Association for Information Systems
AIS Electronic Library (AISeL)

ECIS 2013 Completed Research

ECIS 2013 Proceedings

7-1-2013

An Approach To Design Business Processes Addressing Data Quality Issues

Cinzia Cappiello
Politecnico di Milano, Milano, Milano, Italy, cappiell@elet.polimi.it

Angelica Caro
Universtity of Bio Bio, ChillÃ�n, Chile, mcaro@ubiobio.cl

Alfonso Rodriguez
Universtity of Bio Bio, ChillÃ�n, Chile, alfonso@ubiobio.cl

Ismael Caballero
University of Castilla-La Mancha, Ciudad Real, Castilla-La Mancha, Spain, Ismael.Caballero@uclm.es

Recommended Citation

Cappiello, Cinzia; Caro, Angelica; Rodriguez, Alfonso; and Caballero, Ismael, "An Approach To Design Business Processes Addressing Data Quality Issues" (2013). *ECIS 2013 Completed Research*. 216.
http://aisel.aisnet.org/ecis2013_cr/216

AN APPROACH TO DESIGN BUSINESS PROCESSES ADDRESSING DATA QUALITY ISSUES

Cinzia Cappiello, Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria, Piazza Leonardo da Vinci 32, 20133 Milano, Italy, cinzia.cappiello@polimi.it

Angélica Caro, Department of Computer Science and Information Technologies, University of Bio Bio, Andrés Bello s/n, Chillán, Chile, mcaro@ubiobio.cl

Alfonso Rodríguez, Department of Computer Science and Information Technologies, University of Bio Bio, Andrés Bello s/n, Chillán, Chile, alfonso@ubiobio.cl

Ismael Caballero, Alarcos Research Group-Instituto de Tecnologías y Sistemas de la Información, University of Castilla-La Mancha, Paseo de la Universidad 4, Ciudad Real, Spain, ismael.caballero@uclm.es

Abstract

Guaranteeing high data quality levels is an important issue especially in information-intensive organizations. In fact, the core business of such organizations is based on the use of information for either providing personalized services or understanding and better satisfying customers' requirements. However, poor data quality has negative impacts on almost all the enterprises. Indeed, it often implies customer dissatisfaction, increased operational cost, less effective decision-making, and a reduced ability to make and execute organizational strategies. Improving data quality often requires modifying business processes enriching them with additional activities. Such activities change on the basis of the data quality dimensions to improve. In this paper, we present a methodology to support process designers in the selection of the improvement actions to adopt in the design of business processes in order to satisfy the data quality requirements.

Keywords: Data Quality, Business Process Modeling, BPMN, Data Quality Improvement.

1 Introduction

Data Quality (DQ) is often defined as "*fitness for use*", i.e., the ability of a data collection to meet users' requirements (Wang and Strong, 1996). DQ is a multidimensional and subjective concept since it is usually evaluated by means of different criteria or data quality dimensions and the selection and assessment of the DQ dimensions that better describe users' data quality requirements mainly depend on the context of use (Strong, Lee et al. 1997; Batini and Scannapieco 2006). For this reason, in the literature, there is no general agreement on the identification of the most important data quality dimensions. Anyway, it is possible to distinguish a small set of DQ dimensions that are considered relevant in most of the studies; such set includes the following dimensions: accuracy, completeness, timeliness and consistency (Batini and Scannapieco 2006).

Guaranteeing the highest possible levels of DQ is an important issue especially in information-intensive organizations. In fact, poor data quality exposes organizations to non-depreciable risks (el Abed 2009). It negatively affects organizations activities at all levels: it might reduce the efficiency of the business processes and the effectiveness of the decisions if they are based on unreliable data (Kokemüller, 2011). These problems can be avoided or at least alleviated by adopting suitable improvement actions after the corresponding data quality assessment (Maydanchik 2007). Such actions are diverse since there could be various mechanisms suitable for improving each one of the different identified DQ dimensions and their relationships. Anyway, it is necessary to consider that the adoption of improvement actions might also have other impacts on the organizations' business processes and implies several costs. In fact, sometimes, improvement tasks mainly require the addition of new process activities, the modification of the process flow or even the acquisition of new software products.

Business Process (BP) modeling is a technique, often supported by some graphical notation, which is used by process practitioners to capture, organize and communicate information about business processes (Harmon and Wolf 2011). Enterprises create business process models with the purpose of obtaining a simplified view of reality (Eriksson and Pemkel 2001). This realistic description of a business process allows us to understand and eventually modify a business process with the aim of incorporating improvements into it. Notations for business process description, for instance BPMN, are becoming increasingly more important owing to the fact that the success of modeling is based both on the ability to express the different needs of the business and on the availability of a notation in which these needs can be described (Weske 2007). Several unsatisfactory business process performance could be avoided if the problems related to BP are early detected and possible solutions are included as part of the business process description. Particular case of such problems is related to the data quality. In this context, our research aims to develop a comprehensive methodology to support the process designers when modeling DQ-aware business process as presented in (Caro, Rodriguez et al. 2012; Rodríguez, Caro et al. 2012). With respect to previous work, this paper provides details about the actions that the process designer has to perform in order to detect the most suitable data quality improvement activities and consequently change the business process model.

The paper is organized as follows. Section 2 discusses the related work. Section 3 provides a view on the DQ concepts that are considered relevant for this paper. The complete methodology for the design of DQ-aware business processes is presented in Section 4 while Section 5 provides details about the phases related to the definition of DQ requirements and the selection of related improvement activities.

2 Related work

This section describes some of the existing works related to the modeling of DQ management concerns in business processes. In this sense, DQ management is a relevant aspect that deserves to be considered in order to globally improve the effectiveness of organization's performance (Ballou and Tayi 1996; el Abed 2009). Thus, it is important that business processes are designed in a suitable way in order to satisfy DQ requirements (Dewan, Storey et al. 2012). The most important languages that can be used to model business processes, namely BPMN and UML (Harmon and Wolf, 2012), do not allow process designers to fully specify DQ requirements at a high level.

Although DQ concerns are not new to BP literature, the inclusion of DQ modeling concerns into the design of the BP are relatively new (Ofner, Otto et al. 2012), and only a few contributions highlight the need of addressing DQ issues during the business process modeling at design time. It is important to state that this is precisely the main issue addressed in this paper. As already stated, only some works have already dealt with this issue. For instance, Bagchi et al. in (Bagchi, Bai et al. 2006) introduce a business process modeling framework for quantitative estimation and management of DQ in information systems. Based on this framework, they propose to exploit the structure provided by the business process flows to estimate errors arising in transactional data and the impact of their propagation to the key performance indicators. With respect to these contributions, the approach presented in this paper provides an easier way to express data quality requirements and aims to comprehensively support users in the design of quality-aware business process: from the identification of possible data quality problems to the selection of the most suitable improvement actions.

To identify which DQ dimensions are important for the design of the business process it is important to highlight the contributions provided in (Heravizadeh, Mendling et al. 2009); the work proposed the QoBP framework for capturing the quality dimensions of a process. The framework helps modelers in identifying quality attributes in four quality dimensions: quality of functions, quality of input and output objects, quality of non-human resources and quality of human resources. In particular, they specify eleven DQ attributes for the input and output information objects.

Also, it is important to identify which elements should be included as part of the design of the business process as contributing for the appropriate levels of DQ should guaranteed. In this sense, the Data Excellence Framework, proposed in (el Abed 2009), describes the methodology, processes and roles required to generate the maximum business value while improving business processes using DQ and business rules. In this approach, DQ requirements are specified as business rules. The set of business rules supporting DQ grows over time as part of the process of continuous improvement.

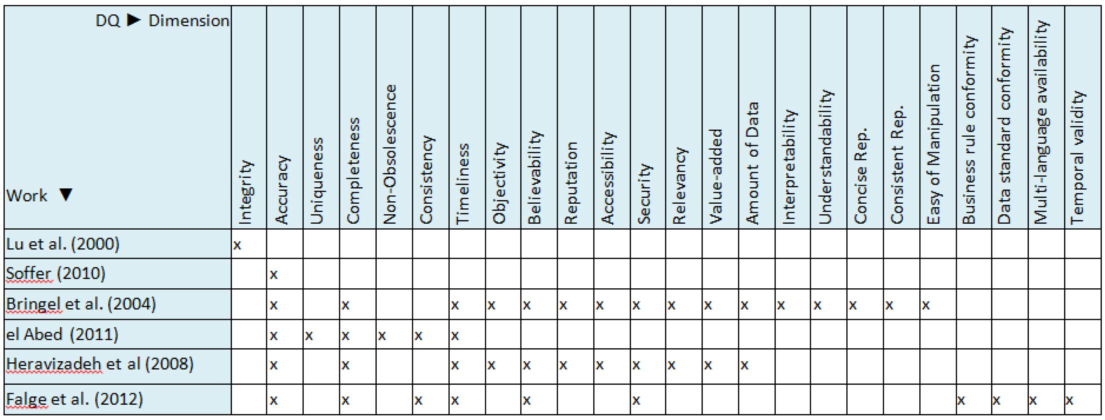
However, many other authors have dealt with data and possible data defects that could appear during data usage in the execution of business processes. In this last sense, the analysis of the relevant literature has been done since two different points of view, which sometimes are dealt jointly: the one related to assure quality of data as meeting requirements, and on the other hand, the one related to the adequate design of business process to guarantee the adequate levels of data quality at execution time.

For example, the work by (Lu, Sadiq et al. 2009) introduces concerns focused on the concept of compliance. Compliance essentially means ensuring that business processes, operations and practices are in accordance with a prescribed and/or agreed set of previously defined norms. Lu et al. consider that a sustainable approach for achieving compliance should fundamentally have a preventive focus, thus achieving compliance by design. Their proposal consists in incorporating compliance issues within business process design methodology to assist process designers. Specifically they propose to model a set of control objectives in the BP that will allow process designers to comparatively assess the compliance degree of their design as well as be better informed on the cost of non-compliance. In this sense, the DQ aspects are to be considered in these control objectives as data integrity checks.

In (Falge, Otto et al. 2012) the authors identified, by mean of a qualitative content analysis on Business Networking case studies, DQ requirements of collaborative BPs in Business Networking. The results showed the combinations of data classes and DQ dimensions that are crucial for the different collaborative BPs in business networks. They used a set of DQ dimensions already defined in the literature, and also defined a newly set of relevant DQ dimensions.

As example of the second point of view, in (Bringel, Caetano et al. 2004) authors propose a business process pattern that can be reused through adaptation in specific organizational scenarios. For this, they define DQ attributes associated with information entities having different meanings that depend on both the business view and the different organizational dimensions.

As an example of approach that bring together both points of view is the proposed by (Soffer 2010), in which the author explores the inaccuracies of data as the situation where the information system does not truly reflect the state of a domain where a process takes place. In the mentioned paper, the potential consequences of data inaccuracy are discussed. The work also provides the bases to support the design of robust processes that avoid problems related to data inaccuracy. Figure 1 summarizes the DQ dimensions considered in the previously cited studies.



DQ ▶ Dimension Work ▼	Integrity	Accuracy	Uniqueness	Completeness	Non-Obsolescence	Consistency	Timeliness	Objectivity	Believability	Reputation	Accessibility	Security	Relevancy	Value-added	Amount of Data	Interpretability	Understandability	Concise Rep.	Consistent Rep.	Easy of Manipulation	Business rule conformity	Data standard conformity	Multi-language availability	Temporal validity
Lu et al. (2000)	x																							
Soffer (2010)		x																						
Bringel et al. (2004)		x		x			x	x	x	x	x	x	x	x	x	x	x	x	x	x				
el Abed (2011)		x	x	x	x	x	x																	
Heravizadeh et al (2008)		x		x			x	x	x	x	x	x	x	x	x									
Falge et al. (2012)		x		x		x	x		x			x									x	x	x	x

Figure 1. Data Quality Attributes identified in BP modelling

Most of the analysed contributions provide interesting results on guaranteeing appropriate levels of data quality levels for one or multiple data quality dimensions during BP modeling but they scarcely considered the impact of the improvement actions on the BP. None of the analysed contributions provide a systematic approach to support the selection of the improvement actions. In order to fulfil this lack, the purpose of this paper is to address this issue sketching the aspects that are relevant to consider in the definition of the methods to adopt for increasing the data quality level.

3 The methodology to design quality-aware business processes

As the main contribution of our research, we have developed a methodology that allows business people to highlight problems associated with data-related element on BP models. In particular, the input process of a methodology has to be described via BPMN notation. This choice has been driven by the fact that among all possible choices, whether languages or notations, for business process modeling, a recent study shows that BPMN (Business Process Model and Notation) is one of the most important and popular standard to modeling business process (Harmon and Wolf 2011). Also, BPMN provides an extensibility mechanism that allows extending standard BPMN elements with additional attributes in a simple way. The following sections provide a description of the concepts that characterized our approach and an overview of the methodology. A more complete description of the phases and of the BPMN artifacts used in our approach can be found in (Caro et al., 2012).

3.1 Data Quality Concepts

Organizations start to understand that high DQ is crucial for the success of their business. Most of the current initiatives focus on the data used for transactional/operational purposes ("running the business"), and analytical purposes ("improving the business"). *Data quality requirements* are usually expressed by defining data quality levels for the different considered data quality dimensions. An organization is characterized by poor data quality if the assessment of DQ dimensions reveals values that are below the required level (Strong et al., 1997b). The specification of data quality requirements is not a trivial issue because not all the people in the organization have the suitable knowledge to identify the DQ dimensions that cause DQ problems that they experience and to specify the needed quality levels. For this reason, our proposal aims to support people in the definition of the data quality requirements starting from the perceived problems in the considered business context. In fact, business process performance can be affected by several problems related with organizational human resources, lack of money, data problems, among others. Business people are aware of these problems: they periodically monitor suitable indicators in order to be aware of the current situation and find strategies to solve current issues and to prevent future ones (Wetzstein et al., 2009). Some example of problems related to data at business process level are lack of security in some sensible data, delays due to data unavailability, failures due to inaccurate or incomplete data, and so on.

In particular, we ask the business people/analysts to consider, at design time, the possible data-related problems that can affect the BP performance and that could imply the failure of it. They can express these possible data quality problems by means of a special mark called *DQ Flag*, which represents a signal in a specific point in the BP model (and on a BPMN data-related element on it) where some data problems may appear and have to be controlled. A DQ Flag is represented in a symbolic language, to be consistent with the simplicity of BPMN (for details refer to Rodriguez et al. 2012). DQ Flag can be defined as a *high level data quality requirement* that should be analyzed and refined to obtain the corresponding *low level DQ requirements*. Low level DQ requirements are described by defining the set of DQ dimensions that can be used to represent the relevant quality aspects of the data involved in the data-related element associated with the DQ Flag. For such dimensions, appropriate quality levels should be guaranteed. For the maintenance or the improvement of DQ levels in the analyzed BP, some *DQ improvement activities* must be considered. Such activities can include activities to insert in the BP flow (e.g., detect and correct activities) or off-line activities that aim to adjust the data contained in the databases without affecting the BP (e.g., data bashing). Such DQ improvement activities are then implemented by *actions, tasks or mechanisms*. In fact, each activity could be executed by using different actions. For example, for guaranteeing completeness, the BP could be modified by including activities that retrieve the missing information by using another source or including procedures that check the data entry activities and avoid the presence of null values. Figure 2 shows graphically these concepts and their relationships.

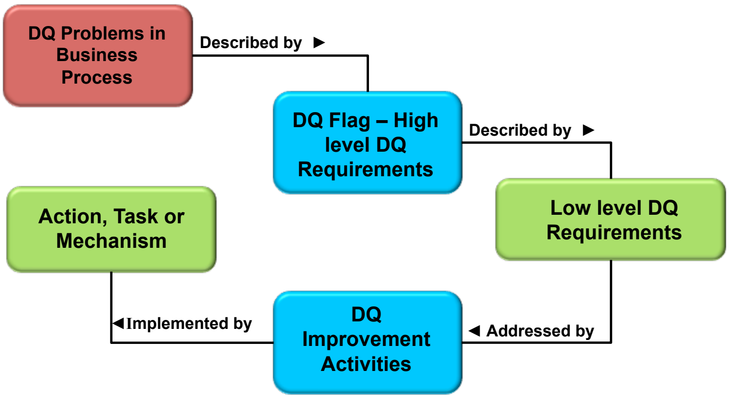

DQ Problems in Business Process
Described by ►
DQ Flag – High level DQ Requirements
Described by ►
Low level DQ Requirements
◄ Addressed by
DQ Improvement Activities
◄Implemented by
Action, Task or Mechanism

Figure 2. Relationships among the data quality concepts that characterize our approach

3.2 Designing quality-aware business processes

Formally, in order to provide a systematic approach for designing data DQ-aware business processes we defined a methodology that starts from the analysis of a process modelled by using BPMN and guides process designers to produce an version of the same process model including enhancements (DQ activities) that guarantee the satisfaction of data quality requirements. The methodology is composed of four steps (see Figure 3), and it follows a top-down approach for the specification of data quality requirements.

The first step, *Data Quality Aware Business Process Modeling*, is devoted to capture *high level DQ requirements* by using the DQ Flags. Such DQ Flags can be associated with data-related BPMN elements for which data quality should be guaranteed for the success of BP. At this stage Business People/Analysts (i.e., people that know the process flow and performance) will also provide documentation about the estimation of the impact (economical wastes or functional errors) of poor data quality on the BP success. The output of this stage is such documentation together with the BPMN model enriched with a set of DQ Flags and the identification of data elements involved in each DQ Flag specification.

The second step, *Data Quality Requirements Specification*, is dedicated to specify the *low level DQ requirements* and involves Business Analysts/Designers and DQ Experts. This stage is composed of several tasks. The first task aims to obtain additional information about the process, e.g., metadata about the flow, the performance and the data in the process (see more details in Section 5). Once these metadata have been collected, the next task consists in the specification of the requirements for each DQ Flag and each data element on it. For this, the relevant DQ dimensions to the context of the BP have to be identified together with their corresponding level of importance (“Low”, “Medium”, “High”); the dependencies between the DQ dimensions associated with each data element (to decide if any can be eliminated e.g. to be incompatible with other) have to be studied. Finally, the dependencies between the data elements in the same BP branch will be studied to refine the approach (e.g., delete some redundant DQ Flag). As a result of this stage, documentation about the DQ Flags and the specification of low level DQ requirements should be generated.

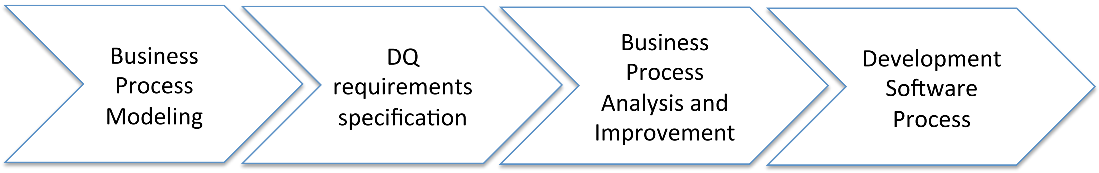

Business Process Modeling
DQ requirements specification
Business Process Analysis and Improvement
Development Software Process

Figure 3. Methodology to model DQ-aware business processes.

The third step, *Business Process Analysis and Improvement*, is devoted to analyse and decide about the most suitable way to accordingly improve the BPMN model. The improvement actions aim to modify the process (e.g., insertion of new activities) in order to minimize the risk due to poor data quality. To decide and select the new activities for DQ improvement workers must consider to use some of the metadata collected at the previous stage such as: a) the level of importance of each DQ Flag for the success of the BP, obtain from data metadata; b) business constraints, obtained from performance metadata; c) probability of use of the data-related element associated with DQ Flag, calculated with the flow metadata; d) DQ Flag Overhead, defined as the implementation costs of the different improvement activities. The workers involved are the DQ Expert and the Business Designer and the output of this stage is an extended BP model with the new activities that guarantee the satisfaction of the DQ requirements, at the BPMN Analytic level.

Finally, the *Development Software Process* step, represents a set of activities that starting from the DQ Requirements specifications will generate artefacts useful to the development process. The idea is that these artefacts represent and deliver the documentation of requirements for the software that will support the business process.

4 From the data quality requirements to the selection of data quality improvement actions

The previous section provided an overview on the entire methodology that we defined for the design of quality-aware business processes. It is worth to note that this methodology is suitable for both the redesign of existing processes and the design of new ones. In the former case, we consider the BPMN model of existing processes and we aim to improve them by addressing data quality issues. In the latter case, data related problems are not yet known but the business people might, anyway highlight the points of the process in which a suitable level of data quality is necessary for guaranteeing business process performance.

In this paper, we mainly focus on the description of the second and third steps of the methodology in order to provide details about the DQ-driven analysis of the BP under study, the DQ requirements specification and also about DQ improvement actions. Figure 4 summarizes the activities to be performed in these steps. Starting from the business process model enriched with the DQ Flags, all the artefacts that can provide additional information about the process and the information exchanged in the process are to be considered. These artefacts include, for example, Data Flow Diagrams and documentation about functional and non-functional requirements. If the process already exists and it is necessary to redesign it in order to meet the specified DQ requirements we can also extract useful knowledge from Process logs (when available). All these sources could be used to enrich the process with additional annotations that allow the process designer to thoroughly analyse the characteristics of the process. Such analysis is conducted in order to evaluate the DQ Flag to consider and to identify the DQ dimensions that have to be associated with an appropriate DQ value. Finally, by exploiting the process annotations and evaluating the low level DQ requirements and the properties of the data quality improvement actions, the process designer will be able to provide, for example, an enriched business process model in which some changes (like new activities that are added or modifications to the process flow) are performed (see for example the activities added in the BP model at the right in Figure 4).

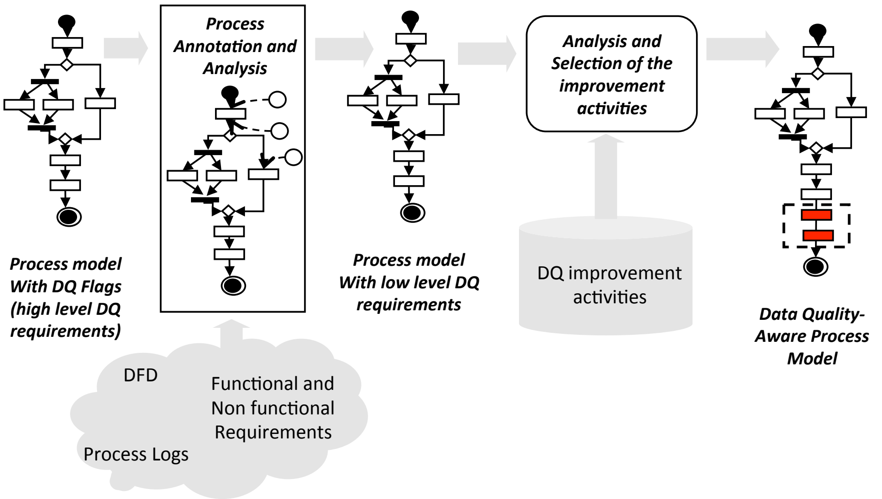


Figure 4. Data Quality-aware design of business processes.

In the next subsection, we provide details about the different steps illustrated in Figure 3. Note that in the description that follows we assume that the process model provides details on the process structure and each activity t_i of the process can be represented as:

$$t_i = \langle d_{ti}, IN_{ti}, OUT_{ti}, r_{ti}, A_{ti} \rangle$$

where: d_{ti} is the activity name; IN_{ti} is the set of activity inputs, and OUT_{ti} is the set of activity outputs); r_{ti} is the *role* (or *actor)* which is responsible of the activity execution; A_{ti} is the set of CRUD actions (Use-Update-Create) performed by the activity on its inputs/outputs. Activity inputs/outputs are expressed as simple attributes or collection of attributes. Moreover, the process structure is enriched with the definition of the process flow in which a generic element f_{ij} is defined as the pair (t_i, t_j) and thus provide information about the link among activities. Finally, the model provides a specification of DQ requirements by means of the set of DQ Flags *DQF*. Note that a DQFlag $dqf_h \in$

DQF can be associated with an activity t_i or a connection f_{ij} when data quality issues affect the communication between two activities.

4.1 Business Process Annotations and Analysis

In order to enable the design of quality-aware business processes, and thus to select the most suitable improvement strategies to apply, it is necessary to add more knowledge to the BPMN process representation. To identify which knowledge has to be added, it is necessary to first analyze the business process. This phase is composed of two sub phases described in the following sections: a) Business process description; b) DQ Flag analysis.

4.1.1 Business process description

The business process annotation decorates the process and its activities with a set of metadata *M* that describe their properties, application data that are relevant to the process, temporal constraints and resources requirements. In this subsection, the metadata used to describe the business process are to be introduced. It is possible to classify the annotation metadata on the basis of the type of information they refer to:

- *Flow metadata*: they provide information regarding the business process control flow, and thus the execution of certain activities in a process. Such metadata are defined at the activity or flow relation level. At the activity level, it is fundamental to describe the activities in terms of its importance in the process. In fact, an activity might be mandatory or optional inside a process. At the flow relation level, a flow relation should be labeled with the execution probability ExP_{ij}. If a flow relation links two task t_i and t_j, ExP_{ij} is equal to 1 while it might be lower than one if the flow relation links a task with a branch. The importance of the activity inside a process is defined by experts while the execution probability is a value that at design time is estimated by experts but at run-time can be easily gathered from process mining applications.
- *Performance constraints*: they refer to performance conditions or constraints within process flows. These metadata can be defined either at the process or at the activity level. In both cases, they store data about temporal conditions (e.g., maximum time that may be needed to respond to a request) and about all the other application KPIs (e.g., availability). Experts define these constraints.
- *Data metadata*: provide information regarding the data used throughout a process. At the activity level the metadata are used to store: the input and output files, the estimated I/O, the data volatility (i.e., permanent or transient information), and the variability in time.

4.1.2 DQ Flag Analysis

The analysis of the DQ Flag aims to identify the exact relevance of the DQ Flag and to validate the position of the DQ Flag in the business process.

The relevance of a DQ Flag $dqf_h \in F$ depends on two main factor: (i) the probability of occurrence of the identified data quality problem and, if available, (ii) the associated costs due poor data quality. The relevance is evaluated in a different ways depending on whether we are redesigning an existing process or we are designing a new one from scratch. In the former case the probability of occurrence will be calculated on the basis of the information available from previous runs. Precisely, it will be calculated as the conditional probability $P_h(A|B)$ where A= *"an error related to the occurs"* and B= *"the activity associated with the data-related element annotated with the DQ flag* dqf_h *is executed"*. Such probability should be then combined with the poor data quality cost associated with a data quality flag dqf_h and the analyzed data quality problem C_{hk} in order to obtain the *data quality risk cost* (RC_{hk}) defined as:

$$RC_{hk} = P_h(A|B) * C_{hk}.$$

If the RC_{hk} is lower than a defined threshold then the DQ Flag is considered irrelevant and a *laissez faire approach* can be adopted. On the other hand if the RC_{hk} is significant (i.e, greater than or equal to the threshold), the analysis of the DQ Flag will continue in order to detect suitable improvement actions.

If we are designing a new process from scratch, then detailed information about the problem occurrence is not available and the process analysts can identify the relevance of the DQ flag just trying to estimate the probability of occurrence and the impact of the problem on the business process.

Once that the DQ Flag relevance has been confirmed, the process analysts is in charge to define if the critical point, in which the root cause of the problem is, matches with the one identified by the DQ Flag. This can be determined by analyzing data dependencies starting from the annotations defined in the previous phase, and in particular by combining the information of the process model together with the information retrieved by the documentation of the system that supports the BP, i.e., DFD (if available). Given two activities t_i and t_j, a *data dependency* holds between them if t_j uses (updates, reads or deletes) at least a business data object that is created or updated by t_j.

Therefore, in this step, the DQ flag should be analyzed in order to identify the set of data that should be associated with the data quality level and check, analyzing the data dependencies, in which activities such data are created or manipulated. Since literature confirms that most of data quality problems rise during such operations and not in reading activities, the process analysts, if necessary, should move the DQ Flag to the crucial points in which the problem might be generated.

In addition, for each activity, it is also necessary to specify the data inputs coming from external data sources, which are used by the activity although they are not derived from previous activities executed within the process. In fact, according to this model, an error in the output data can be consequence of the execution of (one of) the activity(ies) that precede the analysed one, or even can be generated by the analysed activity. This type of error can be classified as self-generated error. In case of a self-generated error occurs, the causes can also be related to the data structure or external processes. In fact, it is necessary to consider that the activity can be influenced not only by the previous activities in the process but also by other external processes that for example might use the same data sources and cause value inconsistencies.

Finally, the DQ Flag analysis also includes a more general process analysis of the process flow. Such analysis aims to detect possibilities of process optimization. In fact, process flow also impacts data quality dimensions. For example, if two sequential activities are independent from each other, they can be executed in parallel by improving the time-related data quality dimensions.

4.2 Definition of low level data quality requirements

The redefined DQ Flags have to be analysed in order to derive the low level requirements, i.e., the data quality dimensions to consider together with their admissible values. Different studies show previous experiences that can help us to build a knowledge base in which it is possible to associate the most common data quality problems with affected data quality dimensions. For example, (i) the fact that goods' deliveries fail usually depends on the "correctness" of the address based on the level of "accuracy" and "completeness" of the data used for the delivery to inaccurate or incomplete data; (ii) the "consistency" dimension can be associated with situations in which different databases containing the same values are used by different activities belonging to the same business process: (iii) a DQ Flag on a message exchanged between different pools in the business process could also require the analysis of "security" related dimensions.

Therefore, the knowledge base to support data quality requirements contains a sort of association rules, such as ar_n: DQ*Problem* $\Rightarrow$ *DQdimension(s)* supported by a *confidence level* C_n that specifies the reliability of the rule calculated on the basis of past observations (or literature empirical studies), This level of confidence can be increased or decreased anytime that the rule is respectively confirmed or contradicted. Note that in the situation in which a data quality problem can be associated with more than one DQ dimensions, the confidence indexes provide an idea of the impact that the different DQ dimensions have on the problem.

In summary, at this stage we assume that on the basis of the position of the DQ Flag and the context of the business process the process analyst identifies the problems and supported by the knowledge base is able to define the set of quality dimension associated with the h-th DQFlag DQ_h on which low level data quality requirements have to be elicitated.

4.3 Analysis and selection of improvement activities

The selection of DQ improvement activities depend on the data quality dimensions for which data need to be improved. Data-related improvement activities that can be considered for the design of data quality-aware business processes include:

- *Data enrichment*: it is about correcting and completing the used data with some values conveniently retrieved from external reference data sources;
- *Laissez*-faire: you do not do anything;
- *Data cleaning*: it involves changes to existing values for data after a comparison between value stored in the database and the real or correct value or any other that can be considered as a certified reference;
- *Data quality monitoring*: automatic or manual procedures that verify that the data values satisfy specific requirements or business rules (e.g., formats or value included in specific intervals). A specific type of data monitoring activities is *Data edits*, automatic procedures that verify the compliance of the data in the moment in which they are inserted in the system.

Process-related improvement activities that we consider for the design of data quality-aware business processes are:

- *Re-execution*: automatic procedures that verify that the inserted data satisfy specific requirements or business rules
- *Workaround*: method, sometimes used temporarily, for achieving an activity or goal when the usual or planned method is not working.

Such actions can be implemented in different ways by using various mechanisms that can be divided in two classes: mechanisms that change the process flow and mechanisms that require the execution of additional activities that do not affect the process model. For example, data enrichment can be performed using two approaches. In fact, the enrichment of internal databases can be performed as a periodic activity getting all the possible data available from external sources at once. The activity is independent on a specific process and all the processes that access to the databases would benefit of such action. On the other hand the data enrichment can be also performed on demand when an error in the process occurs. In this case, the process flow would change, since when an error occurs it is necessary to insert an activity to model the link to external sources. The link between data quality dimension, improvement activities and related mechanisms in our framework are stored in an apposite repository. Table 1 provides an example of the contents of such repository.

Notice that some of the listed DQ activities can be also combined. Monitoring activities should be performed before every cleaning and data enrichment activity. Also from the small examples contained in Table 1, it is possible to notice that for a specific dimension, it is possible to adopt several DQ activities and related DQ mechanisms. The selection of the most suitable DQ activity is mainly driven from the following factors:

- *Effect of the adoption of the DQ activity on the affected quality dimensions:* each data quality activity has a different impact on the improvement of the data quality level of each data quality dimension. For example, data cleaning and data enrichment activities, performed only when the error occurs, surely will improve data quality while the same activities periodically performed correct most of the errors but they do not completely eliminate the possibility to have errors inside the process. Thus, the periodic cleaning could be not sufficient to guarantee the process needs. Furthermore, it is also necessary to consider that several studies show that data quality dimensions are dependent on each other. These dependencies should be evaluated in order to have a correct assessment of the impact of the DQ activities. In fact, dependencies can be highlight also trade-offs, e.g., security vs timeliness and a DQ activity that has a positive impact on a data quality dimension might have a negative impact on another one.

- *Impact of the improvement action of the business process:* the selection of the improvement actions should not affect the business process performance: the adoption of DQ activities should not violate the process requirements. For example, in case of business processes in which the execution time is a crucial variable, all the DQ activities that require the insertion of new activities in the process should be avoided
- *Cost of the implementing the improvement actions compared with the cost of allowing poor quality:* the cost of the improvement activities should be also considered. Some actions would require the involvement of new actors or specific software modules. A cost-benefit evaluation should be carried out in order to evaluate the feasibility of the different initiatives. The alternative to consider is that the first benefit that the investment will have is the elimination of the poor data quality costs. Thus at least the cost of the initiative should not be higher of these costs.

DQ Activities	**Dimensions**	**DQ mechanisms**
Data Enrichment	Completeness	Enrich data sources data retrieved from external sources (periodically or when the error occurs)
Data Cleaning	Accuracy, Consistency	Correct data comparing them with the real-world values (periodically or the error occurs) Correct data comparing them with certified sources (data bashing) (periodically or the error occurs)
Data Monitoring	Accuracy, Consistency, Completeness, Timeliness, Security	Verify/Ensure whether all mandatory items of data have values Check data with vocabulary to find syntactic errors Check if data is delivered in the time required Verify the permission to access the data Register the trace of accesses, creation and changes suffered by data

Table 1. DQ activities, affected dimensions and related DQ mechanisms

5 Conclusions

In this paper we have introduced a top-down to support process designers in the selection of the improvement actions to adopt in the design of business processes in order to satisfy the data quality requirements. The actions will be supported by feasible mechanisms, which once implemented, will assure the adequate levels of data quality at the critical points of the business process. The methodology provides guidance to refine from high level data quality requirements to the reasoned election of the most suitable data quality improvement activities that will assure adequate values for each one of the considered data quality dimensions. In addition, and to quantitatively support the selection of the improvement alternatives, we also introduce some measures that help workers to choose the most efficient ways to improve the business processes taking into account the cost of implementing actions and the benefits of doing so. The main limitation of the proposed approach is that we start from the assumption that the involved actors have a good knowledge of the analysed BP. The lack of information about the analysed BP could affect the effectiveness of the presented methodology

Future work focuses on the implementation of a tool to support the methodology described in this paper. We also aim to show the effectiveness of the methodology by using several case studies.

References

Bagchi, S., X. Bai, et al. (2006). Data quality management using business process modeling, IEEE.

Ballou, D. P. and G. K. Tayi (1996). Managerial Issues in Data Quality. First International Conference on Information Quality (ICIQ'96), MIT, Cambridge, MA, USA.

Batini, C. and M. Scannapieco (2006). Data Quality: Concepts, Methodologies and Techniques. Data-Centric Systems and Applications., Springer-Verlag Berlin Heidelberg.

Bringel, H., A. Caetano, et al. (2004). Business Process Modeling Towards Data Quality Assurance. 6th International Conference on Enterprise Information Systems, Porto, Portugal.

Caro, A., A. Rodriguez, et al. (2012). Designing Business Processes able to satisfy Data Quality Requirements. 17th International Conference on Information Quality (ICIQ). Paris, France.

Dewan, R. M., V. C. Storey, et al. (2012). "Data quality: Setting organizational policies." Decision Support Systems.

el Abed, W. (2009). "Data Governance: A Business Value-Driven Approach."

Eriksson, H. and M. Pemkel (2001). Business Modeling with UML. OMG Press.

Falge, C., B. Otto, et al. (2012). Data Quality Requirements of Collaborative Business Processes. 45th Hawaii International International Conference on Systems Science (HICSS-45 2012), Grand Wailea, Maui, HI, USA, IEEE Computer Society.

Harmon, P. and C. Wolf (2011) "Business Process Modeling Survey." Business Process Trends (http://www.bptrends.com/).

Heravizadeh, M., J. Mendling, et al. (2009). Dimensions of business processes quality (QoBP), Springer.

Lu, R., S. Sadiq, et al. (2009). "On managing business processes variants." Data & Knowledge Engineering **68**(7): 642-664.

Maydanchik, A. (2007). Data quality assessment, Technics publications.

Ofner, M. H., B. Otto, et al. (2012). "Integrating a data quality perspective into business process management." Business Process Management Journal **18**(6): 9-9.

Rodríguez, A., A. Caro, et al. (2012). A BPMN extension for including data quality requirements in business process modeling. 4th International Workshop on the Business Process Model and Notation, Vienna, Austria, Springer.

Soffer, P. (2010). "Mirror, mirror on the wall, can i count on you at all? exploring data inaccuracy in business processes." Enterprise, Business-Process and Information Systems Modeling: 14-25.

Strong, D., Y. Lee, et al. (1997). "Data Quality in Context." Communications of the ACM **Vol. 40, Nº 5**: 103 -110.

Weske, M. (2007). Business Process Management: Concepts, Languages, Architectures, Springer-Verlag Berlin Heidelberg.

Wetzstein, B.; Leitner, P.; Rosenberg, F.; Brandic, I.; Dustdar, S.; Leymann, F., (2009) "Monitoring and Analyzing Influential Factors of Business Process Performance," *Enterprise Distributed Object Computing Conference, 2009. EDOC '09. IEEE International* , vol., no., pp.141,150.