DESIGNING BUSINESS PROCESSES ABLE TO SATISFY DATA QUALITY REQUIREMENTS

(Research-in-Progress)

Angélica Caro, Alfonso Rodríguez
Department of Computer Science and information Technologies
Universtity of Bío-Bío, Chillán, Chile
{mcaro, alfonso}@ubiobio.cl

Cinzia Cappiello
Dipartimento di Elettronica e Informazione – Politecnico di Milano
Piazza Leonardo da Vinci 32, 20133 Milano, Italy
cappiell@elet.polimi.it

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

Abstract: Nowadays, data quality is a fundamental issue to be considered in order to avoid inefficiencies and to fully exploit all the benefits of adopting sophisticated information technology platforms that can support essential activities for business such as decision making, business intelligence and customer services. Business efficiency and effectiveness also depend on the way in which business processes are modeled. A sound modeling of the business processes is becoming a higher priority for business managers and analysts since documenting and understanding business processes support them in the optimization and improvement of the business functions. In this paper we propose a methodology (named BPiDQ) to consider data quality issues in the business process modeling phase to support the design of data quality-aware business processes.

Key Words: Data Quality, Business Process Model, BPMN, Data Quality Requirements.

1. INTRODUCTION

Modern organizations use different strategies to achieve success, sustainability and competitiveness. Most of them concentrate their efforts in adopting sophisticated information technology platforms that can support essential activities such as making decision, business intelligence, and costumer services, among others. However, these platforms per se are not useful if the core business relies on inefficient processes. For this reason, some organizations have focused their efforts on the definition and management of suitable Business Processes (BP) that optimize the procedures, the use of information technology, and the involvement of the human resources.

On the other hand, in order to avoid inefficiencies and to achieve all the benefits of the adoption of advanced information management solutions, high quality data is also needed [1]. Thus, achieving adequate levels of Data Quality (DQ) could be a strategic approach to consider as part of the business process management.

Formally, DQ is often defined as "*fitness for use*", i.e., the ability of a data collection to meet users' requirements [2]. DQ is a multidimensional and subjective concept since it is usually evaluated by means of different criteria, namely DQ dimensions, whose selection of those that better describe users' DQ requirements and the corresponding evaluation largely depends on the context of use.

Guaranteeing high levels of DQ for the data used in tasks at hands is an important issue especially in information-intensive organizations. In general, poor data quality exposes organizations to nondepreciable risks especially when a business process relies on incorrect, incomplete or out-of-date data. Such data quality issues might also imply the complete or partial failure of the business process: e.g., the use of a wrong address for a product delivery, or the delay in communicating the needed information in a process with strict temporal constraints. These consequences can be avoided or at least alleviated by adopting suitable strategies to early tackle the DQ necessities as a proactive attitude facing the occurrence of data quality problems when the BP is executed.

A BP model represents the flow of physical items or informational artifacts through a sequence of tasks and sub-processes that operate on them [3]. In business process modeling, the main objective is to produce a description of the business work in order to better understand the process, and eventually, improve it: for example, the way in which a commercial transaction is carried out. Our idea is to model business processes considering in addition the data quality issues, and consequently, including activities able to minimize the risk associated with data-related errors. To this aim, it is also important to have a suitable notation for modeling the essence of the business as clearly as possible. Among all possible choices, a recent study shows that BPMN (Business Process Model and Notation) is one of the most important and popular standard to modeling business process [4]. Unfortunately, BPMN lacks of the mechanisms to represent data quality concerns. In addition, there is no guide either that allows business people to incorporate data quality requirements into the representation of the business model when this is done by means of BPMN. So, as part of our research-in-progress work, and as the main contribution of this paper, we introduce a methodology named “Business Process including Data Quality view point” (BPiDQ), that aims to provision a methodological approach for the modeling and design of data quality-aware business processes as well as the generation of the corresponding DQ requirements for the software development that support the business processes. This contribution extends our previous work [5], which consists of an extension of BPMN 2.0 that allow business people, in a simply way, to identify the critical points where the DQ is crucial for the success of the BP. BPiDQ aims to support the workers (business analyst/designer, DQ expert and System analyst) to improve the BP by means of some changes or by introducing new activities that guarantee the satisfaction of the DQ requirements and derive DQ use case for the software development. In addition, other necessary artifacts that complement the methodology are introduced.

The rest of the paper is organized as follow. Section 2 discusses the related works. The BPiDQ methodology and its main components are introduced in Section 3. To illustrate the use of the BPiDQ an example is developed in Section 4. Finally, Section 5 gives our conclusions and future works.

2. RELATED WORKS

DQ management has been widely recognized as a relevant aspect that deserves to be considered in order to globally improve the effectiveness of organization’s performance [6]. Thus, it is important that business people are aware of DQ requirements from the earliest stages of the design of a business process, i.e., business process modeling. The most used languages to model business processes, namely BPMN and UML [4], do not allow process designers to fully specify DQ requirements at a high level.

To the best of our knowledge, at the present time, there is only one specific notation to represent DQ issues in business process, allowing the depiction of what its authors named information products maps (IP-MAP) [7]. It permits the specification of business processes by means of a conceptual map and a sort of activity diagrams, in which the efforts corresponding to data quality management are properly addressed by means of some specific constructs [7]. Indeed, BPMN is widely recognized as de facto standard to model business processes [4, 8]. Its expressiveness can be and has been already extended, to support some other concerns of interests. For example, it has been extended to support customer needs related with quality requirements such as time, cost, and reliability [9], to submit/response-style user interaction [10], to specify non-functional properties such as performance and reliability oriented to a characterization of the business process [11], to include Business Activity Monitoring (BAM) relevant concepts in BP models [12], to capture the temporal perspective of business processes [13], to include information coming from sensors and smart devices [14], to model security requirements [15], to represent explicitly legal constraints directly by specific artifacts [16], or to analyze business processes performance [17], to name a few.

However, DQ concerns are not new to BP research area: some existing contributions highlight the need of addressing DQ in the business process modeling during the design time. So, for instance, in [18], Soffer explores the inaccuracies of data, the situation where the information system does not truly reflect the state of a domain where a process takes place. The potential negative consequences of data inaccuracy are discussed. The work provides the bases to support the design of robust processes and avoid problems related to data inaccuracy. Bringel et al. in [19] propose a business process pattern to ensure data quality in an organization. The pattern consists in a business process model that can be reused through adaptation in specific organizational scenarios. For this, they define DQ attributes associated with information entities having different meanings depending on the business view and the different organizational dimensions. The Data Excellence Framework is proposed in [1]. This framework describes the methodology, processes and roles required to generate the maximum business value while improving business processes using data quality and business rules. In this approach, DQ requirements are specified as business rules. The set of business rules supporting data quality grows over time as part of the process of continuous improvement. Bagchi et al. in [3] introduced a business process modeling framework for quantitative estimation and management of data quality in information systems. Based on this framework, they propose to exploit the structure provided by the business process flows to estimate errors arising in transaction data and the impact of their propagation to the key performance indicators.

Also, Heravizadeh et al. in [20] 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.

Finally, the work presented in [21] introduces some 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 preventative focus, thus achieving compliance by design [21]. 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. A DQ aspect considered in these control objectives is the data integrity.

The cited studies consider different DQ dimensions, which are summarized in Table 1.

Table 1. Data Quality Attributes identified in BP modelling.

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 Manip.
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					

3. BPiDQ: A Methodology to Design DQ-Aware Business Processes

A methodology is generally defined as a guideline for solving a problem, with specific components such as phases, tasks, methods, techniques and tools [22]. We propose BPiDQ, a methodology to support the modeling and design of data quality-aware business processes and the generation of DQ requirements for the software development.

BPiDQ uses BPMN as BP modeling language and works with different models in the two first out of three levels of abstraction of BPMN [23]. Such levels are: (i) the *Descriptive level*, which uses the basic set of shapes and symbols that are adequate for the needs of business people seeking to document a process; (ii) the *Analytical level*, in which the full set of shapes and symbols can be used to deal with events and exception handling showing the complexity and depth of the process; (iii) the *Executable level*, which deals with the XML language underneath the shapes.

As shown in Figure 1, BPiDQ is composed of four stages. The first stage (BPiDQ-S1) in the BPMN Descriptive level, starts by introducing high-level DQ requirements into the BP model. In our work we have defined as high-level DQ requirement a mark included into a shape of a BPMN element to highlight a point where the DQ is necessary for the BP success. In the second stage (BPiDQ-S2) the high-level DQ requirements will be refined in order to generate low-level DQ requirements. In our work a low-level requirement is a detailed specification that included among others: the data involved and a set of relevant DQ dimensions. In the third stage, (BPiDQ-S3) in the BPMN Analytic level, the DQ requirements will guide the data quality-aware BP improvement that will imply the addition of new activities or the modification of the process flow. Finally, the fourth stage (BPiDQ-S4) supports the generation of use case diagrams to specify DQ software requirements.

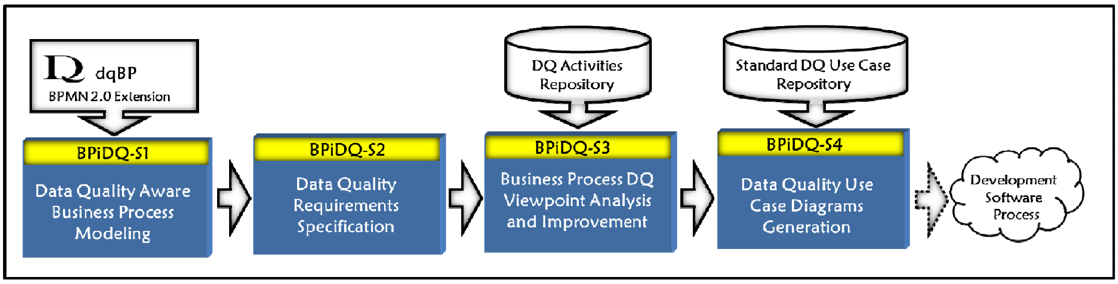

dqBP
BPMN 2.0 Extension
DQ Activities Repository
Standard DQ Use Case Repository
BPiDQ-S1
Data Quality Aware Business Process Modeling
BPiDQ-S2
Data Quality Requirements Specification
BPiDQ-S3
Business Process DQ Viewpoint Analysis and Improvement
BPiDQ-S4
Data Quality Use Case Diagrams Generation
Development Software Process

Figure 1. Methodology to design data quality-aware business processes

Also, Figure 1 shows that BPiDQ uses three basic components to support the stages: (a) dqBP, a BPMN 2.0 extension to include high-level DQ requirements in a BP model, (b) a repository of DQ activities to improve the BP from DQ point of view, and (c) a repository of standard DQ use cases to specify DQ software requirements. Such components will be described in the following sections together with a detailed description of the BPiDQ´s stages.

3.1 Components to support BPiDQ

In the following, three components used in BPiDQ to model and design data quality-aware business process and to generate DQ requirements for the software development are introduced.

3.1.1 dqBP: A BPMN 2.0 extension to support BPiDQ

Various elements of BPMN are used for data representation (e.g., Data object or Message). However, aspects related to data quality cannot be included in this kind of elements using the BPMN language. Thus, to support the first stage of BPiDQ and to fill this gap, we have introduced an extension of BPMN 2.0, named dqBP that enriches the BP modeling with DQ requirements [5]. The high level DQ requirements will be modeled in a BP model by means of a set of flags, named DQ Flags. The DQ Flags may be associated with the BPMN data-related elements (Data Objects, Message, Message flow, Conversation, Data Store, and Activity) to mark that they are susceptible to be linked to special data quality requirements. We have also defined the symbol DQ, coming from merging letters D and Q, to perform the marking of these BPMN elements. Consequently, such symbol must be included into the shape of the BPMN data-related element in order to show that the quality of data in that specific point of the process is crucial for the business. Table 2 shows a description of these BPMN elements and their graphical representation.

Table 2: Representation of BP data- related elements enriched with the DQ flags

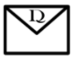
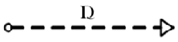
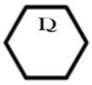
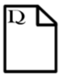
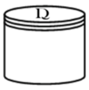
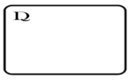
Data-related BPMN element	Graphical Representation	Intended use of the Graphical Representation
Message: Content of a communication between two participants. May be data structured or unstructured.		It represents that data contained in the message might satisfy some DQ requirements for the sake of the business success, e.g. Completeness and Consistency in a drug prescription from the doctor to a patient.
Message flow: It shows the flow of Messages (explicit with a Message or implicit without the Message) between two Participants		It represents that data implicitly contained in the message (the message does not appear in the flow) might satisfy some DQ requirements to develop success- fully the BP, e.g. Timeliness for a credit card authorization from the bank.
Conversation: Logical grouping of Message exchanges (Message Flows) that can share a Correlation. Conversation has the data contents in the messages included on it.		It represents that data in some messages contained in the conversation might satisfy some DQ requirements for the sake of the success the business process, e.g., Security and Accuracy of the data interchanged between a customer and an airline Web application during the flight booking process.
Data Object: Primary construct for modeling data within the Process flow in BPMN. It can represent a singular object or a collection of objects, input data or output data.		It represents that data in the data object might satisfy some DQ requirements to successfully achieve the goals of the business process, e.g. Completeness, Consistency and/or Accuracy of the data required to successfully deliver and ordered package to a customer.
Data Store: It provides the necessary mechanisms for Activities to retrieve or update stored information that will persist beyond the scope of the Process.		It represents that data contained in a data store might satisfy some DQ requirements for the sake of the success of the business process, e.g. Checking the completeness of the data updated about product sale.
Activity: Work that is performed within a Business Process. The activity's work may be the generation/processing of data.		It represents that used/produced data in the activity might satisfy some DQ requirements to the business success, e.g. Checking the Precision and Accuracy of the budget generated as the output of one activity.

Figure 2 shows graphically the extension in BPMN 2.0 and the metamodel proposed to support the specifications of the DQ requirements for each DQ Flag in a BP model. In white color, Figure 2 illustrates some classes from BPMN 2.0: (a) the extension metamodel classes (Definition and Extension) from where is derived our proposal, and (b) a set of BPMN classes related with dqFlag class (our extension). In the same figure, in grey color, the metamodel that will support the derivation of DQ requirements from DQ Flags in the BPMN model is showed. More details about the extension can be found in[5].

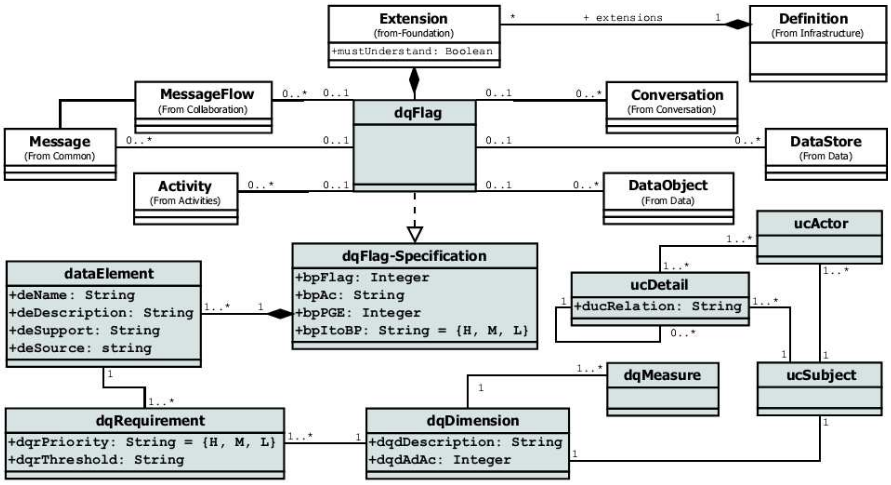


Figure 2. BPMN 2.0 extension and metamodel to obtain of DQ requirements from a BPMN model.

3.1.2 A Repository of DQ Activities to tackle DQ Requirements

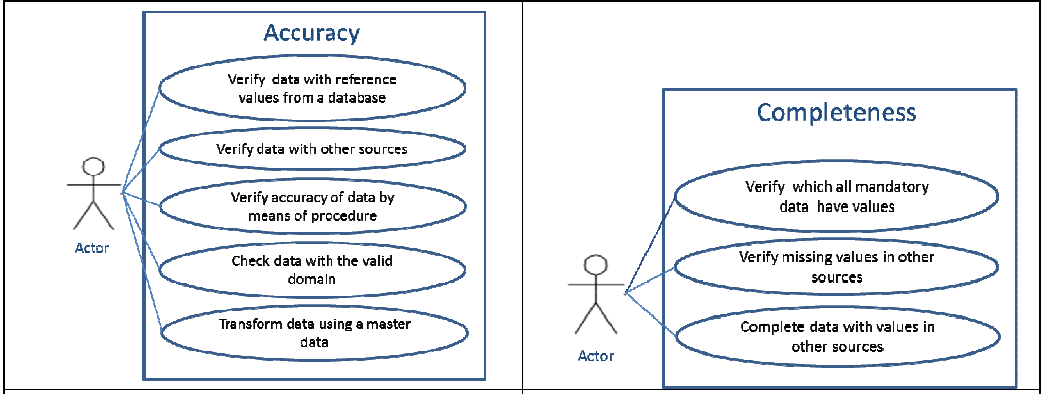
In BPiDQ, the DQ requirements are expressed by means of some DQ dimensions. The DQ field provides different DQ models (generic set of DQ Dimensions) in order to address DQ concerns in different contexts [2]. Anyway, these DQ models have to be interpreted, and adapted to better fit in specific context. In our work we have decided to consider the most referenced DQ dimensions in the BP literature (see Table 1) to define a repository with DQ activities to tackle each one of them. Thereby, BPiDQ aims to enrich the BP model including a set of DQ activities, obtained from the repository, to tackle the DQ requirements. Table 3 shows some commonly used DQ dimensions and some DQ activities to be performed in order to guarantee that DQ requirements are satisfied within the considered Business Process. Note that these activities have been defined in a generic way and they need to be customized on the basis of the analyzed process and its corresponding context.

DQ Dimension	Definition	Improvement Activities	Examples
Accuracy	The extent to which data reflects a real-world view within a context and a specific process [1, 18, 20].	- Determine the data set, which requires accuracy. - Verify data provided against the right domain. - Verify data coming from alternatives sources. - Clean database to achieve the required level of accuracy.	- The price received by the client for a booking hotel must be accurate. - In a medical prescription, the name of the medicines can be confronted with the Vademecum. - The weight of a package to be delivered must be contained within a specific range of values.
Timeliness	The extent to which data are sufficiently updated for the context and a specific process [1, 19, 20].	- Verify if data have the required age for the task. - From different sources, select the one providing data with the age required for the process. - Check if data are delivered within the required time.	- Check if the same data are in different company's source and if it is closer to the right age required, and then take values from this source. - Bank's response to check a credit card must be lower than 5 seconds.
Completeness	The extent to which data have all values necessary for a successful execution of a process in a specific domain and context [1, 19, 20].	- Specify which data are mandatory - Verify/Ensure whether all mandatory items of data have values. - Complete data provided with other sources of data. - Use a procedure to force the delivery of all mandatory data.	- Check if the same data are in different company's and then complete the golden register - To deliver a package, all data about the address and customer identification must be complete.

Table 3. Example of improvement Activities associated with DQ dimensions.

3.1.3 A Repository of Standard DQ Use Cases to tackle the DQ requirements

Taking into account that the BP will be supported by an information system, BPiDQ supports the generation of a set of use cases to represent the DQ requirements for the application to develop. Due to this reason, we have introduced as the third component of BPiDQ a repository that contains a set of standard use cases for each DQ dimension. The use cases have been customized and defined by considering: (a) the definition of each DQ dimension, (b) the set of DQ activities to address each DQ dimension (the previous component), and (c) knowledge previously extracted from existing literature contributions or/and from software developers experience. The idea is that based on these standard use cases the workers could specify a final use case version according to the BP modeled as it is explained in the following subsection. The standard DQ use cases do not consider specific associated actors because they must be specified in the final use case diagram (that will represent the requirements of the application that will support the BP) as «include» use cases for the use cases that will have interaction with the real actors (system´s users). As an example, Table 4 shows some standard DQ use cases for the DQ dimensions for accuracy (part a) and for completeness (part b).


Accuracy
Verify data with reference values from a database
Verify data with other sources
Verify accuracy of data by means of procedure
Check data with the valid domain
Transform data using a master data
Actor
Completeness
Verify which all mandatory data have values
Verify missing values in other sources
Complete data with values in other sources
Actor

(a) DQ dimension of Accuracy

(b) DQ dimension of Completeness

Table 4. Some standard use cases for Accuracy and Completeness DQ dimensions

3.2 BPiDQ´s Stages

In this section the four stages of BPiDQ, detailing the workers involved, component used, input and outputs for each one of them will be described.

3.2.1 BPiDQ-S1: Data Quality-Aware Business Process Modeling

This stage is devoted to capture *high level DQ requirements* at a BPMN Descriptive Level [23]. Such requirements are specified by Business People/Analysts and are graphically expressed by means of a specific mark called DQ Flag. Figure 3 shows graphically this stage highlighting the involved workers, inputs and outputs.

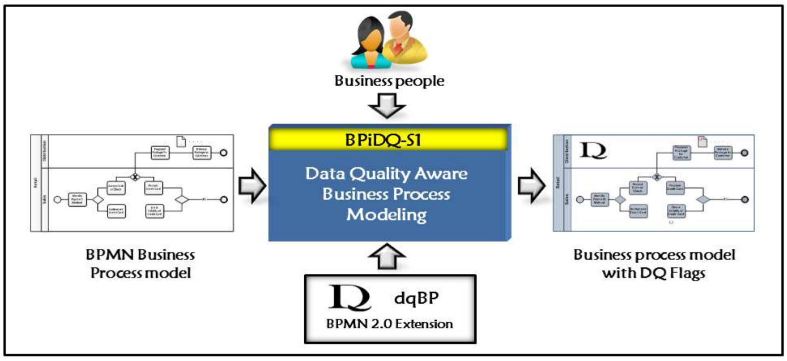

Business people
BPiDQ-S1
Data Quality Aware Business Process Modeling
BPMN Business Process model
Business process model with DQ Flags
dqBP
BPMN 2.0 Extension

Figure 3. BPiDQ-S1: Data Quality-Aware Business Process Modeling

- As showed in Figure 3, the workers start modeling a new BP or analyzing an existing one. The result of this stage is the BP enriched with a set of DQ Flags that highlight the points of the BP where the DQ is considered essential for the business success. The main activities involved in this stage are:
- **BPiDQ-S1.1. Enrichment of BP model with DQ Flags**. Workers model a BP in the traditional way or start analyzing a BP model created previously. Using the dqBP extension, workers place DQ Flags for some data-related BPMN elements where they think that some DQ management activities are necessary to warranty the BP success.
- **BPiDQ-S1.2. Registration of additional information about the BP and DQ Flags**. Some additional information must be registered by means of text annotations. In particular: (a) business people must include the identification of each data element contained in the data-related BPMN elements marked with a DQ Flag, and (b) an estimation of the level of influence of each DQ Flag in the overall success of business process ranged as {"Low", "Medium", or "High"}.

It is important to note that this stage is supposed to be performed by business people. Generally speaking, they are not expert in technical issues, but they are expert in their own business processes. Thus, in this stage, our aim is to provide adequate mechanisms to express in a simple way the DQ necessities for a specific BP.

3.2.2 BPiDQ-S2: Data Quality Requirement Specification

In the second stage, the involvement of Business Analyst/Designer and also the DQ Expert is required. These workers should work together to analyze the modeled BP from a DQ point of view. Figure 4 shows graphically the involved workers, inputs and outputs of this stage.

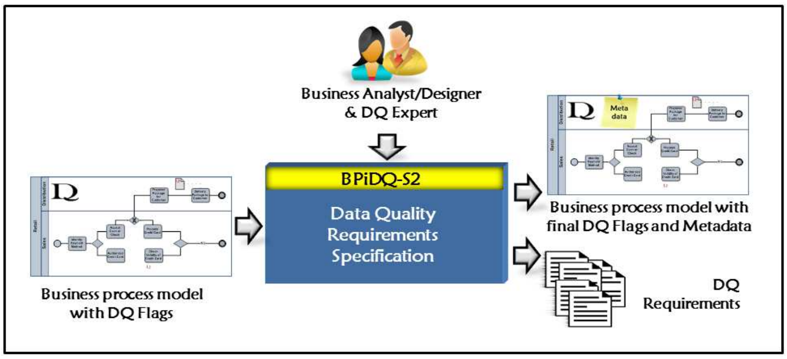

Business Analyst/Designer & DQ Expert
BPiDQ-S2
Data Quality Requirements Specification
Business process model with DQ Flags
Meta data
Business process model with final DQ Flags and Metadata
DQ Requirements

Figure 4. BPiDQ-S2: Data Quality Requirement Specification

Taking as input product the artifact generated in the previous stage (a BP model enriched with DQ Flags), this stage is dedicated to specify *low level DQ*. Workers must review and analyze each DQ Flag in order to make a more refined and complete specification of the DQ requirements related with each one of the DQ Flags. The main activities involved in this stage are:

- **BPiDQ-S2.1. Collection and registration of metadata about the BP and the DQ Flags included on it**. This metadata can be provided by experts and/or by software applications (when the BP is actually implemented and working). We have defined three types of metadata which should be collected:
  - *Metadata about the BP flow* provide information related to the BP control flow, and thus some metadata about the execution of certain activities in a process. For example, it should be useful to know the execution probability of each path on the BP. The range for the probability value is greater or equal than 0 and less or equal than 1. This probability could be estimated by the business analyst from previous executions of the BP, or it may be calculated taking into account the alternative paths drawn by the gateways in the BP.
  - *Metadata about the BP Performance* refer to the performance conditions or constraints within

process flows. This metadata can be defined either at the process or at the task level. In both cases, the metadata store data about temporal conditions (e.g., maximum time that may be needed to respond to a request).

- *Metadata about Data* provide information regarding the data used throughout a process. For example, for each DQ Flag, and for each data element on it, the corresponding metadata must be registered: BPMN element associated with the DQ Flag, DQ Flag's path, previous and posterior activities associated to the DQ Flag, data element description, support (electronic/manual, etc.), source (internal/external), actions of use (use, creation, modification, etc.), the data volatility (i.e., permanent or transient information), and some other points of the BP where the same data is used, to name a few.

- **BPiDQ-S2.2. Specification of the DQ requirements for each data element on a DQ Flag**. For each data element, a set of DQ dimensions along with their corresponding level of importance ranged typically as {"*Low*", "*Medium*", "*High*"} must be identified. The DQ experts must study dependencies between the DQ dimensions associated with each data element to decide if any of them can be eliminated (e.g., to be incompatible with other). This decision must also consider the importance given to the DQ dimensions.
- **BPiDQ-S2.3. Refining the set of DQ Flags in the BP**. Taking into account some of the metadata described previously, the specified DQ requirement, and the cost to satisfy the DQ requirements, the workers must decide the final set of DQ Flags. For this decision the following information is needed:
  - The level of influence of each DQ Flag on the overall success of the BP (registered in the first stage).
  - Probability of execution of the path in which each DQ Flag is placed (obtained from the metadata about the flow).
  - DQ Flag overhead, defined as the ratio between the number of new activities that has been added to tackle with the new DQ requirements (one activity for each DQ dimension) and the total number of activities in the BP. This factor shows the overhead relative of each DQ Flag in the BP.
  - Business constraints (obtained from metadata about performance).

  For example, if a DQ Flag has (i) a grade of influence higher than another one, (ii) a higher probability to be executed than another DQ Flag, and (iii) a medium overhead in the BP, then, the first DQ Flag is considered more important and could have more probabilities to be addressed.

  Finally, the dependencies between the data elements in the same BP branch have to be studied, (for example to eliminate some redundant DQ Flags).

As a result of this stage, the final configuration of DQ Flags for the BP model should be released. Also, the documentation about all DQ Flags (data-related BPMN elements and data elements associated), and the specification of DQ requirements for it (in low level) should be generated.

3.2.3 BPiDQ-S3: Business Process DQ viewpoint Analysis and Improvement

The third stage is devoted to analyze and decide the most suitable way to improve the BPMN model from the DQ point of view. This stage is executed at the BPMN analytic level [23], and the workers involved are the Business Designer and DQ Expert. Figure 5 shows graphically the workers, inputs and outputs of this stage.

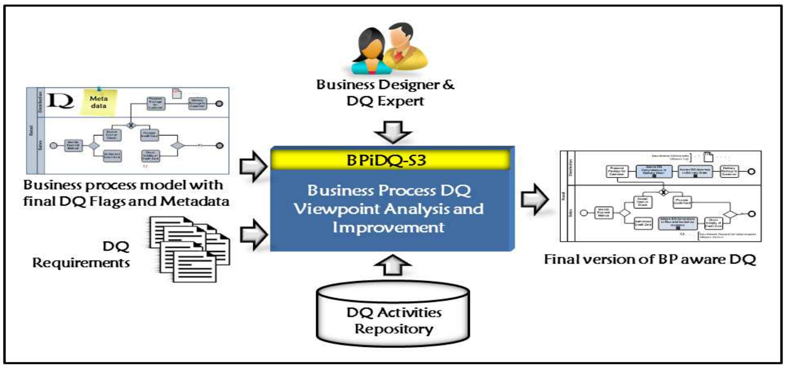

Business Designer &
DQ Expert
Meta data
Business process model with final DQ Flags and Metadata
DQ Requirements
BPiDQ-S3
Business Process DQ Viewpoint Analysis and Improvement
Final version of BP aware DQ
DQ Activities Repository

Figure 5. BPiDQ-S3: Business Process DQ viewpoint Analysis and Improvement

As showed in Figure 5, in this stage the workers generate the final version of the BP model. This version will include the modifications needed to address the DQ requirements and defined taking into account the metadata associate with the BP to decide the best solution. In this stage, the following activities are executed:

- **BPiDQ-S3.1. Selection of the improvement actions for satisfying the low-level DQRequirements**. Considering the DQ requirements in each DQ Flag and some metadata, a set of new activities should be selected to assure the adequate level of DQ for the data in the BP. Thus, the BP model will be enriched with the inclusion of new activities in order to avoid some DQ problems and consequently to minimize the risk due to poor data quality. The activities will be provided from the DQ activities Repository (component of BPiDQ previously explained) that contains a set of DQ activities for each DQ dimension and considering the use of data (creation, use, modification).
- **BPiDQ-S3.2. Improvement of the BP model to satisfy the DQ requirements.** Taking into account the DQ requirements, the flow of the BP and the set of new DQ activities selected to be included in the BP model, workers must study how to change the BP model in order to assure the most appropriate configuration to satisfy the DQ requirements. This means:
    - Generate alternative BP models, which integrate the DQ activities to satisfy the DQ requirements in the BP flow. For example, alternative models depending on the actions to develop where a DQ problem raises may be generated, one of them to abort the execution or another one to develop some actions and follow the execution.
    - Study the BP flow to decide whether it is necessary a redefinition of it in order to satisfy some DQ requirements. For example, if two sequential activities are independent between them, then, they can be executed in parallel in order to improve the time-related data quality dimensions.
    - Evaluate the proposed alternatives and select the most suitable one that better satisfy both data quality requirements and the business objectives. A cost-benefit analysis must be conducted, considering the costs of the implementation, the user satisfaction, the success of the BP, etc.

In this stage, the final version of the BP will be released. This stage works with the BP model at the BPMN analytic level what allows modeling the BP with more details than at the BPMN descriptive level. Thus, considering the granularity of the activities in this level, we have decided to generate the model with two levels of details. The first one will include, for each DQ Flag, a set of collapsed sup-processes. Each one of these sub-processes represents a DQ dimension that must be assured for the data element involved in the DQ Flag. The second, for each one of these DQ collapsed sub-processes, in a lower level of detail, will include an expanded Sub-Process that contains all the activities selected to assure the corresponding DQ dimension.

3.2.4 BPiDQ-S4: Data Quality Use Case Diagram Generation

The common next step for the business process modeling, considering for example an MDA approach, is the development of software to support it. Thus, the fourth stage of BPiDQ represents a first approach toward the definition of requirements for developing applications able to satisfy the specified DQ requirements. For doing so, we provide the support by means of the generation of a set of use cases that represent the requirements related with the activities in the BP that tackle the DQ expressed like DQ Flags. Figure 6 shows graphically the workers, inputs and outputs of this stage.

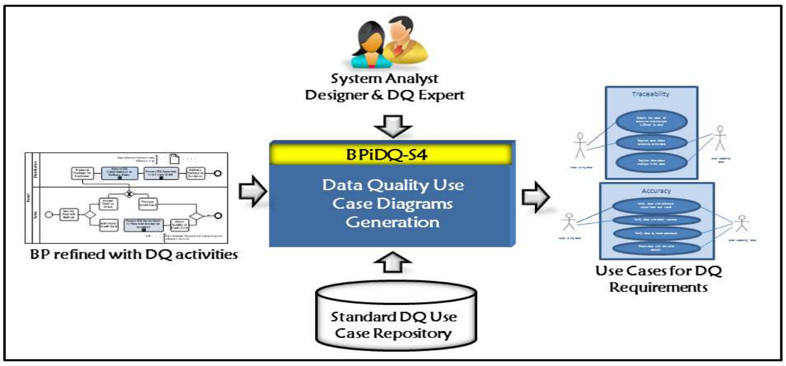

System Analyst
Designer & DQ Expert
BPiDQ-S4
Data Quality Use Case Diagrams Generation
BP refined with DQ activities
Traceability
Accuracy
Use Cases for DQ Requirements
Standard DQ Use Case Repository

Figure 6. BPiDQ-S4: Data Quality Use Case Diagram Generation

Figure 6 introduces the involved workers: System Analyst and DQ Expert. Based on the BP refined with DQ activities and a set of standard use cases (from the repository explained in section 4), these workers will instantiate a set of use case diagrams customized for the specific BP. The following activities are to be executed in this stage:

- **BPiDQ-S4.1. Generation of use case diagrams based on the DQ requirements.** For each DQ collapsed sub-process incorporated in BP, the workers will select the appropriate use cases from the standard use case repository. Indeed, they must select the use case based on the DQ dimension related, and the activities contained in the expanded Sub-Process.
- **BPiDQ-S4.2. Customization of the use cases with the specific BP.** The workers will refine the DQ use cases diagrams generated, customizing the use cases with the BP. From the swimlanes, they can identify the actors. From metadata, they can also identify the data elements to be manipulated, the action developed with the data (creation, elimination or use), the sources of data, etc.

Thus, in this stage a set of use cases diagram available of input for the software development will be produced. Our intention is that the developers can be aware as early as possible of the DQ requirements for the BP, and they have a set of seminal use cases to implement the software considering DQ issues.

The following section illustrates the use of BPiDQ by means of an example.

4. An Example of the Application of BPiDQ

Let us consider the process of payment and delivery of the ordered products. The description of the BP of this example starts with the payment phase. The payment can be processed in two different ways: by credit card or by cash (or check). If payment is made by credit card, it is necessary to ask for card authorization to the «Financial Institution». If the credit card payment is not authorized, then, the process finishes. If the payment is performed by cash (or check), no controls are needed. When the payment is complete, the Distribution Department prepares the package and delivers it to the customer, and after this, the process ends. The remainder of this section is devoted to demonstrate how the proposed methodology can be applied.

In the first stage ("*BPiDQ-S1.Data Quality Aware Business Process Modeling*"), business people must identify which data-related BPMN elements could be susceptible to be linked to DQ Flags. In our example, two DQ Flags are defined. The first one, named DQFlag1, is associated with the Data Object needed as input in the "*Delivery package to customer*" activity (see Fig. 7). This Data Object contains the *Delivery Order* with the customer information necessary to deliver the package (identification, address). The second DQ flag, named DQFlag2, is associated with the Message Flow coming from the Financial Institution pool to Sales lane. This Message Flows contains the *Financial Institution response* a message with the authorization or the rejection to process the payment with customer credit card. As output of this stage, the business process model shown in Figure 7 is generated and enriched with symbol for DQ Flag. In this figure it is possible to see how the data-related BPMN elements have been marked with the special symbol Ꝺ. Also, workers registered the additional information about DQ Flags, data-element identified and their influence in the BP success, in a text annotation artifact.

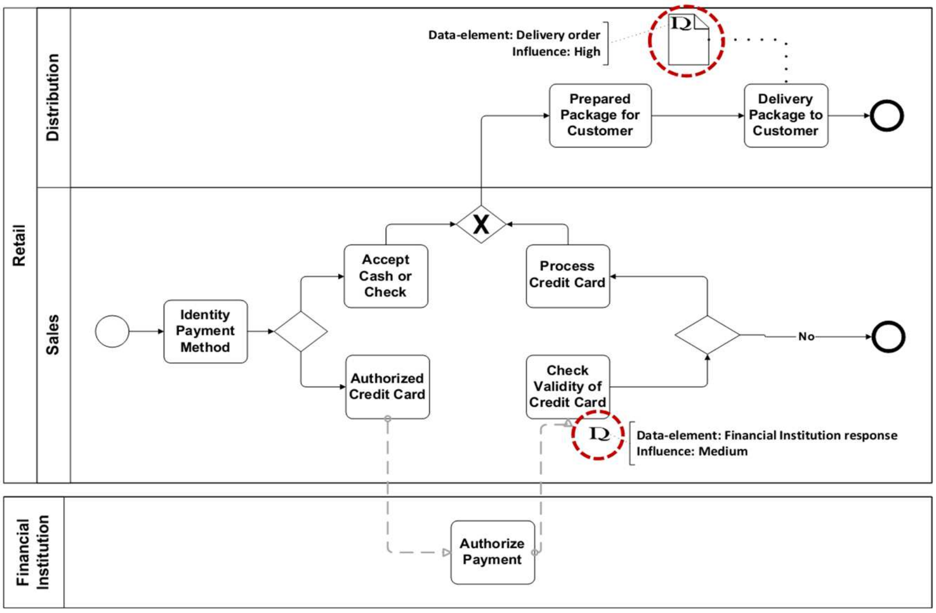


Figure 7. Illustrative example: BPMN model with DQ Flags.

In the second stage ("*BPiDQ-S2. Data Quality Requirements Specification*"), the workers (Business Analyst/Designer and DQ Expert) register metadata about the BP and DQ Flags. They also reviewed each one of the DQ Flags to specify the corresponding low level DQ requirements. For performing the realization of the *DQFlag1*, the definition of a *DQFlagSpecification1* is required. Therefore, they must define the DQ dimensions, and their importance. DQ requirements for *Delivery Order* involve two DQ dimensions, which are considered as relevant for the BP: Accuracy and Completeness. On the other hand, for the *DQFlag2*, the *DQFlagSpecification2* is defined and DQ Requirements for "Financial Institution Response" consider the DQ Dimension Currentness. In addition, for the two DQ Flags the probability of execution and overhead (and some other information) are obtained and/or calculated. Most important details about both DQ flags specifications are shown in Table 5. Taking into account the available information, the workers must decide the definitive set of DQ dimensions for the data elements in each DQ Flag. Besides, they must decide the final set of DQ Flags. In our example, *DQFlag1* has a *High* impact on the success of the business process. Even if they have not any initial knowledge on the process execution, the estimated probability of execution of the delivery action is *75%* because the BP flow shows (taking into account the exclusive gateways) that in some cases the activity related with the DQ Flag may be not executed.

Table 5. DQ Flags specifications

DQFlagSpecification1		
BPMN data element		Data Object
Influence		High
Probability Exec.		75%
Overhead		2/8*100=25%
Data Quality Requirement	Data Elements	
	Name	Delivery Order
	Description	Delivery order (customer information)
	Support	Electronic
	Source	Internal
	DQ Requirements	
	Accuracy	High
	Completeness	Medium

(a)

DQFlagSpecification2		
BPMN data element		MessageFlow
Influence		Medium
Probability Exec.		50%
Overhead		1/8 * 100 = 12.5%
Data Quality Requirement	Data Elements	
	Name	Financial Institution response
	Description	Message from the Financial Institution
	Support	Electronic
	Source	Internal
	DQ Requirements	
	Currentness	High

(b)

The overhead associated with this DQ Flag is *25%* because in order to tackle with the DQ requirements, two new activities must be included in the process (see in grey colour, the new activities in the left side of Figure 8).

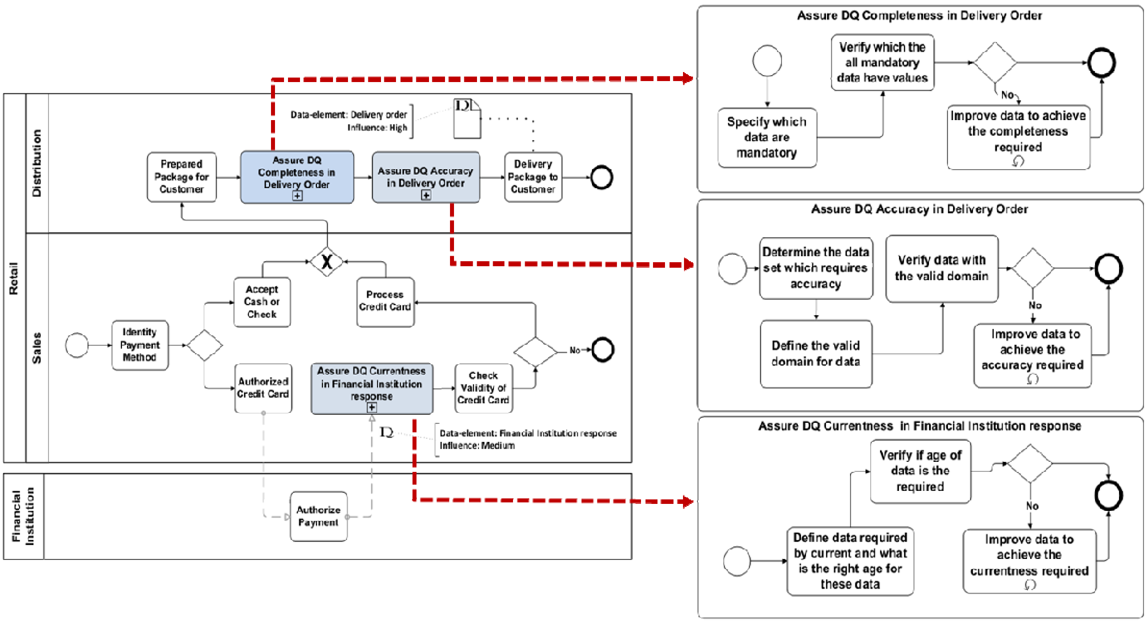

Retail
Distribution
Sales
Financial Institution
Data-element: Delivery order
Influence: High
DL
Prepared Package for Customer
Assure DQ Completeness in Delivery Order
Assure DQ Accuracy in Delivery Order
Delivery Package to Customer
Accept Cash or Check
Process Credit Card
Identity Payment Method
Authorized Credit Card
Assure DQ Currentness in Financial Institution response
Check Validity of Credit Card
No
DL
Data-element: Financial Institution response
Influence: Medium
Authorize Payment
Assure DQ Completeness in Delivery Order
Verify which the all mandatory data have values
No
Specify which data are mandatory
Improve data to achieve the completeness required
Assure DQ Accuracy in Delivery Order
Determine the data set which requires accuracy
Verify data with the valid domain
No
Define the valid domain for data
Improve data to achieve the accuracy required
Assure DQ Currentness in Financial Institution response
Verify if age of data is the required
No
Define data required by current and what is the right age for these data
Improve data to achieve the currentness required

Figure 8. BP model improved.

DQFlag2 has a *Medium* impact on the success of the business process. The probability of requesting the payment authorization is *50%* because when the payment is not performed by credit card the activity related with the DQ Flag is not executed. The overhead associate to this DQ Flag is *12.5%* because to tackle the DQ requirements must be included a new activity in the process (see, in grey colour, the new activity in Figure 8 (left side)). Finally, since the data elements associated with each DQ Flag are crucial for the business process success, the workers decided to implement the improvement actions for both DQ Flags. Note that in this stage the BP is modified including a new activity (collapsed sub-process) for each DQ dimension in each DQ Flag point (BPMN Descriptive Level).

In the third stage ("*BPiDQ-S3. Business Process DQ Viewpoint Analysis and Improvement*"), the business process designer and DQ Expert must decide which specific DQ improvement activities should be adopted. First, and considering each DQ dimension to engage, the use of the data elements and the necessary information recollected, they must select from the repository the most suitable activities. After this, workers must evaluate the possible alternatives to integrate these new activities in the BP. In our example, the activities selected and their flow is showed in Figure 8 (right side). Note that in this stage the collapsed sub-process are replaced for expanded sub-process, considering a more detailed level in the BP model (BPMN Analytic Level). Finally, in the fourth stage (*"BPiDQ-S4. Data Quality Use Case Diagrams Generation"*) the Use Case diagrams which specify the DQ requirements for the software that will implement the improved BP model must be generated. Thus, in our example, the standard use cases for each DQ Flag and the corresponding DQ requirements were firstly selected. After this, the workers customized the use cases the BP modelling. Figure 9 shows the use case diagram generated for the requirements related with the DQFlag1.

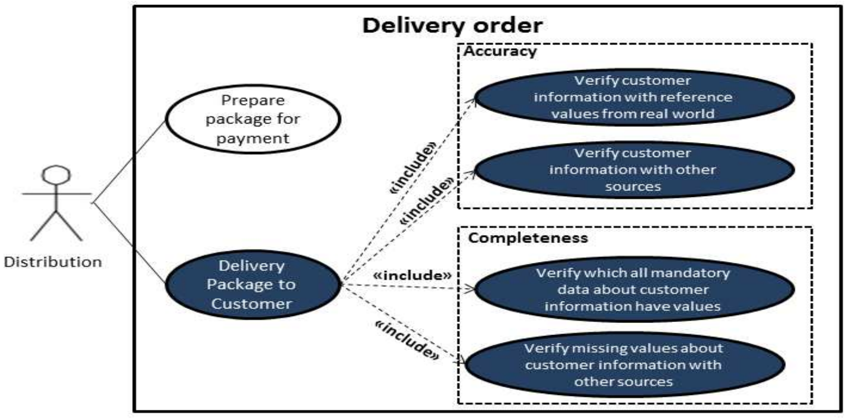

Delivery order
Accuracy
Prepare package for payment
Verify customer information with reference values from real world
Verify customer information with other sources
«include»
«include»
Distribution
Completeness
Delivery Package to Customer
«include»
Verify which all mandatory data about customer information have values
«include»
Verify missing values about customer information with other sources

Figure 9. Use case diagram for the BP model.

The use case diagrams generated in this stage are generic, and they constitute a very first approach towards the software development. In our opinion, although in a simple way, this example demonstrates that our methodology is useful to involve the business people since the earliest definition of the DQ requirements of their business process.

5. CONCLUSIONS AND FUTURE WORK

Poor data quality has severe impacts on the performance of an organization. Most of the organizations are aware about data quality issues, but frequently, they do not have a proactive attitude to address the DQ problems before their apparition. To this aim, in this paper we have presented the BPiDQ methodology that is oriented to support the modelling and design of data quality-aware business process and the generation of DQ requirements for the software development. BPiDQ allows business people to include DQ needs in business process modeling using DQ Flags. Then, for each one of these DQ Flags, BPiDQ allows workers to specify DQ requirements that will drive improvements over the original BP model in order to guarantee the DQ level required. Furthermore, the methodology supports the specification of use cases for the data quality-aware software development. Our future work will focus on three different goals: (a) Conduct some more case studies to obtain the opinion and feedback of the different workers involved, (b) Build a tool to support the methodology allowing the automatic development of some activities, and (c) Refine the methodology stages in order to better support the process improvement.

Acknowledgments

This research is part of the following projects: MECESUP (UBB0704) and IQMNET (TIN2010-09809-E).

REFERENCES

[1] el Abed, W., *Data Governance: A Business Value-Driven Approach.* 2009.

[2] Wang, R. and D. Strong, *Beyond accuracy: What data quality means to data consumers.* Journal of

Management Information Systems; Armonk; Spring. *12*(4). 1996, pp. 5-33.

[3] Bagchi, S., X. Bai, and J. Kalagnanam. (2006). *Data quality management using business process modeling*. pp. 398-405.

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

[5] Rodriguez, A., A. Caro, C. Cappiello, and I. Caballero. (2012). *A BPMN extension for including data quality requirements in business process modeling*. In *4th International Workshop on the Business Process Model and Notation.*

[6] Redman, T., *Data Driven*. 2008: Harvard Business School Press.

[7] Shankaranarayanan, G., R.Y. Wang, and M. Ziad. (2000). *Ip-map: Representing the manufacture of an information product*. In *Fifth International Conf. on Information Quality (ICIQ'2000)*. pp. 1-16.

[8] Recker, J., *Opportunities and constraints: the current struggle with BPMN.* Business Process Management Journal. *16*(1). 2010, pp. 181-201.

[9] Saeedi, K., L. Zhao, and P.R. Falcone Sampaio. (2010). *Extending BPMN for Supporting Customer-Facing Service Quality Requirements*. In *Proceedings of the 2010 IEEE International Conference on Web Services* pp. 616-623.

[10] Auer, D., V. Geist, and D. Draheim. (2009). *Extending BPMN with Submit/Response-Style User Interaction Modeling*. In *IEEE Conference on Commerce and Enterprise Computing*. pp. 368-374.

[11] Bocciarelli, P. and A. D'Ambrogio. (2011). *A BPMN extension for modeling non functional properties of business processes*. In *Proceedings of the 2011 Symposium on Theory of Modeling & Simulation: DEVS Integrative M&S Symposium*. pp. 160-168.

[12] Friedenstab, J.-P., C. Janiesch, M. Matzner, and O. Müller. (2012). *Extending BPMN for Business Activity Monitoring*. In *Proceedings of the 45th Hawaii International Conference on System Sciences* pp. 4158-4167.

[13] Gagne, D. and A. Trudel. (2009). *Time-BPMN*. In *IEEE Conference on Commerce and Enterprise Computing*. pp. 361 - 367.

[14] Gao, F., M. Zaremba, S. Bhiri, and W. Derguerch. (2011). *Extending BPMN 2.0 with Sensor and Smart Device Business Functions*. In *IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises*. pp. 297 - 302

[15] Rodríguez, A., E. Fernández-Medina, and M. Piattini, *A bpmn extension for the modeling of security requirements in business processes*. IEICE transactions on information and systems. *90*(4). 2007, pp. 745-752.

[16] Goldner, S. and A. Papproth. (2011). *Extending the BPMN Syntax for Requirements Management*. In *Business Process Model and Notation*. pp. 142-147.

[17] Lodhi, A., K. Veit, and G. Saake, *An Extension of BPMN Meta-model for Evaluation of Business Processes*. J. Riga Technical University. *43*2011, pp. 27-34.

[18] Soffer, P., *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. 2010, pp. 14-25.

[19] Bringel, H., A. Caetano, and J. Tribolet. (2004). *Business Process Modeling Towards Data Quality Assurance*. In *6th International Conference on Enterprise Information Systems*. pp. 565-568.

[20] Heravizadeh, M., J. Mendling, and M. Rosemann. (2009). *Dimensions of business processes quality (QoBP)*. pp. 80-91.

[21] Lu, R., S. Sadiq, and G. Governatori, *On managing business processes variants.* Data & Knowledge Engineering. *68*(7). 2009, pp. 642-664.

[22] Klein, H.K. and R. Hirschheim, *Choosing Between Competing Design Ideals in Information Systems Development* Information Systems Frontiers. *3*(1). 2001, pp. 75-90.

[23] Silver, B., *BPMN Method & Style: A levels-based methodology for BPM process modeling and improvement using BPMN 2.0*. 2009: Cody-Cassidy Press.