EUROPEAN ACADEMIC RESEARCH
Vol. II, Issue 4/ July 2014

Impact Factor: 3.1 (UIF)
DRJI Value: 5.9 (B+)

ISSN 2286-4822
www.euacademic.org

The Data Quality Management

MUSTAPHA BENMAHANE
Professor
Polydisciplinary Faculty
Chouaïb Doukkali University
Morocco

Abstract:

In a world of increased competitiveness, one of the major challenges that organizations should face is the data quality management. The speed with which data can be cleaned, transformed and integrated into a data warehouse becomes essential for the enterprises competitiveness. The data quality depends firstly on the context in which it was originally typed, but from a decision-making point of view; its interest depends on the use made by a user. This paper has for vocation, according to a theoretical (conceptual) and practical study, to show the importance of the data quality in the operational and strategic decision-making. In this regard, we will start by defining some important concepts that are sometimes combined: data, information and knowledge. Once the data is defined, we'll explain what makes their quality by identifying their characteristics and their dimensions while specifying the costs which are generated by the data quality (costs of the non-quality and quality improvement). Finally, to improve the data quality, we will try to show how it is becoming essential for a company to take into consideration governance and a sustainable program to analyze the level of data quality and make this company more efficient and competitive.

Key words: data, meta-data, Data Quality, Management, information, Performance

1. Introduction

The Data Quality Management (DQM) is a growing concern of companies. It is originally derived from a compliance need of the results of financial institutions; the Data Quality Management extends to all operational functions of companies in different economic sectors. Those latter have become aware that their effectiveness and competitiveness depend widely on the value and the quality of the data and information that they are managing.

The strategic management of a company requires applying information of quality. Thus, the availability of qualified data constitutes not only a major challenge for companies, but also, a competitive advantage in markets where customers have become demanding in an increasing way.

The effective management of basic data (Master Data Management / MDM) constitutes nowadays one of the major challenges for companies. This basic data management consists of organizing all business data, in a way in which theses ones are available – depending on current criteria and required quality - for all operational and analytical applications. The goal is getting a "single version of the truth".

What is the data quality? Why is it important? How to manage the data quality? How to control the quality / identify problems? How to improve the quality? These are some questions that this contribution tries to answer.

In this paper, we distinguish between three concepts that are often confused: data, information and knowledge. We will try to introduce latter common definitions of what we call data quality, clarify the main dimensions and characteristics of the data quality without forgetting to identify the data quality cost (non-quality costs and the quality improvement costs).

This contribution will be ended by an answer of a major issue: how does it become essential for companies concerned with the data quality, to implement a program and governance

in order to analyze and control the data quality level to improve it, thus, contribute to a good decision making.

2. Definition of concepts: data, information, and knowledge

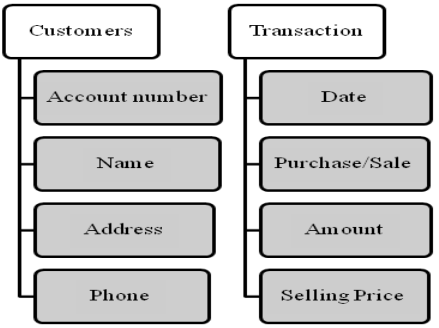
Let's start by defining the concepts of data, information and knowledge.

- Data is a basic description, often encoded, of one thing; it is a description of a business transaction, an event, etc. Data can be stored and classified under various forms: paper, numerical, alphabetical, images, sounds, etc. According to Donald Hawkins, counselor of technology information: "data are statistics facts that can be quantified, measured, counted and stored" (Phillippe Trouchaud, Zouheir Goh, Reda Gomery, 2011).
- The information represents the transformed data in meaningful form for the person who is receiving them: it has a value for its decisions and actions. It is used to identify a data that has been the subject of an interpretation. This information will take its meaning in the context in which it is intended (Martin Goulet, 2012, p.5).
- The treatment of data and information provides the ability to generate knowledge: a way of understanding or learning a problem or an activity. It should be noted that in a context of managing knowledge, data represent the information source, while those latter represents the knowledge source. That is to say, the quality of anything that uses the data source can be directly linked to the quality of these ones. Indeed, if data are in a bad quality, the information will be of poor quality, as a result, we will have a bad knowledge quality in the end.

The general idea is to manage the data as an asset of the company in the same way as its products, its employees, its

customers. It is needed thus, to understand the needs of the clients (the users in this case), create families of data, which means all of the associated data and complete life cycles. We should appoint a data steward that has a similar role to a manager of a product.

Figure 1: Example of data families



3. What is the data quality?

The data quality is a generic term which describes both the characteristics (dimensions) of data (complete, reliable, relevant, up-to-date, coherent...) and all the processes that help in ensuring its dimensions. The purpose is to obtain data without duplication, without spelling mistakes, without omission, without superfluous variation and complied with the defined structure.

Data are said to have a quality if they meet the requirements of their uses. In other words, the data quality depends not only on their use but also on their State. To satisfy the intended use, the data must be accurate, timely and relevant, complete, understandable, and trustworthy.

Data quality is often defined as "employability" (fitness for use), that is to say, the capacity to collect data to meet the user's needs.

So "the data quality in computer science refers to the data compliance in an intended use, in operating procedures,

processes, decision-making, and planning" (J.M. Juran). A data that has a good quality must be valid, accurate, and comprehensive.

In industry, the product quality is properly measured by comparing its effects to the expectations of the customers. A product or a service that meet the customer's needs is therefore of a high quality. It is the same thing for data: data has a quality if it perfectly meets the needs of its users.

We can conclude from these definitions, that the data quality is the degree of the data adequacy to the use that we make.

4. Main features of data quality

If accuracy is considered as one of the necessary conditions for the definition of a data quality, other characteristics or dimensions must also be taken into account. As it is mentioned in the following table, data can be accessed on the basis of their content, their accessibility, their flexibility and their security.

Table 1: Main features of data quality

MAIN CHARACTERISTICS OF DATA QUALITY	
Content quality	- Accuracy - Adequacy - Understanding
Accessibility	- Availability - Ease of access
Flexibility	- Scalability - Consistency with other sources - Translation
Security	- Privacy - Reliability - Traceability - Integrity

Source : Christophe Brasseur (2008, p.38)

It should be noted that if the operational data is of poor quality, those of the pilotage system have all chances to be the same, since the data of these ones are generally derived from the operating system. As a result: a hazardous operational and a wrong piloting of the company.

5. Indicators and measures of a quality

From these theoretical definitions, organizations must create their own operational definitions based on the objectives and priorities of the company, in order to define indicators for each of the dimensions and check by regular measures their evolution in time.

Each dimension may be measured either in a subjective way by collecting the user's perception, or in an objective way through the automatic monitoring of specific indicators.

The following table gives examples of quality indicators according to different criteria.

Table 2: Indicators of the quality of the data

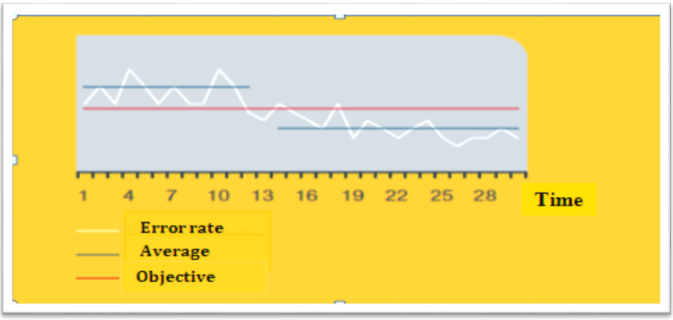
Criteria for data quality	**Features**	**Examples of indicators**
Opportunity	Is the age of data consistent with the needs of the business?	- Date of the collection of data - The last treatment date - Version control
Full/ Completeness	Are all the necessary data available?	- Entirety of the optional value - Number of non-informed values - Number of default values in relative with the average
Consistency	What are the data sources of the conflicting information?	- Plausibility check - Value of the standard deviation
Accuracy	Do the values reflect reality?	- Frequency of changes in value - Reaction (feedback) of clients
Interpretability	Are the data understandable by the users?	- Recovery of user data - Violation of areas
Standardization, conformity	What are the data entered, stored or displayed in a non-standard format?	- Certificate of conformity
Duplication	What are the repeated data?	- Number of duplicated records

Source: INFORMATICA (2008, p. 10)

Once the indicators are defined, it is a must to implement a measurement system which allows monitoring their evolution in time. The quality indicators publication, their target and their evolution allow in defining the potential plans of action that will be implemented to correct a situation.

Another important point, the company needs to develop the means of measurement to assess the data quality before and after the program implementation. For example, the measurement of errors rate within a business process can be one of these indicators.

Figure 2. Measurement of errors rate before and after implementation of the solution:



Source: Christophe Brasseur (2008, p. 37)

In addition to the launch of punctual improvement programs, it seems so profitable to set up good practices of data management in a sustainable way. Based on clear principles that focus on preventive measures and pragmatism, the data management can undoubtedly improve the data quality and make the company more efficient and competitive. Good practices cover the organization, management, consolidation, documentation, monitoring, and audit ability of data.

6. Major consequences of the data non-quality.

In order for a decision to be timely at the level of a company, it must rely on corrected data. Each actor, starting from the

general branch to the point of sale, needs to dispose reliable and accurate information to make appropriate decisions.

However, it should be recognized that bad data inevitably lead to bad decisions, inadequate strategies as well as poor customer service.

The consequences of the data poor quality are numerous and return as expensive to businesses. The impact may spread to the outside, because of the growing place of electronic exchanges. Finally, the image deterioration and the credibility loss are sometimes fatal to those organizations whose data are of poor quality. The most frequent consequences are as follows:

- Customer dissatisfaction.
- Non-compliance of the published figures.
- Devaluation of the company's image.
- Disturbance of the operational functioning.
- Strategy errors.
- Increase in costs.

According to The Data Warehousing Institute, a company of specialized studies, the poor quality of data would cost for American businesses more than 600 billion dollars annually (IT-expert n ° 73 - July/August 2008).

According to a survey (DataFlux, 2011) concerning practices of French companies in the data quality, 90% of respondents acknowledge that the data quality is a critical element in the decision-making, only a third of the companies admitted having confidence in its data.

This study has shown that the obstacles, related directly to the data quality, can also generate:

- An increase of costs (fear evoked by 61% of respondents),
- A decrease in customer satisfaction (59%),
- A decrease of the employee satisfaction (46%)
- A negative impact on the Corporation revenues (42%).

These results confirm that the data quality level of the business has a direct influence on the effectiveness of organizations.

It should be noted that the under-estimation of the data challenge lies in the fact that the company objective seems not to be related to the data production. While the exercise of different professions requires information, the first aim consist in obtaining, depend on the case, a satisfied customer, a higher level of sales, a better product, a better margin, etc. Therefore, data is not set in the foreground of the activity, and are not generally seen as an essential element of competitiveness. Moreover, a large number of managers consider that the investment in a political quality of data is translated only into loss of time and money.

Data recovery of an old application to a new one often generates the quality lack, because policymakers do not always give to it the importance it deserves.

7. The cost of non-quality data

The data quality constitutes a critical challenge for the company at a three stages of their life cycle: when entering, during the transformations and aggregations and during the analysis and presentation of results.

7.1 Capture of the data quality

Raw data feed information systems of businesses. However, the management of their quality is not a subject to rules and standards. This leads to make decisions based on incorrect or misinterpreted data.

Poor quality of data is mainly due to errors in entering the information to the source. Misspellings, wrong codes, incorrect abbreviations, entering in a wrong field are all sources of quality degradation that can have negative consequences for the company.

In addition, accurate data at any given time may become inaccurate as a result of a situation change. For example, client relocation can cause the creation of a new identifier instead of a change of the original plug.

7.2 Operation and analysis of the data quality

In its approach of flexibility, the company is looking for operational efficiencies. The operation of the data quality enables the optimizing of the participation and interactions between all collaborators beyond administrative or technical boundaries. Unfortunately, many companies overlook checking the quality of their data, things which lead them to operate false or erroneous data. This causes many impacts on the management and the performance of the company.

It is always possible to encode the direct cost of the data non-quality. The following table shows the calculation of an investment return of a companion marketing of a telephone operator.

Table 3: Value of quality in a marketing companion

Assumptions		
- Number of delivered brochures	50 000	
- Total cost of the program	600 000 MAD	
- Average profit per sale	2 000 MAD	
- Duplication ratio	**1**	**1.04**
- Response rate	2%	1.92%
- Ratio of home	**1**	**1.11**
- Conversion rate	20%	18.02%
Results		
- Number of response	2 000	2 000/1,04 = 1 923
- Cost per response	300 MAD	300 x 1,04 = 312 MAD
- Number of buyers	2 000 x 20% = 400	1 923 x 18, 02% = 347
- Total profit of the campaign	800 000 MAD	693 077 MAD
- Return on investment	**33.33%**	**15.50%**

Inspired from : INFORMATICA (2008, p. 4)

The company wants to send a brochure to announce a new service to all its customers. The base of its customer contains duplicated records (1.04 duplication ratios) as well as multiple records for the same household (household ratio 1.11).

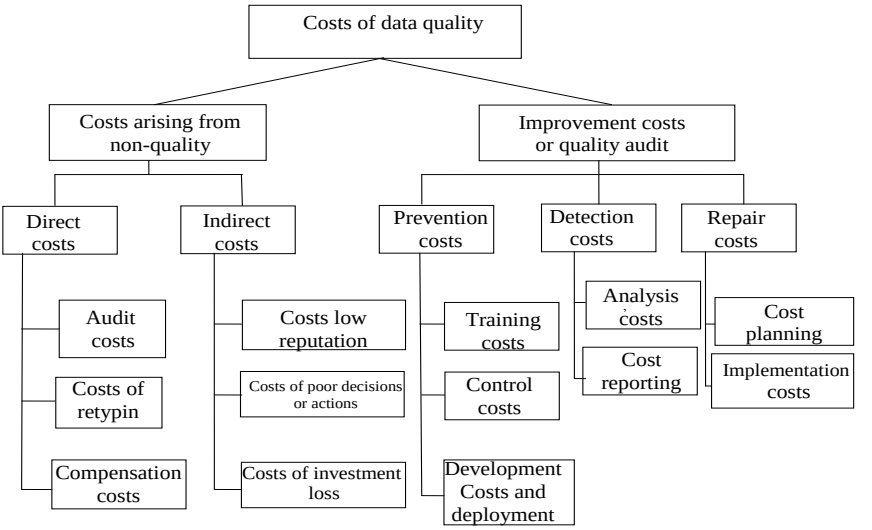
The table above shows that if this business is based on the data quality, the number of the response increases by 4%, the cost-per-response declines of 3.85%, the number of buyers

increases of 15.27% and the full benefit of the companion increases of 15.43%.

Based on the data quality, the return on investment moves from 15.50% (693 077- 600 000) / 600 000) to 33.33% (800 000 - 600,000 / 600 000). Therefore, this simple calculation shows that a marketing operation at the data quality may double its effectiveness.

The data quality generates more costs, the figure below represents a classification tree that gives a synthetic overview of the data quality costs with a distinction at the non-quality costs and costs that maintain and prevent the data quality (Delphine Clements, Brigitte Ladi, Domunique Duquennoy, Andrea Michaud (2008)).

Figure 3: The cost of the quality of the data



Source : Delphine Clément, Brigitte Laboisse, Domunique Duquennoy, Andrea Micheaux (2008, p. 23)

Finally, during the analysis and presentation of the data, the company must ensure its quality. Taking into account the fact that the analytical and decision-making systems of businesses (dashboards, business intelligence tools...) are based on their

transactional systems, the reliability of key performance indicators and data mining predictions entirely depends on the data validity on which they are based. In parallel to the growing importance of valid data in the decision-making in a business, the validation process of these data is getting more and more complex. Data flow constantly in the business, from systems and varied sources in addition to a large number of users.

We can give as an example the case of a large insurance company which had decided to merge its customer databases to have a better understanding of its clients and of the products they bought, in order to improve its services. Before the project, the management thought to have 13 million customers, an estimate which is based on the available information. During the project, teams have discovered many records duplicated in the bases and it had to be reduced of 5 million the number of customers of the company at the end of the project. A risk in the process analysis can therefore be rendered inoperative because of unreliable data.

8. Data Quality Management (DQM) and Master Data Management (MDM): the quality tools

To support data quality management, DQM tools are now available. These tools aim to improve the data quality.

A majority of solutions allow in detecting anomalies and correcting them afterward. It is therefore in the overall a healing mode, although some products work preventively by detecting anomalies in the source in the real time.

In practice, these tools are very effective. They enable in a special way the improvement of the standardization of some data such as (addresses, the detection and elimination of duplicates), and the generation of statistics and reports that facilitate the detection of anomalies.

In addition to these highly specialized tools, we have been assisting from years to a craze for MDM (Master Data

Management) solutions. MDM allows managing a number of basic data called business objects within a repository (mainly blogs / products, customers, suppliers, employees) in order to distribute them to the various applications of the company. Thanks to the common repository, users can manipulate the same version of reality, which obviously facilitates exchanges between applications, and improves the quality of the underlying business flow.

MDM solutions have improved the data quality in an obvious way, thanks to the better controlled database of business information management.

In the field of the management of basic data (MDM), we can distinguish between two types of data:

- Metadata: we mean by here the data used to define the properties of the content data and the useful data. This includes the appropriate matching of different data sources fields, the reference key description, etc.

- Content data (Content Data): for this second category of data, the quality of the content and of the syntax plays an essential role. For example, we can mention the correct address of a contact, a single format of a date, or a free basic data stock for duplicates.

These two types of data are essential and important in the same proportion to the data quality. The rise (or fall) in the quality level of one of them influences directly the quality level of the other.

The following examples provide a view concerning the issues that may arise in the field of the metadata:

- **An example of the field "client status"**

This example illustrates the complexity of the metadata and of the meaning of the contents of the fields.

Meaning in the 1st system:	A customer is any person who has received advertising material
Meaning in the 2nd system:	A customer is any person who is registered on the website.
Meaning in the 3rd system:	A customer is any person who has settled a bill.

- **An example of the field "date formats".**

This example shows that it is important to agree on date formats to be used. Date indications can be displayed in different formats.

1st System:	2010-06-11
2nd System:	10-06-11
3rd System:	June11th, 2010
4th System:	11.6.2010

By observing the **date format** in the system number 2, we can note that there are two possible interpretations: June 10, 2011 or June 11, 2010.

9. The data governance implementation.

The knowledge of the data quality level may be the trigger event that will justify the establishment of quality insurance program that will be used in order to maintain or increase the data quality level next to the different factors that are identified in the data quality evaluation (Martin Goulet, 2012, p. 10).

To improve the data quality, the company should define its governance model, which means its controlling model formalized by people, processes and techniques to facilitate the ability to rely on the data as a major asset for the company.

It should be noted that such a project should not be launched only if there is a real need and a significant challenge for the company. It is therefore recommended to make sure that the relevant processes play a significant role in the competitiveness of the company. It should give also a clear and realistic objective and assess the chances of getting a positive result. The implemented solutions that are part of a program may include technical and/or organizational aspects.

9.1. Roles of the general branch and operational branches

To launch this process, two guarantees of success must be met: the sponsorship of the branch, and the involvement of all stakeholders. We need, in order to convince the general branch to take the time to measure the non-quality impact and especially to demonstrate that the quality is a competitiveness source for the company. Continue then in making evidence to the pedagogy nearby the operational branches.

Thus, the data quality is not a technical problem; it affects primarily the company. For this purpose business, implementing programs to improve the information and data quality is one of the keys to progress. A program must be managed as a project and involved largely here where there are the concerned functional departments (sales, purchases, production, etc).

9.2. The technological solutions

After having convinced the general branch and the trades of the data quality importance, it is time to evaluate the technological solutions. The establishment of a data quality technology must allow:

- Make the diagnosis and assessment of the quality problems.
- Support the efforts of integration on all data sources.
- Computerize errors treatment in the processes of extraction and reloading.

- Define a framework for capturing and managing all of the errors related to the poor data quality.
- Provide a framework to measure the quality indicators evolution over time.
- Provide confidence indicators on the used data quality.

9.3. The functions of the technological tools

Most of the technology solutions of the data quality ingest tools that provide the following data quality features: profiling, standardization, cleaning, matching, parsing, enrichment and monitoring.

Profiling

Data profiling tools analyze the State of the data in databases or files. They collect statistics and information about the data to analyze whether they are of sufficient quality in order to be used in other contexts. They analyze the conformity of data relatively to the standards of the company and to the definitions of these fields (metadata). They identify dependencies with other data sources and evaluate the duplication of information.

Profiling allows obtaining quantitative information that details the strengths and weaknesses of the business data. A special knowledge which can therefore, serve as a starting point for the improvement of these processes. The information obtained from the profiling serve as a foundation for the next phase: standardization.

Standardization

By using the rules defined in trades, standardization and validation tools compute the process of verification and correction of data so that the abbreviations are standardized, the data that are spelled correctly and formatting models that are properly used. They validate the value of the data relatively to an interval of distribution or to a field (for example: validation of addresses according to postal standards).

Cleaning

The cleaning tools allow detecting and correcting (or deleting) corrupt or inaccurate records from a database or a file. Detected errors could be created in heterogeneous application environments, seized in error by a user or corrupted during transmission or storage. The purpose of cleaning is to make the source of data coherent with other sources of the company. Cleaning tools are used retrospectively on the data, differently to standardization and validation tools that are used when the entry of the data.

Matching

The matching tools allow comparing data from various sources. They allow in identifying relationships between data records in order to dice-duplicate them or to carry out treatments by group. They help to identify the records that describe the same entity.

Parsing

The parsing tools allow transforming an entering field that contains multiple data generally in a tree structure used by applications. For example, parsing tools can be used to recognize in a field the data addresses, measurements, quantities or references products.

Enrichment

The recording tools allow adding to records, data from other internal or external sources. Enrichment consists therefore, in the integration of data in order to give more value to the existing data.

Monitoring

The monitoring tools allow identifying and reacting immediately to problems before the data quality degrades. They allow to follow the evolution of the data over time and to determine their possible damage. They identify trends over the

data quality and alert on violations of the rules on the defined data quality.

It is clear that these various tools that manage the various aspects of data quality are not independent of each other.

It should be noted finally that all methods of improving the data quality include a cycle of four stages:

Definition: In this step, the company defines how to measure the data quality to the needs of users. It determines the priority axes of work.

Measurement: It is a must to measure the data quality in the online projects with the strategy of the company and according to criteria and measures defined by users.

Analysis: The organization assesses the impact and the costs of the non-quality for business directions. It prepares also plans of improvement of this quality. The objective is to represent to managers a business case of the improvement project.

Improvement: In this step, the company performs projects of improvement and correction. It implements the measurement tools. It verifies the indicators of achievement and gives back the results to decision-makers.

10. Conclusion

The data quality is a subject that becomes increasingly important within the business. It becomes a must that allows companies to take the full potential of the data that they use in their day-to-day operations as well as in their strategic decision making (Martin Goulet (2012)

The quality of any operational or strategic decisions depends on the data. In the absence of rich data, accurate and reliable transverse to the company, an organization can quickly be found facing misleading findings, erroneous and potentially dangerous.

This article has evoked the concepts of the data quality and its importance in businesses. A poor quality of data is expensive and leads to less relevant decisions, to a poor management of customer relationship, and to a decrease of the reputation and image of the company, which is translated into a loss of its performance and competitiveness.

To improve its data quality, the company must launch a permanent management program. This latter is concerned with the different functions of businesses and computer science in the company. It requires the support of the general branch and of any function implied by the data management. The awareness and training of the teams on the issue of the data quality is therefore essential.

This data governance program must rely on technological solutions that allow acting on all the projects: the standardization of data, profiling, passing through the cleaning to the enrichment.

It is a must to define the rules of management of the company's data. They are enacted to ensure the quality of completeness, compliance, consistency, accuracy, non-duplication and data integrity.

It should be noted that the data quality is not only a technical problem. It also concerns the business functions and therefore must be incorporated to the habits and culture of the companies.

REFERENCES

_____.2011. "Gouvernance des données: comment mettre durablement les données contrôle?" LES FOCUS SOLUCOM.

Brasseur, Christophe. 2008. "La Gestion de la Qualité des Données." IT-expert n°73.

Cappiello, Cinzia, , Chiara Francalanci, Barbara Pernici. 2004. "Data quality assessment from the user's perspective."

Proceeding IQIS '04. Proceedings of the international workshop on Information quality in information systems, Pages 68-73.

Clément, Delphine, Brigitte Laboisse, Domunique Duquennoy, Andrea Micheaux. 2008. "Non Qualité de données & CRM : quel coût ?" QDC.

DataFlux. 2011. "Une Etude sur les Pratiques des Entreprises Françaises en matière de qualité des Données." http://www.cxp.fr/gespointsed/imgbreves /CP_CXP_DataFlux_survey_090611. pdf.

DataFlux. 2011. Livre Blanc. "Cinq étapes-clés pour des données d'entreprise performantes." http://www.sas.com/offices/europe/france/software/documents/DataFlux-Cinq-etapes-cles-pour-des-donnees-d-entreprise-performantes.pdf.

Goulet, Martin. 2012. "Hiérarchiser les dimensions de la qualité des données : analyse comparative entre la littérature et les praticiens en technologies de l'information." Essai présenté au CeFTI, Faculté des Sciences, Université de Sherbrooke, Longueuil, Québec, Canada.

INFORMATICA. 2008. Livre Blanc. "Des Données de Qualité : Exploitez le capital de votre organisation." JEMM research.

Tayi, Giri Kumar and Donald P. Ballou. Eds. 1998. "Examining Data Quality." *Communications of the ACM* 41(2).

Trouchaud, Philippe, Zouheir Guédiri, Reda Gomery. 2011. "Qualité des Données: Quelles vérités dans les Entreprises." Livre Blanc, Crédits – réalisation.

Uniserv Gmbh. 2009. White Paper. "La qualité des données : Un facteur clé pour le succès de la gestion des données de base."

Wittmer, S., V. Ranaivozanany, A. Olympio. 2012. "Gestion des Risques d'Entreprise : Qualité des données, levier de pilotage stratégique." CNP Assurances.