	XML		Yes	
(Becker, McMullen et al. 2007)	Relational		Yes	
(Gomes, Farinha et al. 2007)	Object Oriented			Yes
(Caballero, Verbo et al. 2008)	XML		Yes	

Upon studying the analysis in greater depth, we noted that none of the existing works provide a methodology for obtaining and managing DQ requirements. We had hoped to find a methodology that could, at some point, lead analysts and developers to implement a correct management of data quality requirements from the earliest stages, and throughout the process of an information system's development. This lack consequently

motivates the challenging research goal of depicting a methodology for managing and combining data quality software requirements together with those that remain. On the other hand, and with regard to the technology used, we concluded that since many different kinds of applications could be developed by using different kinds of technologies, some sort of generalization should be used, in order to make different kinds of developments possible. This generalization can be achieved by working with models and metamodels. Therefore, our most important conclusion in relation to this issue is that we should work upon the foundations of Model Driven Engineering, MDE (Bézivin 2004) and Model Driven Architecture, MDA (OMG 2003), in order to make the development of different kinds of development possible by using the same concepts concerning data quality requirements.

5 CONCLUSIONS

Conducting an SR is a highly intensive task in comparison to that of a conventional literature search. However, if the complete protocol of an SR is followed step by step, then a better validation of the results is generated, and the efforts are worthwhile. The main goal of this paper is to show the results obtained and conclusions reached after carrying out an SR to discover how well the management of data quality requirements (at both the methodological and technological levels) is dealt with in specialized literature. After analyzing the obtained results, it is evident that there is a need for new proposals dealing with methodological issues, owing to the scarcity of existing initiatives aimed at this particular area. Furthermore, technological issues must be also dealt with. To do this, we can conclude that MDA foundations might be the best environment in which to carry out research into this area. Due to the benefits it provides, mainly in the generation of diverse models and transformations between different abstraction levels. We can consider the incorporation of elements for management of DQ requirements from the early stage, and propagate them throughout all the development cycle of any kind of software.

ACKNOWLEDGEMENTS

This research is part of the PEGASO-MAGO (TIN2009-13718-C02-01), and DQNet (TIN2008-04951-E/TIN) projects, both of which are supported by the Spanish Ministerio de Educación y Ciencia, ENGLOBAS (PII2I09-0147-8235), and TALES (HITO-2009-14), both supported by the Consejería de Educación y Ciencia of Junta de Comunidades de Castilla-La Mancha.

REFERENCES

Ballou, D. P., R. Y. Wang, et al. (1998). "Modelling Information Manufacturing Systems to Determine Information Product Quality." *Management Science* 44(4): 462-484.

Becker, D., W. McMullen, et al. (2007). A Flexible and Generic Data Quality Metamodel. *International Conference on Information Quality*.

Bernes-Lee, T., J. Hendler, et al. (2001). "The Semantic Web." Scientific American.

Bertino, E., C. Dai, et al. (2009). The Challenge of Assuring Data Trustworthiness. Database Systems for Advanced Applications. *Springer-Verlag. Volume 5463/2009*: 22-33.

Bézivin, J. (2004). "In Search of a Basic Principle for Model Driven Engineering." UPGRADE, Novática. Vol. 2(No.2): 21-24.

Biolchini, J. C. d. A., P. G. Mian, et al. (2007). "Scientific research ontology to support systematic review in software engineering." *Adv. Eng. Inform. 21(2)*: 133-151.

Caballero, I., A. Caro, et al. (2008). "IQM3: Information Quality Maturity Model." *Journal of Universal Computer Science* 14: 1-29.

Caballero, I., E. M. Verbo, et al. (2008). DQRDFS:Towards a Semantic Web Enhanced with Data Quality. *Web Information Systems and Technologies*, Funchal, Madeira, Portugal.

Eppler, M. and M. Helfert (2004). A Classification and Analysis of Data Quality Costs. International Conference on Information Quality, MIT, Cambridge, MA, USA.

Gomes, P., J. Farinha, et al. (2007). A data quality metamodel extension to CWM *Proceedings of the fourth Asia-Pacific conference on Comceptual modelling* - Volume 67 Ballarat, Australia Australian Computer Society, Inc.: 17-26

ISO-25012 (2008). "ISO/IEC 25012: Software Engineering-Software product Quality Requirements and Evaluation (SQuaRE)-Data Quality Model."

Karel, R., C. Moore, et al. (2009). "Forrester's report for Business Process and Application Professionals on Trends 2009: Master Data Management." *Forrester*.

Laudon, K. C. (1986). "Data Quality and Due Process in Large Interorganizational Record System." *Communications of the ACM 29(1)*: 4-11.

Missier, P., S. Embury, et al. (2006). "Quality views: capturing and exploiting the user perspective on data quality." *Proceedings of the 32nd international conference on Very large data bases*-Volume 32.

OMG. (2003). "Common Warehouse Metamodel (CWM) Specification v1.1." October, 2008, from http://www.omg.org/docs/formal/03-03-02.pdf [Consultado el: 29-09-2008].

OMG (2003). MDA Guide Version 1.0.1., Object Management Group: 62.

Scannapieco, M., B. Pernici, et al. (2002). IP-UML: Towards a Methodology for Quality Improvement Based on the IP-MAP Framework. *International Conference on Information Quality*, ICIQ-02.

Shankaranarayan, G., R. Y. Wang, et al. (2000). IP-MAP: Representing the Manufacture of an Information Product. *Fifth International Conference on Information Quality* (ICIQ'2000), MIT, Cambridge, MA, USA.

Strong, D. M., Y. W. Lee, et al. (1997). "Data Quality in Context." *Communications of the ACM 40(5)*: 103-110.

Wang, R., V. Storey, et al. (1995). "A Framework for Analysis of Data Quality Research." IEEE Transactions on Knowledge and Data Engineering 7(4).

Wang, R. Y. (1998). "A Product Perspective on Total Data Quality Management." *Communications of the ACM 41(2)*: 58-65.

Wang, R. Y. and S. Madnick (1993). Data Quality Requirements: Analysis and Modelling. *Ninth International Conference on Data Engineering* (ICDE'93), Vienna, Austria, IEEE Computer Society.

Wang, R. Y., M. Reddy, et al. (1995). "Towards quality data: An attribute-based approach." *Journal of Decision Support Systems 13(3-4)*: 349-372.

APPENDIX

Table 5: Primary Study (Wang and Madnick 1993).

Data Extraction of the Study	
Publication	Richard Wang, Henry Kon, and Stuart Madnick. April, 1993. *Data Quality Requirements Analysis and Modeling*. In: Proceedings of the Ninth International Conference of Data Engineering. Austria.
Objective Results of the Study	
Proposal	The article is focused on: (1) establishing a set of premises, terms and definitions for the management of DQ, and (2) developing a step by step methodology for defining and documenting DQ parameters for users. The requirements analysis methodology proposed by the authors is based on two main approaches: - Specification of *labels* needed for users with the objective of assessing, determining or improving data quality. - Obtaining, from the user viewpoint, the general aspects of DQ non-sensitive to labeling, for example, the features of completeness and response time. A series of views (view of application, view of parameters and quality view) is also proposed which should be included as part of the documentation of quality requirements specification, the authors jointly refer to a list of possible data quality candidates.
Results	Methodology for collecting and documenting data quality requirements.
Model	It uses a "Relational" type of model.
Mentioned Difficulties	There is no definition and standardization of quality dimensions.

Table 6: Primary Study (Becker, McMullen et al. 2007).

Data Extraction of the Study	
Publication	David Becker, William McMullen y Kevin Hetherington-Young. November, 2007. *A flexible and generic data quality metamodel*. In: Proceedings of the 12th. International Conference on Information Quality, ICIQ 2007. U.S.A.
Objective Results of the Study	
Proposal	Analyze and describe three generic metamodels mentioning some of their most important capabilities: Common Warehouse Metamodel (CWM), Data Warehouse Quality (DWQ) and Universal Meta Data Model. The authors propose an architecture and a basic metamodel for DQ, which meets the objectives of adequately providing a flexibility, generality and ease of use of the requirements in situations of potential use. This metamodel adequately represents the information products (IP), data objects, and metrics, measurements, requirements, evaluations and actions of DQ.
Results	Proposal for architecture and a generic metamodel for DQ.
Model	It uses a "Relational" type of model.
Mentioned Difficulties	No mention of any.

Table 7: Primary Study (Wang, Reddy et al. 1995).

Data Extraction of the Study	
Publication	Wang, Richard Y., Reddy, M., Kon, H.. March, 1995. *Toward quality data: An attribute-based approach*. In: Journal of Decision Support Systems. U.S.A.
Objective Results of the Study	
Proposal	The authors propose a quality perspective using labeled data in a cell level with quality indicators, which are objective characteristics of the data and its manufacturing process. Based on these indicators, the user can evaluate the quality of data for a specific application. Additionally, the authors investigate how these quality indicators can be specified, stored, retrieved and processed. They propose a data model based on attributes, query algebra and integrity rules that facilitate cell-level tagging, along with data processing of the application that is augmented with quality indicators.
Results	A methodology for analyzing of data quality requirements based on an entity-relationship model, for the specification of the types of quality indicators to be modelled.
Model	It uses a "Relational" type of model.
Mentioned Difficulties	Study and research object-oriented approach, because the relational model that represents the schema of quality, may be restrictive. An object-oriented approach seems simpler to model the data and its quality indicators, because many of the quality control mechanisms are oriented towards procedures and this approach could manage them without any problem.

Table 8: Primary Study (Scannapieco, Pernici et al. 2002)

Data Extraction of the Study	
Publication	Monica Scannapieco, Barbara Pernici y Elizabeth Pierce. November, 2002. *IP-UML: Towards a Methodology for Quality Improvement Based on the IP-MAP Framework*. In: Proceedings of the Seventh International Conference on Information Quality, ICIQ´02. U.S.A.
Objective Results of the Study	
Proposal	It proposes a UML profile for data quality in order to sustain the quality improvement within an organization. This profile is based on the IP-MAP Framework (Shankaranarayan, Wang et al. 2000), but differs from it, mainly owing to: (1) it specifies the artefacts for production during the improvement process in terms of diagrams drawn using UML elements defined in the data quality profile, (2) it uses the IP-MAP not only to evaluate the quality and think about the improving actions, but also as a schematic means to design and implement improving actions. The IP-MAP is an extension of a Information Manufacturing System (IMS) proposed by (Ballou, Wang et al. 1998), this Framework has the advantage of combining both data analysis and process analysis, with the aim of assessing the quality of the data. The data quality profile consists of three different models: Data Analysis Model, Quality Analysis Model and Quality Design Model. The data analysis model specifies which data are important to consumers because their quality is critical to organizations' success. The quality analysis model consists of modelling the elements that permit the representation of the data quality requirements, a quality requirement can be related to a dimension of quality or features that are typically defined for data quality. The quality design model incorporates the perspective of IP-MAP, which helps in understanding the details associated with the manufacturing process of the information products.
Results	Shows a profile and a methodology for producing UML artifacts designed by the data quality profile.
Model	It uses an "Object Oriented" type of model.
Mentioned Difficulties	None.

Table 9: Primary Study (Wang 1998).

Data Extraction of the Study	
Publication	Richard Wang. February, 1998. *A Product Perspective on Total Data Quality Management*. In: Communications of the ACM. U.S.A.
Objective Results of the Study	
Proposal	This article presents the Total Data Quality Management (TDQM) methodology, whose main purpose is to deliver High quality information products (IP) to information consumers, along with introducing the concepts of TDQM cycle and information products. It explains the stages of the TDQM related to the information products: Definition, Measurement, Analysis and Improvement, with particular emphasis on defining the characteristics of information products and quality requirements of the information. The author also shows a software tool with which to conduct surveys to assess the quality of information, where it may be possible to evaluate a list of quality dimensions defined by the author.
Results	It shows the TDQM methodology and illustrates how it can be put into practice in a wide range of organizations.
Model	It uses a "Relational" type of model.
Mentioned Difficulties	None.

Table 10: Primary Study (Caballero, Verbo et al. 2008).

Data Extraction of the Study	
Publication	Ismael Caballero, Eugenio Verbo, Coral Calero y Mario Piattini . May, 2008. *DQRDFS: Towards a Semantic Web Enhanced with Data Quality*. In: 4th. International Conference on Web Information Systems and Technologies, WEBIST ´08. Portugal.
Objective Results of the Study	
Proposal	This article introduces a new view of the Semantic Web, based on the concept of quantity of data quality (QDQ), where DQ aspects are used as a base to enable machines to process the documents of the Semantic Web for different activities such as information retrieval or filtering of documents. The Semantic Web is an extension of the current Web in which the information is provided with a well-defined meaning, enabling computers and people to cooperate (Bernes-Lee, Hendler et al. 2001). This article has a twofold goal: (1) it shows the readers a brief introduction to DQ, and (2) showing how the DQ fundamentals have been applied with the aim of highlighting the quality of Web documents for the Semantic Web. The first step in order to enable the DQ in the semantic web is to identify the set of elements that need to be studied from the User Requirements Specification for the DQ (DQ-URS). The second step is to identify the DQ dimensions and their related metadata. The third step is to obtain and record the values for the metadata. This information is represented by using XML-type documents.
Results	It shows a proposal of the concept of QDQ oriented towards the Semantic Web.
Model	It uses the XML language (Extensible Markup Language) for its representation.
Mentioned Difficulties	None.

Table 11: Primary Study (Missier, Embury et al. 2006).

Data Extraction of the Study	
Publication	Paolo Missier, Suzanne Embury, Mark Greenwood, Alun Preece y Binling Jin. September, 2006. *Quality Views: Capturing and Exploiting the User Perspective on Data Quality*. In: International Conference on Very large databases, VLDB ´06. Korea.
Objective Results of the Study	
Proposal	This article presents a quality user-centred model and a software environment, which domain experts can use to easily and rapidly code and test their own heuristics quality criteria. As core of the model, they propose the concept of "quality view", similar to customized "*lenses"*, through which the data can be observed. The main contributions of this work are: (1) An extensible semantic model to the user for concepts of quality information in e-science. (2) A process model and a declarative language for specifying abstract views of quality in terms of a few logical operators. (3) An architecture for implementing quality views within many data processing environments.
Results	It proposes a framework for specifying requirements for quality processing by the user, called "quality views".
Model	It uses the XML language (Extensible Markup Language) for its representation.
Mentioned Difficulties	None.

Table 12: Primary Study (Gomes, Farinha et al. 2007).

Data Extraction of the Study	
Publication	Pedro Gomes, José Farinha, Maria José Trigueiros. February, 2007. *A Data Quality Metamodel Extension to CWM*. In: 4th. Asia-Pacific Conference on Conceptual Modelling, APCCM 2007. Australia.
Objective Results of the Study	
Proposal	This paper proposes a metamodel for data quality and data cleaning, both concepts being applicable to the context of data warehouses. This metamodel is integrated with the "Common Warehouse Metamodel" (OMG 2003), providing an extension of this standard towards data quality. It also provides a set of modelling guidelines for the storage of formal specifications of DQ rules. The main purpose of the metamodel is to provide support to profiling and data cleaning activities, with rules that can be established with the aim of detecting data quality problems. It also establishes data cleaning solutions. In relation to data cleaning, a "metadata" holder is provided, with the objective of enabling the ultimate goal of achieving the highest level of automation possible. However, a metadata support is also provided when the user's participation is required in the cleaning process.
Results	It displays a metamodel for quality and data cleaning, both concepts being applied to the context of data warehouses.
Model	It uses an "Object Oriented" type of model.
Mentioned Difficulties	None.