Quality Management of Information Systems - A Human-Centric Point of View

Virginie Thion

▶ **To cite this version:**

Virginie Thion. Quality Management of Information Systems - A Human-Centric Point of View. Databases [cs.DB]. Université Rennes 1, 2019. tel-02407434

HAL Id: tel-02407434
https://inria.hal.science/tel-02407434v1

Submitted on 12 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ DE RENNES 1

HABILITATION À DIRIGER DES RECHERCHES

Université Rennes 1
Spécialité Informatique

par
Virginie THION

QUALITY MANAGEMENT OF INFORMATION SYSTEMS
–
A HUMAN-CENTRIC POINT OF VIEW

Soutenance publique le 10 octobre 2019

David Gross-Amblard, Professeur, Université Rennes 1, France, examinateur
Mohand-Said Hacid, Professeur, Université Lyon 1, France, rapporteur
Óscar Pastor Lopez, Professeur, Universidad Politècnica de València, Espagne, rapporteur
Olivier Pivert, Professeur, Université Rennes 1, Lannion, France, examinateur
Chantal Reynaud, Professeure, Université Paris-Sud XI, France, rapportrice
Philippe Rigaux, Professeur, Conservatoire National des Arts et Métiers, Paris, France, examinateur

Remerciements.

Un grand merci aux collègues rapporteurs de mon manuscrit et membres du jury de cette HDR.

Merci à ceux qui ont cru en moi, et m'ont offert de m'accueillir au sein de leurs équipes : Serena C. et Nicole B. (poste de doctorante), puis Françoise G., Marie-Luce P. et Sylvaine N. (poste d'ingénieur-chercheur à EDF), puis Camille S.-R. et Philippe R. (poste MCF Paris Dauphine), puis Michel S., Michel C. et Philippe R. encore (mutation MCF au CNAM Paris), puis Oliver P. (mutation MCF à l'Univ. Rennes 1, ENSSAT Lannion). Merci à tous les collègues (universitaires et industriels) avec qui j'ai eu le plaisir de collaborer, et qui m'ont tant apporté scientifiquement et humainement.

Mes remerciements tous particuliers vont à quelques chercheurs que j'ai eu la chance de côtoyer durant mon parcours professionnel, remarquables modèles de patience, bonne humeur et constance, qui m'ont si régulièrement apporté leur soutien bienveillant : Nicole B., Sylvaine N., Philippe R. et Olivier P. Merci également aux enseignants du système universitaire (LMD informatique de l'Université Paris XI d'Orsay en l'occurrence) pour leur solide enseignement qui a constitué le socle des connaissances sur lesquelles s'est appuyé mon parcours menant à cette HDR. Je suis heureuse d'avoir été formée à la fac, environnement si précieux de formation de qualité et d'ouverture d'esprit.

Mon parcours n'aurait pas été possible sans le soutien infaillible de ma famille : mon alter ego François et nos petites merveilles Loulou, Chouchou et Mimi, mes parents Claude et Gilles, et mon frère Arnaud. Mille mercis ne suffiraient à exprimer ma tendre gratitude.

À ceux dont la présence au quotidien me permet de maintenir le cap lorsque cela est nécessaire : mon oncle Jean-Louis B., Marie-Noëlle A., Sonia D., Gaëlle M., Sandrine P., Delphine D., Delphine B., Elise et Vincent B., Sébastien L. M., Jean-Paul F., Salima B., Nathalie et Bruno C. Mon affectueuse reconnaissance vous est individuellement témoignée.

Je me dois de saluer quelques uns supplémentaires qui m'ont indirectement soutenue tout au long de la phase de rédaction : Ella Fitzgerald, Christophe André, et Saint-Joseph (cuvée 2016).

$\mathcal{V}$.

À mon oncle Jean-Louis dont la joie de vivre, l'humour et la bienveillance manqueront cruellement à ma vie.

Abstract

This *Habilitation à diriger les recherches* presents a summary of some of my research activities, focusing on my main and favourite research theme: *Quality Management in Information Systems*. I have conducted these activities as a research engineer at EDF Research & Development (Clamart, France) from 2004 to 2008, and then as an Associate Professor in Computer Sciences at Université Paris Dauphine in the LAMSADE laboratory from 2008 to 2010, at CNAM Paris in the Vertigo team of the CEDRIC laboratory from 2010 to 2013 and finally at Université Rennes 1 in the DKM/Shaman team of the IRISA laboratory since 2013.

CONTENTS

INTRODUCTION

In everyday life, we often face the consequences of poor quality in information systems. This may be a letter not delivered to the expected recipient because of an inaccurate address, a misfilled field in an administrative form that blocks the enrolment of our children to activities, a booking not taken into account because of a defective software, an unreachable web site or application, etc. Quality problems are plentiful and may have a lot of negative repercussions on the performance of an organisation, for instance the loss of commercial opportunities (by missing an order or a prospect, or by losing an unsatisfied customer), lawsuits, slower processing operations, higher maintenance costs, etc. As quality problems may have significant impacts and costs, which seriously affect the efficiency of organisations and businesses (English, 1999; Eppler and Helfert, 2004; Batini and Scannapieco, 2016), the quality management of information systems has become a serious issue for the companies and the research community.

The quality management of an information system is a complex subject. The *information system* itself is an intricate concept covering multiple components that interact with each other: databases, software, business processes, and humans. The notion of interaction becomes tricky when it concerns humans, which have their own individual backgrounds, perceptions and ways of thinking. Also, the notion of *quality* is a multidimensional concept whose management is complex. First, eliciting quality requirements is a methodologically delicate problem. Second, the elicited quality criteria may concern many facets of the quality, which may be correlated one with each other, and also concern diverse components of the information system, which may again be correlated one with each other.

As an illustration, let us consider the quality of *data* in the Customer Relationship Management (CRM) database, embedded in the information system of a company. The database contains customers contact details, like the name, the phone number, the postal address and the e-mail address of each customer. Some of these data may suffer from quality problems: some of them may be inaccurate, incomplete, deprecated or imprecise. In the company, various

business users retrieve data from the database, in order to achieve their business goals, which may be different from one user to another.

For an entity that sends the customers' invoices, the focus is (i) on the availability of the invoiced amount, which of course has to be accurate, and (ii) on the postal address, which has to be filled out, up-to-date and sufficiently precise (let us note that a postal address does not necessarily have to be perfectly precise as the post office is often able to deal with some imprecision). For another entity in charge of prospecting the customer portfolio, by e-mail, in order to promote new products, the e-mail has to be filled and exact (we can assume for the illustration that the rest of the customers' information does not really matter for this usage).

This very simple use case shows that the users of the information system have different requirements concerning the quality of the system, depending on their business goals. A user may be concerned by some quality issues for a specific usage, by some other issues for another usage, and they can be completely different for another user. Of course, the quality requirements do not only concern the data of the system, but also the system itself (for instance, they can also concern the quality of its services, like the availability of its access point), and the quality of the system does not only concern the users' satisfaction (for instance, it also concerns the quality of the information system processes). Then the first question that arises is: "Given an information system and its users, which are the quality requirements, and does the system meet them?"

Let us continue our example of the CRM database. We now consider a third entity that uses the database. This entity is composed of the operators in charge of the customer hotline. They must quickly respond to the customers' requests. For this entity, the availability of the database (quick answer of the database management system to data queries) is important. Let us now assume the data freshness is not satisfactory for the first entity (we recall that the data freshness – up-to-date information – is important for the first entity). This leads this entity to ask for an improvement of the freshness, by adding a refreshment process in the system. Is it a relevant measure ? Maybe not because adding a refreshment process on the database could negatively impact the availability of the database, needed by the thirst entity... So, if the system does not meet the quality requirements, another complex question arises: "How to deal with quality problems?"

Roughly speaking, for managing the quality, two main issues have to be considered. The first issue is the *assessment of the quality*, whose goals are to define and examine the quality of the system. The second issue consists in *dealing with quality problems*, either by improving the quality, or by making the usage of the system as robust as possible to quality problems. In these problems, the human being is centric (i) as a stakeholder (internal or external to the system) for whom the quality of the system must be ensured, or (ii) as a "component" of the

information system because humans are involved in its business processes.

This document supporting my *Habilitation à Diriger les Recherches* reports on some of my contributions, on the topic of the *Quality Management of Information Systems – A Human-Centric Point of View*. It is organised as follows.

Chapter 1 (*Preliminary notions*) presents background notions concerning the quality management in an information system. I clarify the concept of *information system*, and introduce the problem of its quality management.

A quality management process includes an assessment stage where the quality is first defined and then measured. Chapter 2 (*Assessing quality*) presents a synthesis of some contributions on the *quality assessment of the data* and of the *quality assessment of the business processes* of an information system.

Once the quality is assessed, the analysis of the quality report allows detecting quality problems, and deciding how to deal with them. Two approaches may be thought of. A first one consists in *improving the quality*, that is to say repairing quality problems or system malfunctions. Another approach consists in *using the information system as it is*, meaning without improvement, taking quality problems into account when the system is used. Chapter 3 concerns these scientific problems. It is divided into two sections. Section 3.1 (*Improving a business process*) presents a synthesis of some contributions on the improvement of a business process (first approach). Then, Section 3.2 (*Quality-aware querying*) presents a synthesis of some contributions for extending query languages in order to improve their usability when queried data have quality problems (second approach).

Chapters 2 and 3 start with a summary of my activities related to the considered topic. I indicate in which scientific projects the topic was investigated, the supervising activities, the institutions I collaborated with, and the associated publications.

Chapter 4 draws a conclusion and presents some perspectives of this work.

CHAPTER 1

PRELIMINARY NOTIONS

In this chapter, I present some background notions underlying the problem of the quality management of an information system. In Section 1.1, I present the notion of *information system*, and then, in Section 1.2, I present the fundamentals of its *quality management*.

1.1 The information system

The notion of *information system* was intensively studied in the literature. There is a consensus for generally defining an information system as a set of components that interact, in a complex environment, in order to support business goals. But things become more difficult when a precise definition has to be laid down. A lot of different formal definitions were proposed in the literature. They differ in the classification and the scope of the elements that compose the system and its environment.

Different points of view. In her PhD thesis, Grim-Yefsah (2012) reviewed about thirty different definitions of the concept of *information system* in the literature. Among these definitions, two general trends emerge. According to their point of view, some authors restrict the information system to a computer-based system only, while some others consider the information system from a more global point of view, including the business organisation and humans in the system.

A typical purely computer-based point of view of the information system is the one of Jessup and Valacich (2008), who propose the following definition : "*Information Systems are combinations of hardware, software and telecommunications networks that people build and use to collect, create, and distribute useful data, typically in organizational settings*." Such a definition restricts the information system to a set of connected programs like front-end

and back-end applications or database management systems, hosted in hardware components. Among the most restrictive points of view that can be found in the literature, Pawlak (2002) defines an information system as being a structured data set: "*An information system is a data table, whose columns are labeled by attributes, rows are labeled by objects of interest and entries of the table are attribute values*."

Avgerou (2001) states that "*what is generally called 'information system' in the jargon of practitioners as well as academics cannot be meaningfully restricted to computer or communications applications within an independently delineated social environment.*" In such a trend, some authors propose a more global point of view that includes human and organisational features in the scope of the information system. This intrinsically includes not only tangible concepts that can be modelled like explicit knowledge, business processes or official social networks, but also intangible concepts like tacit knowledge, skills, intuition, informal social networks or professional affinities carried by humans that belong to the system and interact with it, and in it.

In a process-oriented point of view, Paul (2007) proposes the following definition of an information system. "*The IS is what emerges from the usage that is made of the IT delivery system by the users (whose strengths are that they are human beings not machines). This usage will be made up of two parts: 1. First the formal processes, which are currently usually assumed to be pre-determinable with respect to the decisions about what IT to use. [..] 2. Second the informal processes, which are what the human beings who use the IT and the formal processes create or invent in order to ensure that useful work is done*." This author also insists on the fact that the information system is not a static concept. It constantly adapts to the evolution of business requirements and usages.

With a complementary goal-oriented point of view, Huber et al. (2006) define an information system as "*an organized collection of people, information, business processes and information technology, designed to transform inputs into outputs, in order to achieve a goal*." This definition goes further in the formalization of the components of an information system, identifying types of its components, that is to say the people, the information, the business processes and the technological artefacts.

Some authors consider that human actors are the primary elements of the information system. This is the case for Reix and Rowe (2002), who define an information system as a set of social actors that record and transform tangible concepts through information technologies and procedures[1]. This definition is clearly human-centric. The social agent has her/his own

[1] This is a personal translation for the initial sentence "*Un système d'information est un ensemble d'acteurs sociaux qui mémorisent et transforment des représentations via des technologies de l'information et des modes opératoires.*"

psychological profile, reasoning, understanding and interpretation processes, business goals and context.

For Mason and Mitroff (1973), the information system can even be defined from one person for one usage by the following definition "*An information system consists of, at least, a person of a certain psychological type who faces a problem within some organizational context for which he needs evidence to arrive at a solution, where the evidence is made available through some mode of presentation.*"

According to the previous definitions, there can be an information system without a computer, for instance composed of people that use pens and papers in order to store information, and letters in order to communicate, but there cannot be an information system without a human being.

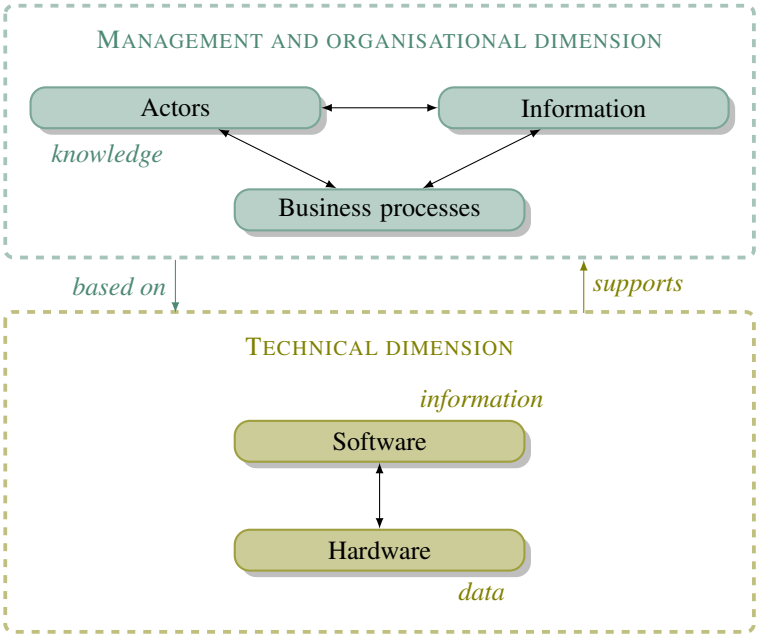


Figure 1.1: The information system (inspired by (Morley et al., 2004))

Alter (1999) distinguishes the management and organisational dimension of the information system from its technical dimension. This widely accepted point of view conducts to consider an information system that has two layers (Morley et al., 2004) that makes explicit the organisational and the technological parts of the information system. This is the definition that we consider in the following. Figure 1.1, inspired by the vision of Morley et al. (2004), is an illustration of this point of view. The *technological* layer is composed of software and hardware components. Data is stored in this layer and retrieved when needed by components of the *management and organisational* layer including the actors. According to a human-centric point of

view of the information system, human actors use their knowledge, communicate with each other, transfer (or not) some information and knowledge. In order to achieve business goals, actors perform complex tasks made of activities that are logically articulated. The activities and their arrangement are formally defined in *business processes* of the information system.

The business processes. The International Organization for Standardization (ISO) (2000) defines a *process* as "*a set of interrelated or interacting activities that transforms inputs into outputs.*" Thomas H. Davenport and James E. Short (Davenport and Short, 1990) define a business process as "*a set of logically related tasks performed in order to achieve a business outcome.*" These definitions are intentionally concise in order to cover a large scale of contexts. According to Porter and Millar (1985), the business processes of a company are distinguished between the *production processes*, which involve the primary activities that allow creating and delivering products to the customers (e.g. factory processes, sales processes), the *support processes*, which provide the inputs and the environment needed by the primary activities (e.g. software development processes, managerial processes).

It is now well understood that the good governance of an information system includes its *Business Process Management* (BPM). The goals of Business Process Management are 1) to align the business processes onto the company business goals and 2) to control and improve the processes of the organisation. If the goals of *Business Process Management* ("why the BPM?") have reached a consensus in the literature, different definitions of this notion ("what is the BPM?") were proposed (Palmberg, 2009). A first movement considers the BPM as a part of managing the whole organization. A second movement considers the BPM as a structured systematic approach to analyse and continuously improve processes. Lee and Dale (1998) proposed a unified vision through the following definition: "*Business Process Management is both a set of tools and techniques for improving processes and a method for integrating the whole organization and it needs to be understood by all employees.*"

More concretely, BPM includes "*concepts, methods and techniques to support the design, administration, configuration, enactment and analysis of business processes*" (Weske, 2012). In order to be analysed, the relevant processes first have to be formalised. This is the modelling. The literature proposes various approaches, metamodels and languages for modelling business processes (Weske, 2012; Morley et al., 2011). The modelling languages make possible to abstract the real world in order to express the relevant elements of a process in the form of a graphical representation usually called *diagram*. In most representations, such a diagram is composed of activities[2]. and their sequencing, performed in order to achieve a business goal.

[2]Let us note that the notion of *activity* itself is also subject to discussion. According to the vision, it may be defined as the most detailed level of work that is formulated, or corresponds to the transformation of an informational entity like an activity that makes an invoice change its state from *unpaid* to *paid* (see the discussion of (Morley et al., 2004) for details). This level of detail does not need to be considered in the following of the document.

The representation of a business process can be *data flow oriented*, meaning that the modelling focusses on the transformation of data (inputs and outputs) across the activities of the process. This approach is usually adopted when the process is modelled in order to be automatically executed. Such a vision (voluntary) minimises interactions with humans.

The notion of business process from the point of view of the Information System community is based on another vision: the process is usually *control flow oriented*, meaning that the modelling focusses on the activities that may be performed by humans being, and their arrangement. The primary goal of this vision is to develop a common understanding of the process that involves different actors (Rosenthal-Sabroux and Grundstein, 2007; Ludäscher et al., 2009), driven by business goals of the company.

Standard languages used for modelling business processes are BPMN (Object Management Group (OMG), 2013) and activity diagrams of UML (Object Management Group (OMG), 2017). Roughly speaking, such languages, used within a methodological modelling process, allow producing an abstract graphical representation of the activities that compose a business process, their sequencing, and the actors that are responsible of the activities.

Figure 1.2 is an example of business process, modelled in activity diagrams UML formalism. It models an accommodation booking process in a hotel. This process involves three actors including two human actors –a customer and a receptionist– and a technological information software –the channel manager–. The initial node (starting point of the process) is the black filled circle. Rectangles having rounded corners are actions (also called *activities*) performed by actors. Each action is placed in the swim lane associated with the actor that performs the action. The other rectangles model input (respectively output) objects received (respectively produced) by actions. Arrows model data or control flows between the elements. The other nodes specify decision points, parallelised flows and end of flows. If needed, the content of the activities can be detailed by a sequencing of actions[3]. In the process, a customer requests for a booking in the hotel (modelled by the first activity in the *Customer* swim lane). Then the receptionist opens a request processing (second activity in the *Receptionist* swim lane) for which the channel manager software checks the feasibility according to the provisional schedule of the hotel (third activity in the *IS* swim lane). The result of the channel manager is sent to the receptionist, who informs the customer. Then there are two cases: either the request of the customer cannot be satisfied (if the hotel is completely booked for the required period), or it can be satisfied, leading to a pre-booking in the system and an offer to the receptionist that communicates this offer to the customer, etc.

[3]This presentation contains only a simplified and incomplete vision of the activity diagram UML formalism, but I do not go further into details as the contributions in this document do not depend on the details of the chosen modelling language.

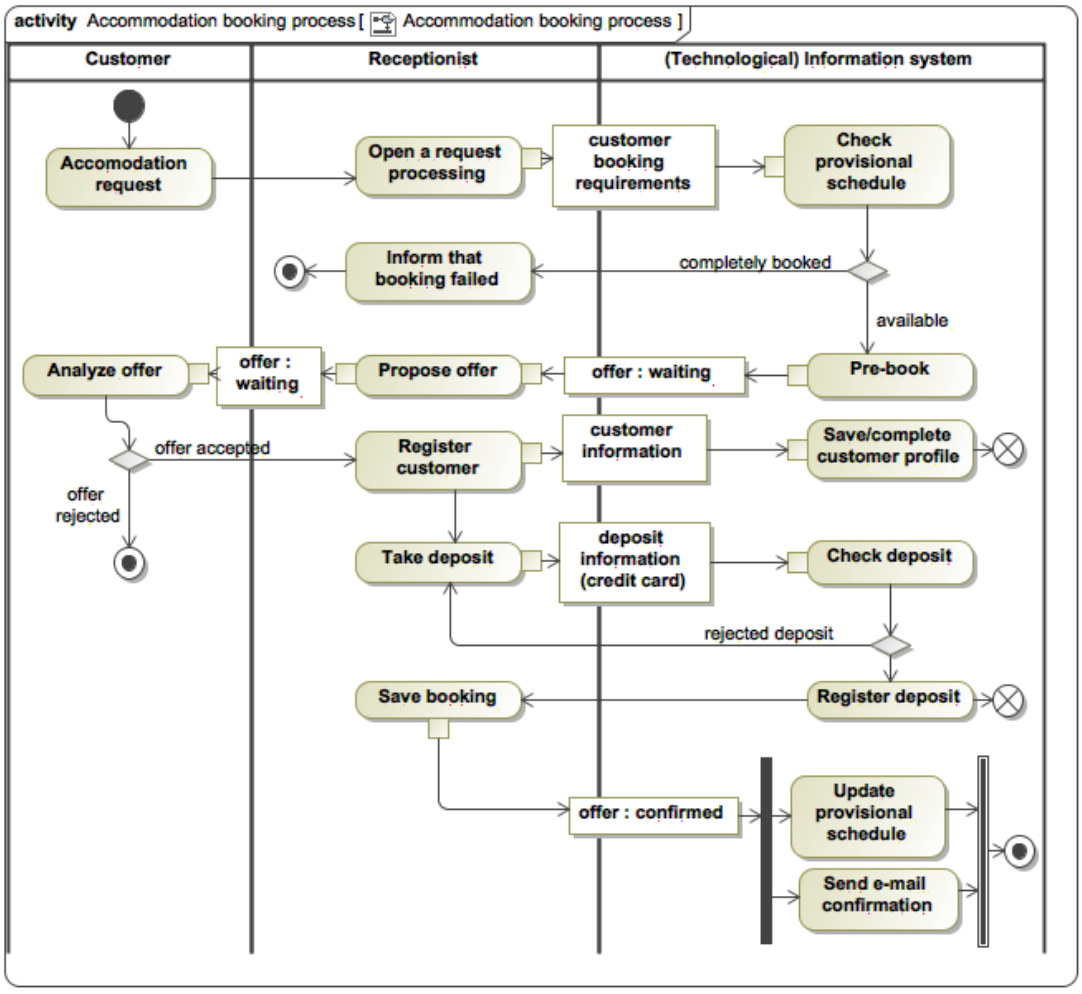


Figure 1.2: A modelled business process

Data may be used in order to perform activities, for instance the information of the customer request -dates, number of persons, type of room-, the provisional schedule of the hotel and the rates of the rooms. The quality and the availability of these data are key factors in the success of the booking process. The efficiency of the channel manager software is also important (must accurately and efficiently answer to queries). But, in most cases, data and software are not the only "inputs" of a business process. We can also go beyond this technological view of the process that only considers encoded information. Indeed human beings (actors) are involved in the business process. They use personal knowledge (skills) in order to perform tasks. For instance, the receptionist knows how to interact with the customer, how to react to dissatisfaction, how to propose relevant alternative solutions according to the customer profile, how to cleverly assign the rooms in order to optimise the customer satisfaction and the hotel occupancy rate, etc. Such behaviour results from the application of her/his personal knowledge.

The notion of knowledge, which is far from being trivial, is presented below.

Data, information and knowledge. Because knowledge is precious for an organisation, its semantics attracted a lot of attention in the scientific community. The definition of the *knowledge* concept itself has been widely discussed. The well-admitted vision consists in distinguishing *data* from *information*, from *knowledge* (Ackoff, 1989; Davenport and Prusak, 1998).

Data are a set of collected observations, measurements or facts, stored on a persistent physical storage (for instance in hard disks or paper files). They are not supposed to have any meaning in themselves but they provide a basic material from which information is produced.

Information is an arrangement of data expressed as a flow of messages (Nonaka, 1994). Information is subject to interpretation from the receiver according to her/his interpretative frameworks and previous knowledge (see (Arduin et al., 2013), on the basis of the theory proposed by Tsuchiya (1993)). By definition, the information is not necessarily persistently stored.

Knowledge is a *justified true belief* (Nonaka and Takeuchi, 1995), that is to say an understood and absorbed information. Knowledge is created and organized by the very flow of information, anchored on the commitment and beliefs of its holder (Nonaka, 1994), meaning that information is converted into knowledge once it is processed in the mind of an individual (Alavi and Leidner, 2001). Knowledge results in the application and the relevant use of information.

These concepts are interdependent as knowledge is a prerequisite for the generation and utilisation of data (Alavi and Leidner, 2001), and information is a necessary medium or material for initiating and formalizing knowledge (Nonaka, 1994).

The notion of *knowledge* itself is subtle, and has been widely studied in the literature. Let us discuss this notion in more detail. Knowledge cannot be considered as an object (Grundstien, 2009) because a part of the knowledge cannot be expressed and thus cannot be easily transferred. Knowledge is often distinguished between the *tacit* knowledge (also called *implicit* knowledge) and the *explicit* one (Polanyi, 1967; Nonaka, 1991).

Explicit knowledge can be codified or formalised (e.g. written or drawn) and articulated since it can be formally and systematically expressed. Knowledge that is made explicit can become some *information*.

Tacit knowledge (Polanyi, 1967; Nonaka, 1991) corresponds to knowledge that cannot be codified like e.g. skills, craft, senses, intuition, physical experiences or "job secrets". These are know-how, action-oriented skills, acquired through practical experience. For instance, after

years and years of cooking, a Brittany top chef knows how to bake the *crêpes bretonnes*[4] having a perfect texture, without being able to explain her/his exact know-how method that stems from a her/his physical experience. Tacit knowledge can only be acquired through practical experience in a relevant context.

Tacit knowledge can also be distinguished between the *individual* knowledge and the *collective* one (Nonaka, 1994). The individual knowledge is owned by a person while the collective knowledge is created and possessed collectively by a group composed of more than one individual. This kind of knowledge is often solicited in innovative processes (for example, scientific research) where a group of persons (researchers in the example) need to integrate the knowledge of individuals in order to solve a problem. Note that collective tacit knowledge is more than the aggregation of individual tacit knowledge of group members as it is created by collective actions (see (Erden et al., 2008) for details).

Let us now turn to the second issue considered in the background notions: the quality management of an information system.

1.2 Managing the quality of an information system

Delone and McLean (1992, 2003) defined the *success* of an information system though a multidimensional model consisting of three inter-dependent levels. Six dimensions of success were proposed, initially classified into three levels: a first technical level, a second level concerning the use of the system, and a third performance-related level.

The first "technical" level is composed of the quality dimensions associated with the *system*, the *information* and the *service*. The quality here includes the quality of *technical* levels of communication and data processes (hardware and software), which makes information available, and the quality of the information itself (the outputs) produced by the system. It also includes the quality of the *service*, which measures the overall support delivered by the service provider (for instance the IS department). The second level contains the *use* and *user satisfaction* dimensions. It measures the intent to which the users use the system and the impact of the produced information in terms the users' and managers' satisfaction and use.

The third level is the *net benefits*. It contains the individual and organisational impacts as using the system impacts individual users in their business outcomes and then collectively impacts the organization outcomes. It also contains consumer impacts and societal impacts.

[4]The *crêpes bretonnes* are traditional pancakes from Brittany. Part of the Breton gastronomy, these very thin pancakes are cooked on a traditional a cast-iron heating plate called *billig*. Cooking them requires hours and hours of practice.

The authors modelled causal relations between the dimensions of success. This model is depicted in Figure 1.3. Each association of the form $d_1 \rightarrow d_2$ means that the quality of the dimension d_1 impacts the quality of the dimension d_2. The Delone and McLean model sets the system and information quality as the basis of the success of an information system (in the first level of the model). In this model, the user satisfaction and her/his use of data and information also clearly appear as key factors of success (the second level). According to a human-centric point of view (see Section 1.1), the user has her/his own psychological profile and knowledge that affect her/his interpretation of the information produced by the information system. The activities are performed though business processes (see Section 1.1).

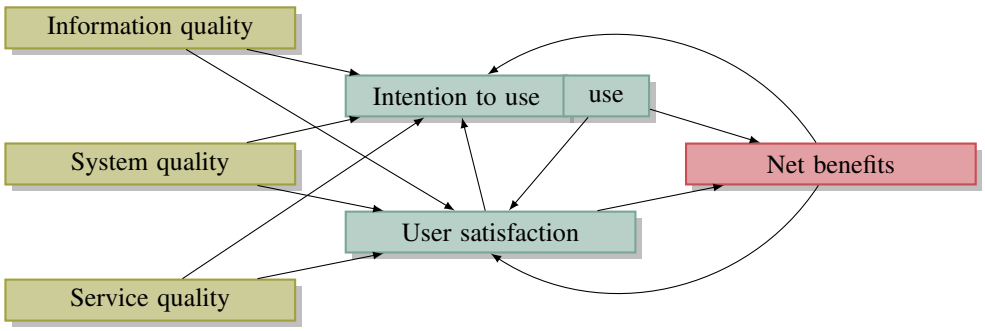


Figure 1.3: Delone and McLean IS success model (Delone and McLean, 2003)

The Delone and McLean model was intensively studied in the literature. Extensions and deeper analyses of the proposed causal relations were performed (Delone and McLean, 2003), and new associations were proposed. Among other results, the influence of the system and information quality (including, among other criteria, the user usability of the system and the data quality) on the individual impacts (e.g. work environment, job performance, quality of work, decision-making performance) was confirmed (Delone and McLean, 2003). This is not surprising as quality problems in information systems generate a multitude of consequences for an organisation, like a loss of revenue or loss of opportunity (missing an order, a prospect, or losing an unsatisfied customer), lawsuits, increased, higher maintenance costs, excess labor costs, etc. This leads to huge costs (of non-quality), which seriously impact the efficiency of organisations and businesses (English, 1999; Eppler and Helfert, 2004; Batini and Scannapieco, 2016).

The impact of the business processes quality (included in the *use* dimension of the Delone and McLean model) over the performance of an organisation is also well recognised and quality standards like the ISO 9001:2008 recommend to continuously control and improve the quality of business processes (International Organization for Standardization (ISO), 2008; Persse, 2006).

In the following, I focus on the problem of managing the quality of the *management and organisational layer* of the information system (see Figure 1.1 page 11), and more specifically on the quality management of its **data** and **business processes**.

The management community initially defined high-level methods for the quality management of production processes (Shewhart, 1980; Deming, 2000), whose initial goal was to ensure customer satisfaction. These methods have been naturally applied for the quality management of other processes, in particular for business processes. One of the prevalent approaches is Total Quality Management framework (TQM) (Oakland, 1989) based on the principle of a continuous improvement of work processes. Such an improvement implements the Deming cycle (Deming, 2000) also known as Shewhart cycle (Shewhart, 1980) or Plan-Do-Check-Act (PDCA), which proposes to continuously control and improve the quality by iteratively executing the four stages: Plan, Do, Check and Act, applied to the context of quality management. The first stage (Plan) consists in defining processes required in order to deliver the expected results. The second stage (Do) consists in implementing the processes previously defined. The third stage (Check) consists in evaluating the results of the execution, checking of they are satisfactory. In the last stage (Act, also sometimes referred to as Adjust), the processes are improved if needed. Then the cycle starts again from the first stage, with a planning based on better initial processes.

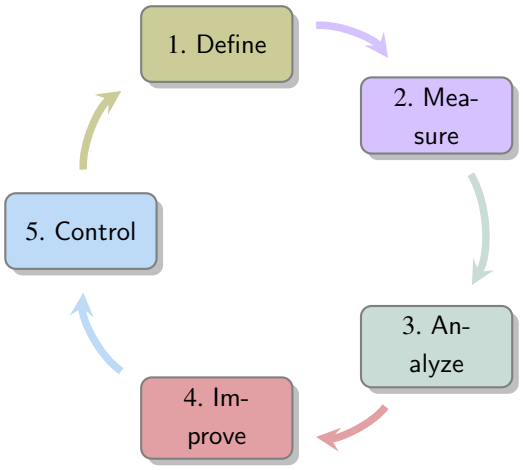


Figure 1.4: The DMAIC cycle

The Six Sigma program proposed an adaptation of PDCA called DMAIC (De Feo et al., 2005), which is applied to the quality management. The acronym DMAIC stands for the five stages of the method: Define, Measure, Analyse, Improve and Control (cycle illustrated in Figure 1.4).

The first stage Define is the quality definition, which consists in eliciting data quality requirements of interest. Concretely, this means choosing a set of measurable criteria of interest called *quality metrics*, and eventually thresholds associated with, that allows measuring in what extent the data fit the quality requirements according to data usages. In the second stage Measure, the quality metrics are measured. In the third stage Analyze, quality results are analysed; this possibly leads to implement improvement actions in the fourth stage Improve. Finally, the effects of the improvement actions are measured in the last stage Control. A comprehensive explanation of each stage and concrete examples implementing the method are presented in the contribution sections of this document (Section 2 and Section 3).

This framework is the foundation of the quality management methods proposed in the literature, which -roughly speaking- specialise it in order to deal with a specific part of the information system, like e.g. for managing *data quality* (see (Batini et al., 2009) for a review of the quality management methodologies dedicated to data quality management).

In the following, each of the contributions that I introduce is positioned according to the DMAIC cycle. The next section, Section 2, deals with the quality assessment, which refers to the stages D, M and A of the DMAIC cycle. Then Section 3 deals with the quality improvement, which refers to the I and C stages of the DMAIC cycle.

CHAPTER 2

ASSESSING QUALITY

This chapter presents a synthesis of some of my research activities concerning quality assessment in information systems (my research activity focused on data and business processes).

Summary of the research activities concerning quality assessment

Projects. This research was conducted in the following research projects: the CNRS Mastodons project called GIOQOSO[a], the *Projet scientifique émergent Univ. Rennes 1* called QUALITY@PANAM[b] and the ANR ARA Masses de Données project called QUADRIS[c] project.

Associated theses and internships. The following PhD thesis and internships participated to this research.

- PhD of Malika Gim-Yefsah (Univ. Paris Dauphine), on the subject *Knowledge management and outsourcing. Technical and managerial contributions for the improvement of a transition process, applied to the outsourcing in a PSTI*[d],
- Master 2 Research internship of Louis Smith (Univ. Paris Dauphine/MODO), on the subject of *The quality management of business processes*,
- Two Master 2 internships Univ. Rennes 1/ENSSAT/INFO, on the subject of *Implementation of quality metrics for a library of digital scores*.

Collaborations. AID (company), EDF R&D (company), ExQI association, CEDRIC laboratory (CNAM Paris), CESR Tours, David (Univ. Versailles Saint Quentin), IReMus Paris, IRISA (Univ. Rennes 1), LAMIH (Univ. Valenciennes).

Associated publications. (Barrau et al., 2016), (Marcal de Oliveira et al., 2012), (Berti-Équille et al., 2006), (Peralta et al., 2009), (Akoka et al., 2007), (Berti-Equille et al., 2011), (Grim-Yefsah et al., 2011b), (Grim-Yefsah et al., 2010b), (Grim-Yefsah et al., 2010a), (Grim-Yefsah et al., 2016) (Besson et al., 2016), (Rigaux et al., 2012), (Besson et al., 2018), (Duquennoy et al., 2007), (Besson et al., 2018), (Fiala et al., 2018), (Foscarin et al., 2018).

[a] GIOQOSO stands for *Quality management of open music scores* (translation of *GestIOn de la Qualité des partitiOns muSicales Ouvertes*). I was co-coordinator of this project.

[b] QUALITY@PANAM stands for *QUALITY focus on oPen dAta maNAgeMent*. I was coordinator of this project.

[c] QUADRIS stands for *QuAlity of Data and multi-souRce Information Systems*.

[d] Translation of the subject (in French) *Gestion des connaissances et externalisation informatique. Apports managériaux et techniques pour l'amélioration du processus de transition : cas de l'externalisation informatique dans un EPST*.

We can consider that the *quality assessment* is composed of the first three stages of the DMAIC cycle, that is to say the Define, the Measure and the Analyze stages (see Figure 2.1).

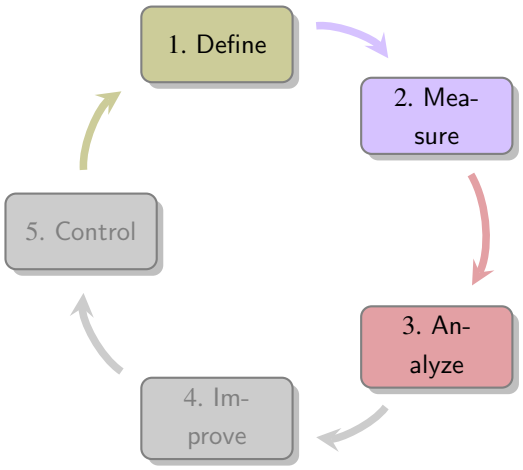


Figure 2.1: Quality assessment in the DMAIC cycle (the D, M, and A stages)

We now give some more details concerning these stages.

Define. The first stage *Define* consists in defining the notion of quality. It is a critical issue as it result constitutes the foundations of the following of the quality management process. The goal of this stage is to identify the quality requirements concerning (a part of) the components of the information system. The contributions that I present in the following of the document focus on the data and the business processes of the information system.

The *quality* is a complex notion, which embraces different semantics depending on the context (Redman, 1996; Batini and Scannapieco, 2016). It is described through a set of quality *dimensions* aiming to categorise the criteria of interest. Classical quality dimensions are the *completeness* (the degree to which needed information is present in the collection), the *accuracy* (the degree to which the measured elements are accurate), the *consistency* (the degree to which the measured elements respects integrity constraints and business rules) and the *freshness* (the degree to which the measured elements are up-to-date).

The quality criteria over a dimension are defined according to a set of *metrics* that allow a quantitative definition and evaluation of the dimension. Examples of quality metrics are "the number of missing meta-data" for the assessment of the *completeness*, and "the number of cities not consistent with the zip code in the postal addresses" for the assessment of the *consistency*. These are very simple examples but the literature proposes a large range of dimensions and metrics, conceptualized in quality models (see the surveys proposed by Batini and Scannapieco (2016) and Zaveri et al. (2016) for details). Of course, not all the existing dimensions

and metrics should be used for evaluating data quality in a given operational context. An important property concerning data quality is that it is defined according to *fitness for use* of data, meaning that quality measurement involves dimensions and metrics that are relevant to a given user for a given *usage*. This means that a user u_1 may be concerned by some quality metrics for a specific usage, by some other metrics for another usage, and they can be completely different than those needed by a user u_2.

This stage mainly raises methodological questions.

Measure. As soon as data quality metrics are elicited, one can consider different processes for their computation, including collaborative ones if the information system allows it. This is the Measure stage of the DMAIC cycle.

The quality measurement of course raises technological issues. The main technological issue is "how to measure the metrics?" including the questions of either using an existing tool (and if so, how to choose it?) or implementing a new quality software (tool or module), where to insert the quality measurement module in the system, etc.

But even with an appropriate assessment tool, the *Measure* stage can lead to reach the limits of the available resources when a large volume of information is considered or when a comprehensive measurement process is too expensive (moreover, this can happen even with a small amount of data). So, the question of the limits of a quality assessment is partly hidden behind the question of the measurement.

Analyze. Analysing the results enables to (partly) answer to the quality questions, and consequently enables to decide whether the measured elements respect the requirements for the given business goal. Quality and business experts seek the causes of quality defaults. This is the Analyze stage of the DMAIC cycle. This analysis is not trivial because quality results are usually incomplete and quality metrics are dependent on each other.

In Sections 2.1 and 2.2, I respectively present some contributions on the quality assessment of data (Section 2.1) and the quality assessment of a business process (Section 2.2). In Section 2.3, I discuss the limits of a quality assessment method.

2.1 Assessing the quality of data

I'll face the truth when I think I can.

Only, Sarah Vaughan, 1963

Positioning. *In this section, I review some contributions on the problem of data quality evaluation, over digital libraries of scores and over multi-sources CRM databases. These contributions are deeply presented in the following publications: (Besson et al., 2016), (Rigaux et al., 2012), (Besson et al., 2018), (Fiala et al., 2018), (Berti-Équille et al., 2006), (Peralta et al., 2009), (Akoka et al., 2007), (Berti-Equille et al., 2011) and (Duquennoy et al., 2007).*

2.1.1 Defining the quality of data

There is a growing availability of music scores in digital format, made possible by the combination of two factors: mature, easy-to-use music editors, including open-source ones, and sophisticated music notation encodings. We are therefore facing emerging needs regarding the storage, organisation and access to Digital Libraries of Scores (DSL). But it turns out that building a DSL, particularly when the acquisition process is collaborative in nature, gives rise to severe quality issues. There are many reasons for this situation. First, encoding formats have changed a lot during the last decades (from HumDrum and MIDI to XML formats MusicXML and MEI). A lot of legacy collections have been converted from one encoding to the other, losing information along the way. Second, the flexibility of music notation is such that it is extremely difficult to express and check quality constraints on the representation. For instance, many of the formats do not impose that the sequence of events in a measure[1] exactly covers the measure duration defined in the music score. Third, scores are being produced by individuals and institutions with highly variable motivations and skills. By "motivation", we denote here the purpose of creating and editing a score in digital format. A first one is obviously the production of material for performers, with various levels of demands. Some users may content themselves with schematic notation of simple songs, whereas others will aim at professional editing with high quality standards. The focus here is on rendering, readability and manageability of the score sheets in performance situations. Another category of users (with, probably, some overlap) is scientific editors, whose purpose is rather an accurate and long-term preservation of the source content (including variants and composer annotations). The focus will be put on completeness: all variants are represented, editor's corrections are fully documented, links are provided to other resources if relevant, and collections are constrained by carefully crafted editorial rules. Overall, the quality of such projects is estimated by the ability of a document to convey as respectfully as possible the composer's intent as it

[1] The *measure* is a part of staff positioned between two adjacent horizontal bars, see Figure 2.2 in page 27.

can be perceived through the available sources. Librarians are particularly interested in the searchability of their collections, with rich document annotations (meta-data). We can finally mention analysts, teachers and musicologists: their focus is put on the core music material, minoring rendering concerns.

Knowing that data quality is *fitness for use* (depends on the context), the first question that has to be tackled is: "How to define data quality in DSL?" In order to answer to this question, dedicated methodological guidelines can be followed like the *Goal Question Metric* (GQM) paradigm Basili et al. (1994), which proposes to define quality metrics according to a top-down analysis of quality requirements. The underlying process of the GQM paradigm is given hereafter.

1. For each user (or each user role) and for each of his/her usage of data, conceptual *business goals* are identified. A business goal specifies the intent of a quality measurement according to a usage of data.

   (Example.) We make this process more concrete by illustrating it on a simple example taken from (Fiala et al., 2018). Let us assume that a business user retrieves music scores in order to *Perform a given algorithm that searches for similar patterns in the parts of a music score*. This is a business goal.

2. Each goal is then refined into a set of operational *quality questions*, which are a first step towards eliciting the quality requirements.

   (Example.) For the running example, the user may express that (i) the results of his/her study is relevant provided that data is complete enough and that (ii) the used algorithm computes relevant results provided that data is accurate enough. Quality questions associated with this use case could then be the following ones.

   (QQ1) Does the data contain all the needed information?
   (QQ2) Are the notes accurate?

3. Each quality question is then itself expressed in terms of a set of quantitative quality metrics with possible associated thresholds (expected values).

   (Example.) The quality question *(QQ1)* could be refined into two more precise quality questions (at the score level). A first "quantitative" quality question could be *Is the time signature available?*. The time signature is the information made of the two numbers that appear after the clef at the beginning of a staff (encoded in a specific tag in the MEI file of the music score). The time signature is $\frac{6}{8}$ in Figure 2.2.

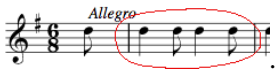


Figure 2.2: Excerpt of a music score

The quality metric *(QQ1/M1)*, defined below, expresses this requirement.

(QQ1/M1) Availability of the time signature. (Boolean result).

A second quantitative quality question associated with *(QQ1)* could be *Does each measure*[1] *cover exactly the expected number of beats?*. For instance, the time signature $\frac{6}{8}$ of Figure 2.2 implies that each measure (except the first one, which is a special case), must have exactly three beats. The circled measure has two quarter notes ♩ (i.e. a value of two beats), and two eighth notes ♪ (i.e. a value of one beat), so it respects the declared time signature. The quality metric *(QQ1/M2)*, defined below, may express this requirement at the music score level.

(QQ1/M2) Number of measures that fit the expected number of beats (defined in the time signature), over the total number of measures.

Assuming that the algorithm is robust up to 10% of malformed measures, then the threshold 0.9 could be associated with this quality metric.

Concerning the quality question *(QQ2)*, it could be refined into a quality metric that measures the syntactic accuracy of the notes, meaning that each note should be an existing one (which belongs to the usual range of notes). A third quality metric could then be *(QQ2/M3)* defined below.

(QQ2/M3) Number of syntactically accurate notes over the total number of notes.

The quality assessment raises scientific challenges because data quality methodologies of literature are designed at a generic level, leading to difficulties for their implementation in a specific context (operational context and available information system and data). Additional context-dependent quality methodologies are then needed (Barrau et al., 2016). In particular, the literature proposes a large range of quality metrics (Batini and Scannapieco, 2016; Zaveri et al., 2016) but such metrics are general ones. Quality metrics that are specific to the data of the considered domain are still needed, more specifically in the context of digital score libraries for which, to our knowledge, only few quality metrics were proposed in the literature.

Contributions. Based on the authors' practical experience and skills, we proposed in (Besson et al., 2016) and (Fiala et al., 2018) a set of quality rules specific to DSL data. About fifty

rules composed the first version of the catalog of Fiala et al. (2018). A data quality rule expresses a possible quality requirement. It may be used either (i) in order to tag the data where a quality problem occurs, or (ii) in order to compute a quality metric associated with a score or a corpus. For instance, the quality rule "*Each note is syntactically accurate, meaning that it is an existing one (which belongs to the usual range of notes)*" expresses the fact that having syntactically accurate notes is a data quality requirement. Such a quality rule can lead to *tag* the notes that are syntactically inaccurate (those that violate the rule) in music scores of interest. It can also lead to *compute a quality metric* in order to assess the quality of a music score according to the rule, like the number of syntactically accurate notes over the total number of notes appearing in the score[2]. The catalog can serve as a basis in order to elaborate users' quality requirements, by choosing relevant quality rules according to specific use cases. In order to classify the rules, we proposed taxonomy that is specific to DSL data. This allows combining the data quality point of view that organises quality rules/metrics according to classical quality dimensions (completeness, accuracy, etc.) and a business point of view that introduces supplementary DSL-dependent levels (Besson et al., 2018) (dissociating the content issues, from the engraving ones, from the metadata ones). The set of quality rules that we proposed is obviously not exhaustive. New quality rules are regularly discovered and added to the framework. The catalog is then subject to evolution and enrichment.

In (Besson et al., 2016), we proposed a methodology for assessing data quality in a digital score library. Our approach defines a generic data model that supports the specification of quality schemas based on quality metrics of Fiala et al. (2018), lets users define their goals with respect to the schema of their DSL, and matches usages against quality evaluation. The implementation of this framework, focusing on the measurement of the DSL quality metrics, is presented in Section 2.1.2.

2.1.2 Measuring the quality of data

The second challenge of the quality assessment concerns the measurement stage, and more specifically the underlying software tools that measure the quality metrics of interest. Two solutions may be thought of: either using an existing software, or developing a new dedicated tool (or module) attached to a data management software.

For the problem of quality management of CRM databases, we proposed in (Duquennoy et al., 2007) a set of guidelines for choosing a quality tool according to the considered quality metrics of interest. Based on our practical experience, we explained what one can expect of a data quality management tool in a CRM context, and we proposed a selection of about sixty quality

[2]By extension, quality metrics at the corpus level may easily be defined by aggregation of the metrics at the score level, for instance the average and standard deviation of the corresponding metric at the score level, computed over the set of scores that belong to the corpus.

metrics that are relevant in such a context. In (Barrau et al., 2016), we reviewed some of the available software tools that support the management of open data. The idea was to estimate the maturity of the existing tools in terms of data quality management. Such contributions propose guidelines for choosing a quality assessment tool amongst the available ones.

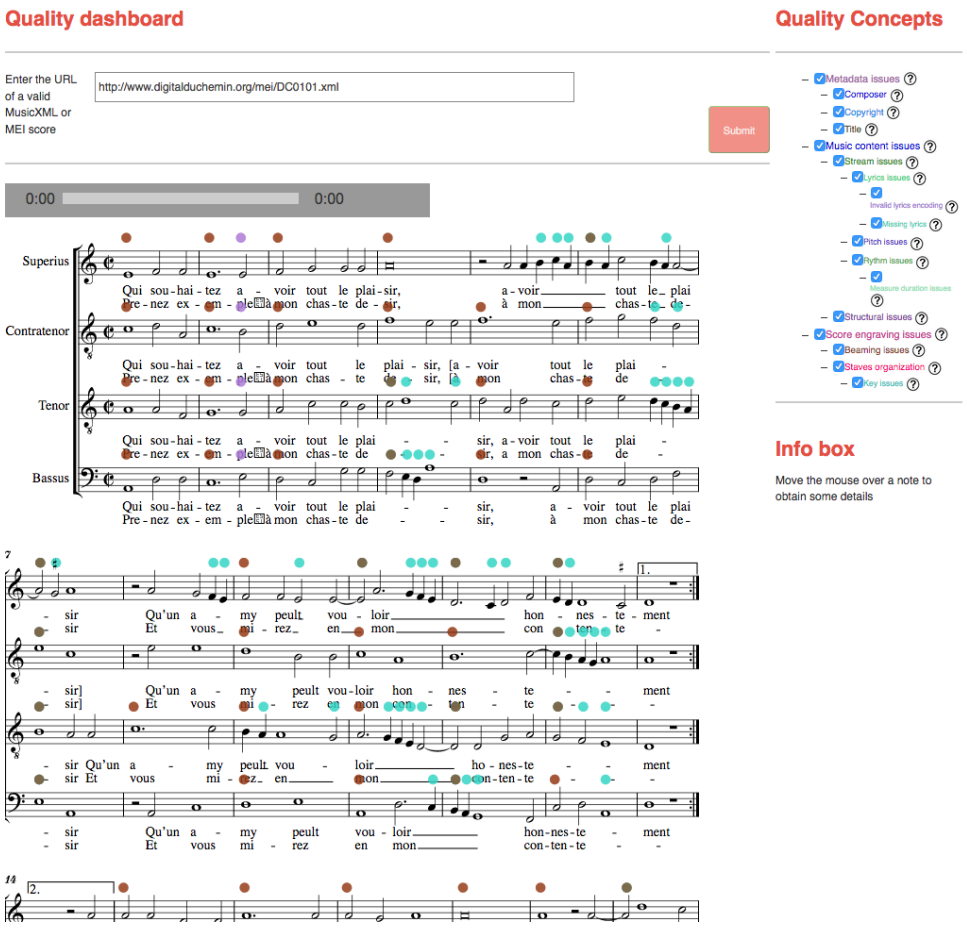


Figure 2.3: Visualisation of quality problems in the NEUMA platform

If no suitable tool exists, then one can consider to implement an *ad hoc* tool or module for managing data quality. This is the approach chosen in the GIOQOSO project, for which a specific tool was developed (Besson et al., 2018) in the NEUMA digital score library platform (Rigaux et al., 2012). The corpora of NEUMA are publicly available, on open access at `http://neuma.huma-num.fr`. Some of the quality rules presented in Section 2.1.1 are currently being implemented in the NEUMA platform in the form of a quality module (Besson et al., 2018; Foscarin et al., 2018) that detects quality problems in the data and tags them (Si-Said Cherfi et al., 2017a,b). A graphical user interface allows their visualisation, as illustrated in Figure 2.3. In such an interface, the user chooses a music score whose quality has to be

checked, her/his data problems of interest (in the right frame in Figure 2.3). After the quality module processing, graphical elements appear in the form of an overprinted layer on the layout of the music score (the coloured points in Figure 2.3) in order to report quality problems that are detected.

(Other works.) As a research engineer at EDF, I also participated to quality assessment studies over data of CRM databases. In this context, we proposed new data quality metrics specific to the electricity business, and participated, in the QUADRIS project[3] (Berti-Equille et al., 2011), to the definition of a multidimensional model that captures a large variety of measures for characterising the quality of data. These contributions are not detailed in this report. Details of these contributions can be found in (Peralta et al., 2009; Berti-Equille et al., 2011; Akoka et al., 2007).

2.2 Assessing the quality of a business process

Positioning. *In this section, I review some contributions on the problem of the quality assessment of a business process, deeply presented in the following publications: (Grim-Yefsah et al., 2010a), (Grim-Yefsah et al., 2010b), (Grim-Yefsah et al., 2011b).*

2.2.1 Defining the quality of a business process

In the current context of increasing competition, organisations are forced to look for new solutions that aim at generating value. Outsourcing is one of the principles adopted by companies that decide to devote internal resources to core business. Outsourcing is a management strategy by which an organisation delegates non-core activities to an external and specialised third party (Willcocks and Kern, 1998). It has been widely adopted in both public and private companies. They mainly outsource support services such as Human Resources management, Finances, or Information System (IS) activities.

In the following, we consider the use case of the IS outsourcing in a French Public Scientific and Technological Institution (PSTI). Willcocks and Kern (1998) define IS outsourcing as delegating to a third party the management of IT/IS assets, resources, and/or activities for required results. Different categorisations of outsourcing were proposed (see the survey of Dibbern et al. (2004) for details). In our use case, the outsourcing concerns the development of a software, performed in project management mode (Lacity and Hirschheim, 1993), meaning that the development of the software is delegated to a service provider, but the in-

[3] `http://quadris.cnam.fr/xwiki/bin/view/QUADRIS/WebHome`

ternal IS Department still manages the project and keeps being the main interlocutor of the business entity.

The outsourcing contract rules for a French public organization, like a PSTI, impose that each outsourced project must undergo a new tendering procedure every three years. Such a procedure may lead to change the service provider, during the project in progress. Changing the provider implies to perform a *transition* stage, which consists in transferring the outsourced project from the outgoing project team to an incoming one. The transfer concerns the documentations, applications, programming codes and knowledge. Knowledge transfer here cannot be relegated to the background as several studies showed that both the transfer of the explicit and the tacit knowledges plays an important role in the success of the transition process (Alaranta and Jarvenpaa, 2010; Beulen et al., 2011; Olzmann and Wynn, 2012).

The *transition* stage is a complex, risky and challenging building block of strategic importance in an outsourced project (Olzmann and Wynn, 2012). Then ensuring a "good quality" of the transition is fundamental. But there is no consensus on the definition of the *quality* of a transition. From a practical point of view, the question is "Which indicators should be measured in order to assess the quality of a transition?" This is the first problem that we tackled in our work. (The second problem concerns the improvement of a transition stage, which is presented in Section 3.1.)

Contributions. We proposed quality metrics for managing the quality of a transition process, and applied these metrics to the quality assessment of a real transition process that is implemented in a French PSTI. The approach that we adopted has the following properties.

- The transition is seen as a business process that can be modelled, analysed and improved.
- The quality management follows the DMAIC approach (see background notions in Section 1.2), using the GQM methodological tool for the definition of the quality (see background notions in Section 1.2).

It is worth noting that such an approach, based on a detailed modelling of the process, is classical in the computer science community but not widespread in the project management community. To our knowledge, it had never been experimented in the context of the management of an outsourced project transition. Our framework also proposes a selection of relevant quality metrics to by used in such a context, including some new quality metrics for assessing the robustness of a business process (which were experimented on a real use case).

Let first present our use case, this is to say the transition process that we consider, modelled as a control flow oriented business process managed by the IS Department (which is an internal entity). It consists of six global activities.

- Activity 1, called *Initialisation*, corresponds to the official beginning of the transition. The institution officially validates the transition and its global goals, duration, provi-

sional schedule and actors involved.

- Activity 2, called *Third Party Maintenance(TPM) ending* consists in inventorying the internal and external documents, applications and programming codes associated with the project to be transferred, and in defining a regulatory framework between stakeholders for the transfer. The TPM ending activity contains seven tasks, including four tasks under the responsibility of the IS Department.
- Activity 3 called the *Transfer planning* consists in the definition of the precise planning of the following of the transition process, aiming at concretely transferring the project from the outgoing team to the incoming one. This activity includes two tasks. The first task is under the responsibility of the outgoing service provider, who defines the planning. The second task is under the responsibility of IS Departement, who possibly adjusts and then validates the planning.
- Activity 4, called *Project transfer*, essentially consists in transmitting documentations, applications and programming codes to the incoming project team. The outgoing service provider writes documents and codes. The IS Department is almost not involved in this activity. This activity includes four tasks (not detailed here).
- During Activity 5, called *Maintenance in cooperation*, the outgoing and incoming service providers assume together a time-limited maintenance of the application. This activity is optional according to the procedure. In practice, for cost or time saving reasons, this activity is often skipped or cut back to the bare minimum. If performed, this activity includes five tasks (not detailed here).
- Activity 6, called *Transmission of responsibility*, is the official ending of the outgoing service provider's job. This activity includes four administrative tasks (not detailed here).

Figure 2.4 is a small excerpt from the modelled transition, focusing on the tasks of the activities 2 and 3 that are under the responsibility of the *IS department* actor (Grim-Yefsah (2012) gives the whole detailed process in her PhD manuscript). One can see that Activity 2 contains four tasks, including the task called *Inventory of the TPM elements*, in which an inventory of all the project elements (documents, codes) that have to be transferred is performed, and a task called *Edition of the TPM ending plan*, in which the procedure for transferring the elements is planned. (The two other tasks are administrative milestones, they are not detailed.)

Following the GQM method, we identified quality goals and questions of interest. Beyond the applicative study itself, this work led us to propose a set of quality metrics for measuring the quality of a transition process, and to define some generic quality metrics that make possible to assess the robustness of a business process to the risk of missing knowledge needed for its execution. I only to survey these contributions, for which details can be found in (Grim-Yefsah et al., 2016).

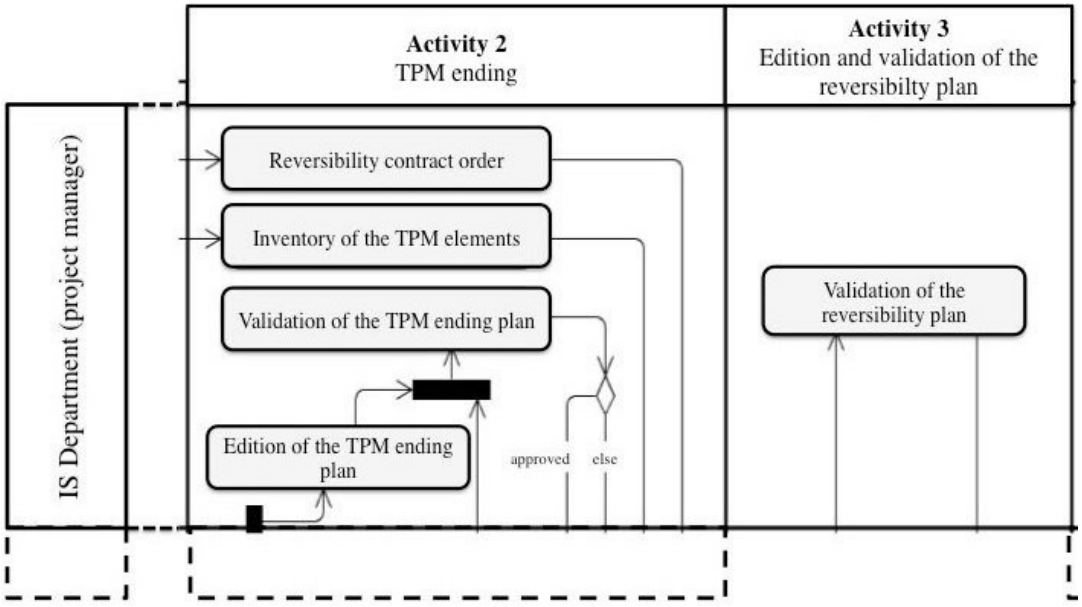


Figure 2.4: Activities 2 and 3 of the transition process restricted to the internal IS department actor

Let me give an intuition of the reasoning that led us to choose the quality metrics presented in the following. At first glance, it appeared that (*Issue i*) the transition stage was difficult to manage by the project manager, and that (*Issue ii*) at the end of the process, the outgoing team missed a part of the knowledge that should be transferred. It was also observed that (*Issue iii*) the execution of the business process seemed to activate a network of informal collaborations implying contributors that did not appear in the formal procedure.

The first observation (*Issue i*) seemed related to the complexity of the process. The second observation (*Issue ii*) seemed to reflect that the knowledge transfer performed in the process was incomplete. The last observation (*Issue iii*) seemed to indicate that a lot of (hidden contributors') knowledge used during the process did not explicitly appear in its modelling. Such a situation weakens the robustness of the process as it could unnoticeably miss contributors' knowledge, especially when it is performed during a period favourable to the absence of persons (holidays, seasonal flu epidemic, corporate reorganisation, etc.). This leads to a strong risk of missing knowledge needed for the execution of the process.

These observations had to be quantified and rationalised (as much as possible). So we proposed some quality metrics that allow defining the quality of the transition.

A first set of metrics used for the quality assessment study concerned the assessment of the complexity of the process (*Issue i*). These quality metrics were taken from the literature, which proposed a lot of relevant metrics for assessing the complexity of a business process (Laue and

Gruhn, 2006; Vanderfeesten et al., 2007). But for *Issue ii* and *Issue iii*, no quality metric of literature allowed evaluating the quality of the knowledge transfer or the robustness of the business process according to "hidden" contributors.

Concerning *Issue ii*, based on the theory of knowledge transfer (Cohen and Levinthal, 1990; Davenport and Prusak, 1998; Argote and Ingram, 2000), we proposed some quality metrics for assessing the quality of the effective knowledge transfer from the outgoing provider to the incoming one (Grim-Yefsah et al., 2016).

Concerning *Issue iii*, we proposed some quality metrics for defining the robustness of a business process, with regard to the risk of losing knowledge that is needed for its execution. These metrics are based on the analysis of social networks (Wasserman and Faust, 1994; Degenne and Forse, 1999) underlying the execution of the process, including formal and informal contributors (Grim-Yefsah et al., 2011b, 2010a, 2011b).

The whole set of metrics were assessed on the transition business use case (Grim-Yefsah et al., 2016).

It is worth noticing that the implemented quality metrics were extremely heterogeneous. They concerned :

- the business process *model* itself independently of any execution, for instance evaluating a part of the process complexity in terms of the number of elements (tasks, edges) appearing in the model;
- for a specific execution:
    - the *target artifacts*, which are the parts of the business process outer environment that we strike to create or alter (Lohrmann and Reichert, 2013) like
        - the resulting autonomy of the incoming provider;
        - the tangible deliverables produced during the business process (outputs);
    - the *resources* of the business process like
        - the execution time of the tasks, the activities or the whole process (time resource);
        - the underlying relationships between persons involved in the execution (human resource).

The quality metrics also concerned the measurements for various aggregation levels: tasks, activities or even the whole process. Moreover, their measurement could be either *objective* (e.g., the execution time of a task) or *subjective* (e.g., some metrics are measured in terms of perceived quality).

The measurement of this diverse set of metrics is presented in the following section.

2.2.2 Measuring the quality of a business process

Concerning the measurement of the metrics, some metrics were measured on the business process model, independently of its execution. This is the ”easy part” of the evaluation, which may be performed by automatic tools based on the modelling of the process (roughly speaking counts/aggregates the number of tasks, connections, etc. of the model).

All the other metrics depended on the execution of the transition business process. The metrics that measure the quality of the knowledge transfer are measured by interviewing the project manager and the incoming service provider. The metrics involving the underlying social networks are much more complex as they included a preliminary modelling of the networks (sociologists helped us for the methodological aspects). The (simplified) approach consists in discovering the social network by interviewing the actors (executors and contributors) of the business process, modelling it as a graph and then analysing its structure (Wasserman and Faust, 1994). Collecting the data that compose the social networks is an expensive task in terms of time and human resources.

In order to make things more concrete, Figure 2.5 presents an excerpt of the report that resulted from the quality assessment of the considered transition business process. In this report table, the columns are activities of the transition business process and lines are the quality metrics assessed over each activity. In the results, the higher the measured value is, the riskier is the activity. For simplicity, I decided not to go into details (see (Grim-Yefsah et al., 2016) for details) that would require to deeply explain the activities and the quality metrics, but one can see for instance that the values for the quality metrics *social_net_size* and *social_net_depth* that measure the scale of the hidden social network activated during the execution of an activity, are high for the *TPM ending* activity, which consists in the inventory of the documents and codes that have to be transferred. Notable values are underlined in Figure 2.5. Moreover, the results of the quality assessment at the *task* level (which are not given here), exhibited some very sensitive tasks in the TPM ending activity meaning that these tasks are much more complex than the official procedure suggested, as the executors of these tasks need additional help. The official procedure does not reflect this.

Metric	Activity	Initialisation	TPM ending	Transfer planning	Maintenance in cooperation	Transmission of responsibility
size	1	4	1	2	1
complex	2	1.4	1.7	1.4	1.4
runtime	1	1	1	10	2
tasks_per_day	1	4	1	0.2	0.5
social_net_size	5	13	4	4	5
social_net_depth	2	3	2	2	2

Kexplicit_transm: 0 (no delay in the delivery of transition documents)
Kexplicit_underst: 2.2 (closer level in the scale: neutral)
Ktacit_underst: 3 (poor tacit knowledge)
autonomy of the incoming service provider: neutral
global_runtime: 20 days (satisfied constraint).

Figure 2.5: Some results of the transition business process quality evaluation

Before going further into the quality management, it is worth noticing some important observations concerning a quality assessment approach.

A first observation concerns the **cost** of a quality assessment. In (Grim-Yefsah et al., 2016), we implemented quality metrics for assessing the quality of a business process that are based on a structural analysis of its underlying social networks of contributors. Discovering the social networks of contributors was performed by interviewing the contributors. Each social network is then modelled and its structure is analysed. Such data collection is expensive in time and human resources. Due to limitations of resources, we could not perform such an assessment for all the tasks.

A second observation concerns the **coverage** of a quality assessment. Some quality features were not totally measurable, or even not measurable at all. For instance, in the evaluation of the transition business process presented in Section 2.2, it is known that the motivation of the outgoing service provider for transferring the project to the incoming team plays a major role in the success of the transition process. This motivation is not always obvious, since the outgoing service provider is not renewed and is about to definitively leave the project. Evaluating this motivation is very difficult. Furthermore, in this work, we limited the evaluation to

tasks managed by the IS Department for not to interfere with tasks that the outgoing service provider independently carries out. This shows us that a quality evaluation is often partial, for intentional and unintentional reasons (Marcal de Oliveira et al., 2012; Si-Said Cherfi and Thion, 2012).

2.2.3 Analyzing the results

The quality assessment study was performed because the quality of the process was poor. Then it was clear, from the beginning of the quality study, that improvement actions should be considered. One of the goals of the Analyze stage here was to identify the defective elements of the transition process that led to lose knowledge during the transition. I review below some conclusions of the analysis.

The quality assessment results exhibited a very sensitive activity of the transition process. This is the *TPM ending* activity, which has some high scores in the results presented in Figure 2.5. This activity includes an inventory of the internal and external documents and codes that have to be transferred [4]. Such a result can be explained by the presence of hidden contributors that are involved in the execution of these tasks. They do not appear as official contributors in the business process but their knowledge is needed. If some contributors are missing, then the quality of deliverables is lower than expected.

The quality assessment results also exhibited the fact that some documents written by the outgoing team were difficult to understand for the incoming service provider. This denotes a poor absorption of knowledge by the incoming service provider. This can be explained by the fact that the outgoing team and the incoming one rarely met during the transition process, since the transition focused on writing and transmitting documents.

At the end of the Analyze stage, some quality problems were exhibited and some causes of these problems were possibly identified too. Then improvement actions may be thought of (the improvement actions that followed this study are discussed in Section 3.1). But before considering the quality improvement, the question of the limits of a quality assessment should to be discussed.

[4] Inside the *TPM ending* activity, at the task level, two very sensitive tasks according to the risk of losing some knowledge needed for executing this activity. This level of detail is not presented here.

2.3 The limits of a quality assessment

Positioning. *I review hereafter some limitations that concern the analysis stage of a quality assessment. Details can be found in the following publications: (Akoka et al., 2007), (Marcal de Oliveira et al., 2012), (Barrau et al., 2016) and (Grim-Yefsah et al., 2016).*

The quality assessment approach presented before has some limitations (some of them were mentioned in Sections 2.1 and 2.2). Being conscious of these limitations is important in order to correctly analyse the results of the assessment.

Amount of result data. The information system presented in Chapter 1 is a complex and multidimensional concept. The quality of an information system concerns different components of the system, for instance its data (schema or content), its software or its processes. The quality of each component is itself defined according to a -possibly large- set of quality metrics.

Researchers and companies attempt to provide formal definitions of information quality enabling (possibly automatic) approaches for quality assessment and improvement. Companies use market software tools in order to compute quality measures from their data, applications, processes and hardware.

At the end, quality metrics and tools that compute these metrics are plethoric, leading to a huge amount of data and metadata that are difficult to analyse.

Interdependencies of quality metrics. In addition to its volume, the results of an information system quality assessment themselves are complicated because quality metrics are interdependent (Delone and McLean, 1992, 2003). Let us discuss the problem of the interdependency of quality metrics.

Database engineers know that having an *expressive data schema* that includes rich integrity constraints positively impacts the consistency of data themselves (so is for the uniqueness and the accuracy data quality dimensions). So there is a dependency between the *expressiveness* quality dimension of the *data schema* and the *consistency* quality dimension of the *data* themselves. Another example concerns an architecture in which a back-end software (for instance retrieving sensors data) provides some data to a database, which may itself be used by a business users' front-end application. In such a configuration, if the back-end software has a poor availability, for instance if it is often down or overloaded, then the database may miss some data required by the end-users. In this example, the quality of the *service* offered by the back-end software (its *availability*) impacts the *completeness* of the database.

The inherent interdependencies of the quality metrics impact the analysis of the quality management because the improvement of a quality dimension (expressiveness of the data schema) may have consequences on other quality dimensions. Sometimes the impact is positive, for instance concerning the example above mentioned, improving the service quality dimension of the back-end application providing data to a database would certainly also improve the completeness quality dimension of the database. This is a favourable situation. But the improvement of one quality dimension may also have negative consequences on other quality dimensions. For instance, increasing the completeness of a database model may lead to add a lot of information in the database and then maybe lead to decrease the understandability of data for some users. Another example is the one of increasing the freshness of data (by adding refreshment processes) that may lead to decrease its accessibility (possibly slower query processing times).

Analysing the results of a quality assessment requires the understanding of the interdependencies between the quality metrics, which is still an open research problem.

Incomplete and approximate results. Let us now consider the coverage of a quality assessment study by considering some quality metrics that cannot be (or only partially be) evaluated.

In the context of data quality assessment, the *semantic accuracy* of data is often not evaluated (or only partially) because of missing referential sources or because of a too excessive cost for retrieving the accurate data. The *freshness quality* dimension is also often difficult to evaluate because of missing meta-data. In the context of a business process quality assessment presented before, the motivation of the outgoing service provider for transferring the project to the incoming team could not be measured, and some other quality metrics were intentionally not evaluated for not to interfere with the execution of the process. In the context of open data, missing provenance meta-data could impede the assessment of the *trustworthiness* quality dimension. Moreover, in real life, cost limitations often lead to measure only a part of the quality metrics.

So the results of a data quality assessment are usually incomplete or approximate. In other words, a quality assessment report does not reflect an exhaustive vision of the data quality.

We can conclude this section by saying that a data quality assessment approach, even if it is needed, has limits. Every stakeholder must be conscious of them all along the quality management lifecycle, and more specifically during the analysis of the quality assessment results, which leads to plan improvement actions. We discuss the improvement issue in the next chapter.

CHAPTER 3

DEALING WITH QUALITY ISSUES: IMPROVING... OR NOT

This chapter presents a synthesis of some of my research activities concerning the problem of dealing with (already detected) quality problems in information systems.

Summary of the research activities concerning quality improvement and quality-aware querying

Projects. This research was conducted in the following research projects: the DGA RAPID project called ODIN[a], the CNRS Mastodons project called GIOQOSO[b] and the *Projet scientifique émergent Univ. Rennes 1* called QUALITY@PANAM[c].

Associated theses and internships. The following PhD thesis and internships participated to this research.

- PhD of Olfa Slama (Univ. Rennes 1), on the subject *Flexible Querying of RDF Databases: A Contribution Based on Fuzzy Logic*,
- Master 2 Research internship of Emmanuel Doumard (Univ. Rennes 1/MRI), on the subject of *Personalised querying of data*,
- Master 2 Research internship of Etienne Scholly (Univ. Rennes 1/MRI), on the subject of *Quality management of NoSQL graph databases*,
- Master 2 Research internship of David Mahéo (Univ. Rennes 1/MRI), on the subject of *Flexible querying of fuzzy graph databases: design and implementation*, *Interrogation flexible de bases de données graphe floues : étude et mise en œuvre*,
- A Master 2 internship of Univ. Rennes 1/ENSSAT/INFO, on the subject of *Implementing a parser for the SUGAR prototype*,
- A Master 2 internship of Univ. Rennes 1/ENSSAT/INFO, on the subject of *Implementing the SUGAR prototype for the flexible querying of graph databases*,
- Two Master 1 internships of Univ. Rennes 1/ENSSAT/INFO, on the subject of *Extending the querying of a graph database management system*

Collaborations. AID (company), CEDRIC laboratory (CNAM Paris), EDF R&D (company), ExQI association, IRISA (Univ. Rennes 1), LIRMM (Univ. Montpellier), Semsoft (company).

Associated publications. (Grim-Yefsah et al., 2011a) (Si-Said Cherfi and Thion, 2012) (Grim-Yefsah et al., 2016) (Pivert et al., 2016d) (Pivert et al., 2016c) (Pivert et al., 2014) (Pivert et al., 2015) (Pivert et al., 2016b) (Barrau et al., 2016) (Marcal de Oliveira et al., 2012) (Rigaux and Thion, 2017) (Castelltort et al., 2018)

Softwares. Fuzzy extensions of graph database querying FUDGE and TAMARI prototypes (details on `http://www-shaman.irisa.fr/shaman-software/`).

[a] ODIN stands for *Open Data Intelligence*.

[b] GIOQOSO stands for *Quality management of open music scores* (translation of *GestIOn de la Qualité des partitiOns muSicales Ouvertes*). I was co-coordinator of this project.

[c] QUALITY@PANAM stands for *QUALITY focus on oPen dAta maNAgeMent*. I was coordinator of this project.

The assessment of the quality (presented in the previous chapter) provides an inventory of the quality problems that the information system suffers of. At this point, stakeholders know if the information system components, for which quality was assessed, fit for use. The conclusion of such a quality assessment is often that some elements miss quality. The logical continuation would then be to improve the quality. But, in real life, things are not so simple: it is often the case that quality is not improved or is only partly improved. There are different possible reasons (not necessarily entirely independent) for that. Let us discuss some of them.

That's impossible. Sometimes, improving the quality at a required level may simply be impossible. An example is the one of a missing recording value in a smart object (e.g. a sensor, a meter), a customer form, etc. If the accurate value cannot be retrieved, it is sometimes possible to mathematically estimate the value (clustering, association rules, etc), but not always with a satisfactory preciseness. It is also often impossible to definitely choose between contradictory information stemming from different sources. In such cases, the quality may be improved but not necessarily as much as wanted.

Did you say "quality"? In some cases, the notion of *quality* itself cannot be defined. This is for instance the case in the context of *open* data. In such context, a supplier provides *open* data, that is to say data that can be used by anyone. So the provider make data available without knowing the users. But the data quality is defined as being the *fitness for use* of data (see Section 2) meaning that the data quality definition reflects the users' point of view. It is difficult for the data provider, who does not know the users, to define the quality of her/his data. From the provider point of view, data quality is unclearly defined and, as a result, difficult to improve.

Contradictory requirements. Quality requirements of different end-users, using the same data, can be contradictory (I refer the reader to the discussion concerning the interdependencies of quality metrics in Section 2.3). In such a case, improving the quality for some users could lead to degrading it for others. Some of the quality requirement inevitably won't be satisfied.

It is too expensive. Last but not least, improving the quality has a *cost*, and the price to pay may be too high, either because the benefit of the improvement does not cover its cost, or simply because resources are missing in order to implement improvement action. For instance, in Section 3.1, one of the possible improvement actions (Improvement 4) was not implemented because it was too expansible in human resources.

Let us deeply discuss the notion of *cost* associated with the quality, which has been intensively studied in the literature (Redman, 1996; English, 1999; Huang et al., 1999; Lee et al., 2006; Madnick et al., 2009; Batini and Scannapieco, 2016).

The cost of the quality is a critical pragmatic issue of the quality management. English (1999) identified three categories of information quality costs:

- the information quality assessment and inspection costs,
- the quality process improvement and defect prevention costs, and
- the non-quality information costs.

The *costs of information quality assessment and inspection* include the costs of the resources (software, human resources) needed in order to provide an overview of the quality of the information used in the company. Checking the quality is an unavoidable action as the company should at least know if the quality level of the information allows the business processes to perform properly. An empiric tracking of observable consequences (or absence of consequences) of non-quality is not sufficient because, even if there is no blatant evidence of negative impact due to quality problems, it may be the case that (i) poor quality information is not immediately obvious (”seems good” but, taking a much closer look, is not) or that (ii) one or more layers in the organisation spend time and effort to reduce the effects of the problems, leading to mask them. Moreover, decision makers, who take decisions based on reports produced from data of the company, must be aware of the reliability of the data.

The *costs of quality process improvement and defect prevention* are rather easy to estimate as they correspond to the means needed for developing and implementing the improvement actions. For a same problem, different actions can be thought of. At a very high level, two kinds of approaches may be adopted: the corrective and the preventive ones. In the context of data quality, Redman (1996) proposed to explain them through the analogy of ”the dirty lake”. He presents the case of a lake that is polluted by two factories. In this analogy, the pollute lake is a database that contains ”dirty” data, and the factories are processes that produce data and dump them in the lake. A way to address the problem of the dirty lake is to cleanse the water itself inside the lake by extracting it, filtering it, processing chemical treatment and putting back the cleaned water in the lake. In terms of data quality, it corresponds to *corrective actions* performed over data themselves (at the level of the lake). Another way to address the problem is to consider the stream that produces data, by reducing the pollutant at the level of the factories. In terms of data quality, this corresponds *preventive actions* aiming at revising the processes that produce the data in order to reduce the quality problems at the end. At first sight, implementing preventive actions seems more relevant, but 1) implementing preventive actions is sometimes impossible, for instance when there is no access to the factories (like in the context open data, in which the users have no control on the way the data are produced), and 2) revising the production processes is usually expensible.

The price of implementing improvement actions may be high, maybe a too high price to pay... So, before wondering *how* to improve the quality, one has to wonder *if* the quality really has to be improved. In other words, does the improvement worth it? Clearly, this is a matter of cost. A straightforward approach, recalled by English (1999) and Lee et al. (2006), consists in estimating the economic trade-offs of undertaking a quality program by estimating its benefits and costs, and then applying the basic value equation *Value=Benefits-Costs*. In terms of quality improvement the concrete underlying question is based on an estimation of the balance between the cost of implementing the quality improvement actions with regard to the benefits of implementing them. The benefits depend on the costs of the effects induced by the non-quality (denoted by the *non-quality information costs* by English (1999)), which should be reduced by the improvement actions. But these effects are difficult to estimate because it comprises a lot of different impacts.

English (1999) distinguishes the costs caused by low data quality between *process failure costs*, which are the costs resulting from the business process that does not perform properly (for instance the recovery costs of upset customers), *lost and missed opportunity costs*, which consist of the revenue not realised because of poor information quality (for instance losing dissatisfied customers or missing prospects), and the *information scrap and rework costs*, which include all the resources spent in order to cope with non-usable information (for instance workers looking for missing information in complementary databases, performing again a processes that failed, fixing software in order to improve their robustness to poor quality).

Two other acknowledged contributions of literature proposed alternative classifications of the costs of information quality. Loshin (2001) distinguishes between the *costs of process improvement* and the *costs caused by low quality*. The latter are divided according to the domain they impact, in tow sub-categories: the costs of *operational impacts*, which are the costs at the level of the operational system (for instance the detection costs, correction costs, the prevention costs, the rework costs), and the costs of *tactical and strategical impacts*, which rather concern the organisation and affect the company at longer term (for instance the costs of lost commercial opportunities, the costs due to delayed decisions). Loshin (2010) also proposes to classify the impacts between the *soft* impacts, which are evident but hard to measure (e.g. impacts of an unhappy customer) and the *hard* impacts, whose effects can be measured (e.g. costs associated with fixing customers' problems).

Eppler and Helfert (2004) propose a classification having two main categories: the *costs of improving and assuring information quality* and the *costs caused by low-quality information*. The costs of improving and assuring information quality include the *prevention costs*, the *detection costs* and the *repair costs*. The costs caused by low-quality information are divided between the *direct* costs and *indirect* ones. The direct costs immediately raise from low quality like the verification costs and the re-entry costs. The indirect costs are much longer-term

impacts, including for instance the costs of a lower reputation and the costs of wrong decisions.

The proposed classifications show that identifying the impacts of poor quality is a difficult problem, because the poor quality of the information impacts different domains of the business, at short or long term, inside and outside of the company. Some costs are quantifiable (the cost of compensating the dissatisfied customers, the cost of implementing a monitoring action). But some other costs are extremely difficult to estimate, particularly the costs associated with long-term external impacts (for instance the cost of a lower reputation). In this context, the delicate decision of improving or not the quality must be taken by the chief executive officers (CEOs) of the company. Not surprisingly, the decision may be to improve, not improve or partly improve (restricted area or budget).

As a consequence, the community that studies the problem of managing the quality of an information system considers two kinds of approaches for dealing with quality problems: either improving the quality or going on performing business tasks by coping with quality problems. (Of course, in practice, the two approaches can be combined as the improvement usually solves only a part of the quality problems.)

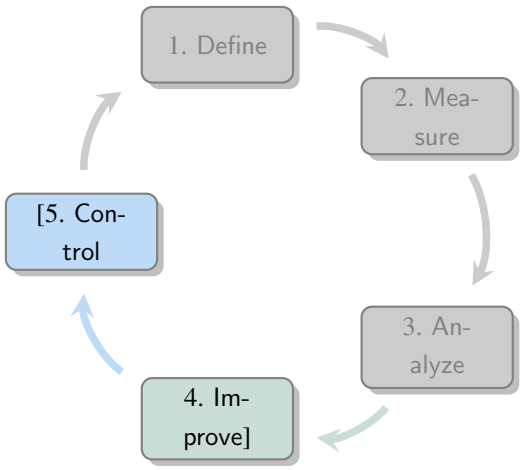


Figure 3.1: Quality improvement in the DMAIC cycle (optional stages 4 and 5)

Improvement in the DMAIC cycle. Let us now consider the improvement of the quality, from the perspective the DMAIC cycle. The *quality improvement* embraces the last two stages of the DMAIC cycle, that is to say the Improve and the Control stages (see Figure 3.1). If no improvement action is decided then these two stages are skipped. The DMAIC method is a cycle so, even is no improvement action is implemented, at least a monitoring of the quality is performed (including adjusting the quality requirements if needed in the Define stage).

In Section 3.1, I present the contributions I participated to, when an improvement action is performed. In Section 3.2, I present the contributions I participated to, when it is decided deal with quality problems at usage time.

3.1 Improving a business process

> I want a little steam on my clothes. Maybe I could fix things up.
>
> I want a little sugar in my bowl, Nina Sinome, 1967

Positioning. *In this section 3.1, I present a contribution on the problem of improving a business process in order to improve the quality of its results. This research is deeply presented in (Grim-Yefsah et al., 2016), (Si-Said Cherfi and Thion, 2012), and (Grim-Yefsah et al., 2011a).*

Let us recall that the *transition* stage is performed in an outsourced project when the service provider changes, leading to the transfer of the project from the outgoing project team to the incoming one. The transfer concerns not only materials (documents and code) but also knowledge. The *transition* is a complex, risky and challenging building block of strategic importance in any outsourced project (Olzmann and Wynn, 2012), whose quality has to be ensured. In Section 2.2, we dealt with the problem of assessing the quality of the transition business process. This lead us to assess the quality of a real transition process. We now consider the problem of improving the transition process.

The literature proposes some approaches for improving a business process. We can mention Becker et al. (2000), who propose a set of guidelines to improve various characteristics of a process model, such as clarity, comprehensibility, or accuracy. Other authors like Mendling et al. (2010) focus on improving the comprehensibility of the models by providing some naming conventions, documentation, and the use of icons or symbols. Other approaches, such as the one introduced by Van Der Aalst et al. (2003), propose a set of best practices encapsulated in reusable and applicable patterns depending on the context. These improvement actions are very general, and need to be adapted and completed when a specific real case is considered.

The literature also proposes best practices in order to achieve a “successful” transition. Olzmann and Wynn (2012) propose a survey of such recommendations. Their analysis focuses

on identifying the critical success factors, for which recommendations are proposed. These proposals concern different stages of the outsourcing process, different involved actors, and address several issues (planning, strategic, operational and financial). But most of the improvement actions are rather general meaning that no practical guidelines are given for their implementation. An example is "Ensure that senior managers from all parties are actively involved in the process".

Contributions. Our first contribution consisted in proposing complementary improvement actions that make possible to improve the knowledge transfer during the transition of an outsourced project (Grim-Yefsah et al., 2011a, 2016). Some of them were used for the improvement of the real transition process that we considered (see Section 2.2.1). I summarise this work hereafter.

The approach that we adopted in (Grim-Yefsah et al., 2016) in order to evaluate the quality of the transition of an outsourced project consists in 1) modelling the transition business process and then 2) assessing the quality of knowledge transfer over the model. This approach allowed exhibiting weaknesses of the process and also locating them in the transition process (located in specific activities and tasks of the business process). To be more concrete, the project manager, with the help of a knowledge manager and a quality expert, analysed the results of the quality assessment (presented in Section 2.2) and, according to the results, identified four possible improvement actions.

As discussed in Section 2.2.2, the quality assessment results exhibited two very sensitive tasks according to the risk of losing some knowledge, in the *TPM ending* activity, where an inventory of the elements (documents, applications and programming codes) to be transferred is performed.

(Improvement 1) As a simple improvement action, the project manager will take a closer attention in the future to the quality of the deliverables produced during the tasks that were identified risky during the assessment stage, especially when these tasks are performed during a period favourable to the absence of employees (holidays, seasonal flu epidemic, corporate reorganisation, etc.)

The quality assessment results also showed that some documents written by the outgoing team were difficult to understand for the incoming service provider. This denotes a poor *absorption* of the knowledge by the incoming service provider. A reason for this is that the outgoing team and the incoming one rarely met during the transition process, since the transition focuses on writing and transmitting documents. The two following improvement actions, focusing on the improvement of knowledge transfer, were then proposed.

(Improvement 2) The Project transfer activity will focus not only on the explicit knowledge transfer but also on the tacit knowledge transfer. The project manager now orchestrates the

activity during which the tangible elements are transferred. He not only ensures the transmission of the documents, but also organises working face-to-face sessions during which the outgoing and the incoming teams, including senior engineers of all parties, share explicit and *tacit* knowledge .

During the *Maintenance in cooperation* activity, the outgoing and incoming service providers assume together the maintenance of the application. This activity is optional according to the procedure. In practice, this activity is often either restricted to the bare minimum, i.e., a short observation phase of the project by the incoming service provider, or simply skipped for cost or time saving reasons.

(Improvement 3) The *Maintenance in cooperation* activity will now become an *exercising Ba*[1] introduced by Nonaka and Konno (1998) taking the form of a workshop in which the outgoing team and the incoming one jointly solve several ongoing incident(s) on the project.

(Improvement 4) Another improvement option would be to enrich the business process procedure by the decomposition of sensitive tasks into sub-tasks where each sub-task models a request to a contributor, which is then made explicit.

In such a modelling, all the contributors (including those that were hidden before) would officially appear in the process. But *Improvement 4* would have complicated the business process procedure, making its management more complex. This situation underlines the problem of the quality metrics interdependencies: improving a quality criterion may lead to degrade another. This impact could be accepted or not by the business actors. In our use case the impact on the complexity was judged unacceptable and the *Improvement 4* was rejected.

We can notice that improving the transition business process implies involving more human resources. Indeed, first, the project manager spends more time on managing the transition. Second, we introduced more face-to-face meetings with all parties, which implies much more investment from the service providers. Therefore, even if it is not surprising, having a higher quality has a cost (the price to pay).

The improvement actions are founded on relevant theories on knowledge transfer, which prove their relevancy and expected effectiveness (see (Grim-Yefsah et al., 2016) and (Grim-Yefsah et al., 2011a) for details). The contributions on which we have relied include:

- the theories of Polanyi (1974) and Alavi and Leidner (2001), who prove that tacit knowledge is needed for the understanding of explicit knowledge;
- the knowledge transfer definition of Davenport and Prusak (1998) (*Transfer = Transmission + Absorption (and Use)*); and

[1] Roughly speaking, a *Ba* is a shared space favourable to individual and collective knowledge advance.

- the theories of Nonaka and Konno (1998) for whom physical and face-to-face experiences are the key to conversion and transfer of tacit knowledge (with the notion of *Ba* previously mentioned).

The result of the improvement actions were measured on the next execution of the transition process, by assessing the same quality metrics while the improvement actions were implemented. This is the Control step of the DMAIC cycle. The control showed an improvement of the quality report (detailed results can be found in (Grim-Yefsah et al., 2016)).

As discussed in Section 3, it is sometimes the case that only a part of the quality is improved. In such situation, one has to deal with quality problems at the time of usage. In the following section, I review some contributions on the problem of dealing with dirty data at the time of usage, when querying data having quality problems.

3.2 Quality-aware querying

> I took the good times, I'll take the bad times.
> I'll take you just the way you are.
>
> Just the way you are, Billy Joel, 1977

Positioning. *I review hereafter some contributions on the problem of introducing quality-awareness in query languages, deeply presented in (Pivert et al., 2016d), (Pivert et al., 2016c), (Pivert et al., 2014), (Pivert et al., 2015), (Pivert et al., 2016b), (Rigaux and Thion, 2017) and (Castelltort et al., 2018).*

The first contribution consists in extending the query evaluation process in order to to attach data quality information with each query answer. This contribution is presented in Section 3.2.1. The second contribution consists in extending the query language with user preferences that improve the language usability. This contribution is presented in Section 3.2.2.

3.2.1 Quality-aware queries for graph-based data

The literature proposes a wide range of metrics for assessing data quality, for different data models including graph-based ones (see the surveys of Batini and Scannapieco (2016) and Zaveri et al. (2016)). These metrics are used to detect quality problems in data and to measure the data quality level. Now, assuming that the quality level is known, a question still raises: "How to take quality information into account at the time of use, when querying data?" This is the problem we considered in the context of graph databases.

Contributions. In (Rigaux and Thion, 2017), we proposed an extension of the graph database querying process that allows introducing quality awareness when querying data. Based on *quality annotations* that denote quality problems appearing in data subgraphs (the annotations may result either from an automatic evaluation of data quality, for instance by computing quality metrics defined in the literature (Kontokostas et al., 2014; Zaveri et al., 2016), or from a human tagging process that is a typical collaborative practice in the context of open data usages (Zaveri et al., 2013; Acosta et al., 2013)) and a *quality vocabulary*, we propose a notion of *quality aware query* based on (usage-dependent) *quality profiles* defined according to the quality vocabulary. Roughly speaking, the framework extends the basic graph pattern queries in order to introduce the computation of quality scores for the answers, according to a given quality profile.

For simplicity, I do not present the theoretical foundations of the work, which can be found in (Rigaux and Thion, 2017). I only give the intuition of the contribution, introducing it by an example.

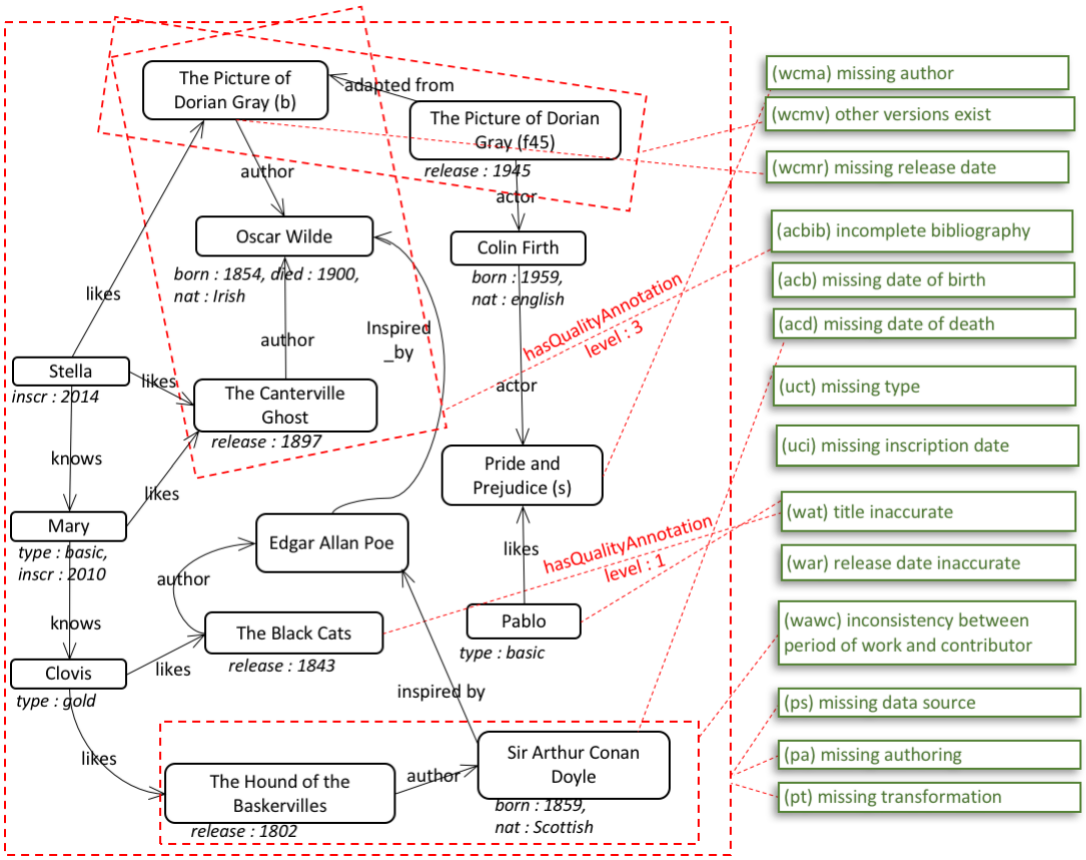


Figure 3.2: Data graph $\mathcal{G}$ (black) and associations (red) of quality annotations (green)

Figure 3.2 illustrates the elements of the data model that we consider: the attributed data graph model[2]. The left black part of Figure 3.2 models data of a social network information system dedicated to literature. It contains nodes denoting users, works of art and artists (authors and actors), and connections between nodes that we expect as being explicit enough for not to detail them.

Quality annotations of a given vocabulary, which denote quality problems, may be attached to data. In Figure 3.2, the annotations appear in green on the right part of the illustration. The attachment is modelled by the red dashed relation that connects annotations (quality problems) to data subgraphs.

The annotations are defined in an adaptable quality vocabulary, which has the form of a taxonomy. For the running example, the taxonomy of Figure 3.3 (limited to the green elements, which are the edges without labels, and the nodes) classically organise the annotations according to quality dimensions of the literature.

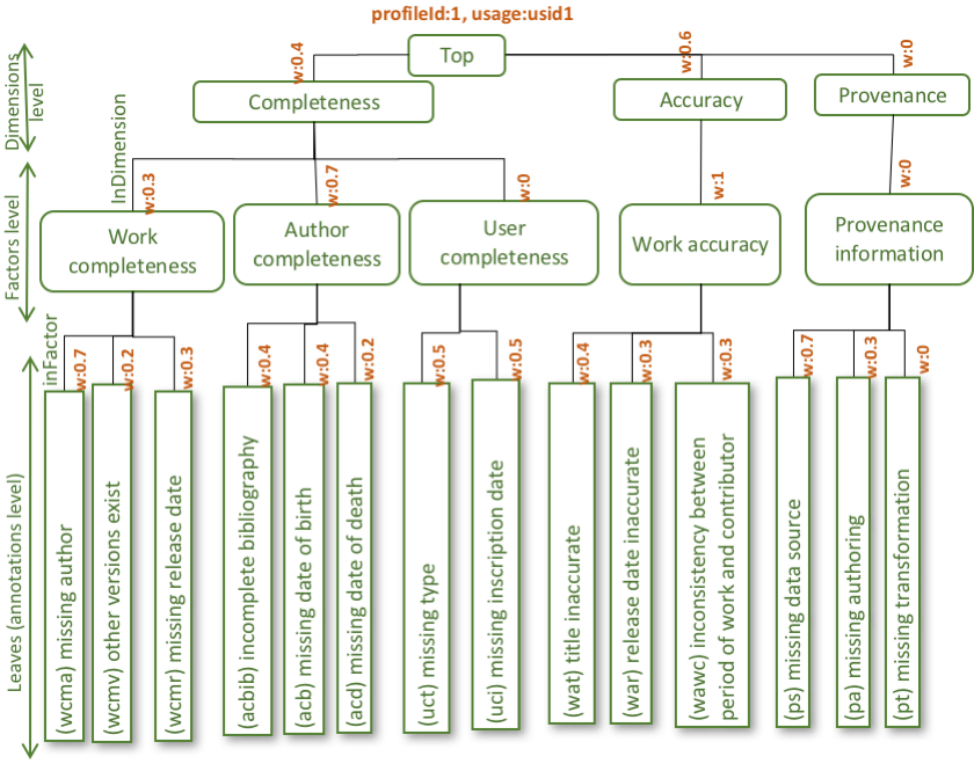


Figure 3.3: Quality taxonomy (green) and quality profile (orange)

[2]See Angles and Gutierrez (2008); Wood (2012); Angles (2012) for a survey presenting the graph data model and its theoretical foundations

Depending on its quality requirements (see Section 1.2), a user can define a quality profile by attaching weights to edges of the quality vocabulary, where a weight defines the degree of interest of a quality element (node of the vocabulary) for the usage. Weights joined to the vocabulary in Figure 3.3 constitute an example of a profile, denoted by *profileId1*.

A classical pattern query for querying graph data is a graph where variables and conditions can occur, which defines the shape that has to be found in the data. Figure 3.4 is a graph pattern query, denoted by $\mathcal{P}_{\text{Clovis}}$, that aims at retrieving works of art (variable `w2`) of authors (variable `a2`) inspired by authors (variable `a1`) of works (variable `w1`) that Clovis (variable `c`) likes.

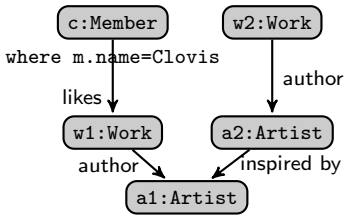


Figure 3.4: $\mathcal{P}_{\text{Clovis}}$

The semantics of the interpretation of a pattern query $\mathcal{P}$ over a graph $\mathcal{G}$ is the set of data subgraphs that "match" $\mathcal{P}$ (see (Gallagher, 2006; Barceló et al., 2014) for the theoretical foundations). The evaluation process consists in binding elements of the pattern in subgraphs of the database. Roughly speaking, if $\mathcal{P}$ is a query, then the answer retrieved by $\mathcal{P}$ is the subgraphs that fit the shape of the pattern. Figure 3.5 presents the answer of $\mathcal{P}_{\text{Clovis}}$ (Figure 3.4) over the graph $\mathcal{G}$ (Figure 3.2).

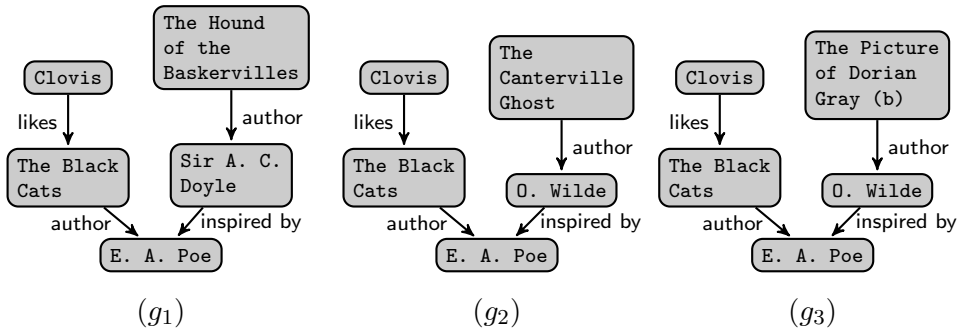


Figure 3.5: Answers of $\mathcal{P}_{\text{Clovis}}$ over $\mathcal{G}$

We extended the classical notion of graph-pattern query by proposing the notion of *quality-aware query*, which takes a quality profile into account in the query. According to the profile and the quality annotations, the result of such a query computes a quality score associated to each retrieved answer.

Query 3.1 is an example of extended query, expressed in a quality-aware extension of the Cypher language. In the CYPHER query language, a graph pattern is defined *à la ASCII art*. The symbol `( )` denotes a node, which may contain information of the form `query_variable:Type` concerning this node. The symbol `-[form]->` denotes a connection between two nodes (or variables) i.e. the form of a path that connects the nodes. Query 3.1 aims at retrieving works of art of authors inspired by authors of works that Clovis likes ($\mathcal{P}_{\text{Clovis}}$), according to the quality profile *profileId1*. Syntactically, the extension consists in adding a clause `QTAWARE` at the beginning of the query, for specifying the considered quality profile and the quality concepts of interest. The semantics of the quality scores (how they are calculated) is defined in (Rigaux and Thion, 2017). The intuition is that, for each subgraph that belongs to the answer, the more suspect is the subgraph according to the quality elements of interest declared in the query, the higher is the quality alert score associated with the subgraph.

```
1 QTAWARE profileId1, Completeness, Accuracy
2 MATCH
3  (c)-[l:likes]->(w1:Book), (w1)-[:author]->(a1),
4  (a2)-[:inspired_by]->(a1), (w2)-[:author]->(a2)
5 WHERE c.name='Clovis'
6 RETURN c, w1, a1, a2, w2
```

Query 3.1: Quality aware Cypher query (Q_{qt})

We showed how to simply extend a generic state-of-the-art algorithm for graph pattern queries evaluation in order to implement quality awareness at evaluation time, and we studied its complexity. We proved that the additional cost for introducing quality-awareness is highly dominated by the cost of the evaluation of the pattern query without quality awareness. In other words, adding the quality awareness has an acceptable cost in terms of query evaluation time.

Implementation-wise, two architectures may be thought of. A first one consists in implementing a specific quality aware query evaluation engine. The advantage of this solution is that optimisation techniques implemented directly in the query engine should make the system very efficient for the query processing. The downside is that quality aware queries may not be evaluated by an independent engine that does not implement the quality aware functionality.

The second solution consists in using a possibly distant classical engine, combined with a dedicated add-on layer. This is the solution that we have chosen. The implementation relies on a query-rewriting derivation mechanism, carried out as a pre-processing and a post-processing

steps.

As a proof-of-concept of the proposed approach, we implemented the open-source prototype TAMARI[3] (for Quality Alerts Management using RabbIthole), which adds quality awareness to the Cypher language (Neo Technology, 2013) for querying a Neo4j graph database (Neo4j, 2019). TAMARI implements a quality add-on layer on top of a classical (non-quality aware) Cypher query engine, by extending the RabbitHole console (RabbitHole project, 2019). The architecture of TAMARI is depicted in Figure 3.6.

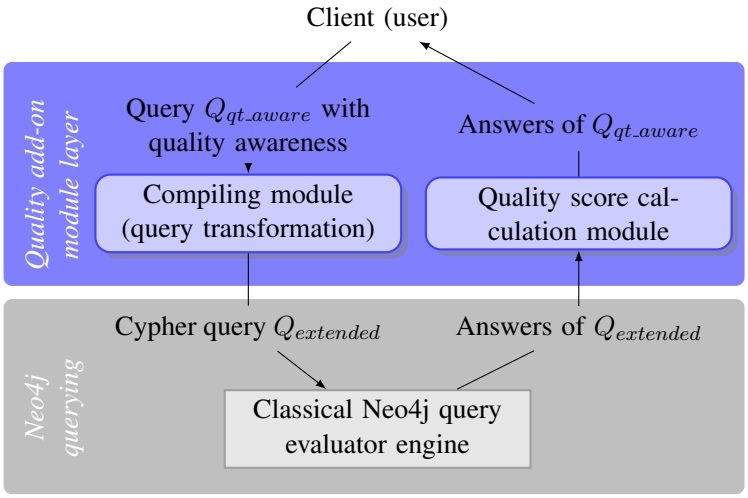


Figure 3.6: Architecture of TAMARI

The add-on layer is composed of two modules. A *Compiling module* transforms the graph pattern Cypher query Q_{qt_aware} into an extended one $Q_{extended}$ that retrieves all the needed information, concerning not only the answers but also the user profile, the quality vocabulary and the association of the vocabulary to the answers. The extended query is then sent to the (classical) Neo4j engine. Based on the answers of $Q_{extended}$, a *Quality score calculation module* calculates the quality alert scores associated with each answer of Q_{qt_aware}[4].

Figure 3.7 is a screenshot of the TAMARI graphical user interface, after the evaluation of Query 3.1. For each retrieved answer, the quality alert scores for the accuracy dimension and the completeness dimensions are given in the sixth and seventh columns respectively.

[3] TAMARI is available at `www-shaman.irisa.fr/tamari`.

[4] Note that, in terms of expressivity, this calculation cannot be expressed in the extended query. A calculator module is then needed.

Figure 3.7: Screenshot of the TAMARI prototype

3.2.2 Flexible query language for graph-based data

In practice, the information system often has some characteristics that make its data management difficult. For instance, the information system can be difficult to use (some possible reasons are a query language that is too complex for a non-expert user, a data schema that is -voluntary- flexible (e.g. NoSQL databases) but intrinsically difficult to understand), the volume of data may be large, and the data themselves may be complex, incomplete, redundant or inconsistent. When querying data of the information systems, these problems may lead to empty answers or, on the contrary, to plethoric answers. In such situations using the information system data is difficult for the end-user, negatively impacting the *usability* quality dimension of the system. This reduces the user satisfaction, even though it is a key factor in the success of an information system (see discussion in Section 1.2). Introducing flexibility in the query language allows improving the usability of the information system and its robustness to quality problems. This is the problem that we consider in the following, for the graph data model.

A way to introduce flexibility in a query language is to allow users to define preferences (Dubois and Prade, 1997; Kießling and Köstler, 2002) in their queries (the preferences take the form of fuzzy conditions in the framework that we proposed). First, such an approach offers a more expressive query language that can be more faithful to what a user intends to say. Second, the introduction of preferences in queries provides a basis for rank-ordering the retrieved items, which is especially valuable in case of large sets of items that answer to a query. Third, a

classical query may also have an empty set of answers, while a less restrictive version of the query might be matched by some items.

Much work has been done about fuzzy querying of *relational* databases (for instance, Bosc and Pivert (1995); Pivert and Bosc (2012) defined a fuzzy extension of the SQL language). Graph databases raise new challenges in terms of flexible querying since two aspects may be involved in the preferences that a user may express: i) the content of the nodes and ii) the structure of the graph itself.

Contributions. Let us first consider the data model (independently from its querying). The classical graph data model is only capable of representing Boolean notions whereas real-world concepts are often of a vague or gradual nature. This is why it may be extended into the notion of a fuzzy graph database[5] where a fuzzy degree is attached to edges in order to express the "intensity" of a gradual relationship (e.g., *likes*, *is friends with*, *is about*). We then first proposed a formal definition of this concept, based on an extension of the attributed graph data model.

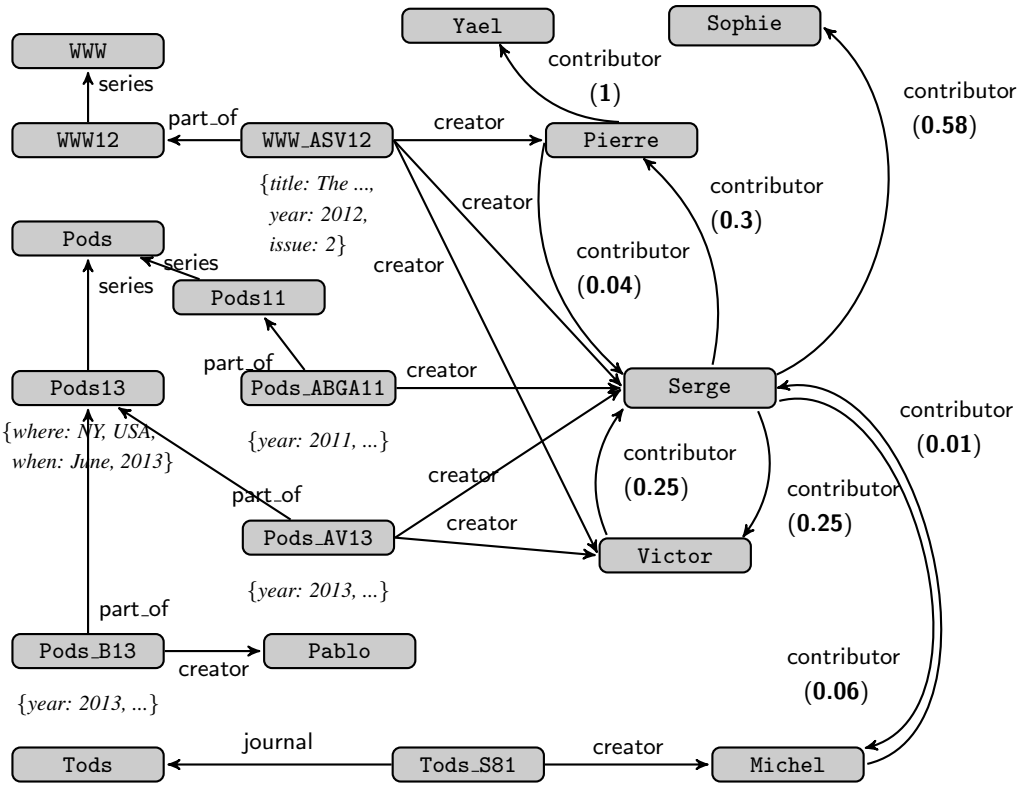


Figure 3.8: A fuzzy data graph $\mathcal{DB}$ inspired by an excerpt of DBLP data

[5]The fuzziness is based on fuzzy set theory introduced by Lofti A. Zadeh in (Zadeh, 1965).

Figure 3.8 is an illustration of a fuzzy data graph inspired by DBLP[6], with some fuzzy edges (fuzzy degree in brackets), and crisp[7] ones (degree equal to 1). In this example, the degree associated with A -contributor-> B is the proportion of journal papers co-written by A and B, over the total number of journal papers written by B. Here, the degree is based on a simple statistical notion, but it could be made more sophisticated by the integration of expert knowledge.

A second contribution concerns the flexible querying of such data. In (Pivert et al., 2014), we proposed an algebra, based on fuzzy set theory and the concept of a fuzzy graph, which can be used to express preference queries on fuzzy graph databases. The preferences concern i) the content of the vertices of the graph and ii) the structure of the graph. This theoretical foundation has led to the definition of a query language, called FUDGE (Pivert et al., 2014, 2015), which is an extension of the CYPHER language used for querying crisp graph databases in a crisp way in the Neo4j graph database management system. For the sake of conciseness, I do not present the theoretical foundations of this work (the algebra), I only briefly introduce the FUDGE language, by an example.

Let us first consider the CYPHER query 3.2, which aims at retrieving information concerning authors (variable au2) who have, among their contributors, an author (variable au1) who published a paper (variable ar1) in WWW and also published a paper (variable ar2) in Pods after 2014 (ar2.year > 2014). For the example, we will extend this query by adding fuzzy preferences.

```
1 MATCH
2   (ar1:Article)-[part_of]->()->[series]->(s1),
3   (ar2:Article)-[part_of]->()->[series]->(s2),
4   (ar1)-[:creator]->(au1:Author),
5   (ar2)-[:creator]->(au1:Author),
6   (au1)-[contributor]->(au2:Author)
7 WHERE s1.id=WWW AND s2.id=Pods AND ar2.year >2014
```

Query 3.2: A CYPHER query

The FUDGE extension allows introducing fuzzy preferences that may concern i) the content of the vertices of the graph and ii) the structure of the graph. The FUDGE query 3.3 aims at retrieving information concerning authors (au2) who have, among their *close* contributors (authors connected with au2 by a *short* path), an author (au1) who published a paper (ar1) in WWW and also published a paper (ar2) in Pods *recently* (specified by the fuzzy condition ar2.year IS recent). The DEFINE clauses allow defining the fuzzy terms $short$ and $recent$, whose definitions are gradual (respectively given in Figure 3.9 and Figure 3.10).

[6] http://www.informatik.uni-trier.de/~ley/db

[7] The term *crisp* means *non fuzzy*.

```
1 DEFINEDESC short AS (3,5),
2 DEFINEASC recent AS (2010,2014)
3 MATCH
4   (ar1:Article)-[part_of]->()->[series]->(s1),
5   (ar2:Article)-[part_of]->()->[series]->(s2),
6   (ar1)-[:creator]->(au1:Author),
7   (ar2)-[:creator]->(au1:Author),
8   (au1)-[(contributor+) | Length IS short]->(au2:Author)
9 WHERE s1.id=WWW AND s2.id=Pods AND ar2.year IS recent
```

Query 3.3: A FUDGE query

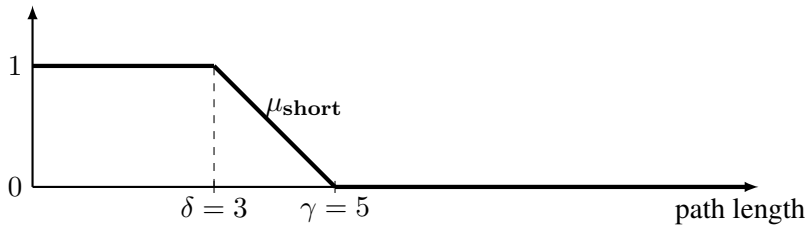


Figure 3.9: Representation of the fuzzy term $short$

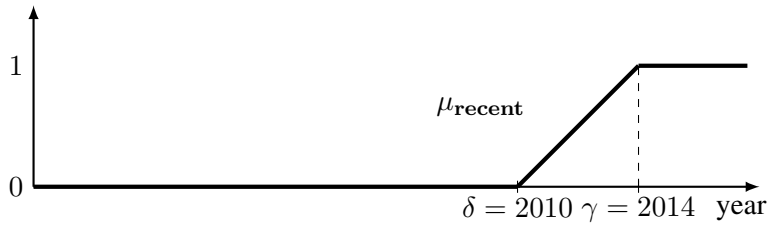


Figure 3.10: Representation of the fuzzy term $recent$

Like for the crisp context, the answer of such a query $\mathcal{P}$ over a graph database $\mathcal{G}$ is the largest set of subgraphs of $\mathcal{G}$ defined by $\{g \in \mathcal{P}(\mathcal{G}) \mid g$ "matches" $\mathcal{P}\}$ (see (Gallagher, 2006; Barceló et al., 2014) for the definition of matching in the crisp context). The fuzzy extension allows introducing a satisfaction degree associated with each answer. This degree reflects the extent to which the answer satisfies the fuzzy query pattern. In terms of contributions, we extended the theoretical foundations of the crisp context to the fuzzy one.

As a proof-of-concept, the FUDGE language was implemented in an open-source prototype called SUGAR[8] (Pivert et al., 2016b). The SUGAR software is based on the Neo4j system

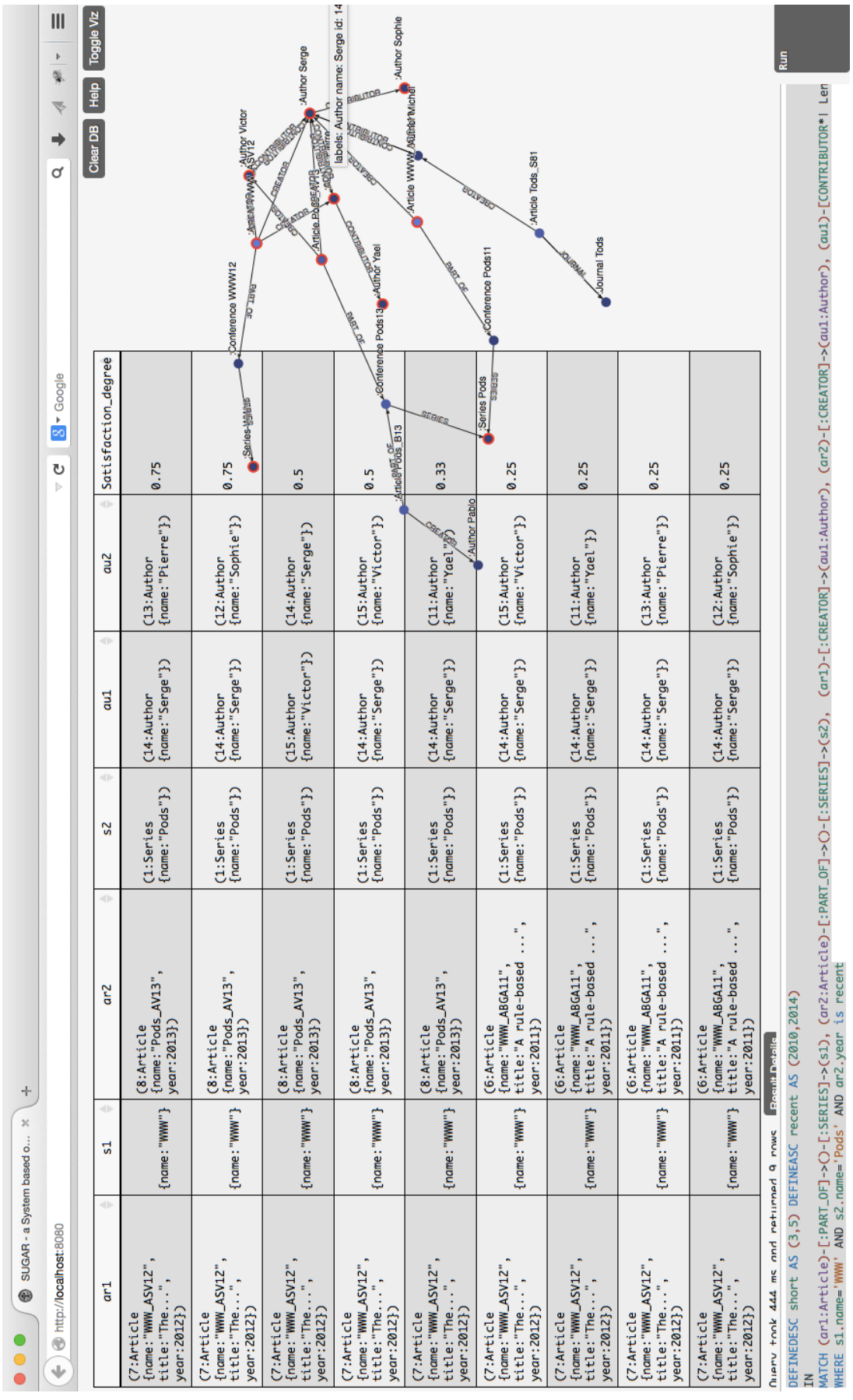
[8] SUGAR is available at `www-shaman.irisa.fr/fudge-prototype`.

(Neo4j, 2019) that implements the CYPHER (crisp) query language. SUGAR extends the interactive Neo4j REPL Console Rabbithole (RabbitHole project, 2019). It takes the form of an add-on layer on top of the classical Neo4j engine. Details of its implementation (how to extend the data model, the modules that allow extending the query language, their optimisation, benchmarks for cost estimation) are presented in (Pivert et al., 2015).

Figure 3.11 on page 61 presents a screenshot of the SUGAR graphical user interface, which contains the final result of the evaluation of Query 3.3 over the database of Figure 3.8. Each line of the result table defines an answer subgraph in terms of the mapping of the query pattern variables into the database. The satisfaction degree associated with each answer subgraph appears on the last column of the table. The answers are ordered by decreasing order of the satisfaction degree, meaning that the most relevant answers appear before.

A similar contribution that considers the flexible querying of fuzzy RDF data model (Pivert et al., 2016d) was proposed. In (Pivert et al., 2016c), we defined the foundations of a flexible query language for RDF, called FURQL, which is a fuzzy extension of the SPARQL query language. The FURQL language was implemented an experimented in a system called SURF (Pivert et al., 2016a) [9].

[9] SURF is available at `https://www-shaman.irisa.fr/surf`



ar1	s1	ar2	s2	au1	au2	Satisfaction_degree
(7:Article {name:"WWW_ASV12", title:"The...", year:2012})	{name:"WWW"}	(8:Article {name:"Pods_AV13", year:2013})	(1:Series {name:"Pods"})	(14:Author {name:"Serge"})	(13:Author {name:"Pierre"})	0.75
(7:Article {name:"WWW_ASV12", title:"The...", year:2012})	{name:"WWW"}	(8:Article {name:"Pods_AV13", year:2013})	(1:Series {name:"Pods"})	(14:Author {name:"Serge"})	(12:Author {name:"Sophie"})	0.75
(7:Article {name:"WWW_ASV12", title:"The...", year:2012})	{name:"WWW"}	(8:Article {name:"Pods_AV13", year:2013})	(1:Series {name:"Pods"})	(15:Author {name:"Victor"})	(14:Author {name:"Serge"})	0.5
(7:Article {name:"WWW_ASV12", title:"The...", year:2012})	{name:"WWW"}	(8:Article {name:"Pods_AV13", year:2013})	(1:Series {name:"Pods"})	(14:Author {name:"Serge"})	(15:Author {name:"Victor"})	0.5
(7:Article {name:"WWW_ASV12", title:"The...", year:2012})	{name:"WWW"}	(8:Article {name:"Pods_AV13", year:2013})	(1:Series {name:"Pods"})	(14:Author {name:"Serge"})	(11:Author {name:"Yael"})	0.33
(7:Article {name:"WWW_ASV12", title:"The...", year:2012})	{name:"WWW"}	(6:Article {name:"WWW_ABGA11", title:"A rule-based ...", year:2011})	(1:Series {name:"Pods"})	(14:Author {name:"Serge"})	(15:Author {name:"Victor"})	0.25
(7:Article {name:"WWW_ASV12", title:"The...", year:2012})	{name:"WWW"}	(6:Article {name:"WWW_ABGA11", title:"A rule-based ...", year:2011})	(1:Series {name:"Pods"})	(14:Author {name:"Serge"})	(11:Author {name:"Yael"})	0.25
(7:Article {name:"WWW_ASV12", title:"The...", year:2012})	{name:"WWW"}	(6:Article {name:"WWW_ABGA11", title:"A rule-based ...", year:2011})	(1:Series {name:"Pods"})	(14:Author {name:"Serge"})	(13:Author {name:"Pierre"})	0.25
(7:Article {name:"WWW_ASV12", title:"The...", year:2012})	{name:"WWW"}	(6:Article {name:"WWW_ABGA11", title:"A rule-based ...", year:2011})	(1:Series {name:"Pods"})	(14:Author {name:"Serge"})	(12:Author {name:"Sophie"})	0.25

Query took 444 ms and returned 8 rows

Figure 3.11: Screenshot of the SUGAR system

CHAPTER 4

CONCLUSION AND PERSPECTIVES

In this report, I have presented some scientific contributions I participated to, concerning the quality management of an information system, with a human-centric point of view.

In a first chapter, I introduced some preliminary notions. I have discussed the notion of *information system* by emphasising its inherent human feature, as an information system is not only composed of software tools handling data, business rules and routines, but also embeds human beings, who organise their job, use the software components (or not), and collaborate by sharing information with each other, each of the humans having her/his own behaviour and individual knowledge. Then I discussed the problem of managing the quality of such an information system, and introduced the DMAIC methodology that I used as a unifying framework within this document.

In a second chapter, I presented some contributions on the quality assessment issue, which corresponds to the D (Define), M (Measure) and A (Analyse) stages of the DMAIC methodology. These contributions focus on the assessment of the data and the business processes of an information system.

In a third chapter, I presented some contributions about handling quality problems. First, I considered the quality improvement (which corresponds to the I (Improve) and C (Control) stages of the DMAIC methodology) of a business process. Then I considered how to deal with data quality problems, when the quality cannot be improved as desired. I focused on the method that consists in introducing quality-awareness at query time (in the context of graph-based data). The final goal of such a method is to improve the usability of the data for an end-user.

These quality management issues open a lot of research perspectives. Some short-term perspectives directly extend the contributions presented in this document. Some other perspectives are longer-term ones. Let us first consider some short-term perspectives that may concern

either the quality assessment issue (more related to the second chapter), or the quality-aware querying (more related to the third chapter).

(Quality assessment) Designing quality metrics. In Section 4, I mentioned that the literature proposes an abundant catalog of quality metrics. But, despite its impressive size, this catalog is usually not sufficient in practice because most of the proposed quality metrics are very general. Additional context-dependent quality metrics are almost always needed. Then the *design of quality metrics* that are complementary to (or more relevant than) the ones proposed in the literature is still an issue.

For instance, in the context of music scores, when checking the availability of a lyric associated with a note, the simplest (but approximate way) is to check that the text is available and that it contains a vowel and possibly a consonant before or after. Some other methods could be thought of (using a phonetizer).[1]

(Quality assessment) Diversifying the applications. When proposing an approach (for quality management here), it is interesting to experiment the contribution in other contexts. In section 2.2.1, I presented some quality metrics for assessing the risk of losing knowledge needed in a business process. This contribution was initially designed, driven by its application to a transition process in an outsourcing project. In terms of use case applications, it would also be interesting to apply these metrics to other business processes in other contexts, for instance to digital score libraries production processes, which require some rare very specific skills.

Another interesting application area is the one of crowdsourcing processes. *Crowdsourcing* is the outsourcing of a piece of work to a crowd of people *via* an open call for contributions (Howe, 2006). Typical tasks that are submitted to a crowd of workers are the tagging of images, or the translation of a piece of text. In this context, the quality of the outputs, produced by heterogeneous contributors having various skills, obviously has to be checked (Daniel et al., 2018). The literature already proposes some quality metrics for assessing the quality of a crowdsourcing system (Daniel et al., 2018). The corresponding quality model includes quality metrics that concern three facets of the crowdsourcing: the tasks proposed in the crowdsourcing system (for instance, checking the usability of their user interface, their cost, the clarity of their description), the data required to perform the tasks (for instance, checking their accuracy, their timeliness), and the people involved in the system (for instance, checking the adequacy of the contributors 'skills to the performed tasks). Some contributions propose to model crowdsourcing processes as a (set of) workflows (Bozzon et al., 2014). As soon as a crowdsourcing

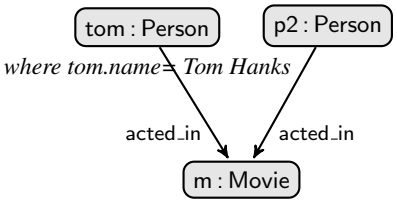
[1] I am grateful to my colleagues of the IRISA/Expression team for the discussions we had concerning this subject.

process is modelled as a set of sequencing tasks, it would be interesting to study in what extent the methodology that we proposed for evaluating the risk of losing knowledge could be applied in such a context.

The contributions that consist in diversifying the application cases would allow transferring the contributions to another application domains, and would also confirm that it is possible to generalise the contributions.

(Quality-aware querying) Fuzzy quality awareness. In Section 3, I presented two approaches for dealing, at query time, with data quality problems that appear in graph-based data. These approaches consist in extending the query language implemented in the database management system, in order to introduce quality-awareness in queries. The first extension allows attaching a quality alert score to retrieved answers. The second extension consists in introducing fuzzy preferences that allow a flexible querying of data. These two approaches could be combined in order to allows a *flexible quality-aware querying* of data. This work is in progress in the Shaman team. We are defining a framework that makes possible to attach data quality information to attributed graphs (extending the attributed graph data model) and to introduce fuzzy preferences concerning data quality (extending the syntax and the semantics of the graph pattern query notion). This theoretical framework may be implemented in order to extend the CYPHER query language. The query 4.1 is an example of such an extended query.

```
DEFINEASC high AS (0.6,0.8)
// not high under 0.6, gradual between 0.6 and
    0.8, definitely high over 0.8
MATCH (tom:Person)-[:acted_in]->(movie:Movie),
      (movie:Movie)<-[:acted_in]-(actor)
WHERE tom.name = 'Tom Hanks'
QTPREF accuracy(m) IS high
QTAWARE accuracy(p2), completeness(p2)
RETURN tom, movie, actor
```



Pattern expressed in the MATCH/WHERE clause

Query 4.1: Query with quality fuzzy preferences

Such a query aims at retrieving the subgraphs that maps the graph pattern expressed in the MATCH/WHERE clause, that is to say the actors (variable p2) who played in a movie (variable m) with Tom Hanks (variable tom denoting a node whose mane is Tom Hanks), for which the accuracy of the movie is high. Associated with the answer, the completeness and the accuracy of the second actor is required.

The expected answer of this query applied to a data graph, is the set of the data subgraphs that match the graph pattern with, for each subgraph, three additional informations calculated

according to the quality metadata attached to the subgraph:

- its satisfaction degree (fuzzy score) according to the high accuracy of the movie that matches the variable m,
- the accuracy degree of the node that matches the variable p_2,
- the completeness degree of the node that matches the variable p_2.

This work is in progress, in collaboration with members of the Shaman team[2].

(Quality-aware graph querying) Diversifying the applications. Among the short-term perspectives, I also mention the *graph-based management of music scores*. This issue consists in 1) modelling music scores and their quality annotation as graphs, and 2) offering a graph-based querying of such data, which offers a convenient way to query the structure of the data. Then the flexible and quality-aware querying methods presented in this document could be applied. Let me go one step further in this suggestion, with an illustration.

Let us consider Figure 4.1, which is a human-readable visualisation of the excerpt of a music score, initially extracted from a MEI dataset available in the NEUMA platform proposed by Rigaux et al. (2012)[3].

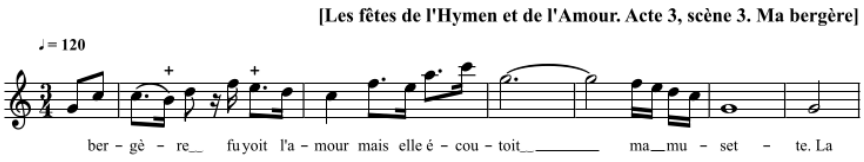


Figure 4.1: Music score

This music score suffers from quality problems. First, over the whole music score, the author of the musical work is missing and the source of the document is also missing. Second, other quality problems occur at the measure [4]level:

- in the measure 0 (the incipit that -voluntary- contains only one beat), a lyric is missing on the first note;
- in the measures 1 and 3, there are unreadable characters (modelled by underscore characters) after the syllables “re” and“toit”;
- in the pair of measures 5 and 6, a beat has been moved from the measure 5 to the

[2]Under submission to the Information Sciences journal. Pivert, O., Smits, G. and Thion, V. Fuzzy Quality-Aware Queries to Graph Databases.

[3]I slightly adapted the content of this music score in order to illustrate the introduced concepts.

[4]A *measure* in a music score is a section of a musical staff that comes between two bar lines.

measure 6, (the time signature $\binom{3}{4}$ indicates that three beats per measure are expected but the measure 5 contains 4 beats, and the measure 6 contains 3 beats), and

- on the last measure, the lyric is not a syllable.

The idea is first to model such data as a graph. The left part of Figure 4.2 is a graphical representation of the measure 1 of the music score considered in Figure 4.1. Elements of interest of the music score, that is to say here measures, notes and lyrics are modelled as nodes. Edges model relevant relationships between the nodes, for instance in order to associate notes with measures and lyrics with notes. The quality annotations can be modelled in the graph based formalism (like proposed in Section 3.2.1) and be associated with data. This is the right part of Figure 4.2, which models some of the quality problems associated with the measure 1 of the music score considered in Figure 4.1.

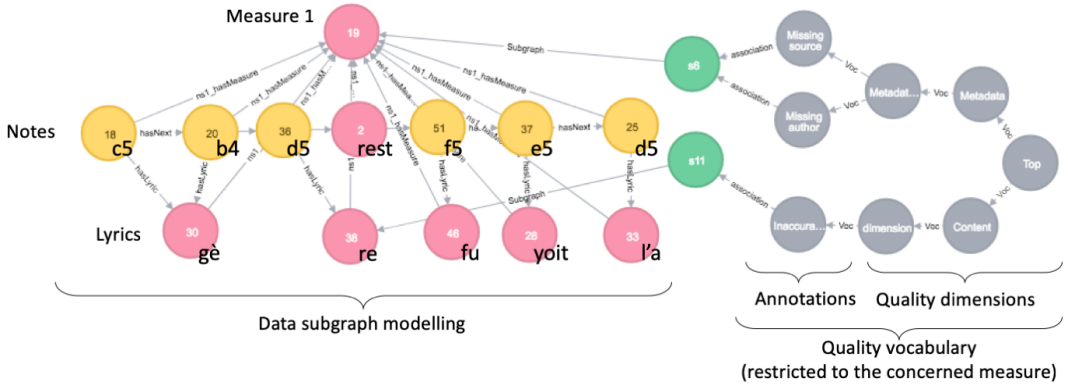


Figure 4.2: Second measure of the music score considered in Figure 4.1 (based on a screenshot of the Neo4j browser software)

Then the flexible and quality-aware extensions of the graph-based query language that we proposed in Section 3.2.1 would aim at retrieving relevant parts of the music score, based on both structural and quality features of the score. An example is Query 4.2, which aims at *retrieving the notes that closely follow a d5*[5] *note and have an associated lyric which is highly accurate*. The information of the completeness of the subgraph pattern that is made of the note and its measure is also needed.

[5]The note *d5* in the ango-saxon notation is the note *ré* (fifth octave) in the Latin notation.

```
DEFINEDESC shortPath AS (2,5) // definitely close if distant of less than 2 notes,
    gradual between 2 and 5 notes, definitely not close if more distant
DEFINEASC high AS (0.6,0.8) // not high under 0.6, gradual between 0.6 and 0.8,
    definitely high over 0.8
MATCH (d5)-[(next+)|Length IS shortPath]->(n),
       (n)-[r1:hasLyric]->(l),
       (n)-[r2:hasMeasure]->(m)
WHERE d5.hasOctave=5 AND d5.hasPitch=d
QTPREF accuracy(l) IS high
QTAWARE completeness((n)-[r1:hasLyric]->(l),(n)-[r2:hasMeasure]->(m))
RETURN n,m,l
```

Query 4.2: Query with quality fuzzy preferences

The intended result is the extraction of music score patterns with their associated quality levels. This work is in progress in collaboration with the CNAM Paris Cedric laboratory and the De Vinci Research Center.

(Quality-aware querying) Extending the considered quality model. Another interesting perspective concerns the quality-aware querying of graph databases, focusing on the considered quality model.

The contributions presented in Section 3.2 are based on a quality vocabulary made of quality metrics (or annotations) organised in taxonomies that classify the quality metrics according to quality dimensions. This is a very simple model. The literature proposes a lot of conceptual quality models for modelling the quality that are much more sophisticated, and go beyond the classification of the quality metrics according to quality dimensions (see the surveys proposed by Batini and Scannapieco (2016) and Radulovic et al. (2018)), by including qualitative information like provenance information concerning the quality values (measurement methods, date, tools, certification, actors, etc.) The quality vocabulary that we proposed could be extended in order to model such complementary information. Such a modelling of quality meta-data raises the problem of offering a user-friendly quality-aware query language that makes possible to use this more complex information.

(Short-term perspective) Assessing the usability of the proposed query languages. Few work has been done in the context of data quality management in attributed graphs. So, relevant quality-driven datasets that could serve as a basis for relevant benchmark studies are still missing in the literature. Such benchmarks would not only allow checking the tractability of the solutions (let us mention that in the contributions proposed in Section 3.2, the tractability is predictable as the theoretical extra cost was studied), but would also allow studying the *usability of the proposed query language*, from the users' point of view.

We can also consider some more general and longer-term perspectives.

(Long-term perspective) Choosing the relevant metrics, in context. The literature proposes a very large range of quality metrics. As a matter of illustration, Batini and Scannapieco (2016) reviewed a wide range of quality dimensions and metrics for data quality, Zaveri et al. (2016) reviewed about seventy metrics dedicated to open data, and Monteiro and de Oliveira (2011) reviewed five hundred metrics for measuring the performance of a software process. This is of course a very useful resource. But there are too many available quality metrics in a sense and too few of them in another sense. Let me explain this observation.

On the one hand, there are too many metrics because, even if guidelines exist in order to identify quality requirements, it is difficult to *extract, amongst the large catalog of available metrics, the metrics that are relevant* for a specific usage (Marcal de Oliveira et al., 2012).

On the other hand, this abundant catalog of metrics is usually insufficient in practice because most of the proposed quality metrics are general. Complementary specific quality metrics are almost always needed. So, in the real life, choosing a relevant set of quality metrics remains a challenging issue (Marcal de Oliveira et al., 2012).

(Long-term perspective) Open data. The publication of open data has become a growing phenomenon, which can be partly explained by regulatory constraints that require from companies and institutions the publication of some of their data. In the context of open data, suppliers provide (open) data that is made available to anyone (without charge). If we focus on some French examples, we can cite the data made available through the French interministerial portal data.gouv.fr[6] managed by the Etalab team[7], and the data published by large companies in the transport sector like the SNCF railway company[8], or electric utility large companies companies like EDF and RTE [9]. Moreover, the development and the standardisation of web semantic technologies (for instance the RDF model and associated tools) encourage the publication of real huge datasets that make possible to experiment the new technologies. For instance, the *Linked Open Data (LOD) cloud* is composed of more than a thousand datasets[10].

The resulting published data sources offer incredible opportunities for the conception of novel applications and tools. But at the same time, the quality of the data provided by these sources can be poor, making their exploitation and usage difficult, sometimes even risky. Quality

[6] `http://www.data.gouv.fr`
[7] `http://www.etalab.gouv.fr`
[8] `https://data.sncf.com`
[9] `https://opendata.rte-france.com`
[10] `http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state`

issues have to be dealt with in this context that raises some new challenges, for the providers and for the users.

First, for a provider, *choosing which data (and meta-data) to publish* is a tricky issue (Barthélémy et al., 2015; Barrau et al., 2016). On the one hand, the provider has to publish as much data as possible in order to generate value (Chignard and Benyayer, 2015), for instance social or innovative. But on the other hand, publishing too much data could lead to lose some competitive advantages over the competitors, or to create privacy breaches (the published data could be cross-matched with data published by someone else, leading e.g. to confidentiality breaches). Forestalling such negative effects is difficult. This is part of the quality dimension referred to as *dissemination control* (Scannapieco and Berti, 2016), and is also related to the problem of evaluating the *costs of quality*.

For the user of open data, a difficult problem is the one of *identifying and choosing relevant data* amongst the available ones, which can belong to the *Deep Web*[11]. There is a need of methodologies that allow identifying available sources and the (combination of the) ones that fit some quality requirements. Some initiatives intend to identify available datasets, in order to create a catalog of available published data (we can mention the *European Data Portal*[12], which proposes a catalog of published European Public Sector Information). This is a first step toward the identification of available data sources. The next step is the recommendation of relevant data to users, according to their requirements in terms of data content and data quality.

Scannapieco and Berti (2016) reviewed some open scientific problems that are specific to the quality management of Web data. Amongst them, the authors identified the problem of the assessment the data *trustworthiness*, which is a quality dimension of great interest for open data as, roughly speaking, "anyone can publish anything". Another correlated important quality dimension is the data *provenance*, which is also a quality dimension of great interest for open data, as trustworthiness may rely on the availability of provenance information. Among other open issues, they also mentioned the problem of *object identification*, which consists in being able to decide if two pieces of data, stemming from different sources, refer to the same real world entity. Another quality issue is dealing with *data inconsistency* that is an inevitable concern when combining multi-source data produced by individuals and institutions with highly variable business fields, culture, motivations and skills. Other problems mentioned by Scannapieco and Berti (2016) are the one of dealing with *evolution and versioning* of data (quality), which becomes even more complex when dealing with the inherent *volatility* of open data (that may have a high temporal variability, for instance for data like stock options or product prices).

[11] The deep web is the part of web data that cannot be reached by traditional search engines, which only index static pages. Deep Web sources store their content in searchable databases that only produce results dynamically in response to direct requests. Such data cannot be indexed by traditional search engines (BrightPlanet, 2001).

[12] `https://www.europeandataportal.eu`

Finally, some research problems, mentioned hereafter, are still largely open, meaning that the current research does not offer real guides to solve them.

(Long-term perspective) Interdependencies of quality metrics. The problem of *identifying the interdependencies* of quality metrics (see discussion in Section 2.3) is still an open problem.

(Long-term perspective) Automatic computation and capitalisation. Associated with any quality management process, the *automatic computation* of the quality metrics is also always a challenge. This automatisation does not only concerns the automatic computation of the quality metrics but also their *capitalisation*, their *visualisation* and their semi-automatic *analysis* that may lead to the *recommendation* of quality improvement actions.

BIBLIOGRAPHY

Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, 16:3–9.

Acosta, M., Zaveri, A., Simperl, E., Kontokostas, D., Auer, S., and Lehmann, J. (2013). Crowdsourcing linked data quality assessment. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 260–276.

Akoka, J., Berti-Equille, L., Boucelma, O., Bouzeghoub, M., Comyn-Wattiau, I., Cosquer, M., Goasdoué-Thion, V., Kedad, Z., Nugier, S., Peralta, V., and Si-Said-Cherfi, S. (2007). A Framework for quality evaluation in data integration systems. In *Proceedings of the International Conference on Entreprise Information Systems (ICEIS)*, page 10.

Alaranta, M. and Jarvenpaa, S. L. (2010). Changing it providers in public sector outsourcing: Managing the loss of experiential knowledge. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*, pages 1–10.

Alavi, M. and Leidner, D. E. (2001). Review: Knowledge management and knowledge management systems: Conceptual foundations and research issue. *MIS Quarterly*, 25(1).

Alter, S. (1999). *Information Systems: A Management Perspective*. Addison Wesley, 3 edition.

Angles, R. (2012). A comparison of current graph database models. In *Proceedings of EEE International Conference on Data Engineering (ICDE) Workshops*, pages 171–177.

Angles, R. and Gutierrez, C. (2008). Survey of graph database models. *ACM Computing Surveys (CSUR)*, 40(1):1–39.

Arduin, P.-E., Grundstein, M., and Rosenthal-Sabroux, C. (2013). From knowledge sharing to collaborative decision making. *International Journal of Information and Decision Sciences*, 5(3).

Argote, L. and Ingram, P. (2000). Knowledge transfer: A basis for competitive advantage in firms. *Organizational Behavior and Human Decision Processes*, 82(1):150–169.

Avgerou, C. (2001). The significance of context in information systems and organizational change. *Information Systems Journal*, 11(1):43–64.

Barceló, P., Libkin, L., and Reutter, J. L. (2014). Querying regular graph patterns. *Journal of the ACM*, 61(1):8:1–8:54.

Barrau, D., Barthélémy, N., Kedad, Z., Laboisse, B., Nugier, S., and Thion, V. (2016). Gestion de la qualité des données ouvertes liées - État des lieux et perspectives. *Revue des Nouvelles Technologies de l'Information*.

Barthélémy, N., Rossin, E., Aldebert, G., Barrau, D., Rougier, G., and Granovsky, F. (2015). Ouverture des données à l'externe : apprentissage. Compte-rendu du groupe de travail ExQI Open Data.

Basili, V. R., Caldiera, G., and Rombach, H. D. (1994). The Goal Question Metric Approach. In *Encyclopedia of Software Engineering*. Wiley.

Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys (CSUR)*, 41(3):16:1–16:52.

Batini, C. and Scannapieco, M. (2016). *Data and Information Quality: Dimensions, Principles and Techniques*. Springer.

Becker, J., Rosemann, M., and Uthmann, C. v. (2000). Guidelines of business process modeling. In *Business Process Management, Models, Techniques, and Empirical Studies*, pages 30–49. Springer-Verlag.

Berti-Équille, L., Akoka, J., Boucelma, O., Bouzeghoub, M., Comyn-Wattiau, I., Cosquer, C., Guisnel, F., Kedad, Z., Nugier, N., Peralta, V., Sisaid-Cherfi, S., and Thion-Goasdoué, V. (2006). QUADRIS : Qualité des données dans les systèmes d'informations multi-sources. Colloque Panorama des Recherches Incitatives en STIC (PaRISTIC).

Berti-Equille, L., Comyn-Wattiau, I., Kedad, Z., Peralta, V., Cosquer, M., Nugier, S., Si-Said-Cherfi, S., and Thion-Goasdoué, V. (2011). Assessment and Analysis of Information Quality: a Multidimensional Model and Case Studies. *Journal International Journal of Information Quality*, 2(4):300–323.

Besson, V., Fiala, D., Rigaux, P., and Thion, V. (2018). Gioqoso, an Online Quality Evaluation Tool for MEI Scores. Music encoding conference 2018 (MEC 2018). Poster.

Besson, V., Gurrieri, M., Rigaux, P., Tacaille, A., and Thion, V. (2016). A Methodology for Quality Assessment in Collaborative Score Libraries. In *Proceedings of the International Society for Music Information Retrieval Conference*.

Beulen, E., Tiwari, V., and van Heck, E. (2011). Understanding transition performance during offshore it outsourcing. *Strategic Outsourcing: An International Journal*, 4(3):204–227.

Bosc, P. and Pivert, O. (1995). SQLf: a relational database language for fuzzy querying. *IEEE Transactions on Fuzzy Systems*, 3:1–17.

Bozzon, A., Brambilla, M., Ceri, S., Mauri, A., and Volonterio, R. (2014). Pattern-based specification of crowdsourcing applications. In *Proceedings of the International Conference on Web Engineering (ICWE)*, pages 218–235. Springer International Publishing.

BrightPlanet (2001). The deep web: Surfacing hidden value. *The Journal of Electronic Publishing*, 7(1).

Castelltort, A., Laurent, A., Pivert, O., Slama, O., and Thion, V. (2018). *Trends and Challenges Related to NoSQL Data Models*, chapter Fuzzy Preference Queries to NoSQL Graph Databases. Hermes Science Publications.

Chignard, S. and Benyayer, L.-D. (2015). *Datanomics – Les nouveaux business models des données*. FYP Éditions.

Cohen, W. M. and Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly*, 35(1):128–152.

Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., and Allahbakhsh, M. (2018). Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, 51(1):7:1–7:40.

Davenport, T. H. and Prusak, L. (1998). *Working Knowledge: How Organizations Manage What They Know*. Harvard Business Review Press.

Davenport, T. H. and Short, J. E. (1990). The new industrial engineering: Information technology and business process redesign. *Sloan Management review*, 31(4):11–27.

De Feo, J., Barnard, W., and Juran Institute (2005). *Institute's Six Sigma Breakthrough and Beyond - Quality Performance Breakthrough Methods*. McGraw-Hill Professional.

Degenne, A. and Forse, M. (1999). *Introducing Social Networks*. SAGE Publications.

Delone, W. H. and McLean, E. R. (1992). Information systems success: The quest for the dependent variable. *Information Systems Research*, 3(1):60–95.

Delone, W. H. and 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.

Deming, W. E. (2000). *Out of the Crisis*. The MIT Press.

Dibbern, J., Goles, T., Hirschheim, R., and Jayatilaka, B. (2004). Information systems outsourcing: A survey and analysis of the literature. *SIGMIS Database*, 35(4):6–102.

Dubois, D. and Prade, H. (1997). Using fuzzy sets in flexible querying: Why and how? In *Flexible Query Answering Systems*, pages 45–60. Springer.

Duquennoy, D., Laboisse, B., Nugier, S., and Thion, V. (2007). An Evaluation Framework for Data Quality Tools. In *Proceedings of the International Conference on Information Quality*.

English, L. P. (1999). *Improving data warehouse and business information quality: methods for reducing costs and increasing profits*. John Wiley & Sons, Inc.

Eppler, M. J. and Helfert, M. (2004). A framework for the classification of data quality costs and an analysis of their progression. In *Proceedings of the International Conference on Information Quality (ICIQ)*.

Erden, Z., von Krogh, G., and Nonaka, I. (2008). The quality of group tacit knowledge. *Journal of Strategic Information Systems*, 17(1):4–18.

Fiala, D., Rigaux, P., Tacaille, A., Thion, V., and The GioQoso members (2018). Data quality rules for digital score libraries. Technical report, HAL.

Foscarin, F., Fiala, D., Jacquemart, F., Rigaux, P., and Thion, V. (2018). GioQoso, an On-line Quality Assessment Tool for Music Notation. In *Proceedings of the International Conference on Technologies for Music Notation and Representation (TENOR)*.

Gallagher, B. (2006). Matching structure and semantics: A survey on graph-based pattern matching. In *AAAI Fall Symposium*, pages 45–53.

Grim-Yefsah, M. (2012). *Gestion des connaissances et externalisation informatique. Apports managériaux et techniques pour l'amélioration du processus de transition : Cas de l'externalisation informatique dans un EPST*. PhD thesis, Univ. Paris Dauphine.

Grim-Yefsah, M., Rosenthal-Sabroux, C., Si-Said Cherfi, S., and Thion, V. (2016). Evaluation and Improvement of a Transition Business Process: A Case Study guided by a Semantic Quality-based Approach. *Information Systems Management*, 33(1):74–87.

Grim-Yefsah, M., Rosenthal-Sabroux, C., and Thion, V. (2010a). Évaluation de la qualité d'un processus métier à l'aide d'informations issues de réseaux informels. *Revue des sciences et technonlogies de l'information (RSTI)*, 15:66–83.

Grim-Yefsah, M., Rosenthal-Sabroux, C., and Thion, V. (2010b). Un premier pas vers l'utilisation d'une analyse structurelle de réseau social pour évaluer la qualité d'un processus métier. In *Proceedings of the conférence InFormatique des ORganisations et Systèmes d'Information et de Décision (INFORSID)*.

Grim-Yefsah, M., Rosenthal-Sabroux, C., and Thion, V. (2011a). Changing Provider in an Outsourced Information System Project: Good Practices for Knowledge Transfer. In *Proceedings of the International Conference on Knowledge Management and Information Sharing*, pages 318–321.

Grim-Yefsah, M., Rosenthal-Sabroux, C., and Thion, V. (2011b). Using information of an Informal Network to Evaluate Business Process Robustness. In *Proceedings of the International Conference on Knowledge Management and Information Sharing*, pages 430–435.

Grundstien, M. (2009). Distinguishing knowledge from information. A prerequisite for elaborating km initiative strategy. In *Proceedings of the International Conference on Knowledge Management and Information Sharing*.

Howe, J. (2006). The rise of crowdsourcing. *Wired magazine*, 14(6).

Huang, K.-T., W. Lee, Y., and Wang, R. Y. (1999). *Quality Information and Knowledge Management*. Prentice Hall.

Huber, M. W., Piercy, C. A., and McKeown, P. G. (2006). *Information Systems: Creating Business Value*. Wiley.

International Organization for Standardization (ISO) (2000). ISO 9000:2000 – Quality management systems – Fundamentals and vocabulary.

International Organization for Standardization (ISO) (2008). ISO 9001:2008 – Quality management systems – Requirements.

Jessup, L. M. and Valacich, J. S. (2008). *Information Systems Today: Managing in the Digital World*. Pearson Prentice Hall, 3 edition.

Kießling, W. and Köstler, G. (2002). Preference SQL – design, implementation, experiences. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 990–1001.

Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., and Zaveri, A. (2014). Test-driven evaluation of linked data quality. In *Proceedings of the International World Wide Web Conferences (WWW)*, pages 747–758.

Lacity, M. C. and Hirschheim, R. (1993). The information systems outsourcing bandwagon. *Sloan Management Review*, 35(1):73–86.

Laue, R. and Gruhn, V. (2006). Complexity metrics for business process models. In *Proceedings of the International Conference on Business Information Systems (BIS)*, pages 1–12.

Lee, R. G. and Dale, B. G. (1998). Business process management: a review and evaluation. *Business Process Management Journal*, 4(3):214–225.

Lee, Y. W., Pipino, L., Funk, J. D., and Wang, R. Y. (2006). *Journey to Data Quality*. MIT Press.

Lohrmann, M. and Reichert, M. (2013). *Understanding Business Process Quality*, pages 41–73. Springer Berlin Heidelberg.

Loshin, D. (2001). *Enterprise Knowledge Management: The Data Quality Approach*. Morgan Kaufmann.

Loshin, D. (2010). *The Practitioner's Guide to Data Quality Improvement*. Morgan Kaufmann.

Ludäscher, B., Weske, M., McPhillips, T. M., and Bowers, S. (2009). Scientific workflows: Business as usual? In *Proceedings of the International Conference (BPM)*, pages 31–47.

Madnick, S. E., Wang, R. Y., Lee, Y. W., and Zhu, H. (2009). Overview and framework for data and information quality research. *Journal of Data and Information Quality (JDIQ)*, 1(1):2:1–2:22.

Marcal de Oliveira, K., Thion, V., Dupuy-Chessa, S., Gervais, M.-P., Si-Said Cherfi, S., and Kolski, C. (2012). Limites de l'évaluation d'un système d'information : une analyse fondée sur l'expérience pratique. In *Proceedings of the conférence InFormatique des ORganisations et Systèmes d'Information et de Décision (INFORSID)*, pages 395–410.

Mason, R. O. and Mitroff, I. I. (1973). A program for research on management information systems. *Management Science*, 19(5):475–487.

Mendling, J., Reijers, H. A., and Recker, J. (2010). On the usage of labels and icons in business process modeling. *International Journal of Information System Modeling and Design*, 1(2):40–58.

Monteiro, L. F. S. and de Oliveira, K. M. (2011). Defining a catalog of indicators to support process performance analysis. *Journal of Software Maintenance and Evolution-Research and Practice*, 23(6):395–422.

Morley, C., Bia-Figueiredo, M., and Gillette, Y. (2011). *Processus Métiers et S.I. – Gouvernance, management, modlisation*. Dunod, 2 edition.

Morley, C., Hugues, J., Leblanc, B., and Hugues, O. (2004). *Processus Métiers et S.I. – Évaluation, modélisation, mise en oeuvre*. Dunod, 2 edition.

Neo Technology (2013). The Neo4j Manual v2.0.0.

Neo4j (Accessed in 2019). Neo4j web site. `www.neo4j.org`.

Nonaka, I. (1991). The knowledge-creating company. *Harvard Business Review*, 69(6):96–104.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. organization science. *Organization Science*, 5(1):14–37.

Nonaka, I. and Konno, N. (1998). The concept of "ba": Building a foundation for knowledge creation. *California Management Review*, 40(3):40–55.

Nonaka, I. and Takeuchi, H. (1995). *The Knowledge-Creating Company – How Japanese Companies Create the Dynamics of Innovation*. Oxford University Press.

Oakland, J. (1989). *Total Quality Management*. Springer.

Object Management Group (OMG) (2013). Business process model and notation (bpmn) – version 2.0.2. `http://www.omg.org/spec/BPMN`.

Object Management Group (OMG) (2017). Omg unified modeling language (omg uml) – version 2.5.1. `https://www.omg.org/spec/UML`.

Olzmann, M. and Wynn, M. G. (2012). How to switch it service providers: Recommendations for a successful transition. *International Journal on Advances in Intelligent Systems*, 5(1 & 2):209219.

Palmberg, K. (2009). Exploring process management: are there any widespread models and definitions? *TQM Journal*, 21(2):203–215.

Paul, R. J. (2007). Challenges to information systems: time to change. *European Journal of Information Systems*, 16(3):193–195.

Pawlak, Z. (2002). Rough sets, decision algorithms and bayes' theorem. *European Journal of Operational Research*, 136(1):181–189.

Peralta, V., Thion, V., Kedad, Z., Berti-Équille, L., Comyn-Wattiau, I., Nugier, S., and Sisaid-Cherfi, S. (2009). Multidimensional Management and Analysis of Quality Measures for CRM Applications in an Electricity Company. In *Proceedings of the International Conference on Information Quality (ICIQ)*.

Persse, J. R. (2006). *Process Improvement Essentials – CMMI, Six Sigma, and ISO 9001*. O'Reilly Media.

Pivert, O. and Bosc, P. (2012). *Fuzzy Preference Queries to Relational Databases*. Imperial College Press.

Pivert, O., Slama, O., Smits, G., and Thion, V. (2016a). A Fuzzy Extension of SPARQL for Querying Gradual RDF Data. In *Proceedings of the IEEE International Conference on Research Challenges in Information Science (RCIS) Posters*.

Pivert, O., Slama, O., Smits, G., and Thion, V. (2016b). SUGAR: A Graph Database Fuzzy Querying System. In *Proceedings of the IEEE International Conference on Research Challenges in Information Science (RCIS) Demos*.

Pivert, O., Slama, O., and Thion, V. (2016c). An Extension of SPARQL with Fuzzy Navigational Capabilities for Querying Fuzzy RDF Data. In *Proceedings of the IEEE International Conference on Fuzzy Systems (Fuzz-IEEE)*.

Pivert, O., Slama, O., and Thion, V. (2016d). SPARQL Extensions with Preferences: a Survey. In *ACM Symposium on Applied Computing*.

Pivert, O., Smits, G., and Thion, V. (2015). Expression and Efficient Processing of Fuzzy Queries in a Graph Database Context. In *Proceedings of the IEEE International Conference on Fuzzy Systems (Fuzz-IEEE)*, page 8.

Pivert, O., Thion, V., Jaudoin, H., and Smits, G. (2014). On a Fuzzy Algebra for Querying Graph Databases. In *Proceedings of the IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 748–755.

Polanyi, M. (1967). *The tacit dimension*. Garden City, N.Y. : Anchor Books.

Polanyi, M. (1974). *Personal Knowledge: Towards a Post-Critical Philosophy*. University Of Chicago Press.

Porter, M. E. and Millar, V. E. (1985). How information gives you competitive advantage. *Harvard Business Review*, 64(4):149–160.

RabbitHole project (Accessed in 2019). RabbitHole. `www.neo4j.com/blog/rabbithole-the-neo4j-repl-console/`.

Radulovic, F., Mihindukulasooriya, N., García-Castro, R., and Gómez-Pérez, A. (2018). A comprehensive quality model for linked data. *Semantic Web*, 9(1):3–24.

Redman, T. C. (1996). *Data Quality for the Information Age*. Artech House, Inc., Norwood, MA, USA, 1st edition.

Reix, R. and Rowe, F. (2002). *Faire de la recherche en systèmes d'information*, chapter La recherche en systèmes d'information : de l'histoire au concept. Vuibert.

Rigaux, P., Abrouk, L., Audéon, H., Cullot, N., Davy-Rigaux, C., Faget, Z., Gavinet, E., Gross-Amblard, D., Tacaille, A., and Thion, V. (2012). The Design and Implementation of NEUMA, a Collaborative Digital Score Library - Requirements, architecture, and models. *International Journal On Digital Libraries (IJODL)*, pages 1–24.

Rigaux, P. and Thion, V. (2017). Quality Awareness over Graph Pattern Queries. In *Proceedings of the International Database Engineering & Applications Symposium (IDEAS)*.

Rosenthal-Sabroux, C. and Grundstein, M. (2007). A knowledge management approach of ict. *Journal of Science*, 24(2):162–169.

Scannapieco, M. and Berti, L. (2016). *Data and Information Quality: Dimensions, Principles and Techniques*, chapter Quality of Web Data and Quality of Big Data: Open Problems, pages 421–449. Springer.

Shewhart, W. A. (1980). *Economic Control of Quality of Manufactured Product*. American Society for Quali-ty Control.

Si-Said Cherfi, S., Guillotel-Nothmann, C., Hamdi, F., Rigaux, P., and Travers, N. (2017a). Ontology-Based Annotation of Music Scores. In *Proceedings of the International Conference on Knowledge Capture (K-CAP)*.

Si-Said Cherfi, S., Hamdi, F., Rigaux, P., Thion, V., and Travers, N. (2017b). Formalizing Quality Rules on Music Notation: an Ontology-based Approach. In *Proceedings of the International Conference on Technologies for Music Notation and Representation (TENOR)*.

Si-Said Cherfi, S. and Thion, V. (2012). *La qualité et la gouvernance des données au service de la performance des entreprises*, chapter Évaluation de la qualité d'un processus métier - Enjeux, cas d'étude et bonnes pratiques, pages 215–240. Hermes Science Publications.

Tsuchiya, S. (1993). Improving knowledge creation ability through organizational learning. In *Proceedings of the International Symposium on the Management of Industrial and Corporate Knowledge*, pages 87–95.

Van Der Aalst, W. M. P., Ter Hofstede, A. H. M., Kiepuszewski, B., and Barros, A. P. (2003). Workflow patterns. *Distributed and Parallel Databases*, 14(1):5–51.

Vanderfeesten, I., Cardoso, J., Reijers, H. A., and van der Aalst, W. (2007). *BPM and Workflow Handbook 2007 - Methods, Concepts, Case Studies and Standards in Business Process Management and Workflow*, chapter Quality Metrics for Business Process Models, pages 179–190. Future Strategies Inc.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.

Weske, M. (2012). *Business Process Management - Concepts, Languages, Architectures, 2nd Edition*. Springer.

Willcocks, L. P. and Kern, T. (1998). IT outsourcing as strategic partnering: the case of the UK Inland Revenue. *European Journal of Information Systems*, 7(1):29–45.

Wood, P. T. (2012). Query languages for graph databases. *SIGMOD Record*, 41(1):50–60.

Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, 8(3):338–353.

Zaveri, A., Kontokostas, D., Sherif, M. A., Bühmann, L., Morsey, M., Auer, S., and Lehmann, J. (2013). User-driven quality evaluation of DBpedia. In *Proceedings of the International Conference on Semantic Systems (I-SEMANTICS)*, pages 97–104.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., and Auer, S. (2016). Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93.