

# Contextual Information Quality Assessment Methodology in Data Processing Using the Manufacturing of Information Approach

by

Mónica del Carmen BLASCO LÓPEZ

THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE IN  
PARTIAL FULFILLMENT FOR THE DEGREE OF  
DOCTOR OF PHILOSOPHY  
Ph. D.

MONTREAL, NOVEMBER 11, 2019

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE  
UNIVERSITÉ DU QUÉBEC



Mónica del Carmen Blasco López, 2019



This Creative Commons licence allows readers to download this work and share it with others as long as the author is credited. The content of this work can't be modified in any way or used commercially.

**BOARD OF EXAMINERS**

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Robert Hausler, Thesis Supervisor  
Department of construction engineering at École de technologie supérieure

Mr. Rabindranarth Romero López, Thesis Supervisor  
Faculty of civil engineering at University of Veracruz

Mr. Mathias Glaus, Thesis Co-supervisor  
Department of construction engineering at École de technologie supérieure

Mr. Mickaël Gardoni, President of the Board of Examiners  
Department of automated production engineering at École de technologie supérieure

Mr. Patrick Drogui, External Evaluator  
Research center water, earth, environment at University of Research

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND PUBLIC

OCTOBER 7, 2019

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE



## **ACKNOWLEDGMENT**

First, I would like to thank my family: my sons Axl and Harold, and my husband Paco, for all your love, patience, comprehension and encouragement. Special thanks as well to my parents Luis y Rocío, for their infinite love, support, comprehension and encouragement, and, especially, to my sister, Gaby, because she taught me that even in the face of life's greatest challenges, we are warriors who never give up.

I would also like to thank my director, my examiners, and the ETS student support team: Dr. Robert Hausler, Dr. Rabindranarth Romero López, Dr. Mathias Glaus, Dr. Rafael Díaz Sobac. Dr. Prasun Lala and Dr. Christine Richard.

Finally, a thank you to all my friends and colleagues who always encouraged me to continue: Audrey, Mag (my big sister), María, Raúl, Diana, Ana María; and to my parents-in law Francisco and Aracelly, my sisters-in-law Jeannie and Landy, and my brothers-in-law Rodolfo (both of them) and Jorge.



# **Méthodologie pour l'Évaluation de la Qualité de l'Information Contextuelle dans le Traitement des données avec l'Approche Manufacture de l'Information**

Mónica del Carmen BLASCO LÓPEZ

## **RÉSUMÉ**

Récents études ont révélé que l'évaluation de la qualité de données (DQ) et de la qualité de l'information (IQ) sont des activités essentielles pour les organisations qui cherchent l'efficacité de ses systèmes de communication et de l'information. Jusqu'à présent, les recherches en ce domaine ont porté essentiellement sur le développement d'approches, de modèles ou de classifications des attributs pour évaluer la DQ or la IQ. Cependant, sur les méthodologies pour l'évaluation de la DQ et de l'IQ dans un contexte spécifique, ils sont difficiles de trouver dans la littérature. Il y a des évaluations portant sur le traitement de documents de bureau en général, mais il y a un manque plus précis pour les formulaires.

Ce projet de recherche porte sur la nécessité d'un outil pour aide dans l'évaluation de la qualité des données entrant dans le système de communication et la qualité de l'information sortant du même système en prenant un formulaire comme le canal et en considérant le contexte dans lequel il est généré. Cette thèse propose une nouvelle méthodologie basée sur : 1) une adaptation de l'approche « Manufacture de l'information », qui considère la perspective du système de communication ; 2) un système de classification existant des attributs de la DQ que les établies en : intrinsèques, contextuelles, représentationnelles, et de l'accessibilité ; et 3) un nouveau modèle conceptuel, lequel fournit les lignes directrices pour le développement de l'outil nécessaire pour l'évaluation des formulaires. L'évaluation se fait considérant uniquement les attributs contextuels précédemment établis : complétude, quantité suffisante de données (ici, appelle suffisance), pertinence (l'accent mis sur le contenu), caractère opportune de l'information (l'accent mis sur le traitement) et la valeur réelle de l'information. Pour présenter l'applicabilité de la méthodologie CIQA (selon son sigle en anglais Contextual Information Quality Assessment) deux cas d'étude ont été présentées.

Les principaux résultats suggèrent que : en considérant la nouvelle représentation des données, ceux-ci peuvent être classifiés en accord à son type (indispensable et de vérification) et composition (simples et composés) ; dans l'une des deux cas d'étude, la quantité des données a été réduite 50 % en raison de l'analyse effectuée, signifiant une amélioration de 15 % dans l'IQ et une meilleure efficacité dans le système de traitement des données. La nouvelle rationalisation et structure du formulaire signifie non seulement une réduction dans la quantité des données, mais aussi une augmentation dans la qualité de l'information produite. Cela nous conduit à la conclusion que la relation entre la quantité de données et la qualité de l'information n'est pas une « simple » corrélation, la qualité de l'information augmente sans une nécessaire correspondance dans la quantité des données. En plus, la conception du formulaire signifie plus que l'aspect esthétique seulement, il signifie spécialement son contenu. En outre, dans le traitement du formulaire, des gains importants peuvent être obtenus en combinant l'évaluation de la qualité de l'information et l'informatisation des processus, pour éviter des problèmes comme l'excès de données, toujours en garantissant la sécurité des données.

**Mot-clé :** Qualité du donné, qualité de l'information, classification de données, manufacture de l'information, représentation de données, évaluation de la qualité de l'information, traitement des données.



# **Contextual Information Quality Assessment Methodology in Data Processing Using the Manufacturing of Information Approach**

Mónica del C. BLASCO LÓPEZ

## **ABSTRACT**

Studies have shown that data-quality (DQ) and information-quality (IQ) assessment are essential activities in organizations that want to improve the efficiency of communication and information systems. So far, research on the evaluation of DQ and IQ has focused on approaches, models or classification of attributes. However, context-specific DQ and IQ assessment methodologies are difficult to find in the literature. While assessment methodologies do exist for office document processing in general, there are none for forms. The focus of this thesis is the need for a context-specific tool with which to assess the DQ input and the IQ output in communication and information systems. The channel analysed for this purpose is the form. This thesis proposes a novel methodology based on: 1) an adaptation of the “manufacturing of information” approach, which adopts the communication-system point of view; 2) an existing DQ classification system that classifies attributes as intrinsic, contextual, representational or accessible; and 3) a new conceptual model which provides the guidelines for assessment of forms. This evaluation only takes into consideration established contextual attributes, such as completeness, appropriate amount of data (here called “sufficiency”), relevance (which emphasises content), timeliness (which emphasises process) and actual value. To present the applicability of the contextual-information quality assessment (CIQA) methodology, two representative forms were used as case studies. The main results suggest that a novel data representation allows data to be classified by type (indispensable or verification) and composition (simple or composite). In one of the two case studies, the data quantity was reduced by 50%, resulting in a 15% improvement of IQ and a more efficient document processing system. The streamlining and new structure of the form led not only to a reduction in data quantity but also to increased information quality. This suggests that data quantity is not directly correlated to IQ, as IQ may increase in the absence of a corresponding increase in data quantity. In addition, the design of the forms requires particular attention to content, not simply aesthetics. Furthermore, in data processing, there could be great benefits in combining IQ assessment and computerization processes, in order to avoid problems such as data overload; of course, data security would need to be considered as well.

**Keywords:** data quality; information quality; data classification; manufacturing of information; data representation; IQ assessment; data processing.



## TABLE OF CONTENTS

	Page
INTRODUCTION .....	1
CHAPTER 1 LITERATURE REVIEW .....	5
1.1 Information and communication technology, its relation and impact on the environment .....	6
1.1.1 Exponential data growth and its impact on the environment.....	6
1.1.2 Impact of data quality within organizations.....	8
1.1.3 Data overload or information overload, an historical debate.....	9
1.2 Information and communication systems .....	11
1.2.1 Relation between information and communication .....	12
1.2.2 Notion of information system .....	12
1.2.3 Notion of communication system .....	14
1.3 Data and information quality assessment and the value of information .....	17
1.3.1 Previous methodologies for information assessment.....	17
1.3.2 Some approaches for the IQ assessment.....	21
1.3.2.1 Manufacturing of information approach.....	22
1.3.3 Main quality attributes for data and information quality .....	26
1.3.3.1 Attributes: Accessibility and representativeness .....	28
1.3.3.2 Attributes: Intrinsic and contextual.....	29
1.3.3.3 Information value.....	31
CHAPTER 2 METHODOLOGICAL APPROACH AND RESEARCH OBJECTIVES .....	33
2.1 Research objectives.....	33
2.2 Manufacturing of Information and Communication Systems: MICS Approach.....	35
2.3 Structure process CD-PI-A .....	38
CHAPTER 3 METHODOLOGY .....	41
3.1 Classification of data [DC] and processing of information [PI] .....	41
3.1.1 Classification of data [CD] .....	41
3.1.2 Processing data into information [PI] .....	45
3.1.2.1 Data-unit value ( <i>duv</i> ) .....	45
3.2 Assessment [A], emphasis on content .....	47
3.2.1 Completeness .....	47
3.2.2 Sufficiency .....	49
3.2.3 Relevance.....	49
3.2.3.1 Relationship DIDV .....	50
3.2.3.2 Relationship RIC.....	50
3.3 Assessment [A], emphasis on process .....	51
3.3.1 Timeliness .....	51
3.3.2 Actual information value .....	55

3.4	Analysis cases .....	57
3.4.1	Analysis case 1.....	57
3.4.2	Analysis case 2.....	59
CHAPTER 4	RESULTS .....	63
4.1	Model [CD]-[PI]-[A] .....	63
4.1.1	Classification of data [CD] and processing data into information [PI] in both analyzed cases.....	64
4.1.1.1	Form F1-00 .....	65
4.1.1.2	Form FIAP-00.....	67
4.2	Assessment [A], content and process.....	67
4.2.1	Content: completeness, sufficiency and relevance .....	68
4.2.1.1	DIDV and RIC relationships.....	71
4.2.2	Process: timeliness .....	72
4.2.3	Actual Information value .....	74
4.3	Comparative analysis.....	76
4.3.1	Re-engineering.....	77
4.3.2	Emphasis on the pertinence of the content .....	79
4.3.3	Emphasis on the pertinence of the process .....	84
4.3.3.1	Timeliness.....	84
4.3.3.2	Actual value of information.....	86
CHAPTER 5	DISCUSSION .....	91
5.1	Contextual Information Quality Assessment (CIQA) methodology .....	91
5.2	Comparison with previous methodologies.....	96
5.2.1	Previous works in contextual analysis .....	98
5.3	Practical and research perspectives .....	99
5.3.1	Practical perspectives.....	99
5.3.2	Research perspective.....	101
CONCLUSION	.....	105
APPENDIX I	TABLE A.....	109
APPENDIX II	PUBLICATION DURING PH. D STUDY .....	111
BIBLIOGRAPHICAL REFERENCES	.....	135

## LIST OF TABLES

	Page
Table 1.1	Summary categories of information system. Source: DeLone & McLean, (1992).....11
Table 1.2	Symbols used in the representation of Manufacturing of information approach. Source: Ballou, (1998) and Shankaranarayanan (2006).....24
Table 1.3	Classification of data/information quality attributes.....27
Table 2.1	The elements of a communication system (CS) corresponding to the manufacturing of information MI approach .....37
Table 3.1	Data unit value (duv) for simple data, corresponding to the weight $w$ (which is related to its content). Dia: Indispensable data for authorization; Dis: Indispensable data for the system; Dv: Simple verification data; Dvv: Doble verification data .....46
Table 3.2	The considered variables in the timeliness evaluation of the emergency scenario corresponding to document processing scenario .....52
Table 4.1	Form F1-00. Form structure and $du$ classification. Ds = Simple data. Dc = Composite data. Dia = Indispensable data for authorization. Dis= Indispensable data for the system. Dv = Simple verification data. Dvv = Double verification data .....65
Table 4.2	Data classification and weighting. Frequency of accumulated data according to information type zone ( $D_{acc}$ ), relative frequency of accumulated data according information type zone ( $D_{rel_{acc}}$ ), information relative value ( $I_{rel}$ ), and accumulated information relative value ( $I_{rel_{acc}}$ ) for the F1-00 form.....66
Table 4.3	Data classification and its weighting, relative information value ( $I_{rel}$ ) and accumulated information value ( $I_{rel_{acc}}$ ) for FIAP-00 .....67
Table 4.4	Completeness assessment for F1-00 at data level ( $C^D$ ), at information unit level [ $C^{IP}(k)$ ] and at document level [ $C^{IP}(K)$ ].....69
Table 4.5	Completeness assessment for FIAP-00 form at data level ( $C^D$ ), at information unit level [ $C^{IP}(k)$ ] and at document level [ $C^{IP}(K)$ ].....70

Table 4.6	Different scenarios for the timeliness assessment according PTs proposed for the form F1-00.....	73
Table 4.7	Different scenarios for the timeliness assessment according PTs proposed for the form FIAP-00. min=minimum, ave=average, max=maximum .....	74
Table 4.8	Actual information value in three different scenarios according users' weight for F1-00 form .....	75
Table 4.9	Actual information value in three different scenarios according user's weight for FIAP-00 form.....	76
Table 4.10	Completeness assessment for F1-01 form. CD= completeness at data level. CIP(k)= completeness at information unit level. CIP(K)= completeness at document level .....	80
Table 4.11	Data classification and its transformation into information, frequency of accumulated data according to information type zone ( $D_{acc}$ ), relative frequency of accumulated data according to information type zone ( $D_{rel_{acc}}$ ), information relative value ( $I_{rel}$ ), and accumulated information relative value ( $I_{rel_{acc}}$ ) for the F1-01 form.....	81
Table 4.12	Results of relations DIDV and RIC for forms F1-00 and F1-01 .....	83
Table 4.13	Timeliness values that can be obtained if the process were constituted only by two exchange stations.....	84
Table 4.14	Comparative analysis of timeliness assessment for both FIAP-00 and FIAP-01 forms .....	85
Table 4.15	Comparative analysis of Actual Information Value for F1-00 and F1-01 forms.....	86
Table 4.16	Comparative analysis of Actual Information Value for FIAP-00 and FIAP-01 forms .....	87
Table 4.17	Contextual Information Quality Assessment for F1 and FIAP forms, versions 00 and 01 .....	88
Table A	Completeness and sufficiency values of F1-00, according to each user profile.....	109

## LIST OF FIGURES

	Page
Figure 1.1	Literature review general framework.....5
Figure 1.2	Classification of information systems according to the operational perspective. Source: Reix, (2002).....13
Figure 1.3	Elements that could impact on the success or fail of information systems. Adapted from DeLone & McLean, (1992,2003).....13
Figure 1.4	(a) a communication system whose main interest is the technical aspect or data transmission, and (b) a communication system whose main interest is the information production process. Source: Shannon, (1948); Fonseca Yerena et al., (2016) .....15
Figure 1.5	Main types of knowledge to develop an IQ methodology. Adapted from Batini & Scannapieco, (2016a).....18
Figure 1.6	Common phases in IQ methodology for IQ assessment. Adapted from Batini & Scannapieco, (2016a).....19
Figure 1.7	Data Quality attributes classification. Adapted from Bovee et al., (2003); Wang & Strong, (1996) .....28
Figure 2.1	Communication system as information-product oriented. Adapted from Ballou, (1998).....36
Figure 2.2	Structure process for the contextual information quality assessment methodology.....39
Figure 3.1	The sigmoid function. Source Chi et al., (2017).....53
Figure 3.2	(a) Timeliness evaluation function considering only resource quantity at t2. (b) Timeliness evaluation function considering only resource arrival time. Source: Chi et al., (2017) .....54
Figure 3.3	F1-00 Form, 8 sections, 32 champs. Retyped from real form. ....58
Figure 3.4	Information manufacturing process for the form F1-00 .....59
Figure 3.5	Structure of the FIAP-00 form.....60
Figure 3.6	Information manufacturing process for the FIAP-00 form.....62
Figure 4.1	Schematic representation of the [CD]-[PI]-[A] model .....64

Figure 4.2	Content composition of FIAP-00. Left bar <i>du</i> input, Right bar IP output of manufacturing of information system.....	72
Figure 4.3	Re-engineering of form F1-00. Here called F1-01 .....	78
Figure 4.4	Proposed re-engineering in FIAP-00 form processing, here called FIAP-01 .....	79
Figure 4.5	Left bar of both graphics: F1-00. Right bar of both graphics: F1-01. Graphic a) Data quantification comparative. Graphic b) Quality produced information .....	82
Figure 5.1	The overall vision of this research in a schematic way .....	93
Figure 5.2	Relationships between the three content contextual IQ attributes: completeness, relevance and sufficiency and the temporal pertinence attribute: timeliness .....	94
Figure 5.3	First practical perspective. Contextual Information Content Measurement for different profiles in the same form .....	100
Figure 5.4	Second practical perspective. A form management matrix to register improvements in information quality for updating documents .....	101
Figure 5.5	Research perspectives emerged from the research conducted. FW = future work .....	102



## **LIST OF ABBREVIATIONS**

AIMQ	Methodology for Information Quality Assessment
CB	Customer Block
CD	Component Data
CD(i)	Completeness at data units level
CD-PI-A	Classification of Data- Processing data into information- Assessment
CIHI	Canadian Institute for Health Information
CIP(k)	Completeness at information product units level
CIP(K)	Completeness at document level
CIQ	Contextual Information Quality
CIQA	Contextual Information Quality Assessment
CIS	Communication and Information System
CS	Communication System
DB	Document level
Dc	Composite Data
DI	Indispensable data
Dia	Indispensable data for authorization
DIDV	Data indispensable/data verification ratio
Dis	Indispensable data for the system
DPB	Document Processing Block
DQ	Data Quality
DS	Data Source

## XVIII

Ds	Simple Data
du	data-unit
dus	Data- units
duv	Data-unit value
duvset	Dataset data-unit value
DV	Data Vendor Block
DV	Verification Data
Dv	Simple verification data
Dvv	Double verification data
DWQ	Data Warehouses Quality
E	Sender
EB	Sender Block
F1-00	Access Form version 00
F1-01	Access Form version 01
FIAP-00	Administrative Information and Budget Form version 00
FIAP-01	Administrative Information and Budget Form version 01
FW	Future work
IC	Intermediate data component
ICT	Information and Communication Technology
II	Indispensable Information
IP	Information-product
IP	Consumer Block

IQ	Information Quality
IQESA	Information Quality Assessment Methodology in the Context of Emergency Situational Awareness
Irel	Information relative value
Irelacc	Cumulative relative information product
ISO	International Organization for Standardization
IU	Pre-processed information
MI	Manufacturing of Information
MICS	Manufacturing of Information and Communication System
OB	Organizational Boundary
P	Process
PB	Processing Block
PT	Processing time
QB	Quality Block
R	Recipient
RAE	Spanish-Academy Real Dictionary
RB	Receiver Block
RIC	Relation information content
RV	Relevance
S	Data/Information Storage
SB	Data Storage Block
SF	Sufficiency

XX

TD	Total Data
TDQM	Total Data Quality Management
TL	Timeliness
TQdM	Total Quality data Management
TQM	Total Quality Management
VA	Actual value of information
VI	Intrinsic Value
VI	Verification Information
Wr	Decision-maker importance's weigh
wu	Work unit

## LIST OF SYMBOLS

%	Percentage
CO <sub>2</sub>	Carbon dioxide
d	day
Kg/person	Kilogram per person
t/yr.	Tons per year



## INTRODUCTION

There has been growing interest in data quality (DQ) and information quality (IQ), due to their relevance to the improvement of the efficiency of information-management tasks (Batini & Scannapieco, 2016a). Additionally, DQ and IQ tools can help mitigate the impact on organizations of the exponential growth (and production) of data (Lee, Strong, Kahn, & Wang, 2002 ; Wang, Yang, Pipino, & Strong, 1998). It has been estimated that low-quality data has cost the United States economy 3.1 trillion dollars per year (IBM Big Data and Analytics Hub, 2016). Globally, this exponential growth of data has been accompanied by new needs and by consequences for humanity and the environment (Clarke & O'Brien, 2012 ; Gantz & Reinsel, 2012 ; Hilbert & López, 2012 ; Lyman & Varian, 2003). The growth in data-processing demand has in turn led to a need for larger communication and information systems (CISs) and higher- capacity data centres (Brunschwiler, Smith, Ruetsche, & Michel, 2009; Ebrahimi, Jones, & Fleischer, 2015; Floridi, 2009; Pärssinen, Wahlroos, Manner, & Syri, 2019). The consequences of this growth have already been demonstrated in data-overload and information-overload studies (Edmunds & Morris, 2000; Eppler & Mengis, 2004). These studies have pointed out that problems such as data overload affect not only individuals but also organizations and societies.

Studies show that processing large amounts of data at the organizational level without sufficient attention to data quality entails consequences such as: 1) client dissatisfaction, due to the reception of unrequested products or services; 2) increased production costs, due to the need for error correction; and 3) decreased employee satisfaction, due to the need to redo procedures (Redman, 1998b). However, quality—as well as quantity—is also a problem. The fact that organizations do not possess tools to evaluate IQ prevents them from monitoring improvements to their data-processing processes (Lee et al., 2002). Thus, since data processing is an essential part of any organization's CIS, the evaluation of office documents, such as application forms (which, obviously, contain data), should be one of the main challenges too.

One way to minimize the impact of data overload in organizations is assessment of the quality of the data input and information output of systems. Several models and approaches to the analysis of DQ have been reported in the literature (Ahituv, 1980; Ballou, Wang, Pazer, & Tayi, 1998; Bovee, Srivastava, & Mak, 2003; Michnik & Lo, 2009; Missier & Batini, 2003; Wang, 1998). With some exceptions (Masen, 1978; Ronen & Spiegler, 1991; Shankaranarayanan & Cai, 2006), most of these proposals use the terms “data” and “information” synonymously, which leads to confusion (Logan, 2012; Meadow & Yuan, 1997). Assessing the quality of information comprises three elements: 1) analysis of object attributes; 2) development of an approach and model; and 3) development of assessment methods and criteria. In the IQ assessment, one of the most common approaches—and one particularly relevant to the research described here—is the “manufacturing of information” approach, also known as the “information as a product” approach (Ballou et al., 1998; Wang, Yang, Pipino, & Strong, 1998). Some studies propose a framework for the classification of attributes that captures the aspects of data quality that are important to data consumers (Ballou & Pazer, 2003; Bovee et al., 2003; DeLone & McLean, 1992; Jarke, Lenzerini, Vassiliou, & Vassiliadis, 1999; Redman, 1998b; Wand & Wang, 1996; Wang & Strong, 1996). One of the most used classifications in DQ assessment is that proposed by Wang and Strong (1996), which groups attributes into four dimensions: 1) intrinsic; 2) contextual; 3) representational; and 4) accessibility-related.

From a communication-systems perspective, the manufacturing of information approach has two particularly noteworthy organizational elements: context and channel. The context provides a reference for communication, while the channel is the collector of the data that flows through the system and is ultimately transformed into information.

With regard to application forms as communications channel, there is a widespread false belief that “anyone can design a form”; those who espouse this belief typically focus on aesthetics and neglect the quality of the data the form will collect (Barnett, 2007). Deficient DQ adds costs to those already outlined above, and a poorly designed document can be considered a collector of mediocre data (Redman, 1998b).



All the considerations outlined above highlight the need for a new instrument with which to assess the quality of the data and the information that flow throughout the communication system. Applications forms flow daily through CISs and are always context-specific. The objective of the research described in this thesis was to fill this need, through the development of a new methodology which can be used to assess the contextual information quality (CIQ) in a data processing system and help improve its efficiency.

The remainder of this thesis is organized as follows. In the first chapter, the research topic is contextualized and the need for this research is emphasized through a review of the literature. In the second chapter, the methodological approach and specific research objectives are presented. In the third chapter, the methodology is described. In the fourth chapter, the results are presented, with some analysis of their implementation. The fifth chapter contains the bulk of the discussion. Finally, the thesis closes with the main conclusions and some recommendations.



## CHAPTER 1

### LITERATURE REVIEW

In order to present an overall picture of what it will be reviewed throughout this chapter, it is shown the following reference scheme (figure 1.1) of the concepts and themes described. A brief overview of the figure. 1.1 can be read as follows: There is a binomial 'data-information' relationship. This relationship has begun to be studied more in detail due to the impact of the data growing, which is reflected in society (at individual and organizational level).

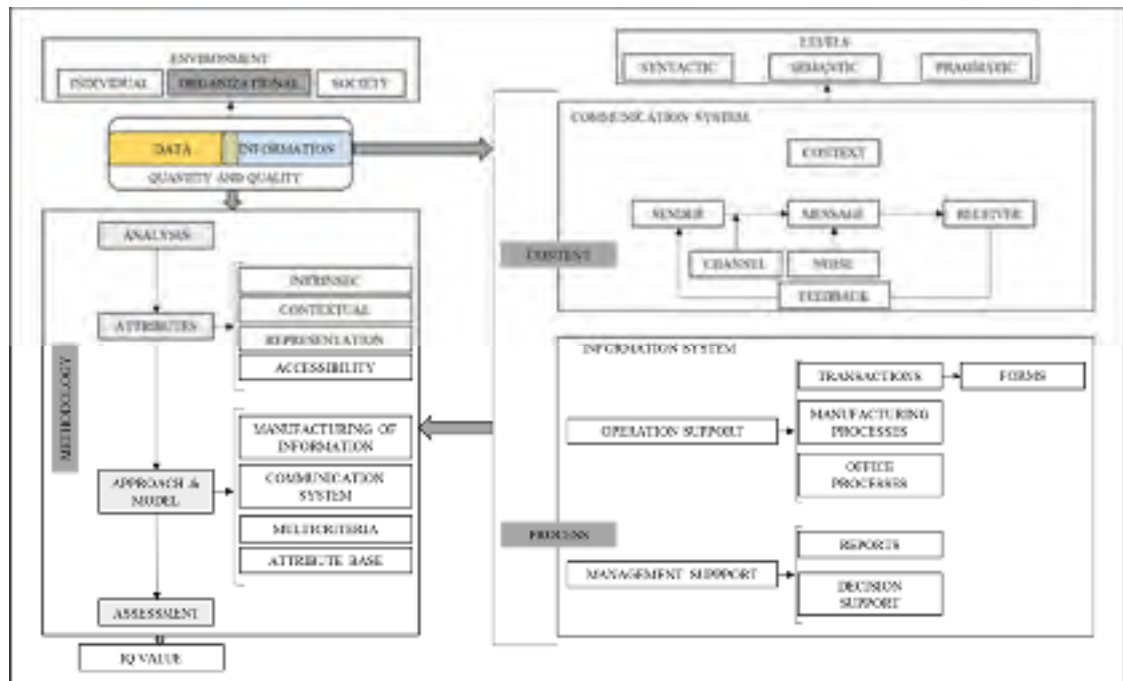


Figure 1.1 Literature review general framework

Data and information quality assessment has been seen as a possible mitigation measure for problems caused by the excess of data production. This binomial, also is part of the communication system, which can be studied from three different levels (syntactic, semantic and pragmatic). Seen as a process, the communication system within an organization is considered as an information system. In order to study, analyze and assess data and information quality within an organization, different methodologies have been developed. These

methodologies could consider different attributes to evaluate, approaches or models. The explanation of each concept and its relation to the research subject is extended below.

## **1.1 Information and communication technology, its relation and impact on the environment**

The passage from the industrial revolution to the information age<sup>1</sup> brings great changes (Floridi, 2009). Changes ranging from passing a real market to a virtual market, hierarchically structured organizations to other organized into large collaborative networks, from access to physical resources to unlimited access to digital resources, from tangible immovable to an intangible information technology (Earl, 2000). Some of these changes, sometimes highly energy demanding, could represent potentially environment damage (Floridi, 2009). For example, inside information technology systems, data centers (DCs) demand about 1.1 to 1.5% of the world's electricity consumption with an estimated annual increase of 15 to 20% (Ebrahimi et al., 2015). That, according to these estimates, by 2020 year, the ecological footprint of what encompasses information and communication technology (ICT) could overcome aviation industry (Floridi, 2009). This growth in the number of DC is due mainly to the increase in demand for data processing (Pärssinen et al., 2019). However, it seems that more attention is given to the technological advances to meet this demand than to the data processing itself.

### **1.1.1 Exponential data growth and its impact on the environment**

Different studies (Clarke & O'Brien, 2012; Gantz & Reinsel, 2012; Hilbert & López, 2012; Lyman & Varian, 2003) have tried to determine how much [data] information (prints, videos, magnetic and optical storage systems, etc.) is managed by mankind in recent times. For example, one study (Lyman & Varian, 2003) showed that the amount of new information [data] stored worldwide increased approximately 30% per year between 1999 and 2002. Certainly,

---

<sup>1</sup> The era in which the retrieval, management, and transmission of information, especially by using computer technology, is a principal (commercial) activity (LEXICO, 2019)

amounts change from one study to another since they use different measurement criteria and measure elements. However, one agreement among all studies is that: (1) the exponential growth of information [data] began with the internet revolution (shortly before the 2000 year), and (2) as the information natural processing grows, the worldwide processing technological capability grows exponentially, too. According to Clarke & O'Brien (2012), in 2010 the digital universe grew 1.2 zettabytes and its prediction for 2020 is that it will be 44 times as large as it was in 2009 (around 35zettabytes).

Regarding environmental effects due to the exponential data growth, we can mention two of the main impacts. The first impact is relating to information and communication technologies (ICT). Several studies have focused their attention on toxic substances and energy consumption caused by the ICT (Schmidt, Ereik, Kolbe, & Zarnekow, 2009). The total worldwide network servers power consumption is equivalent to the total consumption of the Poland economy (Kooimey, 2012). It is estimated that 1.3% of electricity global consumption in 2010 was just for power data centers (Fiorani, Aleksic, & Casoni, 2014). The energy used in these data centers is due to the sum of what they consume: 1) the IT equipment, 2) the cooling systems, and 3) electricity (Fiorani et al., 2014). The second impact is related to documents that flow daily through the organization, whether digital or printed. In both cases, these have environmental impact. For example, from one tree are produced around 80,500 sheets of paper. Early in the 21st century it was required around 786 million trees to ensure annual global paper supply (Lyman & Varian, 2003). Although within organizations has been chosen to digitize much of the documents, the use of paper at global level increases constantly surpassing 400 million t/yr. The global average is 0.055 t/yr. per person, but this distribution is not equitable. A person in North America consumes on average 0.215 t/yr. per person while a person in Africa, consumes on average 0.007 t/yr. per person (Kinsella et al., 2018).

Clarke & O'Brien (2012) have shown that document digitization does not solve either the ecological or the data excess problem. This happens because of a redundancy in both document formats, 52% of data in paper format are digitized, and that 49% of digital documents are

printed. Therefore, the change of document format does not contribute at all to the solution for the environmental damage caused by an excess of data.

### **1.1.2 Impact of data quality within organizations**

The previous section was referred to the impact of the document as a data transport medium within the information system. This section focuses on the content of these documents, the data and specifically in its quality. Since the amount of data starts to be a problem, a possible mitigation solution could be found in the exploration of its quality. However, a lot of information management are still not aware of the impact caused by low-quality data to their companies (Redman, 1998b). It has been estimated that low quality data has cost the United States economy 3.1 trillion of dollar per year (IBM Big Data and Analytics Hub, 2016). Also, other studies suggest that, on average, the financial impact due to low quality data to an organization, round 9.7 million dollars per year (Moore, 2017).

Redman (1998a, 1998b) suggests that some main impacts of low data quality inside organizations can be placed: at operational, tactical, or strategic level. At operational level, the low quality of data leads to unplanned events. Events that ultimately generate an increase in product cost. The low data quality can impact on: (1) the satisfaction's client (2) in the cost of production or (3) in the employee's satisfaction (Redman, 1998b). Some empirical studies estimate that the total cost of low-quality data for a company can be round between 8 and 12% of economic revenues (Redman, 1998a, 1998b). At the tactical level, the impact can be seen in the uncertainty degree generated for the missing accuracy in data collected for supporting the decision-making process. While the decision makers have the most relevant, complete, timely and accurate information, better decisions may take. Finally, the impacts at strategic level can be seeing in the difficulty to execute and end planned tasks due to the compromised ability to make decisions.

### 1.1.3 Data overload or information overload, an historical debate

Despite being two terms that cover different representations and meanings, the terms “data” and “information” are generally used interchangeably. On one side is the data that has been defined as a string of elementary symbols (Meadow & Yuan, 1997) that can be linked to a meaning related to communication and that can be manipulated, operated and processed (Yu, 2015). And on the other side it is the definition of information, a generally accepted definition is a coherent collection of data, messages, or signs organized in a certain way that has meaning or usefulness for a specific human system (Ruben, 1992). The meaning that give relevance to data transformed into information is determined by the context in which it happens. The meaning is defined in terms of what it does, rather than what it is (Logan, 2012). In a document office, the objectives and proceedings grant the meaning and usefulness level to requested data. Indistinct use of these two terms throughout history has generated confusion (Logan, 2012) in some cases and in others, frustration (Meadow & Yuan, 1997). One frustration reason is the inability to compare results between studies about information because they use different terms between them (Hayes, 1993).

Through the historical analysis of the information concept, we could find two great debates (among other less relevant), located at different historical times on the confusion caused by the interchangeable use of these two terms. The first debate was in the middle of the 20th century, concerning the quantification of information. On the one hand, Shannon (1948) presents his mathematical theory of communication. He held the idea that the information could be studied totally independent of its semantic aspect, that it could be measured using a probabilistic model, and that this model could be characterized as “entropy.” On the other hand, Wiener (1948), argued that information is just information, not matter or energy, that cannot be dissociated from its meaning, and that, if it had to be related to the thermodynamic concept of entropy, this last would represent the opposite of information. He, against Shannon, displayed the connection between the concept of organization and information. For him, the information is a measure of the organization degree in the system, and therefore entropy would be a measure of disorder degree in the system.

As an attempt to resolve this discrepancy, a year later, in 1949, Weaver published “his contribution to the communication mathematical theory” (Shannon & Weaver, 1949). In this document, he explained the theoretical basis of the concepts applied by Shannon and Wiener. He explained that both of them worked in the same communication process, but from different scope and perspectives. Also, he introduced the three levels of communication: a) the technical aspect, related to the accuracy of the transmission of symbols between the transmitter and the receiver; b) the semantic aspect, which works with the interpretation of the meaning of the receiver and, c) the level of influence, which refers to the degree of success that the receiver receives the meaning and causes a desirable behavior in him (Shannon & Weaver, 1949). Later, in 1956, Shannon was forced to explain the model scope, given the objectives and possible fields of the application of his theory because several articles of different branches of knowledge used indistinctly his theory. Claiming that his theory was not necessarily relevant to the analysis of phenomena within knowledge areas such as psychology, economics, social sciences (Shannon, 1956) and biotic systems (Logan, 2012). By a broader reference to this regard consult (Henno, 2014; Logan, 2012; Shannon, 1956; Shannon & Weaver, 1949).

The second big debate on the confusing interchangeable use of data and information concepts occurs when data begins to grow exponentially. The problem known as excess of information (information overload) arises with an information growth forecast of 300% between 2005 and 2020 (Gantz & Reinsel, 2012). This problem, according to studies, affects equally to individuals, organizations and society (Eppler & Mengis, 2004). This is the result of the imbalance between processing requirements and information processing capacities (Eppler & Mengis, 2004 ; Galbraith, 1974 ; Tushman & Nadler, 1978). The decision maker begins to experience the consequences of the problem when the amount of information involved in the decision begins to affect the decision-making process (Chewning & Harrell, 1990). Three dimensions of this problem are recognized: (1) affectations on individual skills (2) “too much paper” within organizations, and (3) the widespread affectation of clients' satisfaction levels (Butcher, 1995).



Due to the interchangeable use of the data and information terms, there is not a consensus about if the excess produced by them should be called data overload or information overload. Some researchers (Meadow & Yuan, 1997) argue that to suffer from an excess of information, the message must be received and understood, and not only received; otherwise, the effect of this would be: data overload and would not be, information overload.

## 1.2 Information and communication systems

Turning now to the system where the data and information flow, we have the communication and information systems, which are also referred interchangeably. Considering both as the system that allows us to communicate through the generation, transmission, data distribution and understanding of its result (Beniger, 1988). Literature shows that, although communication and information terms have converged on synonyms (Schement & Ruben, 1993) there are differences that are the basis for delimiting the study system. One of these delimitations is shown in table 1.1, which sum the categories of information system in an organizational context.

Table 1.1 Summary categories of information system. Source: DeLone & McLean, (1992)

<i>Shannon &amp; Weaver (1949)</i>	Technical Level	Semantic Level	[Pragmatic Level] Effectiveness or Influence			
<i>Mason (1978)</i>	Production	Product	Receipt	Influence on Recipient	Influence on System	
<i>DeLone &amp; Mc Lean (1992)</i>	System Quality	Information Quality	Use	User satisfaction	Individual Impact	Organizational Impact

Inside an organization, the communication at different levels (technical, semantic or pragmatic) can be related with the information quality system. Both, information and communication are social constructions that, additionally to be part of a wide phenomenon,

share common concepts (Schement & Ruben, 1993). Following, the relation between information and communication is explained.

### **1.2.1 Relation between information and communication**

Information is considered as an asset (Batini & Scannapieco, 2016a ; Wang, Yang W., et al., 1998). An asset with properties as: (1) be a finite asset, it is not exhausted even if it is consumed (Ballou & Pazer, 1985b); (2) be a symbolic essence, may be interpreted subjectively (Schement & Ruben, 1993); (3) be volatile, it means that its value depends on the time when it comes (Ballou et al., 1998); and (4) be difficult to control it in time (Schement & Ruben, 1993). To establish communication, it is necessary to have: the transmitter, who is the one who sends or transmits the message; the receiver, who receives it; the message, what contains the information to be transmitted; the channel by which the message is sent, and the context that establishes the rules of understanding.

### **1.2.2 Notion of information system**

The information system is a system, automatic or manual, which includes infrastructure, organization, people, machines, and/or organized methods to collect, process, transmit and disseminate the data which represents information for the user (Varga, 2003). The information system by which the communication is done can be classified according to the point of interest. One classification can be from an operational point of view (figure 1.2), which classifies the information systems into two main types: 1) of operational support, which aims to support the company's day-to-day operations and, 2) management support, which is responsible for supporting the decision-making process at the managerial level (Reix, 2002).

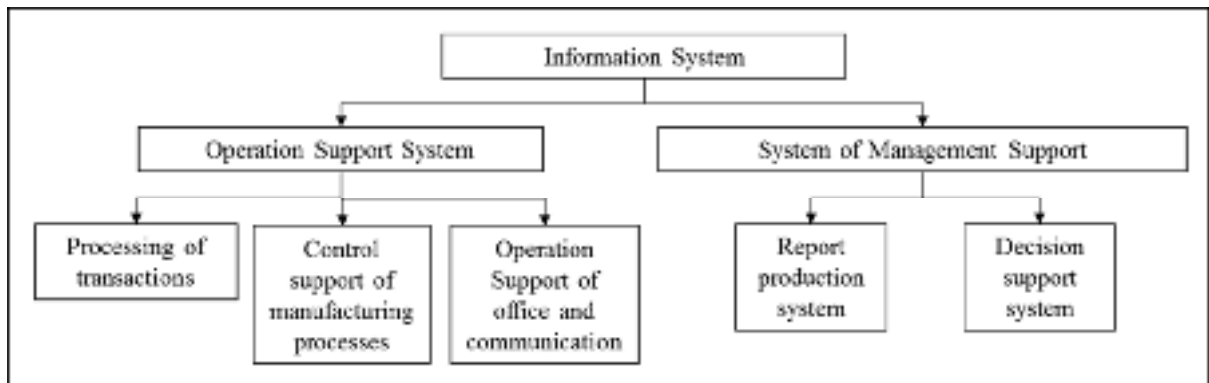


Figure 1.2 Classification of information systems according to the operational perspective.  
Source: Reix, (2002)

Another classification comes from a database perspective, this group has three kinds of information systems: 1) Of distribution, which deals with the possibility of distributing data and applications through a computer network; 2) Heterogeneous, which considers the semantic and technological diversities between the systems used to model data, and 3) autonomous, which has to deal with the degree of hierarchy and coordination rules defined by organizations in relation to information systems (Batini & Scannapieco, 2016b ; Ozsu & Valduriez, 2000). One third classification could be according to the type of activity in which it is dedicated to the organization or a specific department, these can be, for example: education, hospitable, government, administrative, accounting or finance, etc.

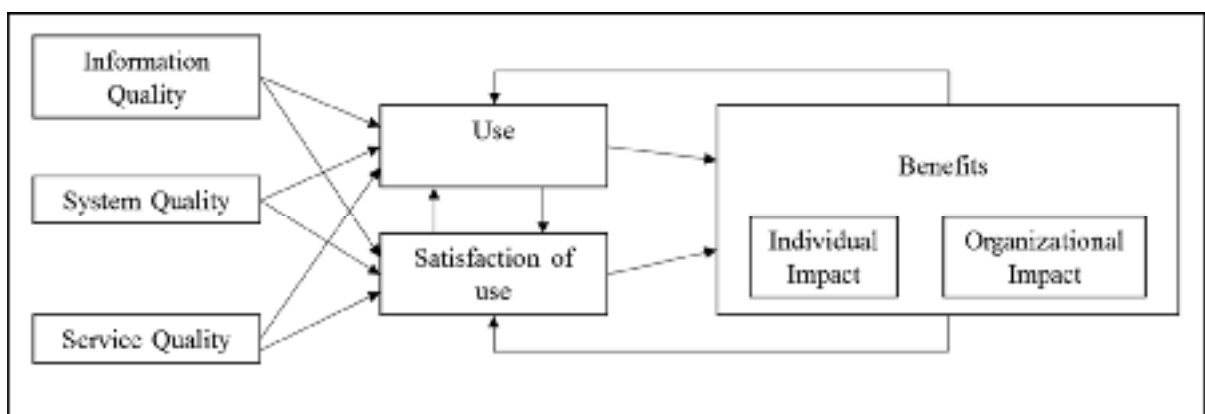


Figure 1.3 Elements that could impact on the success or fail of information systems. Adapted from DeLone & McLean, (1992,2003)

Further the classification, a framework is needed to placing the subject of information quality in relation to the system to which it belongs. Studies have been established in different relationships between information system elements (Figure 1.3): the quality of the system, the quality of the information, the usage and the satisfaction of use (DeLone & McLean, 1992, 2003). This helps to keep in mind that the success (or failure) of an information system does not depend solely on the quality of its content. It depends on information flows, exchanges between different organizational units (Batini & Scannapieco, 2016a) and the relations and influences among different system elements. These relations ultimately will have a positive (or negative) impact on the organization (DeLone & McLean, 1992, 2003).

### **1.2.3 Notion of communication system**

Turning now to the communication system (CS) can be seen as a manufacturing process of information in which signs are produced, transmitted and communicated. Information represented in a symbolic way finds its theoretical basis in the theory of signs (Masen, 1978). According to this theory, there are three levels of analysis,; syntactic, semantic, and pragmatic (Weaver, 1949). The syntactic level refers to the technical issues of sign transmission and its output is measured by the number of signs transmitted between a sender (E) and a recipient (R) (1948). The semantic level drives the relations between signs and the things or qualities that represents. The pragmatic level deals with the relations between signs and their users (Masen, 1978 ; Shannon & Weaver, 1949). There is one narrow relationship between the semantic and pragmatic levels. To reach the pragmatic level, we first have to pass through the semantic level. The semantic level is related to the meaning, which is closely linked to the context in which the communication is addressed. The pragmatic level refers to changes in the receptor's behavior due to the meaning conveyed in the message (Weaver, 1949). Studies (such as Reference (Masen, 1978)) have suggested that the output of an information system at the semantic level can be measured by the number of units of meaning (data signifying something to the recipient) handled by the producing unit during a given period. One remarkable difference between the syntactic and semantic levels is that the last mention appears in the context and the feedback elements. These two elements relate to the communication at the

semantic level, where the meaning gives relevance to the data and transformation into information is determined by the context. The meaning is defined in terms of what it does, rather than what it is (Reading, 2012). Figure 1.4a shows the transmission system (Shannon, 1948), referred to the syntactic level and figure 1.4b represents the communication system related to the information production process (Fonseca Yerena, Correa Pérez, Pineda Ramírez, & Lemus Hernández, 2016).

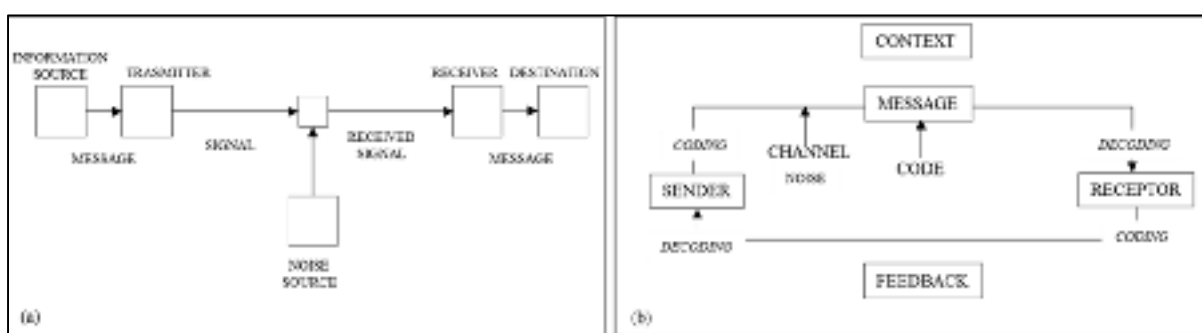


Figure 1.4 (a) a communication system whose main interest is the technical aspect or data transmission, and (b) a communication system whose main interest is the information production process. Source: Shannon, (1948); Fonseca Yerena et al., (2016)

The phenomenon of communication can be defined as the creation of meanings shared through symbolic processes (Ferrer, 1994). Although people involved in the act of communicating have different reference frames related to the time when the communication is giving (Schramm, 1980). The participants intend to achieve something in common through the message they are trying to share (Fonseca Yerena et al., 2016). This reference frame is given in the context in which the communication is made.

The elements of the communication system can be defined in the following way according to Berlo (1976) in (Fonseca Yerena et al., 2016):

1. *The sender, source (encoder)* is who originates the message, it could be any person, group or institution, that generates a message to be transmitted. Also, the sender is who *coding* the message.

2. *The receptor (decoder)* is the person or group of persons to whom the message is directed. The receptor is able to decode the message and respond to the communication.
3. *Noise* is referred as the barriers or obstacles that happen at any time in the communication process. They can be of psychological, physiological, semantic, technical or environmental type.
4. *Feedback* is the way in which occurs interaction or transaction between the sender and the receptor. With it, both parties ensure that the message was transmitted, received and understood.
5. *The message*, it is the content expressed and transmitted by the sender to the receptor. The message is composed of three elements: *the code* (a sign structured system), *the content* (what constitutes the message) and *the treatment* (way to communicate).
6. *The channel*. It is the vehicle by which the message flows, as an application form inside an organization.
7. *The context*. This refers to the physical, social or psychological environment shared by the transmitter and the receiver at the time of communication.

Regarding the channel, the most part of studies about documents (in general) consider only the electronic format (Bae et al., 2004 ; Bae & Kim, 2002 ; Chen, Wang, & Lu, 2016 ; Trostchansky et al., 2011) and with some exceptions (Forslund, 2007 ; Tyler, 2017) the quality of the content is evaluated. Additionally, the design of forms has usually been as something trivial that anyone can do (Barnett, 2007 ; Sless, 2018) regardless of consequences that a bad design can generate for the organization (Fisher & Sless, 1990). The quality of the information should be looked as the matter that in the end will lead to more efficient communication channels in modern enterprises (Michnik & Lo, 2009).

Considering the context, this should be explicit for at least two reasons: (1) because it provides communication efficiency, and (2) because it can be so fundamental that it becomes undetectable (Madnick, 1995). The context may vary mainly for three reasons: 1) because there are geographical differences, i.e. relationships between different countries; 2) due to functional differences, even between different departments within the same organization, some activities

may differ in their way of being realized; and 3) because organizational differences, the same document could have different meaning between different departments (Madnick, 1995).

### **1.3 Data and information quality assessment and the value of information**

Several definitions of quality have been proposed. In the International Organization for Standardization (ISO 8402:1994) this is defined as “the totality of characteristics of an entity that covers its ability to meet established and involved needs” (ISO, 1994). Other definitions are “suitable for use” or “according to the requirements” (Juran, 1989) and “a strategy focused on the needs of the customer” (Deming, 1986). Quality for products has been defined as “fitness for use” (Batini & Scannapieco, 2016a ; Ronen & Spiegler, 1991 ; Wand & Wang, 1996). Quality information is defined as the overall assessment of their fitness for use (Bovee et al., 2003). It depends on the user’s perspective. According to the context, one information could be relevant for one user and no-relevant for another (Bovee et al., 2003). For that reason, the data and information quality should be evaluated according to required attributes for the business. The IQ assessment methodology consider as one stage to establishing the needed attributes according to each organization.

#### **1.3.1 Previous methodologies for information assessment**

An IQ methodology can be defined as “guidelines and techniques that, on the basis of certain incoming information [data] concerning to a reality of interest, define a rational process that uses information [about reality] to assess and improve information quality from [produced by] an organization through stages and decision points.” (Batini & Scannapieco, 2016a). Three types of knowledge are needed in order to develop an IQ methodology: 1) organizational, 2) technological and, 3) of quality. The relations between these three knowledges are represented by arrows in figure 1.5. Please refer to (Batini & Scannapieco, 2016a) if it is required to meet the definition of each element.

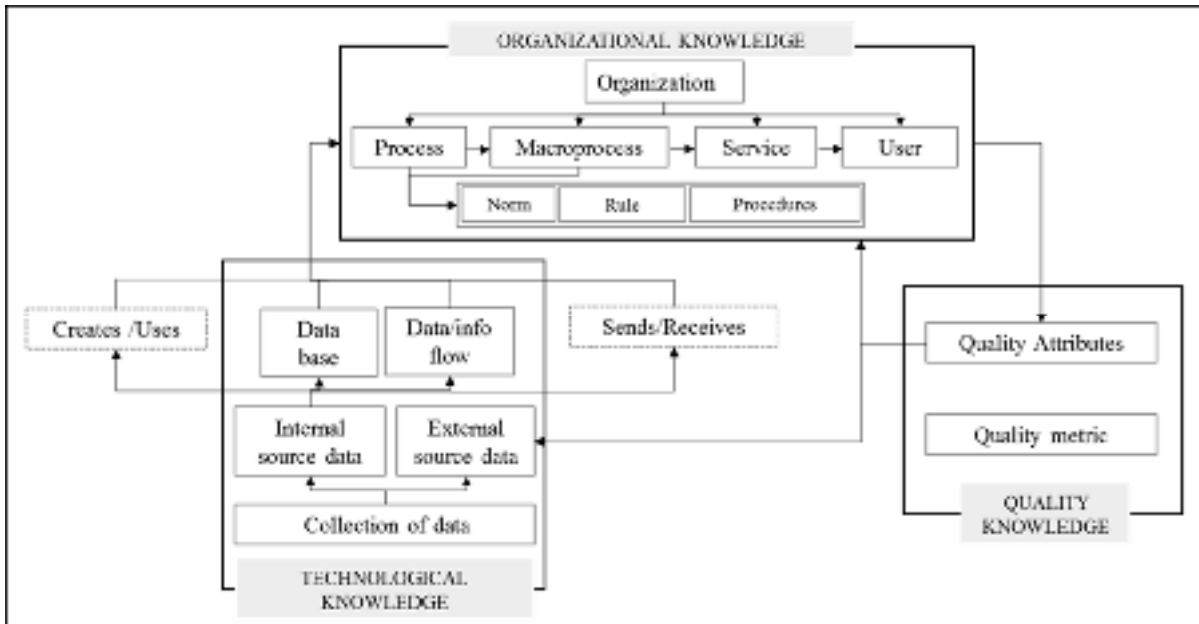


Figure 1.5 Main types of knowledge to develop an IQ methodology. Adapted from Batini & Scannapieco, (2016a)

As it is shown in Figure 1.5 there is a close relation between the (organizational) processes, data collection and quality attributes. For the development of the methodology, it is necessary to considering these three factors and their relations within the organization. A classification of IQ methodologies, according to its stated objective may be of three types: (1) guided by information / guided by the process; (2) of evaluation / improvement; (3) due to a general purpose / private purpose; or (4) intra-organizational / inter-organizational (Batini & Scannapieco, 2016a). However, it can be difficult to categorize a methodology in only one type. We could find a methodology guided by a process <sup>(1)</sup>, with a particular purpose <sup>(2)</sup> and with an assessment aim <sup>(3)</sup>. The phases commonly found in IQ methodologies are represented in Figure 1.6.



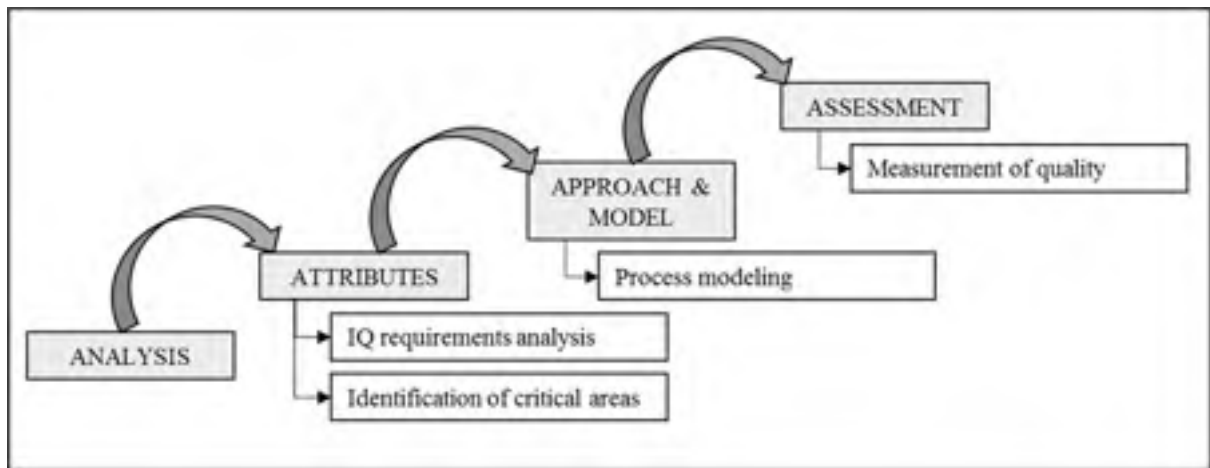


Figure 1.6 Common phases in IQ methodology for IQ assessment. Adapted from Batini & Scannapieco, (2016a)

At first, in the analysis phase, databases, schemas, and existing metadata are examined. Furthermore, interviews are conducted with the staff in charge to understand the architecture and rules related to the subject of analysis. Secondly is the phase devoted to the analysis of IQ attributes, here is carried out a survey among users and administrators on the themes and objectives of expected quality. Followed this it, the information base and the most relevant flows to assess it must be selected. In the next stage, approach and model, the researcher must take a more purposeful stance to design the process model to follow according to the previous analysis carried out. Finally, the quality is assessed based on this model, which can be: objective (based on quantitative measurements) or subjective (from qualitative assessments made by users and administrators (Batini & Scannapieco, 2016a).

There are diverse and valuable methodological proposals to evaluate the IQ (Ballou et al., 1998 ; CIHI, 2017 ; English, 1999 ; Eppler & Muenzenmayer, 2002 ; Jarke, Jeusfeld, Quix, & Vassiliadis, 1999 ; Lee et al., 2002 ; Wang, 1998). All of them have different objectives and scopes. Following the most representative will be commented.

Three different methodologies with different objectives and approaches are CIHI, DWQ and TQdM. The first is the *CIHI methodology* (2017), within the health sector of the Canadian Institute for Health Information. This methodology confirms the importance of data quality by

allowing observe directly the impact of it on the individuals and society that make up. This methodology was conceived considering that, given a better health data, doctors can perform better decisions, which ultimately impact on the overall health of the citizens. The second, the DWQ (Data Warehouses Quality), it regards data warehouse quality, is a methodology designed to assess the quality of this type of data (Jarke, Jeusfeld, et al., 1999), where its main contribution is the consideration of metadata to evaluate large amounts of data. The third, the *TQdM*, Total quality data management (English, 1999), it focuses on data quality of the entire company. This methodology sums that information quality is searching quality in all characteristics of information, the continuous process improvement of all processes of information system and increasing customer and employee satisfaction.

There are other two methodologies using one same approach as the basis, Manufacturing of Information (MI), whereas information is considering as a product (Ballou et al., 1998 ; Shankaranarayanan & Cai, 2006 ; Wang, Yang, et al., 1998). The approach taken as a basis for these two methodologies is born from the analogy existing between the concepts of quality related to product manufacturing and those pertaining to information manufacturing. The concept of manufacturing is seen as the transformation process of data-units (*du*) into information product (IP). Ballou (1998) in his work delved more into the description of the model than in the methodology itself. He proposes a method to evaluate the timeliness, data quality, cost and actual value of information in the manufacturing of information process.

The two methodologies that share the Manufacturing of Information approach are: [1] Total Data Quality Management [TDQM] Wang (1998), and [2] a Methodology for Information Quality Assessment (AIMQ) proposed by Lee et al. (2002). For the case of TDQM methodology, it consists in a survey -based diagnostic instrument for information quality assessment, a software tool to collect data and plot information quality dimensional scores given by information-product (IP) suppliers, manufacturers, consumers and managers. The most relevant of this methodology is the process definition in accordance with the manufacturing of information approach. This process considers phases as: definition (IP characteristics, IQ requirements, MI system), measure IP, analyze IP and improve IP. For the

case of AIMQ methodology, it considers three elements to reach a final assessment. The first element consists of a 2 x 2 matrix that generates the reference of what information quality (IQ) means for information consumers and responsible for handling it. The second element is a questionnaire to measure the IQ accordingly to the selected dimensions by consumers and handlers in the previous section. Finally, the third element consists of two interpretations and evaluation techniques captured by the questionnaire.

Heinrich, Hristova, Klier, Schiller, & Szubartowicz (2018) suggest that quality information measuring criteria must comply at least with certain conditions to carry out its objective which as first instance is to help in the decision-making processes. These criteria are: (1) the existence of a *minimum value* and a *maximum value* to lead the decision-making, a measurement with no maximum or minimum reference can lead to a wrong alternative. (2) the measurement of the data quality value must comply with a scale with intervals, the differences in values may have a meaning. (3) the measurement of data quality should reflect *the context application*, this means that the result must be objective, accurate and valid. (4) the measure of quality must have an *aggregation style*, this means, as well as it could apply to a data-unit it could be applied to the entire dataset in general. (5) these measures must comply with an *economic efficiency for the organization*. The configuration and implementation of quality measures should be guided from an economic perspective that allows both its commissioning and its results evaluation without this represents a major investment for the company.

### 1.3.2 Some approaches for the IQ assessment

Different approaches can be used depending on the analysis scopes. Two main scopes are: (1) a technical analysis (syntactic aspect) of the data transmission process and (2) an information analysis content (practical aspect). Related to the technical analysis, there are proposals that seek to explain the phenomenon for example by thermodynamic laws (Aleksi, 2011 ; Stonier, 1990). Related to the practical aspect, there are proposals that have the as subject of analysis (Ballou et al., 1998 ; Wang, 1998 ; Wang, Yang, Pipino, Strong, et al., 1998) and others that have the information system as subject of analysis (Ahituv, 1980 ; DeLone & McLean, 1992).

For the technical level, it is possible to use physical laws because we talk about a transmission system. But if we would study the semantic aspect of communication considering the information as another physical entity such as matter or energy, the analysis would be more complex. To date, this consideration has been expressed (Stonier, 1990), but it is still not completely accepted by the scientific community for its lack of verification.

For the analysis at the practical level, we can mention four different examples: 1) from a linguistic perspective (Batini & Scannapieco, 2016a) 2) from a multicriteria view (Michnik & Lo, 2009) 3) from the perspective of the communication system (Masen, 1978) and 4) from the analogy with a production system (Ballou et al., 1998 ; Shankaranarayanan, Wang, & Ziad, 2000 ; Wang, Yang, et al., 1998). The last two mention approaches consider that exist a transformation process within the interior of the system.

### **1.3.2.1 Manufacturing of information approach**







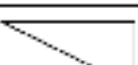

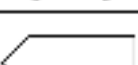
The approach Manufacturing of Information (MI) (Ballou et al., 1998) or Information as a Product (Wang, Yang, et al., 1998) is based on the analogy between a product manufacturing system and the information production system. Wang et al. (1998) make the distinction and point out differences between considering information as a product itself or by a product. In the by a product vision, the emphasis is focused on hardware and software, individual components control, the cost control, quality control implementation within systems and who is in charge is the director of the information technology department. In the case of the information as a product, the priority focuses on delivering quality information products to clients, quality control is over the product (the information) life cycle, which aims to prevent the entrance of garbage, rather than drive the output, and who is in charge of this is a manager in charge of the information-product. On his part Ballou et al. (1998), points out the existing similarities and differences between the information manufacturing process and the product manufacturing. First of all, the outgoing information from the system has a value that is transferred to the client. Secondly, they have similar parts of the system, such as: raw material, storage, assembly, processing, inspection, rework and bagging (formatting)

(Shankaranarayanan & Cai, 2006); Thirdly, the existing theory of what is known as Total Quality Management (TQM) represents a working guide in the production of high quality information-product. And perhaps the most interesting thing about this approach is that it supports a comprehensible assessment of the dimensions of information quality (Shankaranarayanan & Cai, 2006). Some of the limitations of this approach is the raw material nature. The problem generated by the nature of data is that, although they are used, these are not consumed, these can be reused indefinitely. Another fact is that produce multiple copies of the information-product does not generate a representative expenditure. To address this limitation researchers as Lee et al. (2002) makes a distinction between product and service as a final product.

In regards to the name of raw material and final product, Ballou et al (1998), handle the concept data quality; both for incoming data and for intermediate data (those who experience additional processes), and reserves the concept information quality to the final produced product delivered to the client, although he treats units under the same values.

In the literature there is a limited number of proposals which have a representation system for the information manufacturing process. Two proposals were found in this regard (Ballou et al., 1998 ; Shankaranarayanan & Cai, 2006). Ballou (1998), uses only five blocks to represent the model, while Shankaranarayanan (2006) proposed seven blocks to build its representation. In table 1.2 symbols are presented with their respective identification. To make the process representation of the case studies presented later, it was used the symbols proposed by Ballou, for adapting to the objectives set. Only raised a slight variation in the processing block. This is placed as a new block (document block processing) to emphasize document processing, which is concerning to this research.

Table 1.2 Symbols used in the representation of Manufacturing of information approach.  
Source: Ballou, (1998) and Shankaranarayanan (2006)

Symbol	Ballou (1998)	Shankaranarayanan (2006)
	Data Vendor Block (VB)	Data Source (DS)
	Processing Block (PB)	Information System Boundary
	n/a	Process (P)
	n/a	n/a
	Data Storage Block (SB)	n/a
	n/a	Data/Information Storage (S)
	Quality Block (QB)	Quality
	n/a	Organizational Boundary (OB)
	Customer Block (CB)	Consumer Block (IP)

Similarities between representations are briefly described below:

- 1) *Seller, of the source or data issuing block*: Both proposals agree to the use of this symbol to express the source of each piece of data (or dataset) that enter the system.
- 2) *Process block*: in this case, there is a difference between the symbols proposed by Ballou and Shankaranarayanan, as shown in table 1.2. However, the significance remains constant for both. This block represents the manipulation, calculations, or combinations that involve all or part of the raw data entering the system to produce the information-product.
- 3) *Storage block*: is used to represent the data-units (gross or combined) that await in some part of the system to continue processing. The symbol differs between both proposals.

- 4) *Quality or inspection block*: this block represents data quality verification that is used to produce the information-product. The same symbol is used in both proposals.
- 5) *Client block, consumer or information receiver block*: this block is used to represent that the information-product has been completed and delivered to the client. Representation remains the same for both proposals.
- 6) *Information system limits*: this is symbolism incorporated only in Shankaranarayanan proposals. It is used when the data-unit changes of a system (computer) to another (on paper). It is used to express when the document is moved from an information system to another.
- 7) *Limit the business or organizational system limit*: likewise, this representation appears only in Shankaranarayanan proposals. It represents the change between organizational units (departments) that suffer the data-units.

In the case of connectors, this is done through vectors with direction towards the data or pre-processing information goes, placing on top, according to the sequential number of the process. In Ballou case, only data-units (*du*) flow, assuming that these are changing as the process progresses. In Shankaranarayanan case, he considers that data are changing and evolving as they move forward in the process and is represented by what he calls component-data (CD). In our case, as we consider that once data enters the system and begins its processing, they are transformed into information, but that this is converted back to raw material for the next stage, we called that which flows through the process, *pre-processed information*, and is identified as IU.

Since this model is intended to process a document. The document is divided in their main sections, following the example presented (Bae et al., 2004 ; Bae & Kim, 2002). Based on this reference, another element was added to our representation. This helps in the identification of filling the document, but does not represent a drastic change in the essence of the base model. This new representation is a legend ( $d_n, wu_m$ ), where  $d$  is the document,  $n$  is the number of this (if it were the processing of various documents),  $wu$  represents the work unit that is referenced in that section of the processing, and  $m$  is the section number of such document. Another

adjustment made to this representation model is the inclusion of boxes up to the process blocks where the minimum and maximum processing time are pointed out.

### 1.3.3 Main quality attributes for data and information quality

As well as a product out of a production line has associated quality dimensions, the information-product (IP) has quality dimensions, too (Wang, 1998). Dimensions or attributes to assess the quality of the data varies from a proposal to another according to research objectives and the specific context. The attributes mostly shared in the literature on the quality of the data are: *accuracy, completeness, related dimensions to time and consistency* (Batini & Scannapieco, 2016a).

It has been taken two reference frames (Batini & Scannapieco, 2016a ; Wang & Strong, 1996) to give a general overview of the attributes of the data quality (DQ) or information quality (IQ) more named in papers, synthesized in table 1.3. This framework classifies attributes in four main categories: 1) intrinsic, which denotes that data have quality in their own right; 2) contextual, which underlines that data quality must consider the context of the task at hand; 3) representational, which concerns with the data that has to be interpretable, easy to understand; and 4) accessibility, which empathizes that the system must be accessible but secure.

The attribute category that interests us is the contextual since it is related to the semantic aspect of the communication system. A most detailed description of these kinds of attributes is presented in chapter 2 with the presentation of the methodological approach which is an adaptation of the manufacturing of information (MI) developed to perform the IQ assessment.



Table 1.3 Classification of data/information quality attributes.

		Wang & Strong (1996)	Wang & Wand (1996)	Ballou & Pazer (1985, 1995)	Delone & McLean (1992)	Reedman (1998)	Jarke & Vassiliadis (1999)	Bovee, Srivastava, & Mak (2003)	Michnik & Lo (2007)
Intrinsic	Accuracy/correctness	■		■	■	■	■	■	■
	Objectivity	■							■
	Believability	■					■		■
	Reputation	■					■		■
	Consistency			■					
	Freedom from bias /precision/ unambiguous		■						
	Credibility				■		■		
Contextual	Relevance	■			■	■	■	■	■
	Completeness	■	■	■	■	■	■	■	■
	Appropriate amount (sufficiency)	■				■			■
	Timeliness	■	■	■	■	■	■	■	■
	Value-added	■							■
	Usage / Usefulness				■		■		
	Informativeness				■				
Representational	Currency/level of detail				■				
	Interpretability	■				■	■	■	■
	Understandability	■							■
	Concise representation / readable/ clarity	■			■	■			■
	Consistent Representation	■			■			■	■
	Syntax						■		
	Version control						■		
Accessibility	Semantics						■		
	Accessibility	■			■		■	■	■
	Security	■							■
	System availability						■		
	Transaction availability						■		
	Convenience of access								■

The classification given by Wang and Strong (1996) considers four main categories: *Intrinsic*, *Accessibility*, *Contextual* and *Representational*. Considering attributes categories as filters according Bovee et al. (2003) first of all, data must be accessible, that means that it should be possible to access information that could be useful (*accessibility*). And, secondly, it must be interpretable, its meaning should be able to be found (*representativeness*). After this last step, intrinsic and contextual filters must be taken into account. The above description can be read in a descending manner in Figure 1.7.

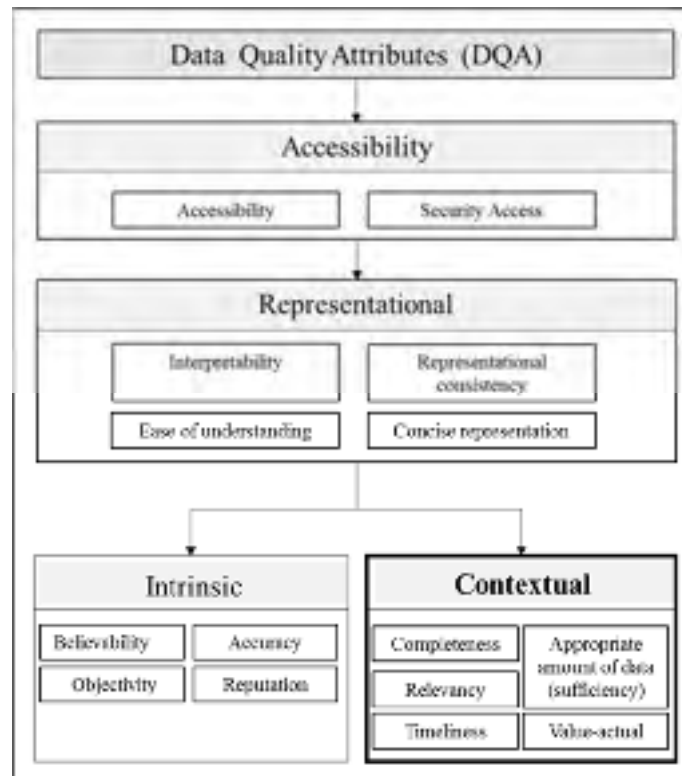


Figure 1.7 Data Quality attributes classification.  
Adapted from Bovee et al., (2003); Wang &  
Strong, (1996)

### 1.3.3.1 Attributes: Accessibility and representativeness

Firstly, it is the accessibility, since if the data potentially leading to information are inaccessible, the rest of the qualities become irrelevant for analysis (Bovee et al., 2003). The representativeness attribute is placed secondly due to, if data is intelligible, it is because they are taking with them a meaning for the receiver (Bovee et al., 2003). This can be understandable because they use the same code within communication, otherwise, there is not understanding and consequently the information processing is not carried out.

For example, in a school enrollment application form given the context, the educational institution is located in an English-speaking country and the responsible for processing these data only understands English language. The application form is answered in Chinese. In this

case, we have access to the user data, but while the form is filling in Chinese, the attribute belonging to the representativeness category, *interpretability*, is not given. This happens because there are different language codes between the sender and the receiver. Therefore, if information is inaccessible and/or non-interpretable, the evaluation of information quality cannot be finalized.

### 1.3.3.2 Attributes: Intrinsic and contextual

After accessibility and representational attributes, it is possible to perform the next data quality assessment either by their intrinsic attributes or their contextual attributes. The intrinsic attributes denote that data have some quality in themselves (Wang & Strong, 1996). These kinds of attributes are the most frequent and widely explored in literature (Ballou & Pazer, 1995 ; Ballou et al., 1998 ; Lee et al., 2002 ; Wang, Reddy, & Kon, 1995 ; Wang & Strong, 1996). Regarding contextual attributes, which mission is to emphasize DQ requirements that must consider the context given for a specific task (Wang & Strong, 1996). These have been very little explored. Only some contextual dimensions have individually been reported (Ballou & Pazer, 1985b) but not explicitly recognized as such. Some investigations (Bovee et al., 2003 ; Kaomea & Page, 1997 ; Lee et al., 2002 ; Wang & Strong, 1996) expressed the need for a new tool for analysis and evaluation of contextual information quality. Also, they recognize that the analysis must be adapted to the task in hand due to the context given (Botega et al., 2016 ; Kaomea & Page, 1997).

Contextual attributes (Wang & Strong, 1996) and work related to them are:

1) *Completeness*, the degree in which data presents enough scope to complete the task that is in charge (Ballou & Pazer, 2003 ; Botega et al., 2016 ; Scannapieco, Missier, & Batini, 2005 ; Shankaranarayanan & Cai, 2006). Completeness can be defined as the degree in which data are extensive enough to meet the scope of the task that is at the moment (Wang & Strong, 1996). An information-product is complete if this includes all data-units that define it (Redman, 1998a). Each of these studies works the completeness attribute in different ways according to each context. Shankaranarayanan & Cai (2006) considered three types of elements that flow

thought the manufacturing of information systems: 1) raw data elements, 2) simple data component, and 3) intermediate data component (IC). Also, they distinguish two types of completeness: 1) Context-independent completeness and 2) context-dependent completeness.

(2) *Adequate data amount*: as the name implies, refers to the amount of data to manage, it must be appropriate and sufficient for the task in hand (Botega et al., 2016 ; Michnik & Lo, 2009). It should reflect the amount of data that is not too little nor too much (Pipino, Lee, Wang, Lowell Yang Lee, & Yang, 2002). forewords it will be identified as “sufficiency” ” (SF), which refers to a quantity of something which is enough (Cambridge, 2019). Michnik & Lo (2009) works with all the attributes classification (Wang & Strong, 1996) but in a general way by a multicriteria analysis without give some specific case.

(3) *Relevance*: the degree to which, data is applicable and useful to perform the task that is in hand (Botega et al., 2016 ; Kaomea & Page, 1997 ; Michnik & Lo, 2009). Some studies (Botega et al., 2016 ; Kaomea & Page, 1997) qualify the relevance of data according to the situation in which it is giving. The relevance parameters are determined by the expert’s criteria. For example, in the case of combat aircraft, the aircraft pilots are those considered as experts (Kaomea & Page, 1997). In the case of the police alert situation, police departments are those considered as experts (Botega et al., 2016).

(4) *Timeliness*: The measure in which data age is appropriate for the task (Ballou & Pazer, 1995 ; Ballou et al., 1998 ; Botega et al., 2016 ; Chi, Li, Shao, & Gao, 2017 ; Kaomea & Page, 1997). Timeliness is an attribute used to reflect data update degree regarding the task in which you are using (Pipino et al., 2002). Another timeliness definition refers to information quality of the information given to the user at the time when it is still susceptible to influence their decisions and which decreases with a time elapse (« Termium Plus, data bank », 2018). The age of data refers to the length of time between it is recollected and it is used (Ballou et al., 1998). Some proposals also link the timeliness concept to the currency and volatility (Ballou et al., 1998 ; Bovee et al., 2003). The currency represents how long the data have been in the system (Wand & Wang, 1996) and the volatility is captured in a way analogous to the shelf

life of the product (Ballou et al., 1998). Ballou (1998) proposes the following as a measure of timeliness:  $Timeliness = \{max [(1-currency/volatility), 0]\}^s$ . Where exponent  $s$  is a sensitivity factor which depends on the task and the analyst judgment. Another methodology to evaluate the timeliness proposes a mathematical model to describe a real phenomenon of emergency-resources scheduling (Chi et al., 2017) which allows performing the evaluation more systematically and directly.

(5) *Added value*: This is defined according to the degree in which data provide a benefit and advantages for their use (Ballou et al., 1998). In this research, added value is considered as the value that is gained once the actual value of information is known. The actual value of the information corresponds to the value that the product has for the consumer ((Ballou et al., 1998 ; Wang, Yang W., et al., 1998). The actual value can be derived from several dimensions (Ballou et al., 1998). In this case, as we will only refer to the contextual attributes, actual value will be in function of the attributes of completeness, sufficiency, relevance, and timeliness.

### 1.3.3.3 Information value

The definition of “value,” either in the RAE (2017a) (Spanish Academy Real Dictionary by its acronym in Spanish) as in the Cambridge Dictionary (2018), it appears with more than 30 different meanings of the term. As the common denominator in all these forms is the dependence of the word in the context in which it is used. It is important to note here that the verb related to this action is to *evaluate*: estimate, appreciate, calculates the value of something (RAE, 2017b).

The word *value* in this thesis is used at different times, each one refers to different scenarios, which, to facilitate the reading of the reader, are described below:

- 1) *Value*: If it refers to the number or amount to a letter or a symbol represents (Cambridge, 2018), will just mention the word *value*. For example, “the IQ relevance value is 0.50.”

- 2) *Data-unit value (duv)*: this refers to the utility degree that rests on a specific data, by their composition and their type of use, given numerical way.
- 3) *Information Value*: As it has been expressed throughout this chapter, the information represents an economic value for the company (Redman, 1998b). From a technical perspective, Hayes (1993) makes the distinction between (1) the obtaining of a *syntactic value of information*, referring to (Shannon, 1948) work and (2) the obtaining of a *semantic value* of information, where he proposes theoretically how to calculate it. In the case of this thesis, the information value refers to information semantic value, even if the method of calculation is different from which Hayes is proposing. Here the information value is relative to the context.
- 4) The last of the meanings of the applicable term *value* refers to *actual value*, referred in Ballou et al., 1998) as the [actual] product value for the consumer, which depends on the intrinsic value (of information), its timeliness and the data quality explained more widely in the development of the thesis.

## **CHAPTER 2**

### **METHODOLOGICAL APPROACH AND RESEARCH OBJECTIVES**

In this chapter the specific objectives of this thesis and the adapted approach used as a guide in the developed methodology will be presented.

#### **2.1 Research objectives**

Researchers in the fields of data quality (DQ) and information quality (IQ) agree that: 1) theoretically grounded methodologies for DQ management are still missing (Wang, 1998); and 2) much more research is needed on the contextual aspects of information quality (Shankaranarayanan & Cai, 2006; Wang & Strong, 1996). This thesis takes into account the following assumptions and under-researched topics in the field:

##### **1) A transmission system is different to communication system**

It will be understood that there is an indissoluble relation between communication and information. If communication happens, so information is required. Otherwise, without information (only data) the system will be a data transmission system (Meadow & Yuan, 1997). Then, if it refers to a communication system, it is known in advance that it is also referring to an information system. Because the interest of this thesis is the semantic aspect related to the communication, it is assumed that the syntactic level of the system works technically well.

##### **2) The representation of data input**

Two studies using the Manufacturing of Information (MI) approach to measure quality have used a data-block (DB) representation (Ballou et al., 1998; Wang, 1998). The first has proposed a logical representation of the flow model (Batini & Scannapieco, 2016c). In this representation, all the entities that flow through the system are treated as physical-information items which can be either elementary or compound entities. The second distinguishes between data and information products (Shankaranarayanan & Blake, 2017). In this research, the (data-

unit) *du* structure is considered to constitute a DB, such as a document. This DB is composed of several data-units, and each *du* can be represented as a function of its particular characteristics for two types of materials: a pure (simple) material, and a composite material (formed from two or more elements).

Also, it is necessary to make a distinction between the three main stages (input, process and output) which passes the material, data-units. In the beginning, data-unit enters as raw material; in the manufacturing process, when they are still being manipulated and experiencing additional processes, they are called pre-processing information; and at the end, the product that comes out and is delivered to the client, is called information-product.

### 3) Input Data is not equal to Output Information

The contextual attributes—completeness, sufficiency, relevance and timeliness—have been measured in the same general way as the rest of the other attributes, such as objectivity, believability, accuracy, and consistency. Previous studies have estimated quality in terms of: 1) the weight given by different information providers (subjects), such as consumers, custodians and managers (Lee et al., 2002; Michnik & Lo, 2009; Wang, 1998); or 2) the weight given to the object in a given context (Botega et al., 2016; Kaomea & Page, 1997). In all cases, unless there is some transformation process, the value given to the data or to the information (depending on the terminology used) remains constant. In this research, the data value should change once the data is transformed into information.

### 4) Content, channel and process

In a manufacturing process, three equally important elements are considered: 1) what is processed (content); 2) how the content is transported (channel, document); and 3) how the content is processed (manufacturing process). For the content, relevance could be related to the characteristics of the raw data which have an impact on the information output. For the channel, there exists an unjustified belief that anyone can build a form (Barnett, 2007; Fisher & Sless, 1990; Sless, 2018), with no special attention to its role as a data collector in the communication and information system. We consider the design of the document that collects



the data as important as the quality of the collected data. We consider fields in the form data collectors. The data-collection process is evaluated in terms of timeliness. Some proposals consider timeliness to be related to the currency and volatility of data (Ballou et al., 1998; Wang et al., 1995). However, we consider that working with volatility (function of time and age) produces a less accurate result. Chi et al. (2017) have proposed a method that measures timeliness as a function of resources and time, which we consider a more accurate definition.

Given this framework, this research had three specific objectives:

- 1) Establish a process structure for the assessment of data- and information-quality related to data processing, based on a new representation of data and context.
- 2) Assess the relevance of information content collected in an application form, and the timeliness of its manufacturing of information process, using in both cases a performance index.
- 3) Perform a comparative analysis of scenarios with and without intervention in the application form (content and process), in order to evaluate the information product value, in light of the previously developed relevance and timeliness indices.

The methodological approach and the process-structure model used as a guide for the research described in the next chapter are presented in the next section.

## **2.2 Manufacturing of Information and Communication Systems: MICS Approach**

The MICS approach adapts both the manufacturing of information and the communication system perspectives for use in document processing. The manufacturing of information (MI) approach establishes an analogy between product manufacturing, a processing system that acts on raw materials to produce physical products (Wang, 1998), and information manufacturing, a process that transforms a set of data units into information products (Ballou et al., 1998). The MI approach starts by using many of the concepts and procedures of product quality-control to solve the problem of producing better quality information outputs (Ballou et al., 1998). This perspective helps to verify the raw material quality through its way to the manufacturing

process and to track the data-units (*dus*) inside the system before they exit it as the information-product (IP) (Shankaranarayanan et al., 2000). The communication system (CS), because it is seen as a manufacturing process for data, can be integrated into the manufacturing of information vision. This representation allows us to perform a kind of preventive control at data level, and a corrective control at information level.

Figure 2.1 shows the representation of communication systems (CSs) merged with the manufacturing of information approach, using an existing symbology (Ballou & Pazer, 2003; Shankaranarayanan & Cai, 2006; Shankaranarayanan et al., 2000). The elements of CSs corresponding to the MI approach are defined in table 2.1.

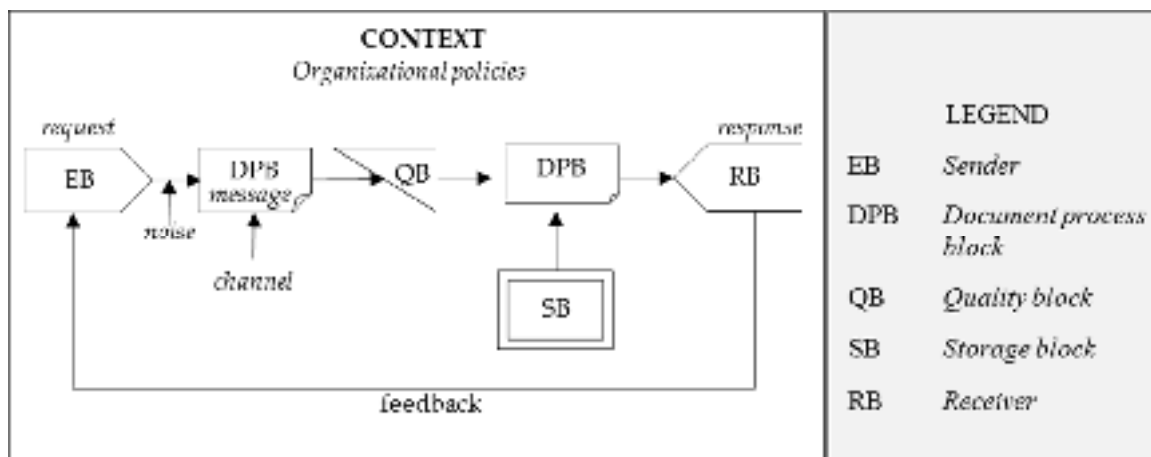


Figure 2.1 Communication system as information-product oriented.  
Adapted from Ballou, (1998)

Table 2.1 The elements of a communication system (CS) corresponding to the manufacturing of information MI approach

Element	Communication System	Correspondence in the MI Approach
<i>Sender (EB)</i>	This is who originates the message and encodes it. The sender can be the person who asks for a service and fills in a form, or a secretary who gathers user data to enter into the system.	This is defined as a vendor block or source block (Ballou, 1998; Shankaranarayanan & Cai, 2006).
<i>Receiver (RB)</i>	This is the person who receives the message. He/she has the capability to decode the message and respond to the communication.	This is defined as the client. The client is the person who receives the information product in the system (Ballou et al., 1998).
<i>Noise</i>	This can be defined as obstacles that arise at any point in the communication process (Fonseca Yereña et al., 2016).	The noise can be represented as irrelevant phrases or fields in a document.
<i>Feedback</i>	This is the method used by both parties to ensure that the message was transmitted, received and shared (Fonseca Yereña et al., 2016).	The feedback is observed once the receiver has received an answer and acted on the basis of the answer.
<i>The Message</i>	This is the content transmitted by the sender to the receiver. The message is composed of three elements: <i>the code</i> (a sign-structured system), <i>the content</i> (the message itself), and <i>the treatment</i> (way to communicate) (Fonseca Yereña et al., 2016).	The message can be placed in three blocks: 1) a document process block (DPB), where the transformation of data into information is carried out; 2) a quality block (QB), where the quality of the content is analyzed; it is expected that the output stream has better quality than the input stream (Ballou et al., 1998); 3) A storage block (SB), where the <i>du</i> set is stored and made available to additional processes (Ballou, 1998; Shankaranarayanan & Cai, 2006).
<i>Channel</i>	This is the vehicle that carries the message. Through this channel, data is collected, processed, and transformed into information.	This can be an office document, for example. In this research, the stored content in the document (application form) is the system of analysis.
<i>Context</i>	This refers to the physical, social, or psychological environment in which the sender and the receiver are located at the time of communication (Fonseca Yereña et al., 2016). The context is a distinctive and necessary element in the CS, since its purpose is to give meaning to communication.	In this case, the context is given by the proceedings and policies of the business.

With some exceptions (Masen, 1978; Ronen & Spiegler, 1991), studies using the MI approach have considered the terms “data” and “information” to be synonyms (Bovee et al., 2003; Lee et al., 2002; Scannapieco et al., 2005; Wang, 1998; Wang & Strong, 1996), which leads to confusion (Logan, 2012 ; Meadow & Yuan, 1997)]. To avoid this ambiguity, this research distinguishes between these two terms on the basis of their moment of processing.

If one views the CS from the perspective of the MI approach, it is possible to distinguish three main stages in data processing:

1. the inputting of raw material (data);
2. the processing period, where data is transformed into pre-processed information. Information is considered pre-processed if the information output from one phase is the raw material for the next phase;
3. the outputting of the finished product (the information product obtained at the output of the system).

### **2.3 Structure process CD-PI-A**

The purpose of the CD-PI-A process is to explore the effectiveness of representing the composition of data in contextual information quality assessment (CIQA). The CD-PI-A structure process takes the MICS approach as its starting point. This process comprises three phases: 1) classification of data [CD]; 2) processing data into information [PI]; and 3) assessment of information quality [A]. Each phase comprises sub-phases. Each phase will be presented in detail in the next chapter.

The foundation of this structure process is the distinction between data and information. In addition, it is assumed that: 1) the communication system is technically adequate; 2) the office document referred to is a form that belongs to an administrative process; 3) the form is the communication channel in the simplest information system (see (Denning & Bell, 2012); and 4) the flow of the form through the organization is dictated by objectives and policies. The [CD]-[PI]-[A] process presented in figure 2.2 is the framework for the contextual quality-information assessment methodology presented in the next chapter.

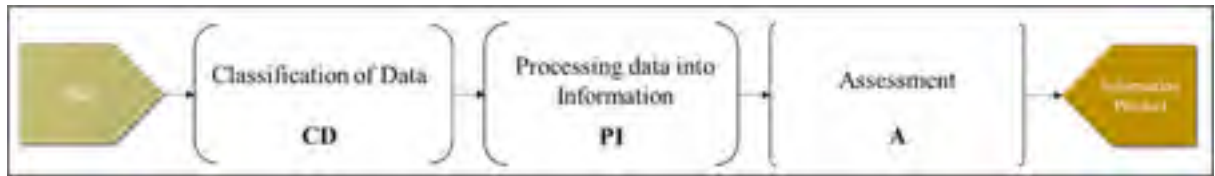


Figure 2.2 Structure process for the contextual information quality assessment methodology



## **CHAPTER 3**

### **METHODOLOGY**

This methodology is based on the phenomenon examination in its real context through two case studies. The case study method uses qualitative and quantitative evidence to examine phenomena of real life [such as reference (Yin, 2002)] to lead us to a better understanding of how and why certain events happen. This method of study is widely used in DQ & IQ investigations (Madnick, Wang, Lee, & Zhu, 2009).

As it is shown in the CD-PI-A structure process, it was broken down into three phases: [CD], [PI] and [A]. In this chapter, in point 3.1 the first two phases [CD] and [PI] will be presented. In point 3.2, the third phase corresponding to the assessment [A] will be presented. In point 3.3 the actual value of information will be described. This value takes into account the two values before obtained in the assessment phase (relevance and timeliness) and the user's criteria.

#### **3.1 Classification of data [DC] and processing of information [PI]**

In this section, in first instance, the data classification system (according content and composition) will be presented. In second instance, the followed system to represent the transformation of data into information will be explained.

##### **3.1.1 Classification of data [CD]**

Classification involves the process of grouping data into different categories according to similar characteristics (Han & Jian Pei, 2012). Data is tagged and separated in order to form the groups. In this case, tags are put onto form fields. The classification is made in accordance with the results of semi-structured interviews with the processors of the form. The processors are considered to be skilled and experienced workers in information product manufacturing.

The fields (data collectors) are each recognized as a unit that will host one datum. We consider two types of data representation criteria. It is assumed that each type is associated with a fixed value. The first criterion is its composition. The composition representation has one sub-classification: 1) *simple (or pure) data*, which considers one symbol to contain only one word, one phrase, one choice box, or, in general, one unit corresponding to one and only one piece of data; and 2) *composite data*, which is a compound of more than one simple piece of data (more extensive explanation below). The second criterion is its content, which corresponds to the degree in which it is placed, according to importance and frequency-of-use scales. Likewise, the content representation has one sub-classification: 1) *indispensable data*, which corresponds to data that is absolutely necessary; and 2) *verification data*, which is used to check the indispensable data. For this second criterion, the order system and the frequency of use are facts dependent on the context. In an office document, the objectives and proceedings, considered as the context, grant the meaning and usefulness levels of the requested data. We denote TD (Total Data) as all incoming data units to the system, classifying them as follows:

1. For their *composition*, the data units can be tagged into two types: 1) simple or 2) composite.

*1.1 Simple (Ds).*  $D_s = \{D_{s_i} \mid i=1, \dots, I\}$ . This is the set of simple data units, where  $D_{s_i}$  is the  $i^{th}$  data unit and  $I$  is the total number of simple *du*s. This type of *du* is composed of one and only one element; such as a name, local identification number, date, signature, and so on. In its transformation into information, the data unit takes the weight value  $w$ . The value of  $w$  is assigned according to the content classification, which is explained via:

$$D_s = w. \quad (3.1)$$

*1.2 Composite (Dc)*  $D_c = \{D_{c_k} \mid k = 1, \dots, K\}$ . This is the set of data unit composites, where  $D_{c_k}$  is the  $k^{th}$  *du* and  $K$  is the total number of composite data units. This type of data unit is a compound of two or more simple data units, which can be, for example, a registration number, social security number, institutional code, and so on. In its transformation into



information, the corresponding weight  $w$  is multiplied by the factor  $x$ , which depends on the number of simple data ( $Ds$ ) units that form the composite data unit:

$$Dc = wx, \quad (3.2)$$

where

$$x = \sum_{s=1}^n Ds. \quad (3.3)$$

2. For *content*, the data units are classified into two types of data representation. These two types of data are indispensable and verification data.

From this classification, the weight value,  $w$ , is assigned. The weight  $w$  is given by the personnel in charge of carrying out the process, since it is assumed that they have the best knowledge of the criteria of data unit importance and the frequencies of use required to process the document. A comprehensive and elaborate case study, presented in Reference (Tee, Bowen, Doyle, & Rohde, 2007), argues that, through the use of interviews and surveys as a method of analysis, it is possible to examine the factors and the levels of influence of data quality in an organization.

This weight captures the relative importance of a data unit within the process in question. We propose the use of a quantitative scale of discrete values, from 4 to 1, to classify the document fields. The field (or *du*) is classified according to the importance degree for the document processing and the frequency of its use, where 4 corresponds to *very important and always used*, 3 to *important and always used*, 2 to *slightly important and not always used*, and 1 to *not at all important and not always used*.

- 2.1 *Indispensable data* (DI),  $DI = \{Dia + Dis\}$ . This type of data unit always appears at some stage in the process and can be one of the following two types:

2.1.1) Authorization (Dia):  $Dia = \{Dia_m | m=1, \dots M\}$ . This is the type of indispensable *du* for authorization, where  $Dia_m$  is the  $m^{th}$  data unit and  $M$  is the total number of indispensable *dus* for authorization. This type of *du* corresponds to the highest value of the weight  $w$ , since it is considered to be a very important *du* for processing. Without this, the system cannot produce the information product. This depends on the approval (or rejection) given by the responsible personnel, according to the policies or organizational procedures.

2.1.2) System (Dis):  $Dis = \{Dis_n | n=1, \dots N\}$ . This is the set of *dus* indispensable for the system, where  $Dis_n$  is the  $n^{th}$  *du* and  $N$  is the total number of indispensable *dus* in the system. This data type is considered to be important. This *du* type is essential within the process and, usually, they correspond to questions such as: who, what, when, where, why, and who authorizes. Without them, the processing of information cannot be completed.

Verification data (DV).  $DV = \{Dv + Dvv\}$ . This *du* type is found frequently during processing; although, in some cases, document processing is carried out without them. This type of *du* can be of two types:

2.2.1) Simple verification data (Dv).  $Dv = \{Dv_s | s = 1, \dots S\}$  This is the simple verification *du* set, where  $Dv_s$  is the  $s^{th}$  *du* and  $S$  is the total number of simple verification *dus*. Some decision-makers consider it necessary to have this kind of unit to make the decision-making process safer (Ackoff, 1967). However, without some of these *dus*, data can still be processed. This type of *du* is sometimes used for processing, and it can be considered slightly important.

2.2.2) Double verification data (Dvv).  $Dvv = \{Dvv_t | t= 1, \dots T\}$ . This is the double verification *du* set, where  $Dvv_t$  is the  $t^{th}$  *du* and  $T$  is the total number of double verification *dus*. This *du* type is rarely used to verify essential data and it may be not at all important to processing but, in some cases, they are still requested.

### 3.1.2 Processing data into information [PI]

In a communication system, there must be a context which serves as a benchmark to determine the pertinence of a *du* in communication. The manufacturing process of information is considered the transformation of raw material, data, into finished products, information. This transformation is represented by the weighting of data after classification (for composition and content).

Data transformation into information leads us to give a value to the data units which are at the intersection of the composition and content classifications. Therefore, the possible resulting sets are of two types: 1)  $D_s \cap Dia$ ;  $D_s \cap Dis$ ;  $D_s \cap Dv$ ;  $D_s \cap Dvv$ , where the value of the data unit (*duv*) corresponds to the weight  $w$  assigned according to the importance and frequency of use criteria mentioned above; and 2)  $D_c \cap Dia$ ;  $D_c \cap Dis$ ;  $D_c \cap Dv$ ;  $D_c \cap Dvv$ , where the *duv* corresponds to the weight  $w$  multiplied by the  $x$  factor. It is clear that all these sets are mutually exclusive.

Finally, at the system exit, information output is the result of the intersections mentioned above and is grouped in the following manner:

1. *Indispensable information (II)*, which is the result of transforming indispensable *du* (simple or composite, catalogued as either for authorization or for the system transformation) into information through its corresponding *duv* assignment.
2. *Verification information (VI)*, which is the result of transforming verification *du* (simple or composite catalogued as either as simple verification or double verification), into information through its corresponding *duv* assignment.

#### 3.1.2.1 Data-unit value (*duv*)

To determine the data unit value (*duv*), the combination of both data classifications (composition and content) must be taken as a reference; that is, for its composition (simple or

composite data), and for its contents (indispensable or verification). Table 3.1 shows the values already mentioned.

Table 3.1 Data unit value ( $duv$ ) for simple data, corresponding to the weight  $w$  (which is related to its content). Dia: Indispensable data for authorization; Dis: Indispensable data for the system; Dv: Simple verification data; Dvv: Doble verification data

Attribute: content	w
Dia	4
Dis	3
Dv	2
Dvv	1

In a form, there is usually more than just one type of data; therefore, it is necessary to calculate the data unit value for the same dataset. This is called  $duv_{set}$  and it is calculated by the following equation, where  $f$  is the frequency of the same type of data.

$$duv_{set} = f(duv). \quad (3.4)$$

The information relative value ( $Irel$ ) for the document, as an information product, will result in a value between 0 and 1, where 0 corresponds to a null value and 1 to the total of the information product contained in the document.  $Irel_i$ , for one type of information will be calculated from the following equation, where  $i$  is the set of same type of data (Dc/Dia, Ds/Dia, Dc/Dis, Ds/Dis, Dc/Dv, Ds/Dv, Dc/Dvv, Ds/Dvv) and  $N$  the total sum of all  $duv_{set}$ .

$$Irel_i = \frac{duv_{set(i)}}{N(duv_{set})} \quad (3.5)$$

The cumulative relative information product ( $Irel_{acc}$ ) calculation is performed according to the following classification:

1. *Information product of the indispensable units (II)*, this type of IP results from indispensable (simple and composite) *du*. It must be ordered as follows: first, the information derived from the authorization type (Dc/Dia, Ds/Dia); and second, for the system (Dc/Dis, Ds/Dis):

$$IIrel_{acc} = \sum Irel(II). \quad (3.6)$$

2. *Information product of the verification units (IV)*. This type results from simple verification and double verification data units. It must be ordered as follows: first, the information that corresponds to Dc/Dv and Ds/Dv; and second, the information that derives from the double verification du (Dc/Dvv, Ds/Dvv):

$$IVrel_{acc} = \sum Irel(IV). \quad (3.7)$$

### 3.2 Assessment [A], emphasis on content

Regarding contextual factors, such as completeness, sufficiency, and relevance, the quality of data and information falls not only on decision-maker but also on the decision task (Shankaranarayanan & Cai, 2006). Decision-makers must be able to weigh the type of data (indispensable or verification) in relation to the decision task.

#### 3.2.1 Completeness

Besides relevance to the decision task, an information product is considered complete if it includes all data units needed by the decision maker for the decision task in hand (Shankaranarayanan & Cai, 2006). It is expected that the output product, at least has the essential parts that constitute it to consider it complete. That means that the completeness is a measure of how complete an IP is in terms of the data units that are included in the IP

(Shankaranarayanan & Cai, 2006). So, in the application form, it must have the indispensable data but also it could have data of the verification type.

We, as some other proposals about completeness (Ballou & Pazer, 1985b ; Shankaranarayanan & Cai, 2006) share the fundamental logic: completeness is a construct that contains both objective and contextual components. We adapted the Shankaranarayanan & Cai (2006) proposal to use it in the measurement of completeness in the application form. Meanwhile they distinguish two types of completeness (context-independent and context-dependent). We consider only one type of data at the entrance previously classified, and only one final information product at the exit. In their method, he first computerized the context-independent completeness and after the context-dependent completeness. For our part, already having the data classified and weighted, we calculate in first the completeness at the data units level  $[C^D(i)]$ , in second, the completeness at the information product unit level  $[C^{IP}(k)]$  and in third, the completeness at the document level (DB),  $[C^{IP}(K)]$ .

According to Shankaranarayanan & Cai (2006), the completeness of the data-unit  $i$  could be defined by simply binary operation as follows:

$$\begin{aligned} C^D(i) &= 0 \text{ if the value of } i \text{ is missing} \\ &= 1, \text{ if the value of } i \text{ is present} \end{aligned} \quad (3.8)$$

The completeness of the information product  $C^{IP}(k)$  at unit level is equal to the multiplication of the  $C^D(i)$  times the information relative value of  $i$  ( $Irel_i$ ), of each unit data considering in the form (described in the point 3.1.2.1)

$$C^{IP}(k) = Irel \times C^D \quad (3.9)$$

The completeness of the data block information product  $C^{IP}(K)$  is equal to the sum of all  $C^{IP}(k)$ .

$$C^{IP}(K) = \sum_{k=1}^n C^{IP}(k) \quad (3.10)$$

### 3.2.2 Sufficiency

The attribute of sufficiency has been recognized as an *appropriate data amount, here called "sufficiency"*. It includes all the data units needed (indispensable) by the decision maker for the decision task. Firstly, the decision-maker needs all the indispensable data for the decision task; secondly, he/she could also need some data units (of verification) to corroborate the first one. The "appropriate" level of sufficiency depends on each decision task. This level should be a value between the accumulated information relative value of indispensable information zone ( $Irel_{acc}(II)$ ) and the total information contained in the document, it is to say: 1.

### 3.2.3 Relevance

Given certain information that can be understood and interpreted by those in charge to process the document, we hope that this is relevant for the purpose for which it was created (Bovee et al., 2003). In the case of document processing, the processors are those considered as the experts who can determine the relevance parameters. Here, the relevance (RV) of the document fields has been linked to the concept of indispensability. Since one data is indispensable, it is necessary. So, in this study, the relevance value corresponds to the set of the information product of the indispensable units ( $Irel_{acc}$ ) described in the section before.

$$RV = Irel_{acc} \quad (3.11)$$

Another used method to express the quality attributes assessment has been the ratio (Pipino et al., 2002). The ratio has been used in free-of error, completeness, and consistency (Ballou & Pazer, 1985a ; Ballou et al., 1998 ; Redman, 1998a).

In order to evaluate the quality of both the data input and the information output, two relationships were developed. These two relationships work as a reference between the real state and the ideal state of the system. They work as an indicator of: a) the sufficiency of the requested data (relationship DIDV); and b) the usefulness of the information gathered through the form (relationship RIC).

### 3.2.3.1 Relationship DIDV

The simple ratio as data indispensable/data verification (DIDV) has been used before, to express the desired outcomes to total outcomes (Pipino et al., 2002). In this case, the ratio DIDV works as a tool to assess the inbound data unit quality considering the quantity of current data. It indicates, in a simple mode, how many of the verification *dus* exist in relation to the indispensable *dus*. Ideally, in order to reduce the extra amount of *dus* in the data processing and, furthermore, produce a better-quality IP, the form should have a smaller amount of verification *dus* in relation to indispensable *dus*. The formal definition of DIDV is:

$$DIDV = 1: \frac{(DV)}{(DI)} \quad (3.12)$$

### 3.2.3.2 Relationship RIC

The relation information content (RIC) allows us to know the quality of the information content at the output of the system once the transformation of *du* into an IP is made. The RIC relation considers not only the content, but also the *du* composition. This relation expresses, in terms of information, what portion of it is relevant to the aim pursued. Given a comparison between two scenarios of the same document, the one with the lower value represents the best option, as fewer requested fields are used to verify the indispensable information. This ratio is calculated from the following equation:



$$RIC = \frac{IVrel_{acc}}{IIrel_{acc}} \quad (3.13)$$

Considering these parameters and the document structure, decision taker can decide if form design, including question format, is the most suitable for the document processing or it can be improved depending on the data amount requested and the quality of expected information.

### 3.3 Assessment [A], emphasis on process

*Relevance* and *timeliness* concepts are tightly linked between them (Ballou et al., 1998 ; Bovee et al., 2003). Meanwhile relevance deals with content, timeliness deals with the process.

#### 3.3.1 Timeliness

Timeliness has been defined as the extent to which the age of the data is appropriate for the task at hand (Wang & Strong, 1996). We found more accurately determining the timeliness evaluation based on the time elapsed in the process than on the volatility of the data. For the Ballou (1998) proposal, we have considered that calibrate the exponent is a task that, in addition to consume valuable time, represents an ability and very specific competences for the responsible to do so. That it would make the evaluation process more laborious than the same information processing itself. For this reason our timeliness assessment takes the proposal of Chi (2017) as a reference.

For the timeliness evaluation: we assume firstly that the data used is updated at the evaluation performing time. This means that the age of data is at its lower value, the data is accurate. Secondly, the processing time (PT)—how long the data units have been within the system (Wand & Wang, 1996)—is taken as a reference of time. The Chi et al. (2017) timeliness evaluation takes the following assumptions presented on the left side of the table 3.2 as a base. On the right side of the table 3.2, we present their correspondence with the information produced in a data processing case.

Table 3.2 The considered variables in the timeliness evaluation of the emergency scenario corresponding to document processing scenario

Variable	Scenario: emergency	Scenario: document processing
$t_0$	The moment when emergency occurs	The time when the process begins
$t_1, t_2$	The times in which the first and the second batches of resources arrive $t_1 < t_2$	$t_2$ is the number of days that the document process really took.
$q_1, q_2$	The corresponding quantities of resources. ( $\tilde{q}_1, \bar{q}_1$ ) The maximum and minimum demand quantities of the first batch related to $t_1$ ; ( $\tilde{q}_2, \bar{q}_2$ ) the maximum and minimum demand quantities of the second batch related to $t_2$ .	In this case, we will consider the whole document as the delivery resource, then $q_1$ or $q_2$ will be equal to 1.
$T_1, T_2$	The arrival time of the first and second batches of resources respectively.	This corresponds to the processing time (PT). We consider as PT the average of the time period that takes (according to records) processing the form.
$u$	The timeliness of emergency resources schedule.	The timeliness of the information produced will be represented as TL.

The sigmoid function (equation 3.14) was selected to construct the function due to its ability to describe some real phenomena. The threshold function is continuous, smooth, strictly monotonic, and centrosymmetric about (0, 0.5). It was transformed as follows to be strictly monotonically decreasing. For more detailed explanation of the sigmoid function transformation see reference (Chi et al., 2017).

$$f(\tilde{x}) = 1 - \frac{1}{1 + e^{-ax}} \quad (3.14)$$

Where  $a$  is the tilt coefficient, and the slope decreases as  $a$  decreases as figure 3.1 shows.

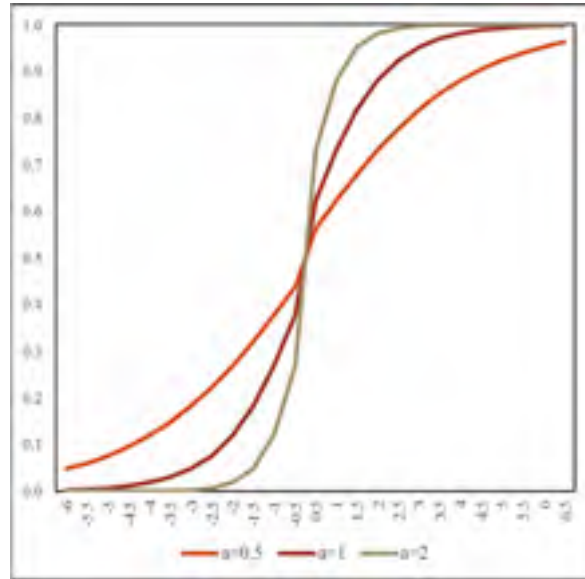


Figure 3.1 The sigmoid function. Source Chi et al., (2017)

The objective of the timeliness of emergency resource schedule ( $u$ ) is to transform the time objective into the impact of time on the emergency response, and converts the resource objective into the influence of resources on the response. In their analysis Chi et al. (2017) consider that the timeliness of each batch of resources is negatively correlated with the arrival time and positively correlated with the quantity of resources that arrive (figure 3.2).

The following formulas (3.15) and (3.16) are the function expressions which correspond to figure 3.2a and 3.2b respectively.

$$u_{21}(q_2) = \frac{1}{1 + e^{-\frac{\ln\left(\frac{1}{\epsilon_{21}} - 1\right)^2}{\bar{q}_2 - \bar{q}_2} \left(q_2 - \frac{\bar{q}_2 + \bar{q}_2}{2}\right)}}, \quad q_2 \geq 0 \quad (3.15)$$

$$u_{12}(t) = 1 - \frac{1}{1 + e^{-\frac{\ln\left(\frac{1}{\epsilon_{12}} - 1\right)^2}{T_1 - t_0} \left(t - \frac{t_0 + T_1}{2}\right)}}, \quad t \geq 0 \quad (3.16)$$

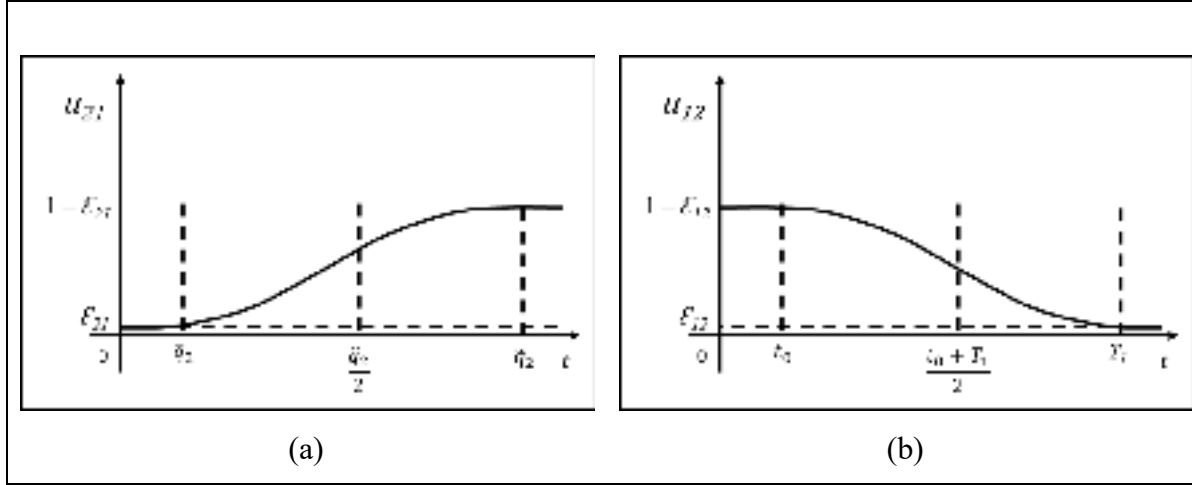


Figure 3.2 (a) Timeliness evaluation function considering only resource quantity at  $t_2$ . (b) Timeliness evaluation function considering only resource arrival time. Source: Chi et al., (2017)

Between the two scenarios presented by Chi et al. (2017), the referenced situation for our case was when no resources arrive at  $t_1$  but the resources arriving at  $t_2$  completely satisfy demands, then, the effect value of the emergency response is expressed as follows:

$$u_2 = u_{21}(\tilde{q}_2) \cdot u_{12}(t_2), \quad q_1 = 0, \quad q_2 = \tilde{q}_2 \quad (3.17)$$

Chi et al. (2017) combine the quantity of received resources  $[u_{21}(q_2)]$  and their arrival time  $[u_{12}(t)]$  by multiplication.

By substituting formulas 3.15 and 3.16 in 3.17 the equation 3.18 is obtained:

$$u_2 = (1 - \varepsilon_{21}) \times \left[ 1 - \frac{1}{1 + e^{-\frac{\ln\left(\frac{1}{\varepsilon_{12}} - 1\right)^2}{T_1 - t_0} \left(t_2 - \frac{t_0 + T_1}{2}\right)}} \right] \quad (3.18)$$

Because  $u_2 < 1$ , timeliness will be better when closer to 1. After this stage, Chi et al. (2017) developed their timeliness evaluation function  $u$  for the emergency resource schedule when the effect value of the emergency response  $u$  is determined only by different values of  $q_1$ . However, as we mention before, the objective of this thesis is interested only in the timeliness

of the process. In our case, the evaluation function is affected only by the arrival time because the quantity of resources provided by each batch, we consider it equal to 1 (one document). So, the equation 3.18 is which we use as a reference.

Following, equation 3.19 is the Chi et al. (2017) adaptation function for the timeliness (TL) of the sub-process  $sp$  at the moment  $t_2$  of the information produced in the manufacturing system concerns to the form. Because  $\varepsilon_{12}, \varepsilon_{21}$  are very small numbers, they were considered as Chi et al. (2017) did, this is equal to 0.01.

$$TL(sp) = (1 - \varepsilon_{21}) \times \left[ 1 - \frac{1}{1 + e^{-\frac{\ln\left(\frac{1}{\varepsilon_{12}} - 1\right)^2}{PT - t_0} \left(t_2 - \frac{t_0 + PT}{2}\right)}} \right] \quad (3.19)$$

Where PT is the average of processing time that usually takes the process;  $t_2$  is the minimum (or maximum, depending analyzed scenario) processing time of the sub-process. As one manufacturing information process can contain more than only one sub-process, the total timeliness value of the process will be equal to the average of sub-process timeliness evaluation which composes this process. N is the total number of sub-process (sp) that integrate the process (P).

$$TL(P) = \frac{\sum_{sp=1}^N TL(sp)}{N} \quad (3.20)$$

### 3.3.2 Actual information value

Since it is working with the user's vision about the product, it is necessary to consider the product's value for the user (in this case, a decision maker) (Wang & Strong, 1996). The approach "manufacturing of information" hypothesized an ideal product with a 100% client satisfaction (Ballou et al., 1998). The stage of assessment considers presenting different scenarios according user weighting for concerned attributes in order to have different

alternatives of the information system regarding the document content and the document processing. The relevance and timeliness are the attributes which the user should weigh in order to have the actual information value ( $V_A$ ). We use the equation proposed by Ballou et al. (1998) and Ahituv (1980). This formula considers that for each client  $C$  the actual value ( $V_A$ ) is a function of the intrinsic value ( $V_I$ ), timeliness (TL) and what for they are the data quality (DQ) that, in this case it is represented by relevance (RV).

$$V_A = f_c(V_I, TL, RV) \quad (3.21)$$

Which brings us to the following functional equation:

$$V_A = V_I\{[w_r \times (RV)^a] + [(1 - w_r) \times (TL)^b]\} \quad (3.22)$$

Where:

- $V_I$  is the information intrinsic value, this value can result from similar analysis such as described here for contextual attributes but that works with information intrinsic attributes. For the moment, due to getting out of the scope of this study,  $V_I$  will take a value of 1.00.
- $W_r$ , is the weight of importance given by the decision maker to the relevant attribute. As the product represents a 100% satisfaction, client weight (according to his expectations) will be divided between the relevance and the timeliness.
- According to Ballou (1998),  $a$  and  $b$  exponents represent client's sensitivity to change in DQ and TL, in our case, both are considered to be equal to 1.

For document analysis, the weight  $w_r$ , proposed by the client who, as Wang and Stuard (1989) points out, works well when the company has a clear understanding of the importance of each attribute in relation to the total of the information. The person responsible for making this assessment in the case of the document should be deep involved in the functioning of the organizational information system and know well the principles and policies governing the company to give appropriate weight to the attributes as better results for the purposes of the institution.

### 3.4 Analysis cases

In this section we will present the two analysis cases used to show how the methodology works. Both of them are application forms, in the first case, the analysis emphasizes on the content assessment and in the second case, the emphasis is done in the process.

#### 3.4.1 Analysis case 1

The presented case corresponds to the processing of a printed application form (here called F1-00) which flows through the CS of a higher-education institution. Its objective, according to institutional policies, is to grant (or deny) access to a certain installation belonging to the institution. The application form can be filled out by an internal user (belonging to the institution) or an external user (as a guest).

The F1-00 application form (figure 3.3) is comprised of 32 fields in total, divided into 8 sections (as shown in Table 4.1). The application form consists of open, closed, and multiple-choice fields to fill out. For this analysis, each field was considered as one data unit. The document must pass through two different departments. In these departments, there are three stations which the document must go through to be processed. A station is understood as the point where *du* is transformed into semi-processed information (IU), since the person who processes the document makes a change to the process. The first station is where the user or the department secretary fills out the application form with the user data. The second station corresponds to the department director, responsible for granting or denying access to the requested installation. Finally, the third station corresponds to the security department which verifies and ends document processing. Semi-structured interviews were conducted with the responsible document processors.

ACCESS REQUESTED FORM

I) IDENTIFICATION

1) Last name

2) First name

3) Tel home

4) Tel office

5) Fax

II) PAID EMPLOYEE

Employee ID

6)

Student ID

7)

8) ☐ Administrative

☐ Help Employee

☐ Professor

☐ Internship

☐ Guest Professor

III) PAID PARTIAL TIME TEACHING

Employee ID

9)

Student ID

10)

11) ☐ Course in charge

☐ Teaching assistant

12) Course

Course dates

13) Beginning

14) End

IV) PAID RESEARCH

15) Employee ID

Student ID

16)

V) STUDENT BY SESSION

17) Student ID

18) ☐ Student club

☐ Research

☐ Others

19) Club name

20) Tutor

21) Other specify

VI) OTHER (UNPAID OR NON-STUDENTS)

22) Temporary ID:

23) ☐ Non-paid internship

☐ Non-paid guest professor

☐ Non-paid research

☐ Other

24) Specify

25) Sponsor

26) Specify reason

VII) REQUESTED LOCALS

27) LOCALS

28) EXPIRATION DATES

The regular opening times is from Monday to Friday from 9:00 to 20:00, Saturday and Sunday from 9:00 to 20:00. Holiday days from 9:00 to 18:00. If you would get access out of these opening times, indicated by marking the next box and specify the reason.

29) ☐ Access out of regular opening times

30) Specify the reason

VIII) AUTHORIZATION

I authorize this requested access to the person above mentioned

31) Signature responsible in charge

32) Date

Figure 3.3 F1-00 Form, 8 sections, 32 champs.  
Retyped from real form.

The information manufacturing process for the form F1-00 is shown in figure 3.4. In this process there are: an information-product (IP) associated with this operation, the granting of access to the applicant (RB1). The input source (EB1) of data-units (*duI*) which could be: 1) the applicant himself (worker, teacher, student or guest) or 2) the secretary who fulfills the application form applicant’s data (DPB1). Once the document is completed and processed,



entered data are verified (QB1) in its same corresponding work unit (*wu1*). Then, it is sent (on a daily basis) to the department's director. Here, the director, based on the processed information so far—semi-processed information— (IU2 and IU3) takes the decision to grant (or deny) access, entering (if that is the case) his signature (IU4) in work unit 2 (DPB2-*wu2*). Finally, the document is forwarded to the security department (DPB3), where the staff in charge (EB3) takes out the document from the system (SB1), verify that all indispensable *du* are there (QB2) to perform processing (DPB4) and use relevant verification *du* to corroborate the indispensable *du* ; if everything is as the procedure indicates, the IP is delivered.

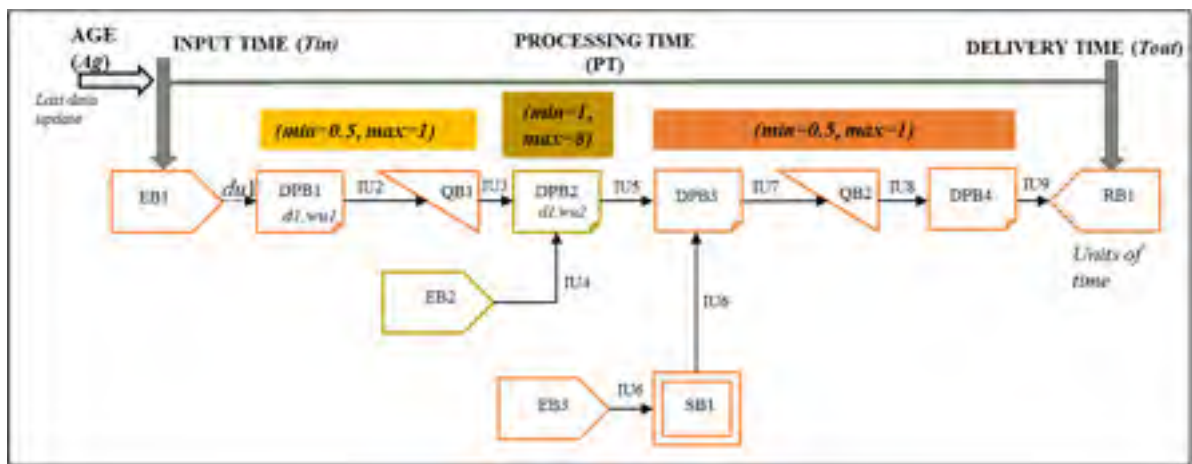


Figure 3.4 Information manufacturing process for the form F1-00

### 3.4.2 Analysis case 2

This case focuses on the processing of a document corresponding to an administrative information and budget form (FIAP-00) that alternates its form on paper and in electronic within the system. This goes through a higher education institution. The form objective is to summarize all administrative and budget information from a research project. The document is an internal communication medium therefore there are no external agents involved in the information-product manufacturing system. The main structure of the form is represented in Figure 3.5. FIAP-00 form is comprised of 79 fields divided into four sections which match

with the work-units (*wu*). The application form consists of open, closed, and multiple-choice fields to fill out.

<i>wu1 = 17 du</i>	
<i>wu2 = 24 du</i>	
<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="padding: 5px; text-align: center;"><i>wu4 = 2 du</i></td> </tr> </table>	<i>wu4 = 2 du</i>
<i>wu4 = 2 du</i>	
<i>wu3 = 36 du</i>	

Figure 3.5 Structure of the FIAP-00 form

The document must pass through five different interchange stations belonging to 3 different departments. In department 1, the first station is where the *agent a* fills out the application form with the project data. In department 2, the second station is where the *professor (P)* fills out the budget data of the project. The form returns to the department 1 where the next two stations are, the third station corresponds to the *agent b* who fills out another project data and the department director, responsible for granting the authorization. Finally, in department 3, the fifth station corresponds to the *finance department* who verifies and ends document processing. Fields are not promptly mentioned for safety reasons.

FIAP-00 information manufacturing system is shown in Figure 3.6. Different data-units (*du*) types have been modeled as *wu1*, *wu2*, *wu3* and *wu4* for their representation in the scheme (figure 3.5). In the first station (DPB1), research project identification data (*wu1*) are fulfilled at FIAP-00 by agent (EB1): name of the project, responsible professor, applicant institution, address, phone, etc. The form is entered to the system and sent to *P* by this same agent. This process stage is represented by DPB2. This task can be done in 1 or 2 days (*subprocess 1*). Immediately, *P* (EB2) enters project budget data (*wu2*) in the form (DPB3) sending the FIAP-

00 again to *agent a* (DPB4). This task, depending on the professor's workload, the processing time of this sub-process can be taken between 2 to fifteen days to be completed (*subprocess 2*). Once received and verified the completeness and accuracy of data at the FIAP-00 (QB1), the form is sent by *agent a* to *agent b* (DPB5). Next, *agent b* enters another project data (*wu3*) concerning the project risk analysis (DPB6). Once this is done, *agent b* sends the form back again to *agent a* (DPB7). This task can take 1 day minimum and 3 days maximum (*subprocess 3*). The *agent a* verifies (QB2) again completeness and accuracy of the data. After this action, *agent a* sends (DPB8) the form to the head of the research department (EB4) to proceed with its authorization. If all agree to the institutional guidelines, the document is authorized (DPB9) and sent for the last time to *agent a* (DPB10). The processing time of this action can take from 1 to ten days (*subprocess 4*). Once the FIAP-00 form is received by *agent a*, he/she verifies for the last time the form (QB3) duly fulfilled to enter it into the system (SB1). This task can take 1 to 2 days to be completed (*subprocess 5*). Once entered into the system, the finance department processes the financial concerning data-units (DPB11) to create the respective project account (RB1). The processing time for this task is 1 to 5 days (*subprocess 6*).

Summarizing, there are 4 work units (*wu*) covering 79 *du*. Into the document these *du* are entered in 4 of the 5 interchange stations where it passes to be processed. As is it assumed that there is a change in the pre-processed information (IU) quality which passes through the quality control blocks, the IU changes in the same way. For example, the IU2 passing by quality block (QB1) is transformed into IU3. So, in this model there are 4 *du* sources (EB), 3 quality blocks (QB), 11 document processing blocks (DPB) and one storage block (SB).

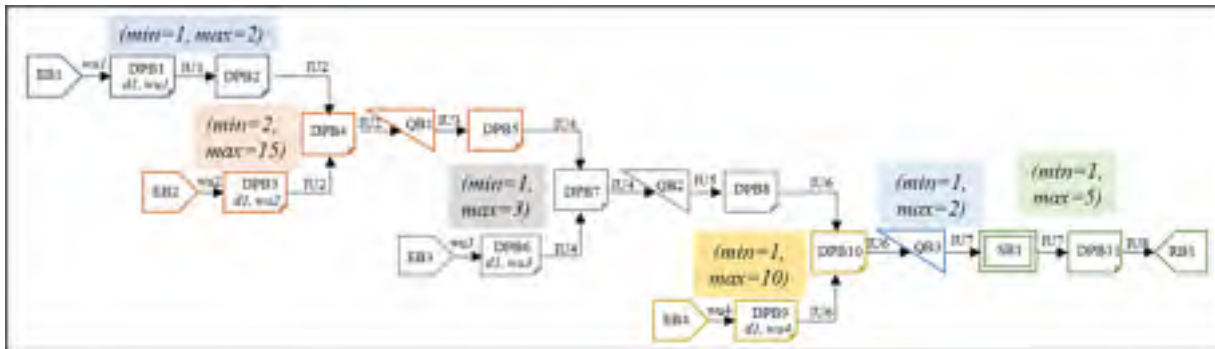


Figure 3.6 Information manufacturing process for the FIAP-00 form

This chapter presented a methodology to determine five quality attributes of contextual information recognized as such by several studies (Wang & Strong, 1996 ; Wang, Yang, et al., 1998): *a) sufficiency, b) completeness, c) relevance, d) timeliness and e) actual value*. Also, the two cases of analysis were presented. In the next chapter will present the results illustrated through these two cases, in order to showing the impact and possible changes that would be generated around its application.

## **CHAPTER 4**

### **RESULTS**

This section will present the results of applying the methodology described in the previous chapter. The methodology was applied in two case studies corresponding to two different forms within a higher education institution. The first case emphasizes in the document's content and the second case emphasizes in the information manufacturing process.

This chapter is structured in the following way: In the first section the model obtained from the methodological process that worked as a guide to conduct the document processing assessment is presented. In this same section the classification of data [CD] and its processing into information [PI] phases are performed in both analysis cases (the two forms). In the second section the assessment phase [A] is presented, making emphasis in both, the content of the document and the processing of the document. Finally, in the third section, after a reengineering proposal of both forms, a comparative analysis is performed in order to make evident the methodology usefulness.

#### **4.1 Model [CD]-[PI]-[A]**

As a result of the methodological process analysis followed to perform the evaluation of the information quality and considering the approach proposed in chapter 2, a schematic model was obtained. This schematic model of the structure process [CD]-[PI]-[A] is presented in figure 4.1.

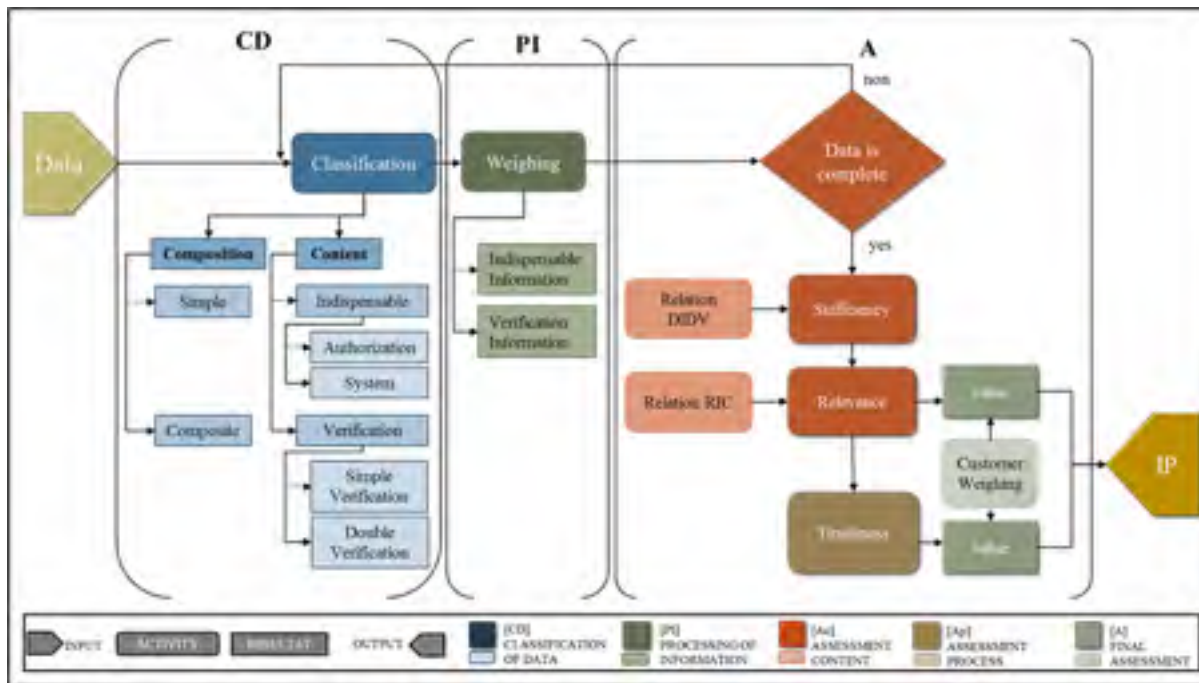


Figure 4.1 Schematic representation of the [CD]-[PI]-[A] model

The first step, once the data have entered the manufacturing of information system, the next step is their classification. Here, data is grouped according to their context into two categories: composition and content. The data representation for their composition can be simple or composite and the data representation for their content can be indispensable or verification. Once data have been classified, they are weighted in order to represent their process into information. After this transformation, the data units, now transformed into information units are evaluated. The contextual attributes chosen to evaluate were completeness, sufficiency, relevance (related to content) and timeliness (related to process). Finally, the actual value of information is obtained by using the customer weighing according to his preference.

#### 4.1.1 Classification of data [CD] and processing data into information [PI] in both analyzed cases

Following, the data classification and the data transformation into information from its specific characteristics of the two cases of analysis will be presented.

#### 4.1.1.1 Form F1-00

From the semi-structured interviews conducted with the responsible document processors, *du*s were classified according to their characteristics described in Section 3.1 and presented in Table 4.1.

Table 4.1 Form F1-00. Form structure and *du* classification. Ds = Simple data. Dc = Composite data. Dia = Indispensable data for authorization. Dis= Indispensable data for the system. Dv = Simple verification data. Dvv = Double verification data

Work unit	Section No.	Section Name	Data ID	Data	Data Classification
1	1	Identification	1	Last name	Ds/Dis
			2	First name	Ds/Dis
			3	Home phone	Ds/Dv
			4	Work phone	Ds/Dvv
			5	Extension phone	Ds/Dvv
	2	Paid Employee	6	Employee ID	Dc/Dis
			7	Student ID	Dc/Dis
			8	Multiple choice 1	Ds/Dv
	3	Paid Partial Time Teaching	9	Employee ID	Ds/Dvv
			10	Student ID	Ds/Dvv
			11	Multiple choice 2	Ds/Dv
			12	Class name	Ds/Dv
			13	Beginning Date	Ds/Dvv
			14	End Date	Ds/Dvv
	4	Paid Researcher	15	Employee ID	Ds/Dvv
			16	Student ID	Ds/Dvv
	5	Student by Session	17	Student ID	Ds/Dvv
			18	Multiple choice 3	Ds/Dv
			19	Club name	Ds/Dv
			20	Tutor	Ds/Dv
			21	Other specify 1	Ds/Dv
	6	Other (Unpaid or non-students)	22	Temporal ID	Dc/Dis
			23	Multiple choice 4	Ds/Dv
			24	Other specify 2	Ds/Dv
			25	Sponsor	Ds/Dv
			26	Reason 1	Ds/Dv
	7	Locals	27	Local numbers	Ds/Dis
			28	Expiration date	Ds/Dis
			29	Access out hours	Ds/Dv
			30	Reason 2	Ds/Dv
2	8	Authorization	31	Signature	Ds/Dia
			32	Date	Ds/Dv

Once the data were classified and organized according to their composition and content (Table 4.2), the *duv* was assigned. In form F1-00, two redundant fields were detected. This

was possibly due to the structure and organization of the form; the two fields were: student ID and employee ID. For our analysis, these two fields were in one instance considered as indispensable data and the rest of the time as double verification data, as it was required only once to carry out the processing.

Table 4.2 Data classification and weighting. Frequency of accumulated data according to information type zone ( $D_{acc}$ ), relative frequency of accumulated data according information type zone ( $D_{rel_{acc}}$ ), information relative value ( $I_{rel}$ ), and accumulated information relative value ( $I_{rel_{acc}}$ ) for the F1-00 form

Information Type	Data Type	$f$	$D_{acc}$	$D_{rel_{acc}}$	$duv$	$duv_{set}$	$I_{rel}$	$I_{rel_{acc}}$
II	Dc/Dis <i>Student ID or Employee ID or Other ID</i>	1			15	15	0.21	
	Ds/Dia <i>Signature</i>	1			4	4	0.05	
	Ds/Dis <i>Last name, first name, locals, expiration date</i>	4	6	0.19	3	12	0.17	0.43
IV	Ds/Dv <i>Home phone, multiple-choice1, multiple-choice2, class name, multiple-choice3, club, tutor, other specify1 multiple-choice4, other specify2, sponsor, raison1, access out, raison2, date</i>	15			2	30	0.42	
	Ds/Dvv <i>Work phone, ext-phone, beginning date, end date, redundant IDs (7 times)</i>	11	26	0.81	1	11	0.15	0.57



#### 4.1.1.2 Form FIAP-00

In the same way that in the first analysis case, semi-structured interviews were conducted with responsible processors. These interviews let classifying the data collectors according to their characteristics described in section 3.1 and presented in Table 4.3.

Table 4.3 Data classification and its weighting, relative information value (Irel) and accumulated information value (Irelacc) for FIAP-00

<b>I Type</b>	<b>Data Type</b>	<b><math>x</math></b>	<b><math>w</math></b>	<b><math>f</math></b>	<b>Dacc</b>	<b>Drel<sub>acc</sub></b>	<b><math>duv</math></b>	<b><math>duv_{set}</math></b>	<b>Irel</b>	<b>Irel<sub>acc</sub></b>
<b>II</b>	Dc/Dis	11	3	1			33	33	0.10	
		9	3	1			27	27	0.08	
		7	3	1			21	21	0.06	
		6	3	2			18	36	0.11	
		5	3	1			15	15	0.04	
		4	3	1			12	12	0.03	
		3	3	1			9	9	0.03	
		2	3	1			6	6	0.02	
	Ds/Dia			2			4	8	0.02	
	Ds/Dis			35	46	0.58	3	105	0.31	0.80
<b>VI</b>	Ds/Dv			33			2	66	0.20	
	Ds/Dvv			0	33	0.42	1	0	0.00	0.20

#### 4.2 Assessment [A], content and process

Three contextual quality attributes have been related to the content assessment: completeness, sufficiency and relevance. One contextual quality attribute has been related to the process assessment: timeliness. The remained fifth contextual quality attribute, the actual value of information sums the relationship between the after-mention attributes with the customer

preferences. Following, results of each attribute will be presented for both analyzed cases, form F1-00 and form FIAP-00.

#### 4.2.1 Content: completeness, sufficiency and relevance

According to exposed in point 3.2.1. the completeness is a measure of how complete an IP is in terms of the data units that are included in the IP (Shankaranarayanan & Cai, 2006). We have: a) completeness at the data level [ $C^D(i)$ ], b) completeness at the information product unit level [ $C^{IP}(k)$ ] and c) completeness at the data block (DB) or document level [ $C^{IP}(K)$ ]. These three evaluations are shown in table 4.4 for F1-00 form and in table 4.5 for FIAP-00 form.

In both cases, F1-00 and FIAP-00 the completeness value that interests is at document level, but this value cannot arise without having been computed the two previous completeness evaluations [ $C^D$  and  $C^{IP}(k)$ ]. In the F1-00 case (table 4.4) the completeness value is equal to 1 because it was considered that the whole form was fulfilled. However, for someone who decide not to responds all fields, this value can be less than one. For instance, if the applicant answers only one time his/her ID, it is to say some double verification data are not answered, the completeness value will be less than one (0.903). In the case of the FIAP form, its completeness value es equal to 1 because all fields are fulfilled and does not exist any double verification data.

Table 4.4 Completeness assessment for F1-00 at data level ( $C^D$ ), at information unit level [ $C^{IP}(k)$ ] and at document level [ $C^{IP}(K)$ ]

Information Type	Data Type	Datum	Irel	C <sup>D</sup>	C <sup>IP</sup> (k)
II	Dc/Dis	Student ID	0.208	1	0.208
	Ds/Dia	Signature	0.056	1	0.056
	Ds/Dis	Last name	0.042	1	0.042
		First name	0.042	1	0.042
		Locals	0.042	1	0.042
		Expiration date	0.042	1	0.042
		IV	Ds/Dv	Home phone	0.028
Multiple-choice 1	0.028			1	0.028
Multiple-choice 2	0.028			1	0.028
Class name	0.028			1	0.028
Multiple-choice 3	0.028			1	0.028
Club	0.028			1	0.028
Tutor	0.028			1	0.028
Specify other1	0.028			1	0.028
Multiple-choice 4	0.028			1	0.028
Specify other2	0.028			1	0.028
Sponsor	0.028			1	0.028
Especify raison 1	0.028			1	0.028
Access out of date	0.028			1	0.028
Especify raison 2	0.028			1	0.028
Date	0.028			1	0.028
Ds/Dvv	Work phone		0.014	1	0.014
	Extension phone		0.014	1	0.014
	Depart date		0.014	1	0.014
	End date		0.014	1	0.014
	Extra ID		0.014	1	0.014
	Extra ID		0.014	1	0.014
	Extra ID		0.014	1	0.014
	Extra ID		0.014	1	0.014
	Extra ID		0.014	1	0.014
	Extra ID		0.014	1	0.014
			C <sup>IP</sup> (K)	1.000	

Table 4.5 Completeness assessment for FIAP-00 form at data level ( $C^D$ ), at information unit level [ $C^{IP}(k)$ ] and at document level [ $C^{IP}(K)$ ]

Information Type	Data Type	Irel	$C^D$	$C^{IP}(k)$
II	Dc/Dis	0.100	1	0.100
	Ds/Dia	0.020	1	0.020
		0.080	1	0.080
		0.060	1	0.060
		0.110	1	0.110
		0.040	1	0.040
		0.030	1	0.030
		0.030	1	0.030
		0.020	1	0.020
	Ds/Dis	0.310	1	0.310
IV	Ds/Dv	0.200	1	0.200
		1.000	$C^{IP}(K)$	1.000

Regarding the attribute of sufficiency, the IP could have a degree of sufficiency. It includes all the data units needed by the decision maker for the decision task. This level corresponds to a value in a range between the accumulated information relative value of indispensable information zone ( $Irel_{acc}(II)$ ) and the total information contained in the document, 1. The case of F1-00 form presents different profiles of applicants: research personal, administrative personal, teaching auxiliary, students, guess and other. All these profiles share the same range of sufficiency [0.43,1] but its value depends each profile characteristic and the data field answered. For instance, for one student who will need local access for research propose, he/she should answer the 6 specific indispensable data plus 5 Ds/Dv data, so the sufficiency value for this type of applicant using the F1-00 form will be 0.69. For more details and the different values of sufficiency according each profile consult the appendix I.

For the FIAP-00 form, the lower range limit of its sufficiency is equal to 0.8. Then, the sufficiency value, according to the verification data type fulfilled can be in the range [0.8,1].

As we mention before, the relevance value corresponds to the set of the information product of the indispensable units ( $I_{relacc}$ ). This means that the relevant (or indispensable) information is the minimum sufficient to process the form. For the form F1-00, the relevance is equal to 0.43 and for the form FIAP-00, the relevance value is equal to 0.80.

#### 4.2.1.1 DIDV and RIC relationships

In F1-00 there are 6 indispensable *du* and 26 verification *du*, which leads to a DIDV 1:4.33 ratio. This is to say that, for each indispensable data that is requested, there are four data units used to verify it. The current structure and design of the form contribute to the generation of data overload in the information manufacturing system.

Regarding the RIC relationship which considers, in addition to the content, the composition that generates this information. The F1-00 form has 0.57 information products of a verification type ( $I_{Vacc}$ ), and 0.43 information products of an indispensable type ( $I_{Iacc}$ ) therefore the RIC is equal to 1.32. Ideally, this value should be equal to or less than 1, because the form should request the same amount or less verification information than that of the indispensable type. This relationship works as an indicator of the relevant information content in the CS.

Unlike the F1-00, the form FIAP-00 has more composite-indispensable data type ( $D_c/D_{is}$ ) than simple-of verification data type ( $D_s/D_v$ ). The DIDV relation for the FIAP-00 results in 1:0.72, this means that there was less than one *du* to verify the indispensable *du* to achieve the process. In the case of the RIC relationship for the FIAP-00 form is equal to:0.24. This means that only one quarter of the fields are used to verify the indispensable information.

Figure 4.2 shows the composition of FIAP-00 form at the entrance in data level and its transformation into information at the exit of the system.

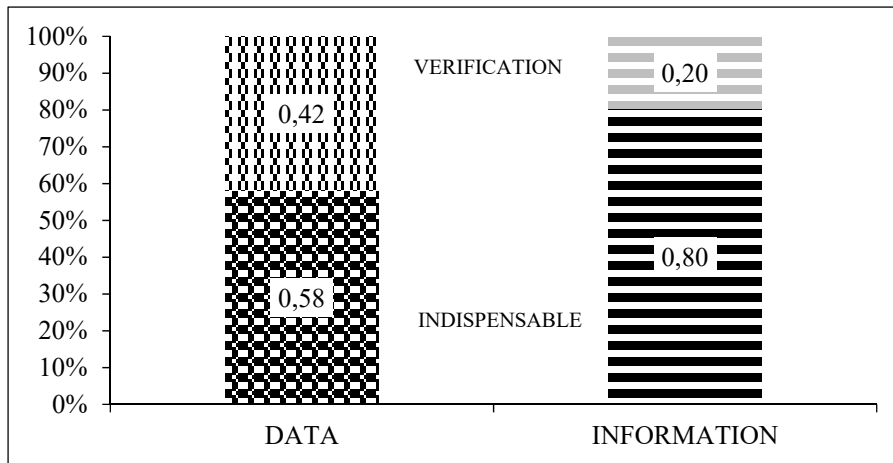


Figure 4.2 Content composition of FIAP-00. Left bar *du* input, Right bar IP output of manufacturing of information system

This change in percentage points occurs mainly because the form FIAP-00 has more composite-indispensable for the system data type than simple-verification type. This produces an increase in the percentage of indispensable information, and therefore, the quality of the information-product.

#### 4.2.2 Process: timeliness

According to the processor's interviews, the F1-00 form information manufacturing process has reached a minimum of 2 days and a maximum of 10 days. Three different scenarios have been proposed considering minimum (min), average (ave) and maximum (max) processing times (PT) to assess the timeliness (TL) of the form processing. The F1-00 processing is comprised of three sub-process. In each scenario, the time of each sub-process is pointing it out in the table 4.6.

Table 4.6 Different scenarios for the timeliness assessment according PTs proposed for the form F1-00

<b>F1-00</b>	<i>in days</i>		<i>in days</i>		<i>in days</i>	
<b>No. subprocess</b>	<b>PTmin</b>	<b>TL(PTmin)</b>	<b>PTave</b>	<b>TL(PTave)</b>	<b>PTmax</b>	<b>TL(PTmax)</b>
1	0.5	0.97	0.75	0.97	1	0.97
2	1	0.97	4.5	0.61	8	0.06
3	0.5	0.97	0.75	0.97	1	0.97
<b>TOTALS</b>	<b>2</b>	<b>0.97</b>	<b>6</b>	<b>0.85</b>	<b>10</b>	<b>0.66</b>

The processing time taken as a reference for the calculation of timeliness was the maximum PT, 10 days. The time when the processing begins ( $t_0$ ) was equal to 0; the time when the data is transformed into information, ( $t_2$ ) was considered as the PT corresponding to each subprocess and scenario (minimum, average or maximum). The timeliness of each scenario was calculated and pointing it out in the table 4.6 aside each considered time.

Because the timeliness value will not be affected by the quantity of resources (data) input (this is evaluated with the completeness and sufficiency attributes) our timeliness evaluation only depends on the PT of the form. The timeliness value decreases meanwhile the processing time increases. The shorter time period, the higher timeliness value we will have. The TL for the process made in a) the minimum time period is equal to 0.97, b) the average time period is equal to 0.85 and c) the maximum time period is equal to 0.66. These values may vary for instance when the process does not take in all sub-process all minimum values or all maximum values, but for practical purposes we considered in the same scenario all minimum, all average and all maximum values.

In terms of the process, for the FIAP-00 case, this involves more sub-processes within the whole process. Following are presented the different sub-processes with their minimum (min), average (ave) and maximum (max) periods of time and their corresponding timeliness (TL) evaluation.

Table 4.7 Different scenarios for the timeliness assessment according PTs proposed for the form FIAP-00. min=minimum, ave=average, max=maximum

<b>FIAP-00</b>	<i>in days</i>		<i>in days</i>		<i>in days</i>	
<b>No. subprocess</b>	PTmin	TL(PTmin)	PTave	TL(PTave)	PTmax	TL(PTmax)
1	1	0.98	1.5	0.98	2	0.97
2	2	0.97	8.5	0.91	15	0.70
3	1	0.98	2	0.97	3	0.97
4	1	0.98	5.5	0.95	10	0.88
5	1	0.98	1.5	0.98	2	0.97
6	1	0.98	3	0.97	5	0.96
<b>TOTALS</b>	7	<b>0.98</b>	22	<b>0.96</b>	37	<b>0.91</b>

The processing time taken as a reference was the maximum PT, 37 days. The time when the processing begins ( $t_0$ ) was equal to 0; the time when the data is transformed into information ( $t_2$ ) was considered as the PT corresponding to each sub-process and scenario (minimum, average or maximum). The timeliness of each scenario was calculated and pointing it out in the row aside each processing time considered in table 4.7. The TL for the process made in a) the minimum time period is equal to 0.98, b) the average time period is equal to 0.96 and c) the maximum time period is equal to 0.91.

#### 4.2.3 Actual Information value

The information value is calculated from the combination of two preceding attributes, the first concerning to the content, *relevance* and the second concerning to the process, *timeliness*. Also, as mentioned before, it is assumed a value of 1.00 either for information intrinsic value information as for a and b exponents which represent clients' sensitivity for both attributes. Three different scenarios of the weight given by the user to both attributes for the F1-00 case of analysis were considered:

- the client considers that both relevance and timeliness weigh the same, therefore  $w_r = 0.50$ ;
- the information relevance attribute weighs twice more than timeliness,  $w_r = 0.67$ ;
- the information timeliness attribute weighs twice more than relevance,  $w_r = 0.33$ .



Obtaining the following values presented in table 4.8. The relevance value corresponds to the  $I_{rel_{acc}}$  of F1-00 (0.43). The timeliness value considering for computing the actual value of information was the average timeliness value, it is to say 0.85

Table 4.8 Actual information value in three different scenarios according users' weight for F1-00 form

Form	$w_r$	RV	$1-w_r$	TL( $PT_{ave}$ )	$V_A$
F1-00	0.5	0.43	0.5	0.85	<b>0.64</b>
	0.67	0.43	0.33	0.85	<b>0.59</b>
	0.33	0.43	0.67	0.85	<b>0.71</b>

Regarding these three values, it is possible observe that the conditions in which information quality may have a higher value (0.71) is when it gives a greater weight of importance to timeliness attribute. Considering the content of the document, the relevance attribute presents a low level on the quality of the information requested. Therefore, the recommendation at this point would be proposing new alternatives, both in the overall document structure as in the design of requested fields.

Three different scenarios in relation to the weight assigned by the client ( $w_r$ ) to relevance (RV) and timeliness (TL) of the information contained in the FIAP-00 form were proposed. The relevance value corresponds to the  $I_{rel_{acc}}$  of FIAP-00 (0.80). The timeliness value considering for computing the actual value of information was the average timeliness value, it is to say 0.96. The three scenarios of user's weight are as follows:

- information relevance attribute weighs three times more than timeliness,  $w_r=0.75$ ;
- the client considers that both attributes weigh the same,  $w_r=0.50$ ;
- information relevance attribute weighs a quarter of the value of information,  $w_r=0.25$ ;

The results are shown below in table 4.9

Table 4.9 Actual information value in three different scenarios according user's weight for FIAP-00 form

Form	$w_r$	RV	$1-w_r$	TL( $PT_{ave}$ )	$V_A$
FIAP-00	0.75	0.80	0.25	0.96	<b>0.84</b>
	0.50	0.80	0.50	0.96	<b>0.88</b>
	0.25	0.80	0.75	0.96	<b>0.92</b>

The higher actual value of information (0.92) is obtained when the user's weight is bigger for the timeliness attribute because it is greater than relevance too. This form, unlike the F1-00 form presents high values in both attributes, which leads to high actual value of information in three different scenarios. However, if the relevance value were the same of F1-00 (0.43) the difference between these three scenarios would be more marked, having  $V_A$  of 0.56, 0.69 and 0.83 respectively. Comparing the second scenario, where the user's weight is the same for both attributes, the difference in  $V_A$  would decrease from 0.88 to 0.69, it is to say 19 points. A greater variation between attributes, more pronounce the difference in  $V_A$  will be. In this case although the  $V_A$  is in general high, we will propose a reengineering in the process in order to see if even there exist some significant change.

### 4.3 Comparative analysis

At a data unit level, in an efficiency assessment of the form we would get a higher value simply by reducing the amount of  $du$ . At an information product level, due to its contextual aspect, it is necessary to follow the DC-PI-A model in order to assess its quality. Considering the results getting in the assessment phase, we propose: 1) for the F1-00 case, a re-engineering mainly in its structure and requested fields design; and 2) for the FIAP-00 case, a re-engineering mainly in its processing. Following we will present the proposed change in the F1-00 structure, and in the FIAP-00 processing.

### **4.3.1 Re-engineering**

As the proceedings for the F1-00 did not establish any set-points regarding extreme security concerns about data gathering, following the document processor's recommendations, we propose a new design for this form, which we call here "re-engineering phase." In the case of F1-00, the new design was called F1-01, which is comprised of three main sections: (I) identification (II) status, and (III) authorization; five fewer sections than the original. Furthermore, the new form is comprised of 16 fields in total. If the document is chosen to be computerized, then the fields are proposed as drop-down menus. If it is chosen to be in paper format, multiple-option questions are proposed. Figure 4.3 shows the proposed F1-01 form.

ACCESS FORM F1-01			
<b>I. IDENTIFICATION</b>			
1)	Last name, First name		2) Telephone
		3) PHONE TYPE	
		home	
		movil	
		office with extension	
You are (make one selection)	4) STATUS 1		5) ID number
	Student		
	Employee		
	Neither student, employee		
6) The ID locals number you are asking are:		7) Period from DD-MM-YYYY TO DD-MM-YYYY	
The regular opentime is from Monday to Friday from 6h30 to 2h00, Saturday and Sunday from 7h30 to 2h00. Holiday days from 7h30 to 18h00			
DIF 0 R	8) If you would get access out of these opentimes, indicate it by marking the next box and specify the reason		
YES	9)		
NON			
<b>II. STATUS</b>			
10)	2A Choose the description that match with your scholar status		
	STATUS 2-A		
	RSCH PNL	Research Personal (paid). No more options to answer	
	EMP	Employee (paid). Continue next section only option 2B-a	
	TEACH	Teaching partial time (paid) Continue next section only option 2B-b	
	STUDENT	Student (valid access only for one enrolled period) Continue section 2B-c	
	ANOTHER	Another (Non-paid, non-student) Option 2B-d	
11)	2B According your election in section 2A make your selection in the next section		
a)	Employee	b)	Teaching partial time
	Administrative		Cours in charge
	Help employee		Auxiliar
	Teaching		12) Cours name
	Internship		
	Guest professor		
c)	Student	Specify	d) Another
	Club	Club name	Internship non paid
	Research	Tutor	Guest professor non paid
	Other	Especify	Research non paid
			Other
13)	Club name/Tutor/Specify		14) Specify other
<b>III. AUTHORIZATION</b>			
15) Signature department in charge		16) Date	

Figure 4.3 Re-engineering of form F1-00. Here called F1-01

For the FIAP-00 case, as the relevance value can be considered high, because 80% of recollected data was clarified as indispensable, no change in the form structure or requested fields were made. However, change was proposed in the processing of the form which is presented in figure 4.4. This change is identified as FIAP-01.

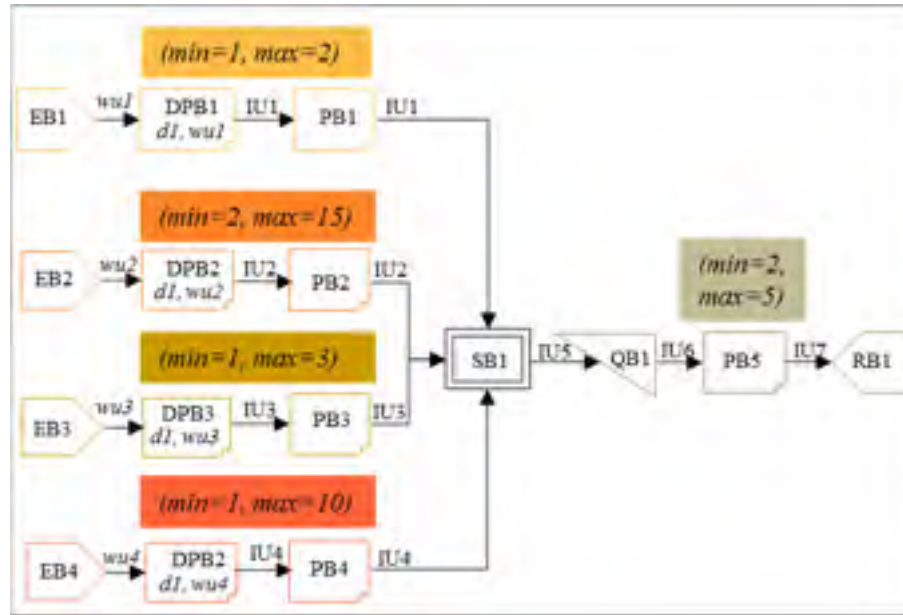


Figure 4.4 Proposed re-engineering in FIAP-00 form processing, here called FIAP-01

Once the results of this comparative analysis are presented, it is possible to observe in an easier way the impact of representing the *du* composition in the assessment of the information quality.

#### 4.3.2 Emphasis on the pertinence of the content

Because the three attributes related to the content, completeness, sufficiency and relevance, are closely connected among them we do not separate each in sub subjects. Table 4.10 shown the completeness assessment of the form F1-01.

Table 4.10 Completeness assessment for F1-01 form. CD= completeness at data level. CIP(k)= completeness at information unit level.  
CIP(K)= completeness at document level

F1-01	Data Type	Datum	Irel	C <sup>D</sup>	C <sup>IP(k)</sup>
II	Dc/Dis	ID student	0.283	1	0.283
	Ds/Dia	signature	0.075	1	0.075
	Dc/Dis	Last name/first name	0.113	1	0.113
	Ds/Dis	Local	0.057	1	0.057
		Expiration date	0.057	1	0.057
IV	Ds/Dv	Contact phone	0.038	1	0.038
		Phone type	0.038	1	0.038
		Status1	0.038	1	0.038
		Status 2-A	0.038	1	0.038
		Out Hours	0.038	1	0.038
		Specify hours	0.038	1	0.038
		Status 2-B	0.038	1	0.038
		Class name	0.038	1	0.038
		Club or tutor or another name	0.038	1	0.038
		Specify another	0.038	1	0.038
		Date	0.038	1	0.038
				C <sup>IP(K)</sup>	1.000

The completeness value at document level  $C^{IP}(K)$  of F1-01 does not change respecting F1-00, if we consider this as a new document independent of its predecessor. However, comparing F1-01 as a new version of the F1-00, considering the F1-00 sum of Irel as a reference, the F1-01 is likewise complete but with 26% less useless data than F1-01.

Because no change was made in form FIAP-00, its completeness assessment does not change either.

Considering the relevance attribute, table 4.11 shows the data classification and its corresponding transformation into information for the F1-01. A total of 100% of the *du* in the F1-00 form was taken as a reference to calculate the F1-01 form. As shown in Table 4.11, in the F1-01 form, five *dus* correspond to indispensable data. These represent 16% (31% of 50%)

of the content that was retained in the document. The 11 remaining *dus* represent 34% (69% of 50%) of the same. In the case of the information products, 58% of the preserved fields represent indispensable information, while 42% remained represent verification information.

Table 4.11 Data classification and its transformation into information, frequency of accumulated data according to information type zone ( $D_{acc}$ ), relative frequency of accumulated data according to information type zone ( $D_{rel_{acc}}$ ), information relative value ( $I_{rel}$ ), and accumulated information relative value ( $I_{rel_{acc}}$ ) for the F1–01 form

Information Type	Data Type	$f$	$D_{acc}$	$D_{rel_{acc}}$	$duv$	$duv_{set}$	$I_{rel}$	$I_{rel_{acc}}$
	Dc/Dis <i>Student ID or Employee ID or Other ID</i>	1			15	15	0.28	
	Dc/Dis <i>Last name / first name</i>	1			6	6	0.11	
	Ds/Dia <i>Signature</i>	1			4	4	0.08	
	Ds/Dis <i>Locals, expiration date</i>	2	5	0.16	3	6	0.11	0.58
IV	Ds/Dv <i>Contact phone, phone type, status1, status 2-A, out hours, specify hours, status 2-B, class name, club or tutor or another name, specify another date</i>	11	11	0.34	22	22	0.42	0.42
	Ds/Dv <i>n/a</i>	-	-	-	-	-	-	-
	<b>TOTALS</b>		16	0.50		53		1.00

With the new streamlining of the form, it is possible to (1) reduce the data requested (2) enhance the information quality produced, and (3) improve the efficiency of the CS. This finding, while preliminary, suggests that a reduction of data does not necessarily mean an improvement in quality of information but a change in the composition of the *dus* do. Additionally, this implies that the quality of information output can increase without necessitating a corresponding increase in the quantity of the data input.

This type of assessment can be considered as a new tool to determine quantitatively the sufficiency level of document filled according to a profile determined. For each profile, this value must be constant, any variation in it can indicate a problem 1) in the filled form or 2) in the form comprehension.

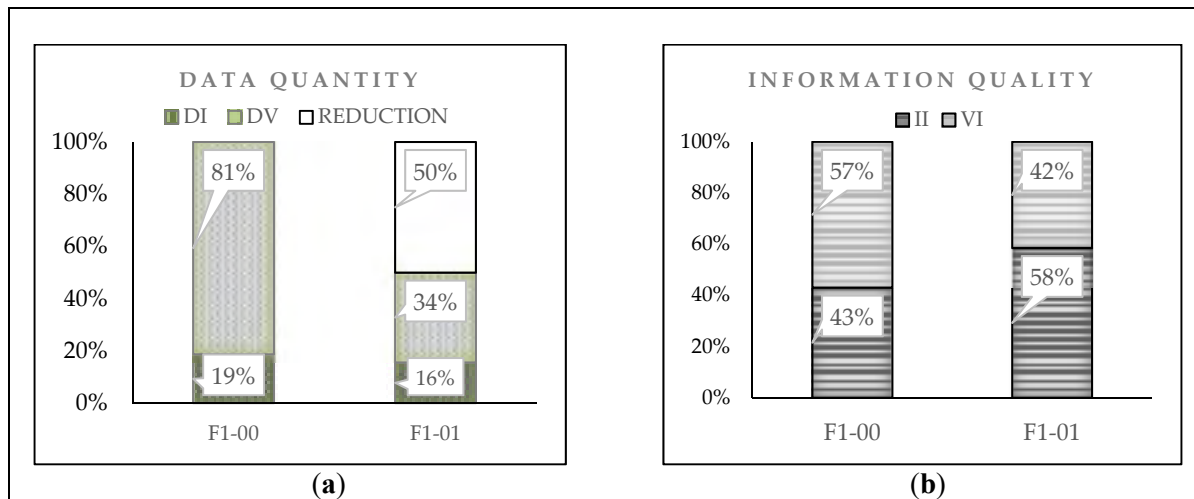


Figure 4.5 Left bar of both graphics: F1-00. Right bar of both graphics: F1-01. Graphic a) Data quantification comparative. Graphic b) Quality produced information

As shown in Figure 4.5, the inbound *du* amount into the system was reduced by 50% in the F1-01 form. This reduction was achieved due to the four major modifications made to the document. In the first place, the redundant fields were eliminated: in the F1-00 form, there were eight different fields asking for the same *du* type. In the second place, in the F1-00 form two *dus* that were considered as indispensable and simple data (first name and last name) were merged in the F1-01 form, becoming only one indispensable composed *du*. The way to convert these *du* from simple to composite (2 *Ds* times w) was by writing in the same field (with a low ink saturation) the format in which it is expected to become the new *du* (last name/first name). In the third place, the computerization of the document considers the possibility of using drop-down menus to select a choice among those already established. The F1-01 form has fewer open fields and more multiple-option fields. Finally, in the fourth place, as a consequence of



this type of menu, now there are more explanatory texts that attempt to clarify and specify to the user the requested *du*.

With regard to the two proposed relationships (DIDV and RIC) to evaluate the *du* input and information output (see Table 4.12), we can mention the following.

Table 4.12 Results of relations DIDV and RIC for forms F1-00 and F1-01

Form	Relation <i>DIDV</i>	Relation <i>RIC</i>
F1-00	1:4.33	1.32
F1-01	1:2.2	0.71

First, the DIDV relation for the F1–00 is equal to 1:4.33 and, for F1–01, this same relation is equal to 1:2.2. In the current study, comparing both results shows that with the new streamlining of the form, the ratio was cut in half. This new design of the form uses only two fields to verify every one. This certainly leads to an improvement in the efficiency of the organization’s information system.

Second, for the RIC ratio, the result of the F1–00 form was 1.32 and the result of the F1–01 was 0.71. Due to the proposed re-engineering, the RIC ratio for the F1–01 is less than 1. This means that there was less information to verify than indispensable information to achieve the process.

The difference in percentage points of the relevant (or indispensable) information quality between the F1–00 and F1–01 forms was 15 points (43% versus 58%). Accordingly, we can infer that the information quality was improved by 15%. What is most interesting is that we pursued the same objective with both forms (the F1–00 and F1–01); both forms achieved the same purpose and captured the same content information and, yet, the second form contained a smaller amount of data and, therefore, a better quality of information.

### 4.3.3 Emphasis on the pertinence of the process

The last two quality attributes will be presented in this section, timeliness (TL) and actual value of information (V<sub>A</sub>). The last quality attribute, V<sub>A</sub>, summarizes the rest of the quality attributes.

#### 4.3.3.1 Timeliness

Regarding the form processing or the manufacturing of information process; first, in the F1-00 form, the original process considers three exchange stations where the main activities are: 1) data recollection, 2) authorization and 3) data verification and processing of information. As it was presented in table 4.6, the second station can lengthen the total processing time (at least 1 day, maximum 8 days). Maybe this station cannot be deleted at all, but considering other elements, the processing time can be reduced considerably, leading to have better timeliness values in general. According to the Pareto law, roughly 80% of cases are similar (I.e. regular students asking for access to some local each scholar period) and 20% can be considered as special cases. Regarding this principle, 80% of similar cases conform to some standards of security verification could be validated by someone who knows the applicant and, the 20% remained must pass through the complete process of three exchange stations. Table 4.13 shown timeliness values that can be obtained if the process were constituted only by two exchange stations. The considered variables taken as a reference were the same as for F1-00.

Table 4.13 Timeliness values that can be obtained if the process were constituted only by two exchange stations

<b>F1-01</b>	<i>in days</i>		<i>in days</i>		<i>in days</i>	
No. subprocess	PTmin	<b>TL(PTmin)</b>	PTave	<b>TL(PTave)</b>	PTmax	<b>TL(PTmax)</b>
1	0.5	0.97	0.97	0.97	1	0.97
2	0.5	0.97	0.97	0.96	2	0.93
<b>TOTALS</b>	1	<b>0.97</b>	1.94	<b>0.96</b>	3	<b>0.95</b>

The information timeliness for the user might be improved, taking the maximum PT, from 0.66 to 0.95, an improvement of 29 percentage points could be obtained. Also, if the information management would assess the timeliness of F1-01 form for a period time, taking this new alternative, he/she would have likely an average of information timeliness high.

Regarding the FIAP-01 processing form, it was found that agent A acted as an information centralizing point, after each update to the document, this was sent back to him/her for verifying and retransmitting it. That means repetitive stages susceptible to changes. Despite the production of this information system is very linear because to have processed 2 (PB2) is necessary to have made process 1 (PB1) and so to have processed 4 to occur (PB4), steps 1, 2 and 3 should have already been made in advance, it was found that it is possible to use the system (SB1) as a buffer and notifier of the document realizations to the participants without necessarily pass through agent A. In figure 4.7 was presented the reengineering proposal made to the process.

The timeliness assessment of this new form processing, FIAP-01, is presented in table 4.14. For comparative purpose the TP taken as reference was that corresponds with FIAP-00. The respective timeliness values of FIAP-00 were included at the table bottom.

Table 4.14 Comparative analysis of timeliness assessment for both FIAP-00 and FIAP-01 forms

<b>FIAP-01</b>	<i>in days</i>		<i>in days</i>		<i>in days</i>	
<b>No. subprocess</b>	<b>PTmin</b>	<b>TL(PTmin)</b>	<b>PTave</b>	<b>TL(PTave)</b>	<b>PTmax</b>	<b>TL(PTmax)</b>
1	1	0.98	1.5	0.98	2	0.97
2	2	0.97	8.5	0.91	15	0.70
3	1	0.98	2	0.97	3	0.97
4	1	0.98	5.5	0.95	10	0.88
5	2	0.97	3.5	0.97	5	0.96
<b>TOTALS</b>	<b>7</b>	<b>0.98</b>	<b>21</b>	<b>0.96</b>	<b>35</b>	<b>0.90</b>
<b>TL values for FIAP-00</b>		<b>0.98</b>		<b>0.96</b>		<b>0.91</b>

As we can observe, the timeliness value does not change significantly either by taking the FIAP-00 or the FIAP-01 way, these values are almost the same meanwhile the processing is carried out between 7 to 21 days. However, a slightly change is observed ones the processing takes the maximum processing time (37 days) using the FIAP-00 way. In this case, the decision maker is who must decide whether realizing the change in the manufacturing of information system worth enough to improving other factors different to timeliness or relevance.

#### 4.3.3.2 Actual value of information

The information value is calculated from the combination of two preceding attributes, the first concerning to the content, *relevance* and the second concerning to the process, *timeliness*. Also, as mentioned before, it is assumed a value of 1.00 either for information intrinsic value information as for a and b exponents which represent clients' sensitivity for both attributes.

Three different scenarios of the weight given by the user to both attributes for the F1-00 and FIAP analysis cases were considered:

- a) the client considers that both relevance and timeliness weigh the same, therefore  $w_r = 0.50$ ;
- b) the information relevance attribute weighs twice more than timeliness,  $w_r = 0.67$ ;
- c) the information timeliness attribute weighs twice more than relevance,  $w_r = 0.33$ .

Obtaining the following values presented in table 4.15 for the F1-00 and F1-01 form and in table 4.14 for FIAP-00 and FIAP-01 respectively.

Table 4.15 Comparative analysis of Actual Information Value for F1-00 and F1-01 forms

Scenario	FORM	RV	Wr	TL	1-Wr	V <sub>A</sub>
a	F1-00	0.43	0.5	0.85	0.5	<b>0.64</b>
	F1-01	0.58	0.5	0.96	0.5	<b>0.77</b>
b	F1-00	0.43	0.67	0.85	0.33	<b>0.57</b>
	F1-01	0.58	0.67	0.96	0.33	<b>0.71</b>

Scenario	FORM	RV	Wr	TL	1-Wr	V <sub>A</sub>
c	F1-00	0.43	0.33	0.85	0.67	<b>0.71</b>
	F1-01	0.58	0.33	0.96	0.67	<b>0.83</b>

The relevance value (RV) of F1-00 was 0.43; and the relevance value (RV) of the F1-01 was 0.58. The timeliness value (TL) considering for computing the actual value of information was the average timeliness value for both forms (F1-00 and F1-01), 0.85 and 0.96 respectively.

The change in quality of the information goes from 12 to 14 percentage points depending upon the weight given by the decision maker. a) 13 points if he/she decides to give the same weight either the content or the process. b) 14 points, in this case, if he/she decides privileges the content more than the process. c) 12 points, if he/she decides that the process weights more than the content.

As it is possible to observe in table 4.16, there is not a significant change in the information value if the information management decides change the manufacturing of information system. This analysis shows that making the change in the process, the information value does not change, or even if it has a slight decrease (-0.01).

Table 4.16 Comparative analysis of Actual Information Value for FIAP-00 and FIAP-01 forms

Scenario	FORM	RV	Wr	TL	1-Wr	V <sub>A</sub>
a	FIAP-00	0.8	0.5	0.91	0.5	<b>0.85</b>
	FIAP-01	0.8	0.5	0.90	0.5	<b>0.85</b>
b	FIAP-00	0.8	0.67	0.91	0.33	<b>0.83</b>
	FIAP-01	0.8	0.67	0.90	0.33	<b>0.83</b>
c	FIAP-00	0.8	0.33	0.91	0.67	<b>0.87</b>
	FIAP-01	0.8	0.33	0.90	0.67	<b>0.86</b>

The relevance value corresponds to the  $I_{rel_{acc}}$  of FIAP-00 is the same for FIAP-01 because no change was made (0.80). The timeliness value considering for computing the actual value of information was the maximum timeliness value for both forms (FIAP-00 and FIAP-01) being almost imperceptible 0.91 and 0.90 respectively.

Finally, to summarize the contextual information quality assessment of the F1-00 and FIAP-00 forms and their respective improvement in the redesign of their content structure and their processing (F1-01 and FIAP-01), table 4.17 is presented. All application forms were considered completed in order to can be processed. We present both quality attributes related to content, sufficiency and relevance, the DIDV and RIC relationships let us perceive at-a-glance, the meaning of the improvement get it. Furthermore, the actual value of information is obtained from the combination of relevance and timeliness quality attributes.

Table 4.17 Contextual Information Quality Assessment for F1 and FIAP forms, versions 00 and 01

FORM	C <sup>IP</sup> (K)	Sufficiency	RV	TL	V <sub>A</sub>
F1-00	1	[0.43,1]	0.43	0.85	0.64
		DIDV 1:4.33	RIC 1.32		
F1-01	1	[0.58,1]	0.58	0.96	0.82
		DIDV 1:2.20	RIC 0.71		
FIAP-00	1	[0.80,1]	0.80	0.90	0.85
		DIDV 1.072	RIC 0.24		
FIAP-01	1	[0.80,1]	0.80	0.89	0.85
		DIDV 1.072	RIC 0.24		

The information value depends on many variables. Here, in the contextual aspect, additionally to the weight given by the user or by the decision maker, two variables were considered, the relevance which deals with the content and the timeliness which deals with temporal pertinence.

The relationship among content attributes is an important point to highlight. A document in general, a form in particular is complete if this has all needed data for the task in hand. This

completeness is composed, in addition to the indispensable data for its processing (given by the relevance value), of the sufficiency data needed according to the user's profile. This, followed by the two relationships developed (DIDV and RIC) provides a perspective of the composition data which built this information.

Additionality to the content, there is the timeliness attributes related to the process which in this study was calculated by a currency proposal (Chi et al., 2017) different to the most adopted in literature (Ballou et al., 1998).

The results of this study imply several benefits for organizations. In the first place, it reinforces the fact that the document has sufficient data for its processing. In the second place, this analysis helps to mitigate problems, such as data overload, that affect the majority of organizations. In the third place, the analysis leads to an improvement in the efficiency of the organization's information system. In the fourth place, it generates a new method for monitoring the quality of the data input and information output.

The F1-00 form possibly contributes to generating the effects of data overload (Edmunds & Morris, 2000 ; Eppler & Mengis, 2004) in workers and to the accumulation of an excess of useless data within the information system. This action, in the end, leads to wastes of material, human, and financial resources. On the contrary, with the use of the F1-01, the organization could contribute to decreasing the data overload of the manufacturing information system, making it more efficient and environmentally friendly.

The analyst of the information system, thanks to this methodology, has a tool to explore a large number of scenarios about the actual value of the information contained in their documents. With this, and following the policies of the company, it is possible to obtain the one which will give to the organization greater benefits. Linking the information with its quality value (actual value) decision takers could have more accurate decisions inside administrative processes.

The complement of some of the comments already issued in this section will be presented in the next chapter. It was considered necessary to leave certain comments aside from the results to emphasize the usefulness of the methodology outlined in each of the stages of the same.



## **CHAPTER 5**

### **DISCUSSION**

Since, as mentioned in the literature review, quality assessments can be negatively impacted by the quantity of data, acting on the quality of data may be a promising strategy. The results of this research show that the application of the contextual information-quality assessment (CIQA) methodology to data processing can help to mitigate both data overload (Edmunds & Morris, 2000) and low-quality data input to the communication system (Redman, 1998b), at both the organizational and operational levels.

The information-quality assessment was based on the new model proposed here [CD]-[PI]-[A], using an approach that links the manufacturing of information with a communication-system perspective. This approach was useful in establishing a new input-data classification paradigm (based on data composition and content), an activity associated with the first stage of the model [CD]. In the second stage of the model [PI], the classification data were weighted. In the third stage [A], the assessment per se was conducted. In this research, the assessment stage was bipartite, emphasizing both content and process. The resultant methodology of this research focused on: 1) how the data is requested in forms; 2) how the data is classified, processed and transformed into information products; and 3) how these information products are contextualized.

The remainder of this chapter presents an overall vision of the content of this thesis, compares this thesis' methodology with previous methodologies and outlines some practical and research-oriented ramifications of the research.

#### **5.1 Contextual Information Quality Assessment (CIQA) methodology**

This research takes a system-communication perspective linked to the manufacturing of information approach (Ballou et al., 1998; Wang, Yang, Pipino, & Strong, 1998) to assess the information quality produced in data processing. It adopts a previously established

classification (Wang & Strong, 1996) of information quality attributes: 1) intrinsic; 2) contextual; 3) representational; and 4) accessibility-related. Only the contextual aspects that included the attributes of completeness, sufficiency, relevance, timeliness and actual information value were analyzed. The first three attributes (completeness, sufficiency, relevance) were strongly associated with document content: the fourth (timeliness) was associated with information processing and temporal pertinence, and the fifth (actual value), was the sum of the other four attributes, with each attribute weighted using user-supplied weighting factors of their importance in the decision-making process.

Figure 5.1 presents a schematic overview of this research. The main subjects with which this research was concerned—data, information, quantity, quality, manufacturing of information system, communication system and data processing—are at the center of this schema. The contextual information-quality assessment (CIQA) methodology is similar to the continuous improvement cycle. The CIQA is composed of 6 stages: definition, data classification, processing of data into information, assessment, and comparative analysis and improvement. The first stage, definition, comprises four sub-stages: 1) the system; 2) the content; 3) the IQ requirements; and 4) the manufacturing of information system. The second stage, data classification, also corresponds to the first part of the [CD]-[PI]-[A] model. Here, the data is classified on the basis of its content and its composition. The data can be indispensable, require verification, or be simple or composite. The third stage, processing data into information, corresponds to the second part of the [CD]-[PI]-[A] model. In this stage, the data previously classified is weighted and transformed into indispensable or verification information types. The fourth stage, assessment, corresponds to the third part of the [CD]-[PI]-[A] model. In this stage, the five contextual quality attributes are evaluated, taking into account the data characteristics analyzed in previous stages. The first three attributes—relevance, sufficiency and completeness—are closely related to each other. Here, two relationships (DIDV and RIC) were developed to indicate in a quantitative way the quality of the document content. The fourth attribute, timeliness, is related to the processing of the form: its value depends mainly on processing time. The fifth attribute, the actual value of the information product, is the sum of each of the four other attributes weighted using user-supplied weighting factors. The fifth

stage, comparative analysis, consists of the analysis of the form's performance in terms of the aforementioned attribute assessments. This comparative analysis is a useful tool for decision support and transactions. Once the analysis is completed, it is advisable to undertake a sixth and final stage: improvement. Here, decision makers make decisions based on the analysis performed in stages 1-5. This six-stage cycle can be followed indefinitely as the organization evolves.

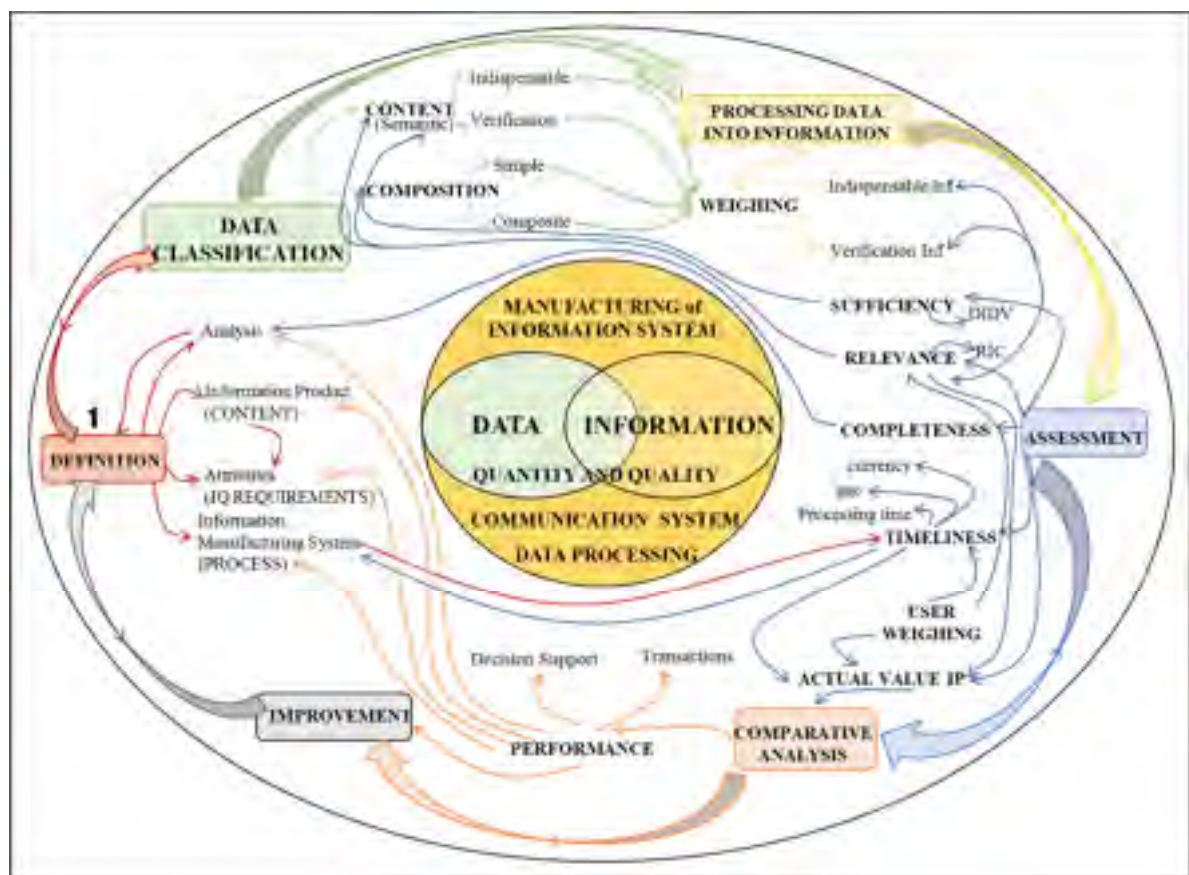


Figure 5.1 The overall vision of this research in a schematic way

The new data representation explored in this research should help to highlight the need for consistent data and information terminology. The results of this investigation show that both the content and the composition of data (among other factors such as the timeliness of the information) are important aspects determining the value of the information, a value that, in the end, will have an impact on the quality of the whole communication and information

system. We found that the data quantification and information-quality assessment are not simply correlated: the quality of information output can increase without any corresponding increase in the quantity of the data input.

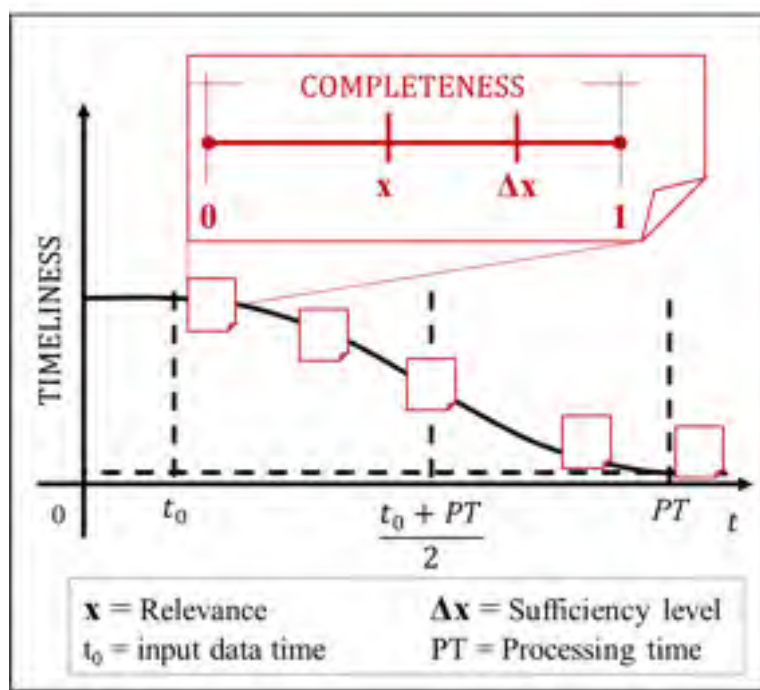


Figure 5.2 Relationships between the three content contextual IQ attributes: completeness, relevance and sufficiency and the temporal pertinence attribute: timeliness

Prior studies (Ballou & Pazer, 2003; Botega et al., 2016; Shankaranarayanan & Cai, 2006) have analyzed the three contextual attributes related to content (completeness, relevance and completeness) independently. Thanks to the data classification of the first stage of the analytical model, this research reveals that they are tightly linked to each other (Figure 5.2). If one considers completeness a line with a range of  $[0,1]$ , where 0 represents an empty document and 1 a completed document. An  $x$  value between this range can be located at any point on this line. This  $x$  value corresponds to the relevance value, any increase of information after the relevant information (represented as  $\Delta x$ ) will be the sufficiency value. At the end we can have the sufficiency level of the information product defined as  $SF^{(IP)} = x + \Delta x$ , where if  $x + \Delta x$  is equal to 1, it will represent at the same time the IP completeness value. In addition to these

values related to content, the value related to the timeliness of the information produced must also be considered. This value decreases as processing time increases, and affects the actual value of the information.

With regard to the role of the channel in the information manufacturing system, it is recommended that, for example, the design of application forms be driven by content not aesthetics: forms should respond to specific business requirements, and reflect the context of their use. In addition, explanatory boxes should be built into the forms to prevent uncertainty by users, who may be either internal or external to the organization. The lack of explanatory boxes could lead to errors, corrections, and unexpected processing delays, with negative consequences for users.

The importance of the process in timeliness assessments was explored using a new mathematical model described in a recent study (Chi et al., 2017). Previous studies have associated timeliness with information volatility (Ballou et al., 1998), leaving the hard task of calibrating exponents to the decision makers. For our part, we considered the mathematical model easier and faster for decision makers. In fact, timeliness value represents a performance index of decision makers' processing times.

The information value may depend on many variables. From a contextual perspective alone, this depends on content pertinence (relevance), temporal pertinence (timeliness) and user weighting. Thanks to the methodology developed here, the information-system analyst can count on a tool to explore a greater number of scenarios based on actual-value information contained in forms. Analysts can use information value, together with organizational policies, to identify the scenario which will give the greatest benefits to the organization.

The comparative analysis demonstrates that the new streamlining of the form results in a 50% reduction of  $du$  in the F1-01 form. This reduction was the result of four major modifications made to the document. First, redundant fields were eliminated: in the F1-00 form, there were eight different fields asking for the same  $du$  type. Second, in the F1-00 form two  $dus$  that were

considered indispensable and simple (first name and last name) were merged, into a single, indispensable, compound (2 *Ds* times w) *du*, in the F1–01 form. The merging of the two simple *dus* was achieved by printing (with a low ink saturation) the expected format of the new *du* (last name/first name). Third, the computerization of the document allows for drop-down menus from which pre-established choices can be selected. The F1–01 form has fewer open fields and more multiple-option fields. Finally, as a consequence of the introduction of this type of menu, there are now more explanatory texts that clarify the requested *dus*.

To the extent possible, our research considered the recent recommendations for data-quality metrics (Heinrich et al., 2018). The CIQA methodology stipulates: 1) minimum and a maximum reference values; 2) contextual specificity; 3) the potential for aggregation (at different analytical levels); and (4) a modest start-up cost that does not jeopardize the economic efficiency of the organization. The only stipulation not respected in this research was the establishment of a scale with intervals. However, we consider that the analyst, once familiar with the methodology, could establish one, using their own decision criteria.

## 5.2 Comparison with previous methodologies

The [DC]-[PI]-[A] model presented in this thesis is based on the manufacturing of information and communication system (MICS) model. This MICS model is distinguished from others—e.g. manufacturing of information or information as product (Ballou et al., 1998; Wang, Yang, Pipino, & Strong, 1998) model and the communication-system model (Masen, 1978)—by the following characteristics:

1. Studies that have used the manufacturing of information approach have generally used the terms “data” and “information” interchangeably, with the same value at the entry to and exit from the system (Ballou et al., 1998; Botega et al., 2016; Kaomea & Page, 1997; Lee et al., 2002; Michnik & Lo, 2009). Very few reports were found that distinguish between these two terms (Masen, 1978; Ronen & Spiegler, 1991; Shankaranarayanan & Cai, 2006), and those that do, do not actually describe how to do so in the field. Considering information and data

equivalent leads to a vision of the data-flow system as a transmission, rather than a communication, system. This research distinguished between these two concepts, in order to avoid misunderstandings. To emphasise the difference between these two concepts, “data” was used to designate input content and “information” was used to designate output content.

2. Masen (1978) considers information an output of a communication system, and presents alternative methods of measuring technical, semantic, pragmatic and functional information. With regard to the semantic aspect, he mentions that information could be measured by the numbers of meaning units transferred between sender and receiver. However, he does not present a method to actually measure this. The method of evaluating the semantic level presented in this thesis considers the information an output of the CS.

3. In contrast to previous reports (Ballou et al., 1998; Shankaranarayanan & Cai, 2006; Shankaranarayanan et al., 2000) that contemplate the document as a data unit, this research considers a document to be a data-block container of several data units, *du*, that are represented according to their distinctive properties (see Chapter 3). The distinction among these *du* is established through a classification based on their composition (simple, composite) and content (indispensable, verification). This representation creates a distinction between data quantification and information assessment. Considering data input and data output separately could be useful in a technical analysis of data transmission. However, the vision of data input and information output implies that in quality information assessments, the finished product has a different value than the raw input material.

Most of the studies of timeliness assessment relate timeliness to volatility and currency (Ballou & Pazer, 1995; Ballou et al., 1998; Islam, 2013). Volatility refers to how long the data remains valid and currency refers to the age of data (Ballou et al., 1998). In his proposal, Ballou (1998) determined the data volatility after performing the timeliness evaluation. However, this data volatility determination depends on analysts’ judgment, which may bias the timeliness result. Other perspectives propose calculating the timeliness value directly from the time variable. There are two examples of this approach. The first one obtains the timeliness value from the

sum of all the times around an event (Botega et al., 2016). However, as Chi (2017) mention, summing resources assumes linear behavior, which does not correspond with the timeliness behavior. In light of this situation, we opted for Chi et al.'s (2017) mathematical model, based on a real case (Chi, 2017) and derived from a sigmoid function, of timeliness. Our form-processing case appears to be analogous to his case. In his case, a high timeliness value corresponds to the arrival of resources in an emergency situation within the range of the opportunity window. In our case, a high timeliness value corresponds to delivery of the form within the minimum acceptable processing time.

### **5.2.1 Previous works in contextual analysis**

Two proposals for contextual-information analysis can be found in the literature (Botega et al., 2016; Kaomea & Page, 1997). Both approaches were developed with reference to emergency contexts, one a military combat operation and the other emergency situations in Brazil.

In the analysis of military combat operations, the author describes the information manufacturing process and the users' needs. From these parameters, he makes system-improvement proposals, in order to deliver relevant information to pilots in a timely and contextualized manner. This is a valuable study, but is not an assessment as such.

In the analysis of the emergency situation, the author adopts the same classification used in this research as his starting point, but directs his analysis specifically to emergency situations. His methodology (IQESA) consists of three steps: 1) collection of data and information requirements; 2) definition of functions and measurement of dimensions; and 3) representation of situational information. As the context of that study was emergency situations, accuracy was measured in terms of syntax. Another evaluated attribute is tolerance for incompleteness, which, unlike in our case, assesses whether a report that does not have 100% of its data completed remains useful. Timeliness is calculated as the sum of time events around the emergency event (the time of the event mentioned in the report, the time it took the report to be processed, and the time required to process information or for objects and attributes to be



found) and the lapse of time (in minutes) since the event happened. Finally, while the IQESA takes into consideration relevance and information consistency, these two attributes do not form a quantitative index, as in our research, but rather are auxiliary factors in the assessment process. Ultimately, however, the IQUESA does reduce uncertainty in decision making in such situations.

### **5.3 Practical and research perspectives**

The information system within organizations is a function of various factors such as information quality, system quality and service quality (DeLone & McLean, 1992). For that reason, information managers should have a "data/information toolbox" which includes the methodology presented here. The following sections present some practical and research perspectives arising out of this research.

#### **5.3.1 Practical perspectives**

The first practical perspective is the development of a contextual information-content matrix for documents. This matrix can help information managers monitor, control and evaluate forms performance. For example, in a form which can be completed by a variety of types of users, this matrix could help determine the adequacy of each profile. Any variation in the range of adequacy for documents could mean that the form has been badly designed or that there has been an error in completing the form. Figure 5.3 shows an example of this matrix for different types of users completing the F1-01 form.

FORM F1-01				CIVIL DATA				TRADING/PT. (T)				STUDENT (S)				ASSOCIATE (AS)				OTHER (O)		RV
				RESEARCH PERSONAL (R)		ADMISSION OTHER (A)		LIBRARY CHANGE (L)		CLUB/RESEARCH OTHER (C)		CLUB/RESEARCH OTHER (C)		TRADING/PT. (T)		OTHER						
Data Type	Datum	n	Int	C <sup>D</sup>	C <sup>D</sup> (k)	C <sup>D</sup>	C <sup>D</sup> (k)	C <sup>D</sup>	C <sup>D</sup> (k)	C <sup>D</sup>	C <sup>D</sup> (k)	C <sup>D</sup>	C <sup>D</sup> (k)	C <sup>D</sup>	C <sup>D</sup> (k)	C <sup>D</sup>	C <sup>D</sup> (k)	C <sup>D</sup>	C <sup>D</sup> (k)			
De/Da	signature	1	0.08	1	0.08	1	0.08	1	0.08	1	0.08	1	0.08	1	0.08	1	0.08	1	0.08	0.58		
De/Da	ID number	15	0.28	1	0.28	1	0.28	1	0.28	1	0.28	1	0.28	1	0.28	1	0.28	1	0.28			
De/Da	Last name first name	6	0.11	1	0.11	1	0.11	1	0.11	1	0.11	1	0.11	1	0.11	1	0.11	1	0.11			
De/Da	Local	1	0.06	1	0.06	1	0.06	1	0.06	1	0.06	1	0.06	1	0.06	1	0.06	1	0.06			
	Expiration date	1	0.06	1	0.06	1	0.06	1	0.06	1	0.06	1	0.06	1	0.06	1	0.06	1	0.06			
De/Da	Contact phone	2	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04			
	Phone type	2	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04			
	Status	2	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04			
	Status 2-A	2	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04			
	Out hours	2	0.01	1	0.01	1	0.01	1	0.01	1	0.01	1	0.01	1	0.01	1	0.01	1	0.01			
	Specify hours	2	0.01	1	0.01	1	0.01	1	0.01	1	0.01	1	0.01	1	0.01	1	0.01	1	0.01			
	Status 2-B	2	0.04	0	0.00	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04			
	Class name	2	0.04	0	0.00	0	0.00	1	0.04	0	0.00	0	0.00	0	0.00	0	0.00	1	0.04			
	Child or adult or auxiliary name	2	0.01	0	0.00	0	0.00	0	0.00	1	0.01	0	0.00	1	0.01	0	0.00	1	0.01			
	Specify mother	2	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.01	1	0.01	1	0.01			
	Date	2	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04			
		53	1.00	SF(RP)	0.85	SF(MA)	0.89	SF(T)	0.92	SF(S)	0.92	SF(AN)	0.92	SF(O)	0.96							
				C <sup>D</sup> (K)																		

Figure 5.3 First practical perspective. Contextual Information Content Measurement for different profiles in the same form

In the same vein, but in another context, the CIQA methodology could be used to produce a form-management (or document-management) matrix. This new matrix can be useful for recording improvements in information quality (as well as data quantity) as a result of document updates. Here, the timeliness of the information and the decision maker's weighting is considered in addition to the information content.

As can be seen in Figure 5.4, the improvement in the actual value of information to the organization due to document updating can be recorded numerically.

FORM MANAGEMENT MATRIX										
F1-00 (REAL)				PT (in dec)	PT (in dec)	PT (in dec)				
				ac 1 = 6	ac 1 = 2	p <sub>0</sub> = 50-50				
				C <sup>0</sup> (R)	RV	SF	IL	V <sup>0</sup> (ac 1)	V <sup>0</sup> (ac 2)	
				1	0.47	[0.43, 1]	0.85	0.96	0.64	0.70
c <sup>0</sup>				22						
Status A	Status B	C <sup>0</sup> (R)	SF profile	IL (0)	IL (2)	V <sup>0</sup> (ac 1)	V <sup>0</sup> (ac 2)			
Employee	Professor FI	12		0.57	0.85	0.96	0.71	0.77		
	Administrative	12		0.57	0.85	0.96	0.71	0.77		
	High employee	12		0.57	0.85	0.96	0.71	0.77		
	Teaching	12		0.57	0.85	0.96	0.71	0.77		
	Intermediary	12		0.57	0.85	0.96	0.71	0.77		
	Guest professor	12		0.57	0.85	0.96	0.71	0.77		
Teaching post at visit	Course in charge	16		0.65	0.85	0.96	0.76	0.82		
	Assistant	16		0.65	0.85	0.96	0.76	0.82		
Student	Club	16		0.69	0.85	0.96	0.77	0.83		
	Research	16		0.69	0.85	0.96	0.77	0.83		
	Other	16		0.69	0.85	0.96	0.77	0.83		
	Intermediary non paid	12		0.65	0.85	0.96	0.74	0.79		
Assistant	Guest professor non paid	18		0.68	0.85	0.96	0.78	0.79		
	Research non paid	18		0.68	0.85	0.96	0.78	0.79		
	Other	15		0.68	0.85	0.96	0.76	0.82		

FORM MANAGEMENT MATRIX										
F1-01				PT (in dec)	PT (in dec)	PT (in dec)	PT (in dec)			
				ac 1 = 6	ac 1 = 2	p <sub>0</sub> = 50-50				
				C <sup>0</sup> (R)	RV	SF	IL	V <sup>0</sup> (ac 1)	V <sup>0</sup> (ac 2)	
				1	0.58	[0.55, 1]	0.88	0.96	0.72	0.77
c <sup>0</sup>				16						
Status A	Status B	C <sup>0</sup> (R)	SF profile	IL (0)	IL (2)	V <sup>0</sup> (ac 1)	V <sup>0</sup> (ac 2)			
Employee	Professor FI	12		0.85	0.85	0.96	0.85	0.91		
	Administrative	13		0.88	0.85	0.96	0.87	0.91		
	High employee	13		0.88	0.85	0.96	0.87	0.95		
	Teaching	13		0.88	0.85	0.96	0.87	0.95		
	Intermediary	13		0.88	0.85	0.96	0.87	0.95		
	Guest professor	12		0.88	0.85	0.96	0.87	0.95		
Teaching post at visit	Course in charge	14		0.92	0.85	0.96	0.89	0.91		
	Assistant	14		0.92	0.85	0.96	0.89	0.91		
Student	Club	14		0.92	0.85	0.96	0.89	0.91		
	Research	14		0.92	0.85	0.96	0.89	0.91		
	Other	14		0.92	0.85	0.96	0.89	0.91		
	Intermediary non paid	14		0.92	0.85	0.96	0.89	0.91		
Assistant	Guest professor non paid	14		0.92	0.85	0.96	0.89	0.91		
	Research non paid	14		0.92	0.85	0.96	0.89	0.91		
	Other	15		0.96	0.85	0.96	0.96	0.96		

Figure 5.4 Second practical perspective. A form management matrix to register improvements in information quality for updating documents

### 5.3.2 Research perspective

The findings of this study have a number of practical and research implications in the field of information management. Some of these implications are commented on below and depicted in Figure 5.5.

The first of these future-work implications (FW1) is the development of new methodologies for the evaluation of IQ, taking into consideration contextual, intrinsic, representational and accessibility-related attributes. To develop a full picture of IQ, the actual value of information must take into consideration the other attribute types. These methodologies could be converted into tools for business management that could be used to design better forms that gather useful, accurate and sufficient data at an opportune time.

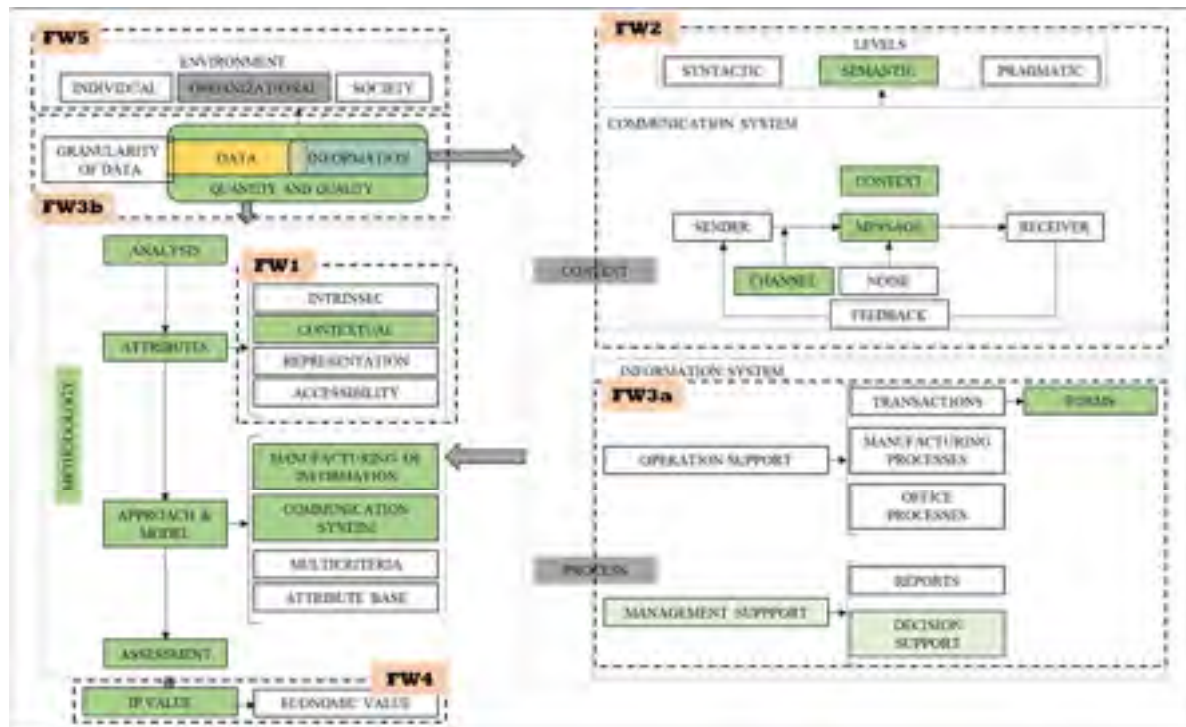


Figure 5.5 Research perspectives emerged from the research conducted. FW = future work

The second future work implication (FW2) is related to the communication system. Here, only the semantic aspect of the information is considered. However, this research could be extended to take into consideration syntactic and pragmatic aspects too. These two aspects can be conceptualized as two layers of the same system. The first layer is the information value gathered from this analysis. The second layer is the measurement of transmitted data in the system—for example, the number of bits obtained by syntactic analysis. It is certainly possible to have the same number of bits in two documents with different actual information values, because of its content and contextual-quality attributes. It would therefore be essential to determine a value which encompasses both information levels without compromising one or the other.

The third future work implication (FW3 a and b) is related to the organizational information system. This thesis only studied one type of office document, application forms. However, information flows in organizational systems comprise more than this type of document (for instance, reports, emails, letters). It would therefore be pertinent to investigate the performance

of different types of document, using the CIQA methodology. Also, different degrees of data granularity should be considered, in order to determine whether this methodology needs some adjustment for use in new cases.

The fourth future work implication (FW4) is based on the use of the same approach taken as a reference, the MICS model, to evaluate the information product—but from an economic perspective. In this economic perspective, the cost of producing information should be considered. This cost calculation should include information-production costs, such as the cost of labour, equipment and infrastructure. The calculated actual value (product of this research), together with the information cost (derived from the results of FW1 and FW2 translated into economic terms), can be used to estimate the profit (or loss) that this good (information) could represent to the company.

Finally, the fifth future work implication (FW5) envisaged is related to the scale of analysis. In this research, only one form was analyzed, but it would also be possible to analyze an entire information system of an organization or a city, as long as the context and specificities of the raw material (data) used to produce the product (information) are taken into consideration.

These implications are only the most representative research avenues. As should be clear, this research is the beginning of a broader research program that would lead, in general, to more efficient and environmentally friendly information-manufacturing systems



## CONCLUSION

The present research was designed to develop a new methodology to assess the contextual-information quality of a processing data. This new methodology is intended to be useful in supporting evident improvements in the efficiency of both the content and the processing of communication channels, as a form. This research had three sub-objectives: 1) based on a new context-oriented representation of data, establishing a process structure for the assessment of the quality of data and information related to data processing; 2) assessing the relevance of information content collected in an application form and the timeliness of the form's manufacturing of information, using a performance index; and 3) comparing the information-product value—in terms of previously developed relevance and timeliness indices—associated with scenarios with and without modifications of the content and processing of an application form.

The first sub-objective of this thesis was to explore a new context-oriented representation of data, for the assessment of the quality of data and information related to data processing. In order to evaluate the effectiveness of this representation, we opted to develop a new approach, the MICS. This approach is an adaptation of both the manufacturing of information and the communication system models. This novel approach to evaluation was very helpful in establishing a data classification method. This classification considered the characteristics of the input components of the system to be raw materials. Based on this approach, a new model to evaluate the information-product quality was developed: the [DC]-[PI]-[A] model. This model has three stages: data classification [DC], processing data into information [PI], and quality assessment [A]. In the first stage, data are classified on the basis of their usefulness and composition. In the second stage, the previously established classification data are weighted in order to process them. In the third stage, the methods for conducting the assessment of the five contextual information quality attributes (completeness, sufficiency, relevance, timeliness and actual value) are developed.

The second sub-objective consisted of assessing the relevance of information content collected in an application form and the timeliness of the form's manufacturing of information process, using a performance index. This sub-objective is in fact the development of the third stage of the [DC]-[PI]-[A] model, the assessment stage. The assessment stage comprises three main parts. The first part is related to the content attributes: completeness, sufficiency and relevance. The second part is related to the process attribute: timeliness. The third part is related to the actual information value, which acts as a summary value of the relevance of the content and the timeliness of the process. In the first part of the assessment stage, the completeness attribute comprises three sub-attributes: 1) completeness at the data level; 2) completeness at the information level; and 3) completeness at the document level. The sufficiency value is dependent on the user's needs and is an indispensable item of information in each user's profile. The relevance value was also considered indispensable, and its weighting was taken into consideration. This research revealed a new relationship between these three contextual attributes of completeness, sufficiency and relevance. The completeness of the document depends on all the data needed for the process. However, the sufficiency of the information depends on the user's needs, which in turn are closely related to information that is relevant (as opposed to necessary for verification). Additionally, two new relationships were found: DIDV and RIC. The DIDV relationship is an indicator of the sufficiency of the input data. The RIC relationship is as an indicator of the relevance of the system's information output.

In the second part of the assessment stage, a mathematical model was adapted for use as a method for the estimation of the information's timeliness (which serves as an index of relevance). In the third part of the assessment stage, once the timeliness value the relevance value and the weights given by the decision maker to these two attributes have been estimated, the actual value of information is obtained. This value turns out to be a useful tool for the decision maker. With this tool, it is possible to demonstrate the improvements made in a form in a quantitative way. The tool also allows the evaluation of different scenarios over different processing times, and provides support for the selection of the most viable scenarios from the point of view of content, processing time, or both.



The third sub-objective consisted of comparing the information-product value—in terms of previously developed relevance and timeliness indices—associated with scenarios with and without modifications of the content and processing of an application form. The analysis revealed that, in this case, the application of the previously developed method and the streamlining of the form led to a 50% reduction of input data. Moreover, the comparison of an original form (F1–00) and a redesigned form (F1–01) revealed that the relevance of the information could be improved by 15%. Both forms achieved the same purpose and captured the same information, but the redesigned form contained less data and, therefore, had better quality of information. Additionally, it was shown that using more data of a composite type (form FIAP–00) can result in information channels of higher quality within the CS.

The results of this investigation show that both the content and the composition of data (among other factors, such as the timeliness of the information product) are important determinants of the value of the information, a value that, in the end, has an impact on the quality of the entire communication and information system. We found that the relation between data quantification and information-quality assessment is not a simple, positive correlation: the quality of information output can increase without there necessarily being any corresponding increase in the quantity of data input.

This new model of data and information evaluation should help to highlight the need for consistent data and information terminology. In the information era, it is no longer feasible to continue using these two terms as synonyms. Once the distinction between the two is made clear, users can treat their data in a more conscientious and responsible way.

This study shows that previously established attributes should be considered in a new classification system. This new classification should be applied at the moment of analysis of the process. If analysis occurs at the beginning of the process, the entities must be treated as data and be evaluated with data-quality attributes (in this case, sufficiency). If analysis is at the end of the process, the entities must be treated as information and be evaluated with an information-quality attribute (in this case, relevance).

Additionally, this study has raised important questions about the nature of the design and processing of forms. With regard to the design of forms, content should take precedence over aesthetics: forms should respond to particular organizational business requirements, with context determining meaning. Completing a forms-based process usually requires authorization of some sort. This authorization, typically issued by heads of departments, are generators of significant bottlenecks, due to these individuals' workloads. Perhaps the solution to these bottlenecks is to give document managers more autonomy (again, without compromising data security or organization). This presupposes a degree of responsibility and conscientiousness on the part of these individuals, but they may well find this in their own best interests, as it would lead to higher quality and more sustainable information and communication systems.

This research was limited to the contextual attributes of relevance, completeness, sufficiency, timeliness and actual value, and did not consider other attributes, such as those related to representativeness. There are several lines of research emanating from this work. These include: application of the methodology to a more complex real case; analysis of the role of the rest of the attributes (intrinsic, representational, and accessibility-related); and analysis of other communication aspects (syntactic, pragmatic). We hope this research represents a breakthrough in the field of information management.

## APPENDIX I

### TABLE A

In reference to point 4.2.1:

A 1 Completeness  $c^p(k)$  and sufficiency ( $sf$ ) values of F1-00, according to each user profile

F1-00			EMPLOYEE				TEACHIN G PT		STUDENT		ANOTHER		ANOTHER	
			RESEARCH PERSONAL		ADM/ HE/ TCH/ INSH/ GP		COURS IN CHARGE/ TCH AUX		CLUB/ RESEARCH / OTHER		INSH NP/ GP NP/ RSCH NP		OTHER	
Data Type	Datum	Irel	C <sup>D</sup>	C <sup>IP</sup> (k)	C <sup>D</sup>	C <sup>IP</sup> (k)	C <sup>D</sup>	C <sup>IP</sup> (k)	C <sup>D</sup>	C <sup>IP</sup> (k)	C <sup>D</sup>	C <sup>IP</sup> (k)	C <sup>D</sup>	C <sup>IP</sup> (k)
Dc/Dis	Student/Emp/other ID	0.21	1	0.21	1	0.21	1	0.21	1	0.21	1	0.21	1	0.21
Ds/Dia	Signature	0.06	1	0.06	1	0.06	1	0.06	1	0.06	1	0.06	1	0.06
Ds/Dis	Last name	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04
	First name	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04
	Locals	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04
	Expiration date	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04	1	0.04
Ds/Dv	Home phone	0.03	0	0.00	0	0.00	1	0.03	1	0.03	1	0.03	1	0.03
	Multiple-choice 1	0.03	0	0.00	1	0.03	0	0.00	0	0.00	0	0.00	0	0.00
	Multiple-choice 2	0.03	0	0.00	0	0.00	1	0.03	0	0.00	0	0.00	0	0.00
	Class name	0.03	0	0.00	0	0.00	1	0.03	0	0.00	0	0.00	0	0.00
	Multiple-choice 3	0.03	0	0.00	0	0.00	0	0.00	1	0.03	0	0.00	1	0.03
	Club	0.03	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	Tutor	0.03	0	0.00	0	0.00	1	0.03	1	0.03	0	0.00	1	0.03
	Especify other1	0.03	0	0.00	0	0.00	0	0.00	1	0.03	0	0.00	1	0.03
	Multiple-choice 4	0.03	0	0.00	0	0.00	0	0.00	0	0.00	1	0.03	1	0.03
	Especify other2	0.03	0	0.00	0	0.00	0	0.00	0	0.00	1	0.03	1	0.03
	Sponsor	0.03	1	0.03	0	0.00	1	0.03	1	0.03	1	0.03	0	0.00
	Especify raison 1	0.03	0	0.00	0	0.00	0	0.00	1	0.03	0	0.00	0	0.00
	Access out of date	0.03	1	0.03	1	0.03	1	0.03	1	0.03	1	0.03	1	0.03
	Especify raison 2	0.03	1	0.03	1	0.03	1	0.03	1	0.03	1	0.03	1	0.03
Date	0.03	1	0.03	1	0.03	1	0.03	1	0.03	1	0.03	1	0.03	
Ds/Dvv	Work phone	0.01	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	Extension phone	0.01	1	0.01	1	0.01	0	0.00	0	0.00	0	0.00	0	0.00
	Depart date	0.01	0	0.00	0	0.00	1	0.01	0	0.00	0	0.00	0	0.00
	End date	0.01	0	0.00	0	0.00	1	0.01	0	0.00	0	0.00	0	0.00
	Extra ID	0.01	0	0.00	1	0.01	0	0.00	1	0.01	0	0.00	0	0.00
	Extra ID	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	Extra ID	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	Extra ID	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	Extra ID	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	Extra ID	0.01	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	C <sup>IP</sup> (K)	1.00	SF	0.57	SF	0.57	SF	0.68	SF	0.69	SF	0.63	SF	0.68
ADM	Administrative				TEACHING PT				Paid partial time teaching					
HE	Help employee				TCH AUX				Auxiliary teaching					
TCH	Professor				INSH NP				Non-paid internship					
INSH	Internship				GP NP				Non-paid guess professor					
GP	Guess professor				RSCH NP				Non-paid research					



## APPENDIX II

### PUBLICATION DURING PH. D STUDY

#### EXPLORING WHETHER DATA CAN BE REPRESENTED AS A COMPOSITE UNIT IN FORM PROCESSING USING THE MANUFACTURING OF INFORMATION APPROACH. INFORMATION.

Monica Blasco-Lopez <sup>1,2</sup>, Robert Hausler <sup>1</sup>, Rabindranarth Romero-Lopez <sup>2</sup>, Mathias Glaus <sup>1</sup>  
and Rafael Diaz-Sobac <sup>3,4</sup>

<sup>1</sup> Station Expérimentale des Procédés Pilotes en Environnement, École de Technologie Supérieure, Université du Québec, 1100, rue Notre-Dame Ouest Local A-1500, Montréal, QC H3C 1K3, Canada;

<sup>2</sup> Unidad de Investigación Especializada en Hidroinformática y Tecnología Ambiental, Facultad de Ingeniería Civil, Universidad Veracruzana, Lomas del Estadio s/n, Zona Universitaria, Xalapa 91000, Mexico;

<sup>3</sup> Instituto de Ciencias Básicas, Universidad Veracruzana, Av. Luis Castelazo Ayala, s/n. Col. Industrial Animas, Xalapa 91190, Mexico; [radiaz@uv.mx](mailto:radiaz@uv.mx)

<sup>4</sup> Universidad de Xalapa, Carretera Xalapa-Veracruz- Km2. No.341, Col. Acueducto Animas, Xalapa 91190, Mexico

This article was published in the Information Journal on April 26, 2019. Information 2019, 10, 156; [doi.org/10.3390/info10050156](https://doi.org/10.3390/info10050156)

**Abstract:** Data and information quality have been recognized as essential components for improving business efficiency. One approach for the assessment of information quality (IQ) is the manufacturing of information (MI). So far, research using this approach has considered a whole document as one indivisible block, which allows document evaluation only at a general level. However, the data inside the documents can be represented as components, which can further be classified according to content and composition. In this paper, we propose a novel model to explore the effectiveness of representing data as a composite unit, rather than indivisible blocks. The input data sufficiency and the relevance of the information output are evaluated in the example of analyzing an administrative form. We found that the new streamlined form proposed resulted in a 15% improvement in IQ. Additionally, we found the

relationship between the data quantity and IQ was not a “simple” correlation, as IQ may increase without a corresponding increase in data quantity. We conclude that our study shows that the representation of data as a composite unit is a determining factor in IQ assessment.

**Keywords:** data quality; information quality; data input; information output; data classification; manufacturing of information; information products; composite data; data representation; IQ assessment

## 1. Introduction

Data quality (DQ) and information quality (IQ) are recognized by business managers as key factors affecting the efficiency of their companies. In the U.S. economy alone, it is estimated that poor data quality costs 3.1 trillion U.S. dollars per year [1]. In order to obtain better information quality, researchers have suggested considering data as a product, and have established the manufacturing of information (MI) approach [2], where data are input to produce output data [3–9] or output information [10–12].

The concept of quality for products has been defined as “fitness for use” [5,13–17]. Meanwhile, for information products (IP), this definition applies only for “information quality” (not for the information alone), because it depends on the perspective of the user. According to the context, one piece of information could be relevant for one user and not relevant for another [16]. For that reason, data and information quality assessment should be evaluated according to required attributes for the business. Some desirable attributes are accuracy, objectivity, reputation, added value, relevancy (related to usefulness), timeliness (related to temporal relevance), completeness, appropriate amount of data (here called “sufficiency”), interpretability, ease of understanding, representational consistency, accessibility, and access security [6,16–21]. Although extensive research has been carried out in this field, data units (du) have always been represented as indivisible blocks (file, document, and so on). No single study exists that represents a du in a different way.

For the DQ and IQ assessment, for our part, we consider that the du structure constitutes a data

block (DB), such as a document. This DB is composed of several dus, and each du can be represented according to its particular characteristics for two types of materials: the first being a pure (simple) material, and the second being a composite material (formed from two or more elements). These characteristics relate to the attributes of sufficiency and relevance and, thus, could have some impact on the IQ assessment of the information products (IP). Relevance has been related to the concept of usefulness [6,16,22], and sufficiency is related to having a quantity of data that is good enough for the purposes for which it is being used [6], not too little nor too much [23]. Both attributes are closely interconnected. The sufficiency of data is a consequence of counting only the relevant information in the system [6]. In order to have relevant information, the document should ideally have only a sufficient quantity of data.

Therefore, the aim of this paper is to explore the effectiveness of representing the data as a composite unit, rather than as an indivisible data block, as has been previously considered. This paper conducts research by the model CD-PI-A (classification of data, processing data into information, and assessment), which is developed to class data, weigh it, and assess the information quality. Data quality is considered to be a dependent factor of (1) the degree of usefulness of the data and (2) the data composition. The applicability of this model is presented through the processing analysis of two organizational forms. These forms are considered as the communication channel which contains requested data. The message is communicated between a sender and a recipient. Once the message is received, the data is transformed into information. The policy, proceedings, and regulations of the organization constitute the context in which communication is done.

In summary, the main contributions of this paper are as follows:

1. The results suggest that this new representation of the data input should be considered in the evaluation of information quality output from a communication system (CS). With the application of the CD-PI-A model developed here, we show that it is possible to pursue and achieve the same objective with two different documents. Thus, it is possible to capture the same information content with a smaller amount of data and produce a better quality of information;

2. This new representation and model for evaluating data and information should help highlight the necessity of the consistent use of data and information terminology;
3. This study shows that, for the already established attributes, a new classification should be considered, according to the moment when the analysis process is made;
4. From the applicability of the CD-PI-A model, we found that the quality of information output can increase without necessarily having a corresponding increase in the quantity of data input.

The remainder of this article is organized as follows: in Section 2, the main case of analysis, an application form is presented. Then, in Section 3, the CD-PI-A model is developed. In Section 4, the results, and its respective discussions are presented. Finally, in Section 5, we present our main conclusions and perspectives for further research.

## **2. Case of Analysis**

The presented case corresponds to the processing of a printed application form (here called F1-00), which flows through the CS of a higher-education institution. Its objective, according to institutional policies, is to grant (or deny) access of a certain installation belonging to the institution. The application form can be filled out by an internal user (belonging to the institution) or an external user (as a guest).

The F1-00 application form is comprised of 32 fields in total, divided into eight sections (as shown in Table 1). The application form consists of open, closed, and multiple-choice fields to fill out. For this analysis, each field was considered as one data unit. The document must pass through two different departments. In these departments, there are three stations that the document must go through to be processed. A station is understood as the point where the document is transformed into semi-processed information (IU), since the person who processes the document makes a change to the process. The first station is where the user or the department secretary fills out the application form with the user data. The second station corresponds to the department director responsible for granting or denying access to the requested installation. Finally, the third station corresponds to the security department that verifies and ends



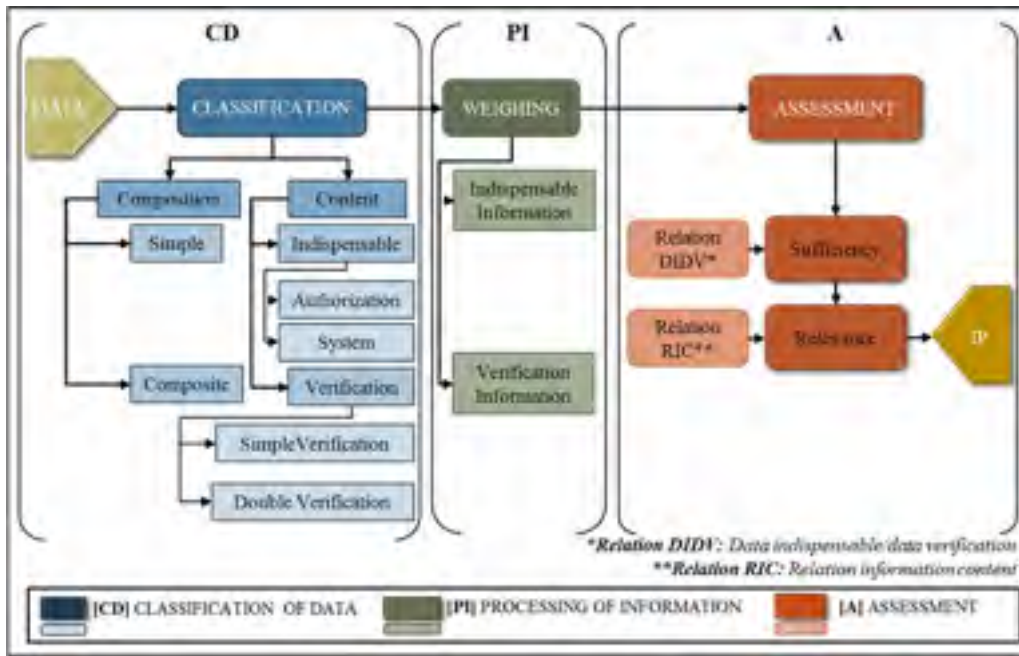
document processing. Semi-structured interviews were conducted with the responsible document processors. From these interviews, *du* were classified according to their characteristics (as will be described in Section 3) and are presented in Table 1.

**Table 1.** (Form F1-00). Structure and *du* classification according to their characteristics. *Ds* = simple data; *Dc* = composite data; *Dia* = indispensable data for authorization; *Dis*= indispensable data for the system; *Dv* = simple verification data; and *Dvv* = double verification data.

Section No.	Section Name	Data ID.	Data	Data Classification
1	Identification	1	Last name	Ds/Dis
		2	First name	Ds/Dis
		3	Home phone	Ds/Dv
		4	Work phone	Ds/Dvv
		5	Extension phone	Ds/Dvv
2	Paid Employee	6	Employee ID	Dc/Dis
		7	Student ID	Dc/Dis
		8	Multiple choice 1	Ds/Dv
3	Paid Partial Time Teaching	9	Employee ID	Ds/Dvv
		10	Student ID	Ds/Dvv
		11	Multiple choice 2	Ds/Dv
		12	Class name	Ds/Dv
		13	Beginning date	Ds/Dvv
		14	End date	Ds/Dvv
4	Paid Researcher	15	Employee ID	Ds/Dvv
		16	Student ID	Ds/Dvv
5	Student by Session	17	Student ID	Ds/Dvv
		18	Multiple choice 3	Ds/Dv
		19	Club name	Ds/Dv
		20	Tutor	Ds/Dv
		21	Other specify 1	Ds/Dv
6	Other (Unpaid or non-students)	22	Temporal ID	Dc/Dis
		23	Multiple choice 4	Ds/Dv
		24	Other specify 2	Ds/Dv
		25	Sponsor	Ds/Dv
		26	Reason 1	Ds/Dv
7	Locals	27	Local numbers	Ds/Dis
		28	Expiration date	Ds/Dis
		29	Access out hours	Ds/Dv
		30	Reason 2	Ds/Dv
8	Authorization	31	Signature	Ds/Dia
		32	Date	Ds/Dv

### 3. Model of Information Quality Assessment: CD-PI-A

The purpose of the model CD-PI-A is to explore the effectiveness of representing the composition of data in information quality assessment. This model is comprised of three phases: (1) classification of data [CD], (2) processing data into information [PI], and (3) assessment of information quality [A], as shown in Figure 1.



**Figure 1.** The classification of data, processing data into information, and assessment (CD-PI-A) model.

Regarding the CS from the context of the MI approach, it is possible to distinguish three main stages in the data processing: (1) the raw material at the entrance (data); (2) the processing period, where data is transformed into pre-processed information. It is considered to be pre-processed as the information that passes from one phase will be the raw material for the next phase, until the end of the process; and (3) the finished product—the information products obtained at the output of the system.

This model initially considers the distinction between the data and information concepts. Here, data has been defined as a string of elementary symbols [24] that can be linked to a meaning related to communication and can be manipulated, operated, and processed [25], and information [26,27] has been defined as a coherent collection of data, messages, or signs,

organized in a certain way that has meaning in a specific human system [28]. In addition, we assume that (1) the communication system works technically well, (2) the office document referred to is a form that belongs to an administrative process, (3) this form is the communication channel in the simplest information system (see reference [29]), and (4) the form flows inside an organization according to its objectives and policies.

### *3.1. Classification of Data (CD)*

Classification involves the process of grouping data into different categories according to similar characteristics [30]. Data is tagged and separated in order to form the groups. In this case, tags are put onto form fields. The classification is made in accordance with the results of semi-structured interviews with the processors of the form. The processors are considered to be skilled and experienced workers in information product manufacturing.

The fields (data collectors) are each recognized as a unit that will host one datum. We consider two types of data representation criteria. It is assumed that each type is associated with a fixed value. The first criterion is its composition. The composition representation has one sub-classification: (1) simple (or pure) data, which considers one symbol to contain only one word; one phrase; one choice box; or, in general, one unit corresponding to one and only one piece of data; and (2) composite data, which is a compound of more than one simple piece of data (more extensive explanation below). The second criterion is its content, which corresponds to the degree in which it is placed, according to importance and frequency-of-use scales. Likewise, the content representation has one sub-classification: (1) indispensable data, which corresponds to data that is absolutely necessary; and (2) verification data, which is used to check the indispensable data. For this second criterion, the order system and the frequency of use are facts dependent on the context. In an office document, the objectives and proceedings, considered as the context, grant the meaning and usefulness levels of the requested data.

We denote TD (total data) as all incoming data units to the system, classifying them as follows:

1. For their composition, the data units can be tagged into two types: (1) simple or (2) composite.

(1) Simple ( $D_s$ ).  $D_s = \{D_{s_i} \mid i = 1, \dots, I\}$ . This is the set of simple data units, where  $D_{s_i}$  is the  $i$ th data unit and  $I$  is the total number of simple *du*s. This type of *du* is composed of one and only one element, such as a name, local identification number, date, signature, and so on. In its transformation into information, the data unit takes the weight value  $w$ . The value of  $w$  is assigned according to the content classification, which is explained via

$$D_s = w. \quad (1)$$

(2) Composite ( $D_c$ )  $D_c = \{D_{c_k} \mid k = 1, \dots, K\}$ . This is the set of data unit composites, where  $D_{c_k}$  is the  $k$ th *du* and  $K$  is the total number of composite data units. This type of data unit is a compound of two or more simple data units, which can be, for example, a registration number, social security number, institutional code, and so on. In its transformation into information, the corresponding weight  $w$  is multiplied by the factor  $x$ , which depends on the number of simple data ( $D_s$ ) units that form the composite data unit:

$$D_c = wx, \quad (2)$$

where

$$x = \sum_{s=1}^n D_s. \quad (3)$$

2. For content, the data units are classified into two types of data representation. These two types of data are indispensable and verification data.

From this classification, the weight value,  $w$ , is assigned. The weight  $w$  is given by the personnel in charge of carrying out the process, since it is assumed that they have the best knowledge of the criteria of data unit importance and the frequencies of use required to process the document. A comprehensive and elaborate case study, presented in reference [31], argues that, through the use of interviews and surveys as a method of analysis, it is possible to examine the factors and the levels of influence of data quality in an organization.

This weight captures the relative importance of a data unit within the process in question. We propose the use of a quantitative scale of discrete values, from 4 to 1, to classify the document fields. The field (or *du*) is classified according to the importance degree for the document processing and the frequency of its use, where 4 corresponds to very important and always used, 3 to important and always used, 2 to slightly important and not always used, and 1 to not at all important and not always used.

(1) Indispensable data (DI),  $DI = \{Dia + Dis\}$ . This type of data unit always appears at some stage in the process and can be one of the following two types:

- Authorization (Dia):  $Dia = \{Dia_m | m = 1, \dots, M\}$ . This is the type of indispensable *du* for authorization, where  $Dia_m$  is the  $m$ th data unit and  $M$  is the total number of indispensable *dus* for authorization. This type of *du* corresponds to the highest value of the weight  $w$ , since it is considered to be a very important *du* for processing. Without this, the system cannot produce the information products. This depends on the approval (or rejection) given by the responsible personnel, according to the policies or organizational procedures.
- System (Dis):  $Dis = \{Dis_n | n = 1, \dots, N\}$ . This is the set of *dus* indispensable for the system, where  $Dis_n$  is the  $n$ th *du* and  $N$  is the total number of indispensable *dus* in the system. This data type is considered to be important. This *du* type is essential within the process and, usually, it corresponds to questions such as who, what, when, where, why, and who authorizes. Without them, the processing of information cannot be completed.

(2) Verification data (DV).  $DV = \{Dv + Dvv\}$ . This *du* type is found frequently during processing; although, in some cases, document processing is carried out without it. This type of *du* can be of two types:

- Simple verification data (Dv).  $Dv = \{Dv_s | s = 1, \dots, S\}$ . This is the simple verification *du* set, where  $Dv_s$  is the  $s$ th *du* and  $S$  is the total number of simple verification *dus*. Some decision-makers consider it necessary to have this kind of unit to make the decision-making process safer [32]. However, without some of these *dus*, data can still be processed. This type of *du* is sometimes used for processing, and it can be considered slightly important;
- Double verification data (Dvv).  $Dvv = \{Dvv_t | t = 1, \dots, T\}$ . This is the double verification *du* set, where  $Dvv_t$  is the  $t$ th *du* and  $T$  is the total number of double verification *dus*. This *du* type is rarely used to verify essential data and it may be not at all important to processing but, in some cases, they are still requested.

### 3.2. Processing Data into Information (PI)

1. In a communication system, there must be a context that serves as a benchmark to determine the pertinence of a *du* in communication. The manufacturing process of information is considered the transformation of raw material (data) into finished products, information. This transformation is represented by the weighting of data after classification (for composition and content).
2. Data transformation into information leads us to give a value to the data units that are at the intersection of the composition and content classifications. Therefore, the possible resulting sets are of two types: (1)  $D_s \cap D_{ia}$ ;  $D_s \cap D_{is}$ ;  $D_s \cap D_v$ ;  $D_s \cap D_{vv}$ , where the value of the data unit (*duv*) corresponds to the weight  $w$  assigned according to the importance and frequency of use criteria mentioned above; and (2)  $D_c \cap D_{ia}$ ;  $D_c \cap D_{is}$ ;  $D_c \cap D_v$ ;  $D_c \cap D_{vv}$ , where the *duv* corresponds to the weight  $w$  multiplied by the  $x$  factor. It is clear that all these sets are mutually exclusive.

Finally, at the system exit, information output is the result of the intersections mentioned above and is grouped in the following manner:

1. Indispensable information (II), which is the result of transforming indispensable *du* (simple or composite, catalogued as either for authorization or for the system transformation) into information through its corresponding *duv* assignment.
2. Verification information (VI), which is the result of transforming verification *du* (simple or composite catalogued as either as simple verification or double verification) into information through its corresponding *duv* assignment.

#### Data Unit Value (*duv*)

To determine the data unit value (*duv*), the combination of both data classifications (composition and content) must be taken as a reference; that is, for its composition (simple or composite data) and for its contents (indispensable or verification). Table 2 shows the values already mentioned.

**Table 2.** Data unit value ( $duv$ ) for simple data, corresponding to the weight  $w$  (which is related to its content). Dia: indispensable data for authorization; Dis: indispensable data for the system; Dv: simple verification data; Dvv: Doble verification data.

Attribute Content	$w$
Dia	4
Dis	3
Dv	2
Dvv	1

In a form, there is usually more than just one type of data; therefore, it is necessary to calculate the data unit value for the same dataset. This is called  $duv_{set}$ , and it is calculated by the following equation, where  $f$  is the frequency of the same type of data:

$$duv_{set} = f(duv). \quad (4)$$

The information relative value ( $Irel$ ) for the document, as an information product, will result in a value between 0 and 1, where 0 corresponds to a null value and 1 to the total of the information products contained in the document.  $Irel_i$ , for one type of information, will be calculated from the following equation, where  $i$  is the set of same type of data (Dc/Dia, Ds/Dia, Dc/Dis, Ds/Dis, Dc/Dv, Ds/Dv, Dc/Dvv, Ds/Dvv) and  $DT$  the total sum of all  $duv_{set}$ .

$$Irel_i = \frac{duv_{set(i)}}{DT(duv_{set})}. \quad (5)$$

The cumulative relative information products ( $Irel_{acc}$ ) calculation is performed according to the following classification:

Information products of the indispensable units (II): this type of IP results from indispensable (simple and composite)  $du$ . It must be ordered as follows: first, the information derived from the authorization type (Dc/Dia, Ds/Dia); and second, for the system (Dc/Dis, Ds/Dis):

$$IIrel_{acc} = \sum Irel(II). \quad (6)$$

Information products of the verification units (IV): this type results from simple verification and double verification data units. It must be ordered as follows: first, the information that corresponds to Dc/Dv and Ds/Dv; and second, the information that derives from the double verification  $du$  (Dc/Dvv, Ds/Dvv):

$$IVrel_{acc} = \sum Irel(IV). \quad (7)$$

### 3.3. Assessment (A)

The last stage of the CD-PI-A model corresponds to the assessment. In order to evaluate the quality of both the data input and the information output, two relationships were developed. These two relationships work as a reference between the real state and the ideal state of the system. They play the role of an indicator of (a) the sufficiency of the requested data (relationship DIDV) and (b) the usefulness of the information gathered through the form (relationship RIC).

#### 3.3.1. Relationship DIDV

The simple ratio as data indispensable/data verification (DIDV) has been used before, to express the desired outcomes to total outcomes [23]. It has been used to evaluate the free-of error, completeness, and consistency [2,33,34]. In this case, the ratio DIDV works as a tool to assess the inbound data unit quality considering the quantity of current data. It indicates, in a simple mode, how many of the verification *dus* exist in relation to the indispensable *dus*. Ideally, in order to reduce the extra amount of *dus* in the data processing and, furthermore, produce a better-quality IP, the form should have a smaller amount of verification *dus* in relation to indispensable *dus*. The formal definition of DIDV is as follows:

$$DIDV = 1: \frac{(DV)}{(DI)}. \quad (8)$$

#### 3.3.2. Relationship RIC

The relation information content (RIC) allows us to know the quality of the information content at the output of the system once the transformation of *du* into an IP is made. The RIC relation considers not only the content but also the *du* composition. This relation expresses, in terms of information, what portion of it is relevant to the aim pursued. Given a comparison between two scenarios of the same form, the one with the lower value represents the best option, as fewer requested fields are used to verify the indispensable information. This ratio is calculated from the following equation:

$$RIC = \frac{IVrel_{acc}}{IIrel_{acc}}. \quad (9)$$



#### 4. Results and Discussion

Once the data were classified and organized according to their composition and content (Table 3), the *duv* was assigned. In form F1–00, two redundant fields were detected. This was possibly due to the structure and organization of the form; the two fields were student ID and employee ID. For our analysis, these two fields were in one instance considered as indispensable data and the rest of the time as double verification data, as it was required only once to carry out the processing.

**Table 3.** Data classification and weighting. Frequency of accumulated data according information type zone ( $D_{acc}$ ), relative frequency of accumulated data according information type zone ( $Drel_{acc}$ ), information relative value ( $Irel$ ), and accumulated information relative value ( $Irel_{acc}$ ) for the F1–00 form.

Information Type	Data Type	$f$	$D_{acc}$	$Drel_{acc}$	$duv$	$duv_{set}$	$Irel$	$Irel_{acc}$
II	Ds/Dia	1			4	4	0.05	
	Signature							
	Dc/Dis	1			15	15	0.21	
	Student ID or employee ID or other ID							
	Ds/Dis	4	6	0.19	3	12	0.17	0.43
IV	Last name, first name, locals, expiration date							
	Ds/Dv	15			2	30	0.42	
	Home phone, multiple choice 1, multiple choice 2, class name, multiple choice 3, club, tutor, other—specify 1, multiple-choice 4, other specify 2, sponsor, raison 1, access out, raison 2, date							
	Ds/Dvv	11	26	0.81	1	11	0.15	0.57
	Work phone, ext-phone, beginning date, end date, redundant IDs (seven times)							

As shown in Table 3, in F1–00 there are six indispensable *du* and 26 verification *du*, which leads to a DIDV 1:4.33 ratio. This is to say, that for each indispensable data that is requested, there are four data units used to verify it. The current structure and design of the form contributes to the generation of data overload in the information manufacturing system. In this case, the data quality attribute of sufficiency is, consequently, not achieved. Unless a security information criterion exists, this relation can be improved by making the relation between different factors shorter. If a security information aspect is not what led to this ratio of 1:4.33, it is necessary to consider form re-engineering in the structure and field composition to request such data. If the organization continues to use the present form, it will continue to contribute to data overload problems in the system.

Regarding the RIC relationship, which considers, in addition to the content, the composition that generates this information, the F1–00 form has 0.57 information products of a verification type (IVacc), and 0.43 information products of an indispensable type (IIacc). According Equation (9), the RIC is equal to 1.32. Ideally, this value should be equal to or less than 1, because the form should request the same amount or less verification information than that of the indispensable type. This relationship works as an indicator of the relevant information content in the CS.

Due to the results of both relationships, it is strongly recommended that the form is re-designed. In this case, we present an alternative.

#### *4.1. Re-Engineering*

As the proceedings for the F1–00 did not establish any set-points regarding extreme security concerns about data gathering, following the document processor's recommendations, we propose a new design for this form. The new design was called F1–01, which is comprised of three main sections: (I) identification, (II) status, and (III) authorization; five fewer sections than the original. Furthermore, the new form is comprised of 16 fields in total.

If the document is chosen to be computerized, then the fields are proposed as drop-down menus. If it is chosen to be in paper format, multiple-option questions are proposed. At a data unit level, in an efficiency assessment we would get a higher value simply by reducing the amount of *du*. At an information product level, due to its contextual aspect, it is necessary to follow the DC-PI-A model in order to assess its quality. Once the results are obtained, it is possible to observe the impact of representing the *du* composition in the assessment of the information quality.

Table 4 shows the data classification and its corresponding transformation into information for the F1–01. A total of 100% of the *du* in the F1–00 form was taken as a reference to calculate the F1–01 form.

As shown in Table 4, in the F1–01 form, five *du*s correspond to indispensable data. These represent 16% (31% of 50%) of the content that was retained in the document. The 11 remaining *du*s represent 34% (69% of 50%) of the same. In the case of the information products, 58% of the preserved fields represent indispensable information, while 42% remained as verification information.

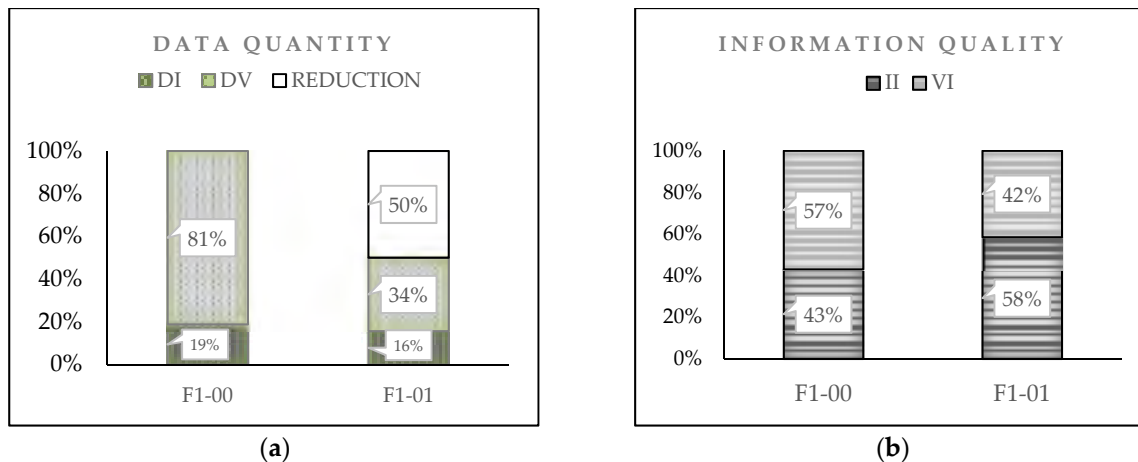
**Table 4.** Data classification and its transformation into information, frequency of accumulated data according information type zone ( $D_{acc}$ ), relative frequency of accumulated data according information type zone ( $D_{rel_{acc}}$ ), information relative value ( $I_{rel}$ ), and accumulated information relative value ( $I_{rel_{acc}}$ ) for the F1–01 form.

Information Type	Data Type	$f$	$D_{acc}$	$D_{rel_{acc}}$	$duv$	$duv_{set}$	$I_{rel}$	$I_{rel_{acc}}$
II	Ds/Dia	1			4	1	0.08	
	Signature							
	Dc/Dis	1			15	15	0.28	
	Student ID or employee ID or other ID							
	Dc/Dis	1			6	6	0.11	
	Last name/first name							
VI	Ds/Dis	2	5	0.16	3	6	0.11	0.58
	Locals, expiration date							
	Ds/Dv	11	11	0.34	22	22	0.42	0.42
	Contact phone, phone type, satatus 1, status 2-A, out hours, specify hours, status 2-B, class name, club or tutor or another name, specify another date							
	Ds/Dvv	-	-	-	-	-	-	-
	n/a							
TOTALS			16	0.50		53		1.00

With the new streamlining of the form, it is possible to (1) reduce the data requested, (2) enhance the information quality produced, and (3) improve the efficiency of the CS. This finding, while preliminary, suggests that a reduction of data does not necessarily mean an improvement in quality of information but a change in the composition of the *du*s do. Additionally, this implies that the quality of information output can increase without necessitating a corresponding increase in the quantity of the data input.

As shown in Figure 2, the inbound *du* amount into the system was reduced by 50% in the F1–01 form. This reduction was achieved due to the four major modifications made to the document. In the first place, the redundant fields were eliminated: in the F1–00 form, there

were eight different fields asking for the same *du* type. In the second place, in the F1–00 form two *du*s that were considered as indispensable and simple data (first name and last name) were merged in the F1–01 form, becoming only one indispensable composed *du*. The way to convert these *du* from simple to composite (2 Ds times w) was by writing in the same field (with a low ink saturation) the format in which it is expected to become the new *du* (last name/first name). In the third place, the computerization of the document considers the possibility of using drop-down menus to select a choice among those already established. The F1–01 form has fewer open fields and more multiple-option fields. Finally, in the fourth place, as a consequence of this type of menu, now there are more explanatory texts that attempt to clarify and specify to the user the requested *du*.



**Figure 2.** (a) Data quantification comparison and (b) quality of produced information for the two forms. Left bar of both graphics: F1–00. Right bar of both graphics: F1–01.

With regard to the two proposed relationships (DIDV and RIC) to evaluate the *du* input and information output (see Table 5), we can mention the following.

**Table 5.** Results for the relations data indispensable/data verification (DIDV) and relation information content (RIC) of the forms F1–00 and F1–01.

Form	Relation DIDV	Relation RIC
F1–00	1:4.33	1.32
F1–01	1:2.2	0.71

First, the DIDV relation for the F1–00 is equal to 1:4.33 and, for F1–01, this same relation is equal to 1:2.2. In the current study, comparing both results shows that with the new

streamlining of the form, the ratio was cut in half. This new design of the form uses only two fields to verify every one. This certainly leads to an improvement in the efficiency of the organization's information system.

Second, for the RIC ratio, the result for the F1-00 form was 1.32 and the result for the F1-01 was 0.71. Due to the proposed re-engineering, the RIC ratio for the F1-01 is less than 1. This means that there was less information to verify than indispensable information to achieve the process.

The difference in percentage points of the relevant (or indispensable) information quality between the F1-00 and F1-01 forms was 15 points (43% versus 58%). Accordingly, we can infer that the information quality was improved by 15%. What is most interesting is that we pursued the same objective with both forms (the F1-00 and F1-01); both forms achieved the same purpose and captured the same content information and, yet, the second form contained a smaller amount of data and, therefore, a better quality of information.

Below is presented another applicability case where the requested fields have a different characterization in their classification.

#### *4.2. Another Example of Applicability*

The form FIAP-00 has as objective to recollect and summarize all needed data for a research project within an educational institution. The FIAP-00 alternates its format on paper and in electronic within the CS. The document is an internal communication medium; therefore, there are no external agents involved in the information-product manufacturing system. FIAP-00 application form is comprised of 79 fields in total divided into four sections. The application form consists of open, closed, and multiple-choice fields to fill out. The document must pass through five different interchange stations belonging to three different departments. In department 1, the first station is where the agent a fills out the application form with the project data. In department 2, the second station is where the agent b fills out the budget data of the project. The form returns to department 1 where the next two stations are; the third station

corresponds to the agent c who fills out another project data; and agent d corresponds to the department director, responsible for granting the authorization. Finally, in department 3, the fifth station corresponds to the agent e, who verifies and ends document processing. In the case of the FIAP-00 form, fields are not promptly mentioned for safety reason but in Table 6 their classification and transformation into information phases are presented.

Once the data were classified and organized according to their composition and content (Section 3), the  $duv$  was assigned. In form FIAP-00, no double verification data were detected. Table 6 shows the data classification with its corresponding weighting, information relative value ( $I_{rel}$ ), and accumulated information relative value ( $I_{rel_{acc}}$ ), from Equations (4)–(7). Because there are different composite data in the form and to make clearer the data transformation into information process, in Table 6 two columns were added: factor  $x$ , which corresponds to Equation (3), and weight  $w$ , which correspond to Table 2.

**Table 6.** Data classification and its transformation into information for the form FIAP-00.

Information Type	Data Type	Factor $x$	$w$	$f$	$D_{acc}$	$D_{rel_{acc}}$	$duv$	$duv_{set}$	$I_{rel}$	$I_{rel_{acc}}$
II	Ds/Dia		4	2			4	8	0.02	
		11	3	1			33	33	0.10	
		9	3	1			27	27	0.08	
		7	3	1			21	21	0.06	
	Dc/Dis	6	3	2			18	36	0.11	
		5	3	1			15	15	0.04	
		4	3	1			12	12	0.03	
		3	3	1			9	9	0.03	
		2	3	1			6	6	0.02	
	Ds/Dis		3	35	46	0.58	3	105	0.31	0.80
VI	Ds/Dv		2	33			2	66	0.20	
	Ds/Dvv		1	0	33	0.42	1	0	0	0.20
	TOTALS				79	1.00		338	1.00	1.00

Unlike the F1-00, the form FIAP-00 has more Dc/Dis than Ds/Dv type. The DIDV relation for the FIAP-00 results in 1:0.72; this means that there was less than one data to verify the indispensable information to achieve the process. In the case of the RIC relationship, the FIAP-00 form is equal to 0.24. This means that only one quarter of the fields are used to verify the indispensable information. In the FIAP-00 case, to have more Dc/Dis types, it helps to have a

higher quality information channel in the CS. The combination of these findings provides some support for the conceptual premise that the data representation as either simple or composite in the information quality assessment is relevant.

The results of this study imply several benefits for organizations. In the first place, it reinforces the fact that the document has sufficient data for its processing. In the second place, this analysis helps to mitigate problems, such as data overload, that affect the majority of organizations. In the third place, the analysis leads to an improvement in the efficiency of the organization's information system. In the fourth place, it generates a new method for monitoring the quality of the data input and information output.

The F1-00 form possibly contributes to generating the effects of data overload [35,36] in workers and to the accumulation of an excess of useless data within the information system. This action, in the end, leads to wastes of material, human, and financial resources. On the contrary, with the use of the F1-01 or FIAP-01, the organization could contribute to decreasing the data overload of the manufacturing information system, making it more efficient and environmentally friendly.

#### *4.3. Comparison with Previous Work*

The CD-PI-A model presented in this paper is distinguished from others models that use the manufacturing of information or information as a product [2,7] approach as a reference according to the following characteristics:

1. Reports that had used the manufacturing of information approach generally used the terms data and information interchangeably, giving them the same value at the entrance and at the exit of the system [2,21,37-39]. Very few reports were found that made a distinction between these two terms [5,12], and those that did were only at a conceptual level. The fact of addressing the information at the same level of data leads us to consider the system by which the flow of data acts more like a transmission than a communication system. In this paper, we established, to the extent possible, the distinction between these two concepts in

order to avoid misunderstandings and to be consistent with the proposal. The criterion to underline the difference between these two concepts was to use the terms according to the processing moment in which they were applied.

2. With regards to the proposal of reference [12], where information was considered as an output of a communication system, different alternatives for measuring the information were presented. Three levels of information were considered: technical, semantic, and pragmatic, and a fourth level, the functional, was also added. Regarding the semantic aspect, it was mentioned that the information could be measured by the numbers of meaningful units between the sender and receiver. However, a method to carry it out was not presented. For our part, we propose a method to evaluate the semantic level, which considers the information as an output of the CS.
3. Additionally, in contrast to previous reports [2,11,40] that considered the document as a data unit, this research considers one document as a data block container of several data units, *dus*, that are represented according to their distinctive properties. The distinction among these *dus* is established through a classification, in accordance with their composition and content. This representation creates a distinction between data quantification and information assessment. Furthermore, it considers that data input and data output could be useful in a technical analysis of data transmission. However, the vision of data input and information output implies that, in the quality information assessment, the finished product has a different value than the initial raw material.

## 5. Conclusions

The present study was designed to explore the effectiveness of representing data as composite entities rather than indivisible blocks in the manufacturing of information domain, in order to assess the quality of information produced.



In order to evaluate this effectiveness, the authors opted to integrate a communication system vision into the manufacturing information approach in order to establish a new data classification method that considered the context in which this information was produced.

Based on this approach, a new model to evaluate the information product quality was developed: the DC-PI-A model. This model uses three stages: data classification (DC), processing of data into information (PI), and quality assessment (A). In the first stage, data are classified according to their usefulness and composition. In the second stage, the previous classification data are weighted in order to process them. In the third stage, in order to conduct the assessment, two relationships are proposed. These relationships work as indicators of the attributes mentioned below.

The relationship DIDV works as an indicator of the sufficiency of the input data. In an investigation, with the application of this relationship and the new streamlining of a form, 50% of the input data to a system was reduced. The relationship RIC works as an indicator of relevance of information output of the system. In our case, the comparison between the original form F1–00 and the re-designed form F1–01 showed that the quality of information, in relation to its relevance, could be improved by 15%.

We pursued the same objective with different forms (F1–00 and F1–01), where both forms achieved the same purpose and captured the same information content, yet the second form contained a smaller amount of data and, therefore, had better quality of information. Additionally, it was shown that by using more composite type data (FIAP–00) it can be possible to have higher information quality channels within the CS.

The results of this investigation show that both the content and the composition of data (among other factors) are important aspects of determining the value of the information; value that, in the end, will have an impact on the quality of the whole communication and information system. We found that the relation between data quantification and information quality evaluation is not just a “simple” positive correlation. The quality of information output can

increase without there necessarily being any corresponding increase in the quantity of the data input.

This new representation and model for evaluating data and information should help to highlight the necessity of consistent use of data and information terminology. In the information era, it is not possible to continue to use these two terms as synonyms. Once delimiting this distinction, users can treat their data in a more conscious and responsible way.

This study shows that the attributes already established should be considered as a new classification. This new classification should be applied at the moment of the process when the analysis is made. If it is at the beginning of the process, the entities must be treated as data and have to be evaluated with data quality attributes (in this case, sufficiency). If it is at the exit of the system, the entities must be treated as information and have to be evaluated with an information quality attribute (in this case, relevance).

Additionally, this study has raised important questions about the nature of the design of forms. This should be a matter of content more than an aesthetic issue. Inside an organization, the forms should respond to the particular business requirements, where the context determines the meaning.

The scope of this study was limited to exploring only two attributes of quality: sufficiency and relevance. Further work will need to be done to determine more accurate information values from this same approach. We wish to include other attributes, such as accuracy, completeness, or timeliness. Additionally, including the syntactic and pragmatic levels of information would be valuable. Likewise, as one external reviewer suggested, the inter-connection between the DB concept, here presented, and the data granularity linked with different types of documents may be of interest.

The findings of this study have a number of practical implications in the field of information management. One example of these implications would be the development of new

methodologies to evaluate the IQ. These methodologies could be converted into tools for business management. These tools would be used to design better forms that gather useful and sufficient data. All these changes would lead us, in general, to have more efficient and environmentally friendly information manufacturing systems.

We hope our study exploring the effectiveness of representing data as composite units will introduce some guidelines for further research and will inspire new investigations in the same field but at a more detailed level.



## BIBLIOGRAPHICAL REFERENCES

- Ackoff, R. L. (1967). Management Misinformation Systems. *Management Science*, 14(4), B-147-B-156. <https://doi.org/10.1287/mnsc.14.4.B147>
- Ahituv, N. (1980). A Systematic Approach toward Assessing the Value of an Information System. *MIS Quarterly*, 4(4), 61. <https://doi.org/10.2307/248961>
- Aleksi, S. (2011). Thermodynamic Aspects of Communication and Information Processing Systems. Dans *13th International Conference on Transparent Optical Networks* (pp. 1-4). Stockholm, Sweden : IEEE.
- Bae, H., Hu, W., Yoo, W. S., Kwak, B. K., Kin, Y., & Park, Y. T. (2004). Document configuration control processes captured in a workflow. *Computers in Industry*, 53(2), 117-131. <https://doi.org/10.1016/j.compind.2003.07.001>
- Bae, H., & Kim, Y. (2002). A document-process association model for workflow management. *Computers in Industry*, 47(2), 139-154. [https://doi.org/10.1016/S0166-3615\(01\)00150-6](https://doi.org/10.1016/S0166-3615(01)00150-6)
- Ballou, D. P., & Pazer, H. L. (1985a). Modeling data and process quality in multi-input, multi-output information systems. *Management Science*, 31(2), 123-248. <https://doi.org/doi.org/10.1287/mnsc.31.2.150>
- Ballou, D. P., & Pazer, H. L. (1985b). Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. *Management Science*, 31(2), 150-162. <https://doi.org/10.1287/mnsc.31.2.150>
- Ballou, D. P., & Pazer, H. L. (1995). Designing information systems to optimize the accuracy-timeliness tradeoff. *Information Systems Research*. <https://doi.org/10.1287/isre.6.1.51>
- Ballou, D. P., & Pazer, H. L. (2003). Modeling completeness versus consistency tradeoffs in information decision contexts. *IEEE Transactions on Knowledge and Data Engineering*, 15(1), 241-244. <https://doi.org/10.1109/TKDE.2003.1161595>
- Ballou, D. P., Wang, R., Pazer, H., & Tayi, G. K. (1998). Modeling Information Manufacturing Systems to Determine Information Product Quality. *Management Science*, 44(4), 462-484. <https://doi.org/10.1287/mnsc.44.4.462>
- Barnett, R. (2007). Designing Useable Forms : Sucess Guaranteed. Retrieved from [http://c.ymcdn.com/sites/www.bfma.org/resource/resmgr/Articles/07\\_46.pdf](http://c.ymcdn.com/sites/www.bfma.org/resource/resmgr/Articles/07_46.pdf)
- Batini, C., & Scannapieco, M. (2016a). *Data and Information Quality. Dimensions, Principles and Techniques*. Springer. (S.l.) : Elsevier B.V. [https://doi.org/10.1007/978-3-319-24106-7\\_Library](https://doi.org/10.1007/978-3-319-24106-7_Library)

- Batini, C., & Scannapieco, M. (2016b). Introduction to Information Quality. Dans *Data and Information Quality* (pp. 1-19). (S.l.) : (s.n.). <https://doi.org/10.1007/978-3-319-24106-7>
- Batini, C., & Scannapieco, M. (2016c). Models for Information Quality. Dans *Data and Information Quality*. (S.l.) : (s.n.). <https://doi.org/10.1007/978-3-319-24106-7>
- Beniger, J. R. (1988). Information and Communication. The new Convergence. *Communication Research*, 15(2), 198-218.
- Berlo, D. K. (1976). El proceso de la comunicacion. *Journal of Communication*. <https://doi.org/10.1111/j.1460-2466.1976.tb01898.x>
- Botega, L. C., de Souza, J. O., Jorge, F. R., Coneglian, C. S., de Campos, M. R., de Almeida Neris, V. P., & de Araújo, R. B. (2016). Methodology for Data and Information Quality Assessment in the Context of Emergency Situational Awareness. *Universal Access in the Information Society*, 889-902. <https://doi.org/10.1007/s10209-016-0473-0>
- Bovee, M., Srivastava, R. P., & Mak, B. (2003). A conceptual framework and belief-function approach to assessing overall information quality. *International Journal of Intelligent Systems*, 18(1), 51-74. <https://doi.org/10.1002/int.10074>
- Brunschwiler, T., Smith, B., Ruetsche, E., & Michel, B. (2009). Toward zero-emission data centers through direct reuse of thermal energy. *IBM Journal of Research and Development*, 53(3), 11:1-11:13.
- Butcher, H. (1995). Information overload in management and business. Dans *IEE Colloquium Digest* (pp. 1-2). London.
- Cambridge, D. (2018). Meaning of « value » in the English Dictionary. Retrieved from <https://dictionary.cambridge.org/dictionary/english/value>
- Cambridge, D. (2019). Sufficiency. Retrieved from <https://dictionary.cambridge.org/dictionary/english/sufficiency>
- Chen, J., Wang, T. T., & Lu, Q. (2016). THC-DAT: a document analysis tool based on topic hierarchy and context information. *Library Hi Tech*, 34(1), 64-86. <https://doi.org/10.1108/LHT-07-2015-0074>
- Chewning, E. G., & Harrell, A. M. (1990). The effect of information load on decision makers' cue utilization levels and decision quality in a financial distress decision task. *Accounting, Organizations and Society*, 15(6), 527-542. [https://doi.org/10.1016/0361-3682\(90\)90033-Q](https://doi.org/10.1016/0361-3682(90)90033-Q)
- Chi, H., Li, J., Shao, X., & Gao, M. (2017). Timeliness evaluation of emergency resource scheduling. *European Journal of Operational Research*, 258(3), 1022-1032. <https://doi.org/10.1016/j.ejor.2016.09.034>

- CIHI. (2017). *CIHI's Information Quality Framework (White Paper)*. (S.l.) : (s.n.). Retrieved from <https://www.cihi.ca/en/submit-data-and-view-standards/data-and-information-quality>
- Clarke, R., & O'Brien, A. (2012). The Cost of Too Much Information: Government Workers Lose Productivity Due to Information Overload. ... *Government Insights, Iron Mountain*, (February 2012). Retrieved from [http://www.emea.ironmountain.com/Elq/Federal-Government/~/\\_media/D0CF180AE56E439F998EB5595D91EF83.pdf](http://www.emea.ironmountain.com/Elq/Federal-Government/~/_media/D0CF180AE56E439F998EB5595D91EF83.pdf)
- DeLone, W. H., & McLean, E. R. (1992). Information Systems Success: The Quest for the Dependent Variable. *Information Systems Research*, 3(1), 60-95.
- DeLone, W. H., & McLean, E. R. (2003). The DeLone and McLean model of information systems success: A ten-year update. *Journal of Management Information Systems*, 19(4), 9-30. <https://doi.org/10.1080/07421222.2003.11045748>
- Deming, W. E. (1986). *Out of the Crisis*. Cambridge : MIT Press.
- Denning, Peter; Bell, T. (2012). The information paradox. *American Scientist*, Nov-Dec, 470-477. [https://doi.org/10.1007/978-3-540-74233-3\\_20](https://doi.org/10.1007/978-3-540-74233-3_20)
- Earl, M. J. (2000). Toutes les entreprises font de l'information. Dans *L'Art du Management de l'information. Gérer le savoir par les technologies de l'information* (p. 373). Paris : Les Echos.
- Ebrahimi, K., Jones, G. F., & Fleischer, A. S. (2015). Thermo-economic analysis of steady state waste heat recovery in data centers using absorption refrigeration. *Applied Energy*, 139, 384-397. <https://doi.org/10.1016/j.apenergy.2014.10.067>
- Edmunds, A., & Morris, A. (2000). The problem of information overload in business organisations: a review of the literature. *International Journal of Information Management*, 20(1), 17-28. [https://doi.org/10.1016/S0268-4012\(99\)00051-1](https://doi.org/10.1016/S0268-4012(99)00051-1)
- English, L. P. (1999). *Improving data warehouse and business information quality methods for reducing cost and increasing profits*. New York : Wiley.
- Eppler, M. J., & Mengis, J. (2004). The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines. *Information Society*, 20(5), 325-344. <https://doi.org/10.1080/01972240490507974>
- Eppler, M. J., & Muenzenmayer, P. (2002). Measuring Information Quality in the Web Context : A survey of state-of-art instruments and an application methodology (Practice-Oriented). *Proceedings of the Seventh International Conference on Information Quality (ICIQ-02)*, 187-196. <https://doi.org/10.1.1.477.4680>
- Ferrer, E. (1994). *El lenguaje de la publicidad*. México : Fondo de Cultura Económica.

- Fiorani, M., Aleksic, S., & Casoni, M. (2014). Hybrid optical switching for data center networks. *Journal of Electrical and Computer Engineering*, (January). <https://doi.org/10.1155/2014/139213>
- Fisher, P., & Sless, D. (1990). Information design methods and productivity in the insurance industry. *Information Design Journal*, 6(2), 103-129. <https://doi.org/10.1075/idj.6.2.01fis>
- Floridi, L. (2009). The information society and its philosophy: Introduction to the special issue on « the philosophy of information, its nature, and future developments ». *Information Society*, 25(3), 153-158. <https://doi.org/10.1080/01972240902848583>
- Fonseca Yerena, M. del S., Correa Pérez, A., Pineda Ramírez, M. I., & Lemus Hernández, F. (2016). *Comunicación Oral y Escrita* (Segunda). México D.F. : Pearson Education de México.
- Forslund, H. (2007). Measuring information quality in the order fulfilment process. *International Journal of Quality and Reliability Management*, 24(5), 515-524. <https://doi.org/10.1108/02656710710748376>
- Galbraith, J. R. (1974). Organization design: An information processing view. *Interfaces*, 4(3), 28-36. <https://doi.org/10.1287/inte.4.3.28>
- Gantz, J., & Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the future, 2007*(December), 1-16. <https://doi.org/10.1098/rspl.1860.0124>
- Han, J., & Jian Pei, M. K. (2012). *Data mining: concepts and techniques* (3rd ed). Boston, Mass. : (s.n.).
- Hayes, R. M. (1993). Measurement of information. *Information Processing & Management*, 29(1), 1-11. [https://doi.org/10.1016/0306-4573\(93\)90019-A](https://doi.org/10.1016/0306-4573(93)90019-A)
- Heinrich, B., Hristova, D., Klier, M., Schiller, A., & Szubartowicz, M. (2018). Requirements for Data Quality Metrics. *Journal of Data and Information Quality*, 9(2), 1-32. <https://doi.org/10.1145/3148238>
- Henno, J. (2014). Grounded multi-level computations. Dans N. Thalheim, Bernhard; Jaakkola, Hannu; Kiyoki, Yasushi; Yoshida (Éd.), *Information modelling and knowledge bases XXVI* (pp. 140-151). Amsterdam : IOS Press BV. <https://doi.org/10.3233/978-1-61499-472-5-140>
- Hilbert, M., & López, P. (2012). How to measure the world's technological capacity to communicate, store, and compute information, part I: Results and scope. *International Journal of Communication*, 6(1), 956-979. <https://doi.org/10.1126/science.1200970>



- IBM Big Data and Analytics Hub. (2016). Extracting Business Value from the 4 V's of Big Data. Retrieved from <https://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data>
- Islam, M. S. (2013). Regulators of timeliness data quality dimension for changing data quality in information manufacturing system (IMS). *3rd International Conference on Digital Information Processing and Communications, ICDIPC 2013*, 126-133. Retrieved from <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84978664322&partnerID=40&md5=c123318bb1a1cecac9a443df710a0c23>
- ISO. (1994). Australian / New Zealand Standard Quality management and quality assurance — Vocabulary ISO 8402:1994.
- Jarke, M., Jeusfeld, M., Quix, C., & Vassiliadis, P. (1999). Architecture and quality in data warehouses: an extended repository approach, *Information Systems. Information Systems*, 24(3), 229–253.
- Jarke, M., Lenzerini, Vassiliou, Y., & Vassiliadis, P. (1999). *Fundamentals of Data Warehouses*. (S.l.) : Springer Verlag.
- Juran, J. M. (1989). *Juran on Leadership for Quality*. New York : Free Press.
- Kaomea, P., & Page, W. (1997). A flexible information manufacturing system for the generation of tailored information products. *Decision Support Systems*, 20(4), 345-355. [https://doi.org/10.1016/S0167-9236\(96\)00067-X](https://doi.org/10.1016/S0167-9236(96)00067-X)
- Kinsella, S., Baffoni, S., Anderson, P., Ford, J., Leithe, R., Smith, D., ... Blacksmith, S. (2018). *The State of the Global Paper Industry 2018*. <https://doi.org/10.1016/j.joen.2014.06.003>
- Koomey, J. (2012). Growth in Data Center Electricity use 2005 to 2010. *Analytics Press.*, 1-24. <https://doi.org/10.1088/1748-9326/3/3/034008>
- Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: A methodology for information quality assessment. *Information and Management*, 40(2), 133-146. [https://doi.org/10.1016/S0378-7206\(02\)00043-5](https://doi.org/10.1016/S0378-7206(02)00043-5)
- LEXICO. (2019). information age. Retrieved from [https://www.lexico.com/en/definition/information\\_age](https://www.lexico.com/en/definition/information_age)
- Logan, R. K. (2012). What is information?: Why is it relativistic and what is its relationship to materiality, meaning and organization. *Information (Switzerland)*, 3(1), 68-91. <https://doi.org/10.3390/info3010068>
- Lyman, P., & Varian, H. R. (2003). « *How much information* » 2003.

- Madnick, S. (1995). Integrating information from global systems: Dealing with the “on-and off-ramps”; of the information superhighway. *Journal of Organizational Computing and Electronic ...*, (March 2015), 37-41. <https://doi.org/10.1080/10919399509540243>
- Madnick, S., Wang, R. Y., Lee, Y. W., & Zhu, H. (2009). Overview and Framework for Data and Information Quality Research. *ACM Journal of Data and Information Quality*, 1(1), 1-22. <https://doi.org/10.1145/1515693.1516680.http>
- Masen, R. O. (1978). Measuring Information Output a communication systems approach. *Information and Management*, 1, 219-234. [https://doi.org/dx.doi.org/10.1016/0378-7206\(78\)90028-9](https://doi.org/dx.doi.org/10.1016/0378-7206(78)90028-9)
- Meadow, C. T., & Yuan, W. (1997). Measuring the impact of information: defining the concepts. *Information Processing & Management*, 33(6), 697-714.
- Michnik, J., & Lo, M. C. (2009). The assessment of the information quality with the aid of multiple criteria analysis. *European Journal of Operational Research*, 195(3), 850-856. <https://doi.org/10.1016/j.ejor.2007.11.017>
- Missier, P., & Batini, C. (2003). A Multidimensional Model for Information Quality in Cooperative Information. *Proceedings of the Eighth International Conference on Information Quality (ICIQ-03)*, 25-40. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.1.5368>
- Moore, S. (2017). How to Create a Business Case for Data Quality Improvement. Retrieved from <https://www.gartner.com/smarterwithgartner/how-to-create-a-business-case-for-data-quality-improvement/>
- Ozsu, T., & Valduriez, P. (2000). *Principles of Distributed Database System*. New York : Springer Science & Business Media.
- Pärssinen, M., Wahlroos, M., Manner, J., & Syri, S. (2019). Waste heat from data centers: An investment analysis. *Sustainable Cities and Society*, 44(July 2018), 428-444. <https://doi.org/10.1016/j.scs.2018.10.023>
- Pipino, L. L., Lee, Y. W., Wang, R. Y., Lowell Yang Lee, M. W., & Yang, R. Y. (2002). Data Quality Assessment. *Communications of the ACM*, 45(4), 211. <https://doi.org/10.1145/505248.506010>
- RAE. (2017a). Diccionario de la lengua española. Edición del Tricentenario. Actualización 2017. Retrieved from <http://dle.rae.es/?id=bJeLxWG>
- RAE. (2017b). Diccionario de la lengua española. Edición del Tricentenario. Actualización 2017. Retrieved from <http://dle.rae.es/?id=H8KIdC6>
- Reading, A. (2012). When information conveys meaning. *Information (Switzerland)*, 3(4), 635-643. <https://doi.org/10.3390/info3040635>

- Redman, T. C. (1998a). *La qualité des données à l'âge de l'information*. (S.l.) : Paris, InterÉditions.
- Redman, T. C. (1998b). The impact of poor data quality on the typical enterprise. *Communications of the ACM*, 41(2), 79-82. <https://doi.org/10.1145/269012.269025>
- Reix, R. (2002). *Système d'information et management des organisations*. Paris : Vuibert.
- Ronen, B., & Spiegler, I. (1991). Information as inventory: A new conceptual view. *Information & Management*, 21(4), 239-247. [https://doi.org/10.1016/0378-7206\(91\)90069-E](https://doi.org/10.1016/0378-7206(91)90069-E)
- Ruben, B. (1992). *Communication and Human Behavior*. New York : (s.n.).
- Scannapieco, M., Missier, P., & Batini, C. (2005). Data Quality at a Glance. *Datenbank-Spektrum*, 14(January), 6-14. <https://doi.org/10.1.1.106.8628>
- Schement, J. R., & Ruben, B. (Edited by). (1993). *Between Communication and Information. Information & Behavior*. New York : Routledge.
- Schmidt, N. H., Ereik, K., Kolbe, L. M., & Zarnekow, R. (2009). Towards a procedural model for sustainable information systems management. *Proceedings of the 42nd Annual Hawaii International Conference on System Sciences, HICSS*. <https://doi.org/10.1109/HICSS.2009.468>
- Schramm, W. (1980). *La Ciencia de la comunicación humana*. México : El Roble.
- Shankaranarayanan, G., & Blake, R. (2017). From Content to Context: The Evolution and Growth of Data Quality Research. *Journal of Data and Information Quality*, 8(2), 1-28. <https://doi.org/10.1145/2996198>
- Shankaranarayanan, G., & Cai, Y. (2006). Supporting data quality management in decision-making. *Decision Support Systems*, 42(1), 302-317. <https://doi.org/10.1016/j.dss.2004.12.006>
- Shankaranarayanan, G., Wang, R. Y., & Ziad, M. (2000). IP-MAP: Representing the Manufacture of an Information Product. *Proceedings of the 2000 Conference on Information Quality*, (May 2014), 1-16.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Thecnical Journal*, 27(3), 379-423.
- Shannon, C. E. (1956). The bandwagon (Edtl.). *IRE Transactions on Information Theory*, 2(1), 3-3. <https://doi.org/10.1109/TIT.1956.1056774>
- Shannon, C. E., & Weaver, W. (1949). The Mathematical Theory of Communication. *The mathematical theory of communication*, 27(4), 117. <https://doi.org/10.2307/3611062>

- Sless, D. (2018). Designing Documents for People to Use. *She Ji: The Journal of Design, Economics, and Innovation*, 4(2), 125-142. <https://doi.org/10.1016/j.sheji.2018.05.004>
- Stonier, T. (1990). *Information and the Internal Structure of the Universe*. (S.l.) : Springer-Verlag London. <https://doi.org/10.1007/978-1-4471-3265-3>
- Tee, S. W., Bowen, P. L., Doyle, P., & Rohde, F. H. (2007). Factors Influencing Organizations to Improve Data Quality in their Information Systems. *Ssrn*, 47(June 2006), 335-355. <https://doi.org/10.1111/j.1467-629X.2006.00205.x>
- Termium Plus, data bank. (2018). *Bank, Government of Canada's terminology and linguistic data*. Retrieved from [http://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-eng.html?lang=eng&i=1&srchtxt=timeliness&index=alt&codom2nd\\_wet=AE#resultrec\\_s](http://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-eng.html?lang=eng&i=1&srchtxt=timeliness&index=alt&codom2nd_wet=AE#resultrec_s)
- Trostchansky, D. J., Sánchez, G., Dibarboure, P., Bado, J., Castiñeiras, B. S., & Sarutte, S. (2011). Historia clínica para trauma . Registro hospitalario específico para pacientes traumatizados . Un recurso para países en desarrollo. *Rev Med Urug*, 27(1), 12-20.
- Tushman, M. L., & Nadler, D. A. (1978). Information-Processing as an Integrating Concept in Organizational Design. *Academy of Management Review*, 3(3), 613-624.
- Tyler, J. E. (2017). Asset management the track towards quality documentation. *Records Management Journal*, 27(3), 302-317. <https://doi.org/10.1108/RMJ-11-2015-0039>
- Varga, M. (2003). Zachman framework in teaching information systems. *Proceedings of the International Conference on Information Technology Interfaces, ITI*, 161-166. <https://doi.org/10.1109/ITI.2003.1225339>
- Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11), 86-95. <https://doi.org/10.1145/240455.240479>
- Wang, R. Y. (1998). A Product Perspective on Total Data Quality Management. *Communications of the ACM*, 41(2), 58-65. <https://doi.org/10.1145/269012.269022>
- Wang, R. Y., Reddy, M. P., & Kon, H. B. (1995). Toward quality data: An attribute-based approach. *Decision Support Systems*, 13(3-4), 349-372. [https://doi.org/10.1016/0167-9236\(93\)E0050-N](https://doi.org/10.1016/0167-9236(93)E0050-N)
- Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, 12(4), 5-33. <https://doi.org/10.1080/07421222.1996.11518099>
- Wang, R. Y., & Stuard, M. E. (1989). The Inter- Database Instance Identification Problem in Integrating Autonomous Systems. Dans *Proceedings of the 5th International Conference on Data Engineering (ICDE 1989)*, (pp. 46-55). Los Angeles, California, USA.

- Wang, R. Y., Yang, L., Pipino, L. L., & Strong, D. M. (1998). Manage Your Information as a Product. *Sloan Management Review*, 39(4), 95-105. Retrieved from <http://search.ebscohost.com.ezproxy.unal.edu.co/login.aspx?direct=true&db=bth&AN=887820&lang=es&site=ehost-live>
- Wang, R. Y., Yang W., L., Pipino, L. L., Strong, D. M., Lee, Y. W., Pipino, L. L., & Strong, D. M. (1998). Manage your information as a product. *Sloan Management Review*, 39(4), 95-105. Retrieved from <https://sloanreview.mit.edu/article/manage-your-information-as-a-product/>
- Weaver, W. (1949). The Mathematics of communication. *Scientific American*, 181(1), 11-15.
- Wiener, N. (1948). *Cybernetics or control and communication in the animal and the machine* (Second). Cambridge : The MIT press.
- Yin, R. (2002). *Case Study Research: Design and Methods*, 3rd ed. Thousand Oaks, CA : SAGE Publications.
- Yu, L. (2015). Back to the fundamentals again. *Journal of Documentation*, 71(4), 795-816. <https://doi.org/10.1108/JD-12-2014-0171>

