Vol. 05, No. 2 (2023) 199-212, doi: 10.24874/PES05.02.002

Proceedings on Engineering Sciences

www.pesjournal.net

APPROACH FOR EVALUATING DATA QUALITY PROJECTS

Maqboul Jaouad[1]
Bounabat Bouchaib

Received 18.11.2022.
Accepted 03.04.2023.
UDC – 336.532.1/2

Keywords:

Evaluation of quality projects, Multi-criteria analysis, Cost-benefit analysis, Cost-effectiveness analysis, Data Quality, Prediction of the completeness of the business object, business processes collaboration.

A B S T R A C T

Data in our era is characterized by overwhelming volume, variety tends towards unstructured architecture and very high and continuous speed of production and sharing. These characteristics lead organizations to demand a level of quality from their data that must meet the needs and requirements of those requesting their service.

Despite the existence of numerous quality procedures, a reference method is essential for the evaluation of data remediation projects. It is in this perspective, this article proposes an approach aimed at helping in the choice of the most profitable scenario according to the gains and benefits expected during the evaluation of data quality projects.

The approach assesses the positive impact of process quality and data quality, as well as the complexity of its implementation. It is based on a cost-benefit analysis and a cost-effectiveness analysis as well as on a multi-criteria analysis for the classification of the processes and subsequently of the projects according to their weight of importance. The approach focuses on data. It is also interested in the prioritization of key processes and their collaboration between different process managers. The approach has been put into practice in the health sector for the identification and strengthening of important processes and objects, eligible to be the subject of data quality improvement projects.

The approach has been applied to the healthcare sector for the identification and strengthening of important business processes and objects, eligible to be the subject of data quality improvement projects.

1. INTRODUCTION

The development of software systems and computer equipment have radically evolved, leading to the emergence of new technologies, such as the sixth-generation technology of standards for mobile telephony (6G), Bluetooth v5.2 and Wi-Fi 6. These new technologies produce an immense quantity of data that can be structured or unstructured data at lightning speed, which led to the adaptation of big data in most companies.

In the era of large companies such as Google, Apple, Facebook, Amazon and Microsoft (GAFAM), data

Maqboul Jaouad
Email: jaouad_maqboul@um5.ac.ma

represents real capital considered as the oil of the 21st century (Simon & De Prato, 2015). Data-driven business analysis forms a basis for innovation and agility in today's business environment (Chen & Siau, 2011; Kiron, Kruschwitz, Haanaes, & Von Streng Velken, 2012). This is how data should be enhanced by digital information systems and their use will have a strong impact on the daily life of businesses and further weaken traditional ones.

Data is an asset that must be collected, controlled, defended, maintained, improved and shared to make relevant decisions to readjust strategy throughout the existence of any organization to remain competitive by retain customers.

Satisfaction offered to customers is the result of a quality of services and data, these two pillars encourage organizations having a clear enterprise architecture (EA) to design and evaluate their data and their business processes. Indeed, EA offers a high-level overview of an entity's business and IT systems and their interrelationships (Tamm, Seddon, Shanks, & Reynolds, 2011). It defines how the organization responds to future problems and facilitates communication betwixt these different levels, identifying the causes of any deterioration in quality.

There are several benefits derived from using an enterprise architecture affecting decision-making and strategy execution. In this context that the identification of EA structures and areas impacted by deteriorating data quality must be continuously improved.

The design of information systems is essentially founded on Enterprise architecture used. It is made up of four domains: business, data, application and infrastructure (Baldwin, Beres, & Shiu, 2007). These areas are interconnected, so the deterioration of the quality of one lead to the deterioration of the other. Data architecture is thus the most important domain, given its influence on the decisions and on the strategy of any organization (Capirossi & Rabier, 2013). The application domain undergoes changes due to the amount and type of data that evolves every day, while the infrastructure, business and application domains are stable after implementation. From this point, it is clear and obvious the importance of quality and its cost which weighs on the budget of companies; and quality improvement typically focuses on business processes and data (Batini, Cappiello, Francalanci, & A, 2009). The company is therefore faced with the problem of the quality of its processes and its data, how will they be chosen? and on what basis?

Reference (Belhiah, B, Bounabat, & Achchab, 2015) talks about the added value of choosing quality projects at minimum cost based on cost-benefit analysis (CBA). The impact of data quality on effectiveness within organizations has not been discussed, with the aim of justifying the improvement of data within non-profits using cost-effectiveness analysis (CEA).

Most companies don't have the experience to estimate costs, so it's best to replace the cost with the complexity of setting up data quality. It is easy to assess the complexity of improving data since it is part of the skills of employees in any organization. According to (Batini & Scannapieco, Data and Information Quality Dimensions, Principles and Techniques, 2016), there are several methodologies to assess and improve data quality, but they do not address the following points:

- How to estimate the quality of business objects before improvement?
- How to minimize the cost of data quality to convince managers of its usefulness?
- How to help decision-makers choose data quality projects and at what cost?

Thus, it is essential to have a system to predict and calculate the value of data quality dimensions, in addition to a collaborative system to save money due to data quality fees. For this purpose, the company must have the tools and approaches to choose project of quality to improve it at a reasonable cost and are there methodologies that can answer all these questions?

2. BACKGROUND AND RELATED WORK

2.1 Data quality methodologies

Definition of methodology

A data quality methodology is a set of guidelines and techniques that takes input information about a particular fact of interest and establishes a logical process for utilizing that information. The aim is to measure and improve the quality of an organization's data from contributions to decisions (Batini & Scannapieca, Data Quality Concepts, Methodologies and Techniques, 2006).

Inputs and outputs of a methodology

The Entries refer to all knowledge types shown in Figure 1, plus the available budget, if it’s mentioned. According to (Batini & Scannapieco, Data and Information Quality Dimensions, Principles and Techniques, 2016), methodologies of data quality may be classified by various criteria:

- Data or process oriented.
- Measurement vs improvement.
- Intra-organizational vs inter-organizational.
- General purpose vs specific use.

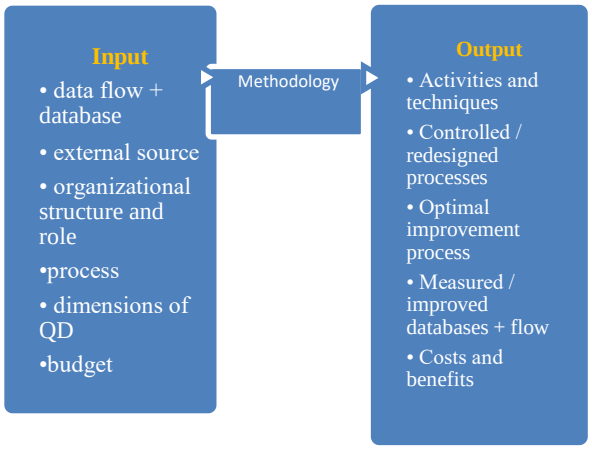

Figure 1. Inputs and outputs of a data quality methodology

There are numerous methods that are developed specifically to a particular field and in particular cases like the Quality Assessment of Financial Data (QAFD) (Amicis & Batini, 2004) and the information Quality Measurement (IQM) (Eppler & Muenzenmayer, 2002). Some frameworks are only relevant for specific domains, such as census, healthcare, or financial; from the above, there are two methodologies that govern quality measures, one is process oriented and the other is data driven founded on data quality dimensions. Given the production rate and flow of data, the data-driven approach is more important than the process.

The dimensions addressed by the methodologies

The dimensions of data quality are input elements of a methodology. Most definitions and calculations of the dimensions of data quality lean toward the value of data rather than patterns. Most researchers (Scannapieco & Catarci, Data quality under a computer science perspective, 2002) focus on completeness, accuracy, timeliness and consistency, this work will start with data completeness and generalize to others, including accuracy and timeliness data considering their importance to organization.

2.3 Assessing and improving data quality within methodologies

Methodologies comparison

To compare and analyze methodologies that deal with data quality, there are perspectives to use:

- The stages of the methodology.
- Techniques adopted when improving and assessing data quality.
- The dimensions chosen by the said methodology.
- The cost types associate to quality of data.
- The data types considered.
- The entire process used that create and update data; the Table 1 summarizes the comparison of theses methodologies.

According to Table 1, the Total Data Quality Management (TDQM), Total Information Quality Management (TIQM) and COLDQ methods deal with a countable number of quality dimensions and are not scalable to others; also, for The Datawarehouse Quality (DWQ) and Cost-effect Of Low Data Quality (COLDQ) don't handle assignment of process and data responsibilities like Comprehensive methodology for Data Quality management (CDQ).

Table1. Comparison of methodologies in the data quality assessment and improvement phase.

Methodology	Extensible to other dimensions	Analysis of data quality requirements	Cost evaluation	Assignment of process responsibilities	Assignment of data responsibilities	Identification of critical areas
TDQM			√	√	√	√
DWQ	√	√	√		√	√
TIQM		√	√	√	√	√
COLDQ		√	√			√
CDQ	√	√	√	√	√	√

Methodologies and costs

References (Lesca & Lesca, 1995; Redman, 1996; English, 1999; Huang, Lee, & Wang, 1999) state that data quality costs are significant, while few studies which really demonstrate how to identify, categorize, and measure these costs. How to determine causal links betwixt data quality failing and business processes, thereby quantifying monetary and non-monetary impacts of quality.

Costs may be determined as "resources given up attaining a given goal or the monetary impact of some procedures or their absence" (Eppler & Helfert, 2004). Table 2 shows the extent to which the costs of data quality are incorporated and considered in the methodologies.

Table 2. Treatment of cost in methodologies

methodology	Budget constraints	Improvement costs	Classification of costs	Cost-benefit analysis
CDQ	√	√	√	√
COLDQ		√	√	√
TIQM	√	√	√	√

Methodologies like Cost-effect Of Low Data Quality (COLDQ) does not treat budget constraints, even the Comprehensive Methodology for Data Quality management (CDQ) don't care about Cost-effectiveness analysis, there is other areas are not considered, such as collaboration of processes to minimize cost when assessing data quality. In addition, the dimension value prediction component is not covered. The improvement process is founded primarily on the evaluation of quality through costs, especially the costs of effectiveness which target non-profit organizations.

3. APPROACH DQPEF DATA QUALITY PROGRAM EVALUATION FRAMEWORK

As mentioned earlier, there is an insufficiency of quality project evaluation within the methodologies. This article proposes an approach called Data Quality Program Evaluation Framework (DQPEF) which contributes alongside other methodologies, using cost-effectiveness analysis, an approach widely applied in the health sector and for organizations non-profit, in addition to cost-benefit analysis for a financial assessment of data quality projects. The DQPEF approach uses the multi-criteria approach, which is a strong tool to aid decision-makers in their decisions when selecting projects. Finally, it contains two complementary stages: 1) the collaboration of business processes that have common objects to minimize costs; 2) prediction of completeness values of objects before data improvement.

3.1 General structure of the DQPEF approach

The DQPEF approach is founded on a cost-benefit and cost-effectiveness analysis of data quality projects. The DQPEF approach is founded on the data qualimetric tree breaking down each dimension into several factors, then broken down into criteria, whose costs (tangible, intangible, etc.) may be measured (see Figure 2).

Qualitative tree for data quality assessment

The qualimetric tree, Figure 2, breaks down data quality into dimensions, each dimension broken down into factors (cost, benefit, and effectiveness). The factors themselves broken down into financial and business impact and tangible and intangible costs.

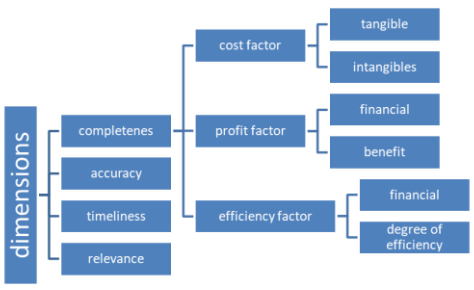

Figure 2. Decomposition of data quality assessment factors and criteria

The tree addresses all the dimensions that can be improved and then broken down into factors of cost, benefit and effectiveness. In the present research work, "completeness" is the dimension of application.

Completeness belongs to the contextual category when classifying dimensions of quality (Wang & Strong, 1996), and according to (Maqboul & Bounabat, 2017) completeness is defined as the extent to which data is not missing and is broad and deep enough to achieve the task.

After having chosen the dimension, the next stage is the construction of the criteria, the evaluation will be carried out on the: (i) cost, (ii) benefit, (iii) effectiveness of data quality. Intending to quantify the impact of quality founded on three factors, a cost-benefit analysis is used to cover the financial benefits, likewise a cost-effectiveness analysis to assess performance, and recommend the multi-criteria analysis.

Cost Benefit Analysis

Reference (Livermore & Revesz, 2013) expresses the benefits and costs in Cost-benefit analysis of an intervention in monetary units. the cost-benefit analysis determines the share of the company's resources that may be allocated to achieve the objective.

The result may be presented in two ways: as a net monetary gain or loss or as a benefit / cost ratio. Cost-benefit analysis, like cost-effectiveness analysis, aims to directly compare various interventions (Cellini & Kee, 2015). Thus, studies that describe cost-benefit analysis should compare the costs and benefits of an intervention and their alternatives.

The main practical problem with cost-benefit analysis is to assess benefits, such as saving lives or relieving pain, in monetary units. Cost-benefit analysis may incorporate the widest range of effects across the widest range of interventions and programs (both inside and outside the health sector), but it is often controversial, as it requires evaluating the benefits, including death and disease, in monetary terms (Wit, et al., 2007).

Cost-Effectiveness Analysis

Cost-effectiveness analysis (Muennig & Bounthavong, 2016) expresses the impacts of interventions in natural units, such as deaths, illnesses or burns averted and the costs of these interventions in monetary units. Cost-effectiveness analysis aims to provide information on the relative effectiveness of alternative interventions that serve the same purpose.

The result of such an analysis is a ratio that reproduces the differences in costs versus the differences in the effectiveness of this intervention compared to other

interventions that serve the same objective (Drummond, Sculpher, Torrance G, O'Brien, & Stoddart, 2006).
Cost-effectiveness analysis is the simplest type of economic evaluation to account for differences in results. The main advantage of cost-effectiveness analysis is that measuring benefits in natural units simplifies the analysis and is often more intuitive to study users.

The disadvantages are the inability to compare evaluations of effectiveness betwixt interventions that produce different outcomes and the need to focus on a single outcome of an intervention even when one intervention generates several distinct benefits (Bertram, et al., 2016).

Multicriteria Analysis

Multi-criteria analysis (MCA) ranks adaptation options against criteria that can be weighted depending on their importance, and the sum of the weights is employed to rank the proposed alternatives.

Multi-criteria analysis is a complementary approach to CBA. This is a two-stage decision process:
The first stage identifies the objectives and seeks the trade-offs betwixt these objectives in different ways.
The second stage seeks to find the “best” policy by assigning weights (scores) to the different objectives.
The MCA allows both qualitative and quantitative data to be taken into account in the ranking of options. For example, MCA is able to take into consideration elements such as feasibility, fairness and acceptability, which may often be difficult to quantify. This approach makes it possible to calculate the score of each alternative from the ratings and weights assigned which characterize the importance of the criterion.

The MCA deals with structuring and solving decision and planning problems involving multiple criteria. Its objective is to accompany decision-makers faced with these issues.

The difficulty of the problem comes from the presence of more than one criterion. There is no longer a single optimal solution to a problem that may be obtained without incorporating preference information.

There are several variations of MCA, among which, the Weighted Sum Model (WSM) multicriteria analysis (Fishburn, 1967), it is the most widely used approach, especially in one-dimensional problems. If there are "x" alternatives and "y" criteria then, the best alternative is the one that satisfies the following expression:

$$A_i^{WSM-Score} = \sum_{j=1}^{y} w_j a_{ij} \text{ , for i = 1, 2, …, x} \qquad (1)$$

Where a_(ij)¨ is the score assigned to criterion ‘i’ in alternative ‘j’ and wj is the weight assigned to criterion ‘j’.

Evaluation of data quality programs: DQPEF

The DQPEF approach relies primarily on multi-criteria analysis (MCA) to select the key processes within the organization and when choosing the optimal project of data quality assessment. Second on cost-benefit analysis (CBA) (Livermore & Revesz, 2013) and cost-effectiveness analysis (CEA) (Muennig & Bounthavong, 2016) for the classification of data quality projects.

The factors of cost, benefit and effectiveness are the result of the decomposition of the objectives predefined by the decision-makers and which are in relation to the chosen dimension "completeness". Each factor will be broken down into criteria by assigning a weight to it intending to adapt to any organization or situation. The assessment will be done in two stages:

- The first stage is to quantify the impact and cost of processes within the organization, with the objective of reducing all processes in stage 2.
- The second stage concerns the assessment of the impact and cost of data quality handled by processes chosen in stage 1.

DQPEF contribution to other methodologies

The approach makes it possible to address other dimensions such as accuracy, relevance, etc. Extra to being a decision support tool by offering a collaborative system to optimize the cost of data quality and the prediction of data dimensions to justify to decision makers and comfort.

In Table 3, a comparison of the DQPEF approach with other methodologies, the DQPEF approach complements the other methodologies cited in Table 1, by the additional phases concerning the prediction of the dimensions of the data quality, likewise a system of process collaboration for the reduction of the cost of the quality of the data.

Table 3. DQPEF comparted to other methodologies

Methodology	Cost evaluation	cost-benefit analysis	Cost-effectiveness analysis	Extensible measures on other dimensions	Treat responsible for the process	treat responsible for the data	Choice of quality projects	Predicting the value of quality	Process collaboration
CDQ	√	√		√	√	√			
COLDQ	√	√							
TIQM	√	√			√				
DQPEF	√	√	√	√	√	√	√	√	√

The DQPEF approach makes it possible to achieve the objectives defined by the decision-makers, then the selection of the target quality dimensions. These objectives will be accomplished by quantifying the criteria of the cost, efficiency and benefit factors of data quality. The next phase is to recognize the services and business processes to improve who is founded on the next stages.

Identification: (i) financial objectives such as cost reduction, increase in profitability, limitation of charges, economic objectives such as modification of the relationship with the customer and non-economic objectives differentiation from competitors; (ii) the evaluation process is open to all dimensions, they must be identified to construct the evaluation criteria ;(ii) criteria which combine in cost, benefit and effectiveness to meet the expectations described in the objectives; (iv) the users who will manage the business objects and processes and who will monitor their quality and monitor quality anomalies; (v) of the organization's key Processes for the evaluation of the costs and the impacts of their quality.

Measurement: after having identified the objectives, dimensions, criteria and business processes, the approach allows: (i) to quantify the costs, benefits and performance gained by the quality of each business process; (ii) to determine the processes that drive the project data quality assessment through MCA; (iii) automatically measure the quality dimensions value of business object; (iv) measure the benefits, efficiency and cost of all business objects managed by key business processes.

Analysis: After having evaluated the business processes and data, the approach allows decision-makers to order the data quality projects by their apport to the objectives, by assigning weights to the criteria predefined in stage 1. The approach helps them. decision-makers to analyse the projects likewise the key processes and business objects that participate to achieve the objectives.

Improvement: The recommendations made by the process in stage 3 allow analysts and business managers to launch improvement projects, then check the dimensions values of processes and objects after improvement.

3.2 Prediction of data quality dimension values

Deep learning (DL) is a prediction tool to estimate the dimensions of data quality in the DQPEF approach. Second, using Shapley's value to incentivize business leaders to collaborate on data quality cost reduction.

DL is a type of artificial intelligence derived from machine learning, where it can learn on its own, unlike algorithmic programming. It is basically a neural network greater than or equal to three layers, they simulate the mechanism of the human brain, allowing it to "learn" from large amounts of data that represents market data, data from the human brain. internal organization and other useful information to find the relationships that bind them to provide precise results such as classification or prediction of data (Fister, Mun, Jagric, & Jagric, 2019).

Evaluating data quality is the upstream stage of improvement, justifying it to decision-makers. It therefore becomes essential to have a tool for predicting data quality values, in support of those for evaluation already developed and developed.

Artificial neural networks

Artificial neural networks (ANN) represent structures for information processing connected to each other from the input layer to the output layer (Gallo, 2015) by simulating the physiological structure of human brain structures (McCulloch & Pitts, 1943). They are systems capable of modifying their internal structure in relation to a function objective. They are particularly suited to the resolution of nonlinear type problems, being able to reconstruct the fuzzy rules which govern the optimal solution of these problems.

ANN model

In a simple model, the first layer is the input layer, followed by a hidden layer and finally an output layer. Each layer can contain one or more neurons. Models grow more complex to solve increased problems by increasing the number of hidden layers and neurons in each hidden layer.

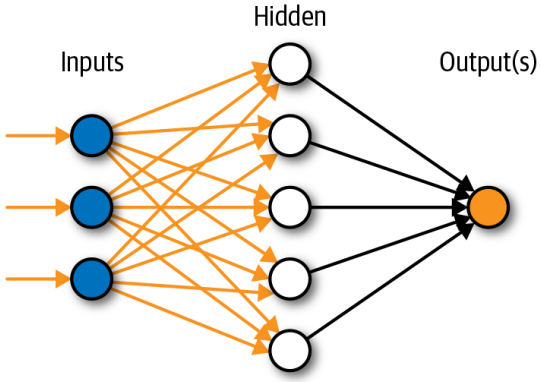

Figure 3. Artificial neural network

Deep learning

DL is an artificial neural network of several intermediate hidden layers. These intermediate layers make it possible to deal with complex problems; without them, the system solves simple calculations. The results of one layer serve as input to the next, leading to complex decision making (Sharif & Gursoy, 2018).

The DQPEF approach implements the DL to predict the values of the dimensions from the factors of complexity, benefit and efficiency, because it is complicated to find a correlation between the criteria of the factors and the dimensions.

Using deep learning in DQPEF

In stage (3) of the DQPEF approach consisting in the quantification of the criteria and sub-criteria of the evaluation of the quality of the business object. It is therefore wise to have a tool to predict the approximate value of the quality of the object or business process using a regression algorithm (Maqboul & Bounabat, 2020). DL is very effective for such high dimensionality problems, since it may deal with complex relationships betwixt variables, sets of categories and complex functions relating to input to output variables.
There are two types of machine learning:

1) supervised learning, which is adapted by the DQPEF approach, this algorithm is guided with prior knowledge of the output values, this model adjusts the difference betwixt the results obtained and those expected.
2) contrariwise, unsupervised learning does not use labelled data and it's hard to calculate the output values with certainty.

The generic RNA architecture of the suggested prediction of the DQPEF approach is founded on the RNA model (see Figure 3), the number of layers and neurons is validated by experience and tests:

1) The number of input neurons represented by the criteria of the factors of complexity, benefit and efficiency.
2) According to the work (Maqboul & Bounabat, 2020) implemented in the approach and the tests made on the RNA model, two hidden layers was the most precise in terms of prediction of completeness which is 0.25 in case 7 of Table 4.
3) Each layer contains 10 neurons.

Table 4. Comparison of the precision of the RNA model adopted in the use case

case	Number of hidden layers	layer 1	layer 2	layer 3	layer 4	Completeness
1	0					0.2835
2	1	16				0.2827
3	2	16	8			0.2772
4	3	16	8	4		0.302
5	4	16	8	4	2	0.3164
6	1	10				0.2693
7	2	10	10			0.2499
8	3	10	10	10		0.2582
9	4	10	10	10	10	0.2548

Table 4 shows that the neural network with two hidden layers, each containing 10 neurons, was the most accurate for the business object completeness value.
The purpose of the RNA model is to predict the value of the dimension, so the final layer of the neural network is going to have a single neuron and the value it returns is the prediction of completeness in this present thesis work.

3.3 Principle of collaboration in the DQPEF approach

Coalitions are widely used in multi-agent systems to perform collective tasks (Norman, et al., 2004), the formation of coalitions is an incentive for cooperation. It allows individuals to come together to jointly achieve common goals within a period.

The cost of improving quality becomes very significant as data quality approaches perfection. Business leaders seek to increase the quality of services and objects at lower cost. The objective is to introduce this principle of cost sharing between business processes.

Theory of cooperative games

The theory of cooperative games may be utilized in the case where the actors may obtain more advantages by cooperating than by remaining alone, it consists of two elements: (i) a set of players and (ii) a function showing the value created by a subset of players.

In a game (Peleg & Sudhölter, 2007), the issue of forming coalitions is one of the major issues of game theory, both in cooperative and non-cooperative games. The grand coalition produces a large surplus that the partial collations will eventually occur after some negotiation (Grabisch & Funaki, 2011).

The value of the grand coalition should be allocated to players individually, founded on the contribution of each player (Norman, et al., 2004). Research in such games has focused on a "fair" arbitration rule, which keeps within account what each player can do for themselves. (Without the help of the other player) and what all the players may accomplish together.

J. Nash in 1950 (Rullière, 2000) developed an axiomatic model, assuming a set of good deals, Nash seeks a result which will be Pareto-optimal and individually rational and which furthermore satisfies certain technical conditions.

Nash shows that the only arbitration rule that satisfies these axioms is the rule that chooses the market that maximizes the product of player utility increments. It is a cooperative game for the actors to form coalitions to optimize the cost of their own operations to achieve an acceptable quality. They may, through cooperation, realize gains in the form of reduced costs rather than gains.

Shapley value

The Shapley's value allows cost sharing in cooperative game theory founded on so-called incremental costs. The Shapley value of player 'i' in the game given by the characteristic function V is the part of the surplus that must be allocated (Chalkiadakis, Elkind, & Wooldridge, 2011).

This is a weighted average of player i's contributions to reach the possible coalition. For example, for example, consider a game with three players, p1, p2, p3 and p4. Suppose player i1 is the first player in the game, i2 is the second player to join the game and i3 the third to join the game and i4 is the last. Player p1 receives a cost C ({p1}), player p2 receives a cost C ({p1, p2}) - C ({p1}), player i3 a cost C ({p1, p2, p3}) - C ({p1, p2}) and player p4 receives C ({p1, p2, p3, p4}) - C ({p1, p2, p3}).

The Shapley value assumes that the order of arrival is random and that the probability of a player joining the first, second, or third in a coalition is the same for all players. The cost allocated to a player 'i' in a game comprising a set N of players is given by:

$$\varphi i(N) = \left(\sum_{\substack{S \subseteq N \\ i \epsilon S}} \frac{(N-|S|)!(|Z|-1)!}{N!} * [v(S) - v(S \setminus \{i\})] \right) \quad (2)$$

|N| and |S| respectively, the total number of players and that belonging to the coalition S.
Shapley imposes four axioms to be satisfied (Efficiency, Symmetry, Zero player and Additivity).

Efficiency: the actors precisely distribute the resources available to the grand coalition among themselves. Namely, efficiency:

$$\sum_{i \epsilon N} \varphi i(N) = v(N) \quad (3)$$

Symmetry: the players 'i', 'j' ∈N are said to be symmetrical with respect to the game v if they make the same marginal contribution to any coalition, i.e., for each S ⊂N with i, j ∉S,

$$v(S \cup i) = v(S \cup j) \quad (4)$$

In another way if the players i and j are symmetrical with respect to the game v, then φi (v) = φj (v)

Null player: if i is a dummy player, it does not import any contribution to the coalition, i.e.,

$$(S \cup i) - (S) = 0 \quad (5)$$

for all S ⊂N, then φi (v) = 0.

Additivity: φ (v + w) = φ (v) + φ (w), where the game v + w is defined by

$$\varphi i (v + w) = \varphi j (v) + \varphi i (w) \quad (6)$$

A player i who participates in two games whose characteristic functions are v and w, having the same players. The sum of the Shapley values in the two sets and the sum in the two separately.

Application of Shapley value on DQPEF

In the DQPEF approach, the value of shapley comes into play, during collaboration betwixt business managers to diminish the cost of improving business objects shared by the processes. The business service managers will be players, to find coalitions. Shapley's axioms will be verified founded on Figure 4.

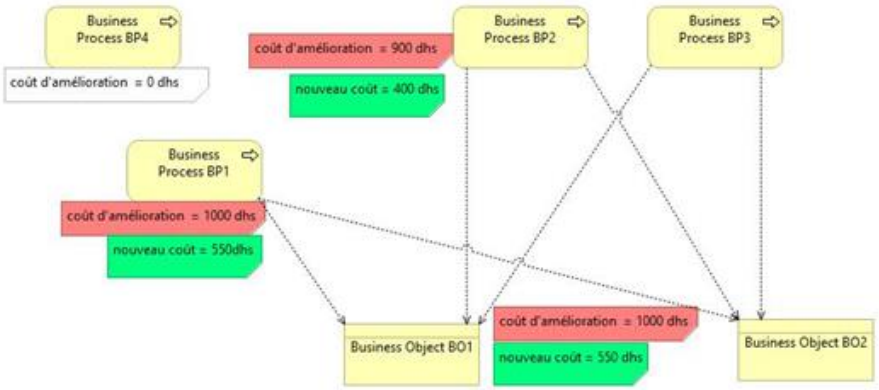

Figure 4. Coalition within the DQPEF approach

Efficiency axiom

The sum of the shares of each process should equal the total gain of the coalition of all the processes, which allow minimizing the cost of each process. During the evaluation, each manager evaluates the complexity, the benefit and the effectiveness of the quality of each of the business objects handled by the business processes. The process BP1 accesses two object BO1 and BO2, the same for BP2 and BP3, since the three processes access both objects, they will all three build a coalition, the cost sharing of data improvement may be worked out according to two approaches:

The cost of the coalition will take the maximum proposed among the sorting processes. assign costs for all possible coalitions which is difficult to set up, adapted by the DQPEF approach.

φBP1 (N) + φBP2 (N) + φBP3 (N) = v (N)

We have v(N) = 1500 according to the example of Figure. 5 And φBP1 (N) + φBP2 (N) + φBP3 (N) = 550 + 400 + 500 = 1500.

Symmetry Axiom

Either two business processes BP1 and BP3 try to make a coalition N and BO1 a business object handled by these two processes, then they perceive the same cost: Cost (BP1, BO1) = Cost (BP3, BO1) then φBP1 (v) = φBP3 (v).

According to the second approach in the efficiency axiom we have: v (N) = 550. Replacing player BP1 by BP3 will not change anything since their contribution is the same and their contribution is φBP1 (v) = φBP3 (v) = 550.

Null player a*xiom*

Let i = 1 ... 3 be three business processes that access the same object "Bo1," and S is the coalition they build except the BP4 process. Since the BP4 process does not access the business object then the cost of the quality of BO1 will be zero φBP4 (v) = 0 and φi (S U BP4) = φi (S).

Additivity Axiom

Let be a BP1 process which participates in two games having the same processes and whose characteristic functions are v and w. The sum of the values of Shapley in the two games and the value of Shapley in the game defined from the sum of the winnings of the two games (v + w).

According to Figure 4, the three processes BP1, BP2 and BP3 access the two objects BO1, BO2, then the cost of the process BP1 in the two sets v and w in parallel will be the sum of the cost of two separate sets. Indeed, the business managers will embark on estimating the cost of the quality of objects separately from the games and when working together to share the cost of several data with the same managers, they will proceed to the sum of the costs.

The set v contains the processes BP1, BP2 and BP3 and the object BO1; for the game w, it has BP1, BP2 and BP3 and the object BO2; then φBP1 (v + w) = φ BP1 (v) + φ BP1 (w).

4. USE CASE OF THE DQPEF APPROACH - APPLICATION TO THE HEALTH SECTOR

4.1 Application of the DQPEF approach to the health sector

The DQPEF approach helps in making decisions on the choice of the most profitable data quality projects founded on the weight given by physicians to the objectives. The objective is to quantitatively measure the contribution of key processes to the achievement of objectives, then to qualify the impact and complexity of improving the data used by the processes. This use case is small, as a validation support for the DQPEF approach in the field.

The Case study

The Trauma, Orthopedics and Physiotherapy Center is the subject of the use case, providing services such as consultation, diagnosis, treatment, surgery and rehabilitation. The three disciplines share the same clients and the same information system, as a result physician are in mutual collaboration to improve their data to benefit more and to minimize costs.

In the field of health, trauma is the study of injuries and those caused by accidents or violence on a person, likewise surgical therapy and repair of damage. It is often seen as a subset of the surgery known as accidental surgery.

As for orthopedics, it focalized on the care of the musculoskeletal system. It is a specialty that treats pathologies of the musculoskeletal system (joints, muscles, tendons and nerves) and bones. The orthopedic surgeon takes care of many problems: fractures, osteoarthritis, bone and joint infections, scoliosis, polydactyly, etc.

Physiotherapy is one of the medical professions that uses kinesiology founded on the prescription of exercises, mobilization, electrical or physical agents and health education. It treats acute or chronic pain, movements and impairments resulting from injuries, traumas, or diseases typically of musculoskeletal, cardiovascular, respiratory, neurological and endocrinological origin.

Physiotherapy tries to improve the physical functions of a patient through the diagnosis, prognosis, reduction, rehabilitation, prevention of disease.

In this example, two quality projects will be launched in a general quality program, the first project will be without collaboration betwixt business process managers and will duplicate with only one difference that of the collaboration betwixt business process managers who share objects trades in the second project.

Identification of aspects of non-quality

The center receives clients as new or old patient in trauma and physiotherapy, when creating the patient, the secretary made errors in entering information and other anomalies, the following was observed in their platform:

- prescriptions and certificates with missing or incorrect information.

- reports without a patient name, including incorrect dates.
- patients' bills do not reflect reality, the amounts for trauma consultations are those for physiotherapy.
- the system is demanding the opposite.
- duplications in the reference tables.
- appointments with incorrect or empty dates.
- lack of customer information, for sending emails, debriefing or canceling appointments or paying bills.

Possible causes of non-quality

The quality problems in the center's platform were discussed with both parties (trauma doctor and physiotherapist), to resolve the sources of the problems to satisfy the stakeholders; the causes discussed are:

- entry errors and forgetting given the load on the secretary requested by the two doctors.
- record printing processes do not always result in complete patient data.
- failure to update patient information by the secretary.
- a lack of a verification process for the data entered.
- Failure to terminate a patient's treatment process by the doctor to empty the waiting room.
- The bad management of the doctors towards the secretary.
- The pressure on the secretary causes him to recreate a client while he is existing.
- In this use case, the choice of the quality dimension is the completeness of the data and the business processes.

4.2 Alignments of DQPEF on the center platform

Definition of the objectives, dimensions and criteria of the data quality assessment

The life cycle of the DQPEF approach is founded on the Deming wheel illustrated in Figure 5.

The objectives were established during the meeting with the doctors and IS managers, the result is to: (1) give satisfaction to the patient; (2) increase income; (3) minimize the cost without having to call on the application developer each time; (4) increase the efficiency of the team in terms of daily operations and in terms of response time, the process tool helps to archive the objectives in order to share them with IT and business managers.

Figure 5. Life cycle deming

Definition of the objectives and dimensions of the data quality assessment

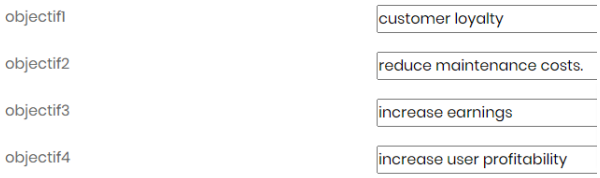

Figure 6. Objectives set during the meeting in DQPEF apps

The work focalized on the completeness of the data because the Clinic practitioners do not care about the accuracy or the updating of the data for the beginning of this work, it is a work which will be started. after finishing that one but have complete data to satisfy the patient as well the doctors.

Identifications of business processes and the users in charge of them

This stage of the process also identifies the key business processes within the medical practice, by classifying them, to choose those which will contribute to the quality of the data. These processes will be managed by business managers to maintain their information and their execution value. The processes chosen for the assessment are exposed in Table 5.

Table 5. Business process targeted by clinic practitioners.

Business process	Description
Patient management	Create patient edit patient delete patient's patient list search for patients
Request an appointment	Request an appointment
Request an order or certificate or report	Request a prescription. Request a medical certificate. Apply for an exemption certificate. Request a certificate of competence request a report.
Payment	Payment of consultation fees and therapist session

Choice of key business processes

The life cycle of the DQPEF approach is founded on the Deming wheel illustrated in Figure 5. After having configured the factors of complexity, efficiency and benefit of business processes, it is the turn to classify the processes according to the criteria by assigning them weights.

This stage of the process also identifies the key business processes within the medical practice, by classifying them, to choose those which will contribute to the quality of the data. These processes will be managed by business managers to maintain their information and their execution value. The processes chosen for the assessment are shown in Table 5.

Based on the key processes, the process of evaluating the improvement of the quality of business objects will be launched, to quantify the cost, benefit and effectiveness of the improvement.

Collaboration of business processes

Among the additional stages of the DQPEF approach, the collaboration in the case study is that of those in charge of the 'patient search' and 'appointment verification' business processes because they use the same 'patient' business object. So, it is wise to cooperate to decrease the cost of improving the business object which benefits all business processes. Figure 7 illustrates this collaboration to decrease data quality costs.

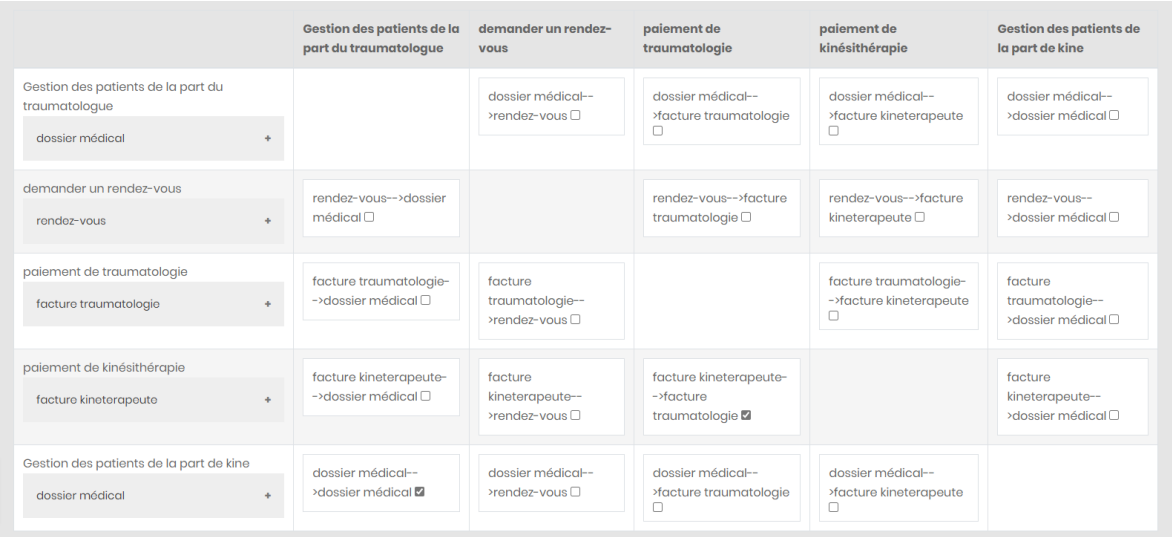

Figure 7. Collaboration grid for business process managers

Prediction of the completeness of business objects

This stage is optional for the user of the approach, to predict the completeness of the object on the basis of previous improvements validated by the management of the company, this prediction is founded on the regression by a neural network with two hidden layers in this use case since there is not enough history of evaluations of business objects, however the approach relies on a configurable and time-extensible tool for changing the number of layers and activation functions in order to have an increase in the accuracy of predictions of the value of a dimension of data quality .

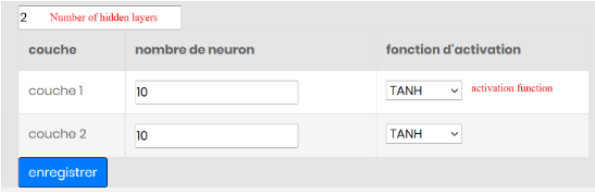

Figure 8. Configuration of the neural network

Figure 9 describe the network, by the number of layers and the type of activation function to measure the completeness of medical file object of patient.

Figure 9. Calculating the completeness of the business object

Analysis and classification of data quality projects

The classification result of the quality assessment projects is presented in Figure 10, the project founded on collaboration obtains a higher ratio of CBA and CEA than that without collaboration since the cost of the first project has decreased.

According to Figure 10, the project founded on collaboration proposes a reduction in the costs of business objects, which increases the rank of this project.

The two projects of the same improvement program focus on the same business processes, with the same costs of improvement, which leads to collaboration betwixt the doctors of the trauma and physiotherapy practice to achieve the objectives defined at the start of the process and minimize costs within the firm.

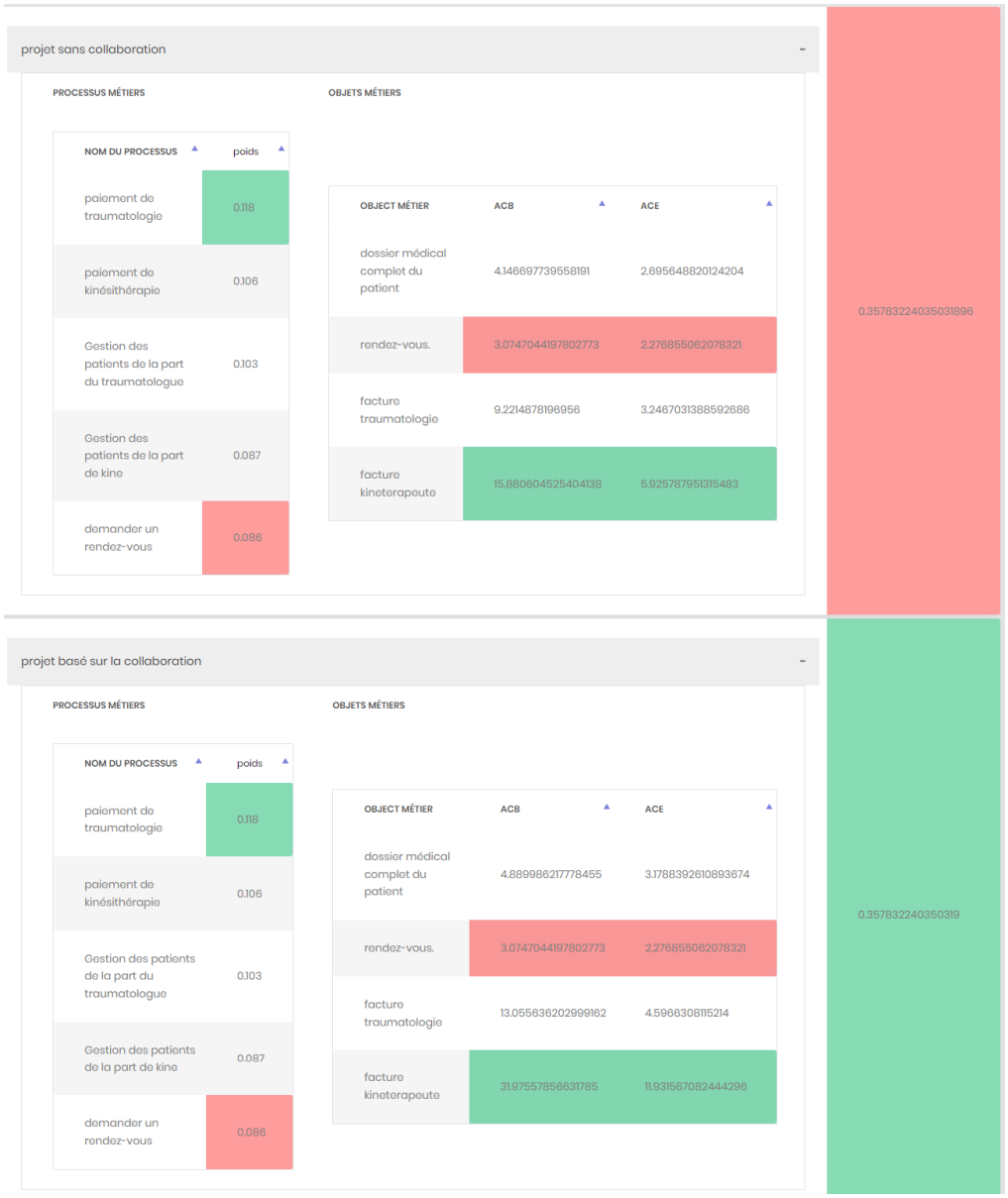

Figure 10. Classification of quality projects according to MCA

5. DISCUSSING

The DQPEF approach allows decision makers and managers of processes and business objects to choose projects of beneficial quality in terms of positive impact on the company at a reduced cost, the strength of this approach lies in integration with other quality methodologies to facilitate decision-making using ACM, quantifying the cost and impact of a quality project, this approach is also based on ACB and ACE in order to deal with any type of business and in any field application to quantify the cost of quality.

This approach consists of two complementary phases, one for the prediction of data completeness and the second is that of the collaboration of process managers to reduce the costs of improving common business objects.

6. CONCLUSION

This literature review explores the contribution of methodologies to the data improvement and evaluation phase. To compare methodologies and find gaps to fill.

The present research work and the fruit of these observations in order to appropriate a DQPEF approach to assess the quality of the data.

The DQPEF approach is founded on a multi-criteria analysis as a decision support system when classifying key processes and quality projects depending on the weight assigned by decision makers. It also includes two approaches, cost-effectiveness analysis and cost-benefit analysis. The first allows evaluating the complexity and the interest of the quality of the data, while the ACE focuses on its efficiency and its complexity to gratify any type of organization (profit or non-profit).

The approach is founded on a Deep Learning algorithm founded on previous validated and approved projects, to predict the quality of the business object, during the evaluation justifying the interest of improving the data quality. The approach includes a complementary phase where there may be collaborations betwixt the managers of the services who have common objectives for the improvement of common objects to minimize the total cost of the improvement of the quality.

The development of a JEE DQPEF platform is the fruit of current work, it is founded on the MVC model whose programming language is JAVA. The tool is put to the test in the healthcare field to help physicians choose improvement projects and identify important processes and objects in the practice.

7. PERSPECTIVE AND FUTURE WORK

The present work concerns the quantification of the factors of complexity and the impact of the completeness of the data. It makes it possible to identify future research avenues:

- The generalization of the approach for all dimensions such as precision, updating, updating, relevance, accessibility and consistency.
- The automation of the stage of calculating the values of the quality dimensions according to the previously chosen processes.
- The automation of collaboration betwixt the dimensions of quality without the intervention of process managers within subsidiaries of the same organization.
- The introduction of the principle of centrality to find the most significant dimensions in order to pay more attention to these dimensions of quality.

References:

Baldwin, A., Beres, Y., & Shiu, S. (2007). Using assurance models to aid the risk and governance life cycle. *BT Technology Journal, 2*5, 128-140. doi:10.1007/s10550-007-0015-7

Batini, C., & Scannapieco, M. (2016). *Data and Information Quality Dimensions, Principles and Techniques.* (S. I. Publishing, Éd.) doi:10.1007/978-3-319-24106-7

Batini, C., Barone, D., Mastrella, M., Maurino, A., & Ruffini, C. (2007). A Framework And A Methodology For Data Quality Assessment And Monitoring. *Proceedings of the 12th International Conference on Information Quality*, MIT. Cambridge

Batini, C., Cappiello, C., Francalanci, C., & A, M. (2009, july). Methodologies for data quality assessment and improvement. *ACM Computing Surveys, 41*(3), 1-52. doi:10.1145/1541880.1541883

Belhiah, M., B, M., Bounabat, B., & Achchab, S. (2015). Towards a Context-aware Framework for Assessing and Optimizing Data Quality Projects. *International Conference on Data Management Technologies and Applications*, (pp. 189-194). Colmar, Alsace, France. doi:10.5220/0005557001890194

Bertram, M. Y., Lauer, J. A., De, J. ,., Edejer, T., Hutubessy, R., Kieny, M.-P., & Hill, S. R. (2016). *Bull World Health Organ, 94*(12), 925–930

Capirossi, J., & Rabier, P. (2013). An Enterprise Architecture and Data quality framework. Dans P. Benghozi, K. D., & F. Rowe, *Digital Enterprise Design and Management* (Vol. 205). Berlin: Springer. doi:https://doi.org/10.1007/978-3-642-37317-6_7

Cellini, S., & Kee, J. (2015). Cost-Effectiveness and Cost-Benefit Analysis. In Dans K. Newcomer, H. P. Hatry, & J. S. Wholey, *Handvook of patical program evaluation*.

Chen, S. Siau. (2011). Impact of Business Intelligence and IT Infrastructure flexibility on Competitive Performance: An Organizational Agility Perspective. in ICIS,

Drummond, M. E., Sculpher, M. J., Torrance G, W., O'Brien, B., & Stoddart, G. L. (2006). Methods for the economic evaluation of health care programmes, 3rd ed. *Journal of Epidemiology and Community Health, 60*(9), 822–823.

English, L. (1999). *Improving Data Warehouse and Business Information Quality*.

Eppler, M., & Helfert, M. (2004). A classification and analysis of data quality costs. *Proceedings of the Ninth International Conference on Information Quality* (ICIQ-04).

Fishburn, P. (1967). Additive Utilities with Incomplete Product Set: Applications to Priorities and Assignments. *Journal of the Operations Research Society of America*. doi:10.1287/opre.15.3.537

Fister, D., Mun, J., Jagric, V., & Jagric, T. (2019). deep learning for stock market trading: a superior trading strategy. *International Journal on Non-Standard Computing anf Artificial Intelligence, 29*, 151-171

Grabisch, M., & Funaki, Y. (2011). A coalition formation value for games with externalities

Gallo, C. (2015). *Artificial Neural Networks: tutorial. Dans Encyclopedia of Information Science and Technology* (éd. 3rd Ed, Vol. 10)

Huang, K., Lee, Y. & Wang, R. (1999). *Quality Information and Knowledge.*

Kiron, D., Kruschwitz, N., Haanaes, K. and Velken, I. von Streng. (2012). Sustainability Nears a Tipping Point. *MIT Sloan Management Review, 53(*2), 69-74

Lesca, H., & Lesca, E. (1995). Gest*ion de l'information, qualité de l'information et performances de l'entreprise.*

Livermore, M., & Revesz, R. (2013). *The Globalization of Cost-Benefit Analysis in Environmental Policy*.

Maqboul, J., & Bounabat, B. (2017). Towards a Completeness Prediction Based on the Complexity'and Impact. *International Conference on Information Technology and Communication Systems(ITCS 2017)*, 640, pp. 108-116

Maqboul, J., & Bounabat, B. (2018). An Approach of data-driven framework alignment to knowledge base. P*roceedings to LOPAL '18*. Rabat

Maqboul, J., & Bounabat, B. (2020). Data Completeness Prediction by Deep Learning. *The 3rd International Conference on Modern Research in Engineering, Technology And Science.* Stockholm

Maqboul, J., & Bounabat, B. (2021). From Prediction of the Improvement of the Quality towards an Equitable Sharing of the Cost of the Improvement between Business Processes. Inter*national Journal of Recent Technology and Engineering (IJRTE), 8*(5).

McCulloch, W., & Pitts, W. (1943). A logical calculus of the ideas immanent innervous activity. *Bull Math Biophys, 5*, 115-133

Muennig, P., & Bounthavong, M. (2016). Cost-effectiveness analyses in health a practical approach. John Wiley & Sons.

Norman, T. J., Preece, A. D., Chalmers, S., Jennings, N. R., Luck, M., Dang, V. D., . . . Fiddian, N. J. (2004). Agent-based formation ofvirtual organisations. *International Journal of Knowledge Based Systems, 17*(2), 103-111.

Peleg, B., & Sudhölter, P. (2007). Int*roduction to the Theory of Cooperative Games.* Dans Theory and Decision Library.

Redman, T. (1996). *Data quality for the information age.* Artech House, Inc.

Rullière, J.-L. (2000). L'indétermination Et La Méthode De John F. Nash. *Revue économique, 51*(5) doi:10.2307/3503086

Scannapieco, M., & Catarci, T. (2002). Data quality under a computer science perspective. *Archivi & Computer, 2*, 1-15

Sharif, M. H., & Gursoy, O. (2018). Parallel Computing for Artificial Neural Network Training using Java Native Socket Programming. *Periodicals of Engineering and Natural Sciences (PEN), 6*(1), pp. 1-10

Simon, J., & De Prato, G. (2015). Is data really the new "oil" of the 21st century or just another snake oil? Looking at uses and users (private/public). 26th European Regional Conference of the International Telecommunications Society. San Lorenzo de El Escorial

Tamm, T., Seddon, P. B., Shanks, G., & Reynolds, P. (2011, Mars). How does enterprise architecture add value to organisations? doi:10.17705/1CAIS.02810

Wang, R. Y., & Strong, D. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems, 1*2(4), 5-33

De Wit, G. A., Verweij, A., Van Baal, P. H. M., Vijgen, S. M. C., Van den Berg, M., Busch, M. C. M., ... & Schuit, A. J. (2007). Economic evaluation of prevention; further evidence. *RIVM rapport 270091004.*

Maqboul Jaouad
ENSIAS
Mohammed V University
Rabat, Morocco
jaouad_maqboul@um5.ac.ma
ORCID 0000-0001-6987-971X

Bounabat Bouchaib
ENSIAS
Mohammed V University
Rabat, Morocco
bounabat@ensias.ma