# Process-Driven Data Quality Management Through Integration of Data Quality into Existing Process Models

## Application of Complexity-Reducing Patterns and the Impact on Complexity Metrics

The authors highlight two options to integrate data quality into existing process models: within-model integration and across-model integration. Within-model integration allows to enhance existing process models with data quality information by integrating data quality checks. Across-model integration provides a new process model with an information product-centric perspective, linking it to existing models. The authors examine the integration approaches' impact on the original models' complexity and patterns for complexity reduction. Gaps in extant research limit the assessment of process model complexity when integrating data quality. There are no generic patterns for reliably decreasing complexity of process models. Compacting and modularization have the highest potential to control complexity while integrating data quality into process models.

**DOI 10.1007/s12599-013-0297-x**

### The Authors

**Dipl.-Wirt.-Inf. Paul Glowalla**
**Jun.-Prof. Dr. Ali Sunyaev** (✉)
Information Systems and Information Systems Quality
University of Cologne
Albertus-Magnus-Platz
50923 Köln
Germany
glowalla@wiso.uni-koeln.de
sunyaev@wiso.uni-koeln.de

Received: 2012-01-26
Accepted: 2013-07-05
Accepted after two revisions by Prof. Dr. Buxmann.

This article is also available in German in print and via http://www.wirtschaftsinformatik.de: Glowalla P, Sunyaev A (2013) Prozessgetriebenes Datenqualitätsmanagement durch Integration von Datenqualität in bestehende Prozessmodelle. Anwendung von komplexitätsreduzierenden Mustern und ihr Einfluss auf Komplexitätsmetriken. WIRTSCHAFTSINFORMATIK. doi: 10.1007/s11576-013-0391-1.

**Electronic Supplementary Material**
The online version of this article (doi: 10.1007/s12599-013-0297-x) contains supplementary material, which is available to authorized users.



### 1 Introduction

Data quality is crucial to organizational success due to the increasing amounts and diversity of data processed by organizations (Madnick et al. 2009, pp. 2, 4; Otto 2011, p. 241; Glowalla and Sunyaev 2012, 2013). Poor data quality is estimated to cost a company 10–20 % of its revenue (Redman 2004). Data quality management is a major concern across organizations and is predicted to gain further importance in the light of increasing amounts and diversity of data, improved analysis capabilities, and business process integration (e.g., Capgemini 2013; Forrester Research 2011; Kurzlechner 2011). However, it is difficult to systematically assess costs that are caused by poor data quality since they depend on the context in which the data is used as well as on the impact of direct and hidden costs of operational and strategic activities and decisions (Haug et al. 2011, pp. 170, 188).

To assess and sustainably improve data quality within organizations, process-driven data quality management (PDDQM) techniques should be applied (Batini et al. 2009, p. 5). PDDQM aims at redesigning processes that create or modify data. Hence, data and data quality should be considered in the context of the business processes they are processed in (Ofner et al. 2012, pp. 1036–1037).

PDDQM requires process modeling (Batini et al. 2009, p. 16), which is widely used to increase awareness of and knowledge about business processes (Recker et al. 2010, p. 501). The increasing interest of researchers and practitioners in business process management leads to a proliferation of a wide range of process modeling languages (Ko et al. 2009; Recker et al. 2009). Each process modeling language emphasizes different aspects of processes (Recker et al. 2009, p. 335). Consequently, organizations may maintain hundreds of process models which are accessible to and customized by non-modeling experts (Rosemann 2006,



Published online: 08 November 2013

p. 254). Since process models are mainly used for communicating processes (Bandara et al. 2005, pp. 348, 353; Dehnert and van der Aalst 2004, pp. 289–290), the model's understandability to model readers is crucial.

Our objective is to examine possibilities to visibly integrate data quality across the plethora of existing process models, that is, instantiations of process modeling languages. We aim to support the communication and understanding of data quality in the context of business processes, especially since research on process modeling rather focuses on formal modeling aspects (Mendling et al. 2010, p. 127). If data quality requirements can be understood across stakeholders, for instance, by data collectors who are no experienced modelers, data quality can be improved throughout processes (Lee and Strong 2003; Rosemann 2006, p. 253). A prominent process-driven data quality perspective is to treat input data as raw material that is processed to the final information product (IP) (Thi and Helfert 2007, pp. 8–9; Wang 1998, p. 59; Sect. 2.1). However, different process models have different foci and – regardless of current process modeling languages (Sect. 2.2) – organizations apply process modeling languages in different ways. We apply a two-step approach, first aiming to answer the following research question.

RQ1: *What varying applications of process modeling languages for PDDQM can be derived from extant research?*

While integrating data quality, existing process models need to remain understandable. Therefore, complexity needs to be controlled. We address this issue in our second research question concerning the applications of process modeling languages which were identified by answering RQ1.

RQ2: *How can data quality be integrated into existing process models while simultaneously controlling model complexity?*

Based on Webster and Watson (2002), we conducted a keyword-based literature review in journals and conference proceedings (Sect. 3). Our contribution is threefold. First, we provide a synthesis of process models' varying applications for PDDQM. Second, based on the identified relevant literature, referred to as primary studies, we propose two approaches to integrate data quality into existing process models. Third, we assess the impact of integrating data quality on process model complexity and evaluate the applicability of extant complexity reduction patterns.

The remainder of the article is structured as follows. In Sect. 2, we introduce PDDQM, process modeling, and our focus on process model complexity. We describe our research methodology in Sect. 3. In Sect. 4, we answer RQ1 by presenting the primary studies structured with respect to the applied process models, and by identifying model characteristics including the integration of data quality within process models. We answer RQ2 in Sect. 5 by examining complexity reduction patterns to control model complexity when integrating data quality into process models. Finally, we discuss our results in Sect. 6, followed by a summary and outlook in Sect. 7.

## 2 Background

In Sect. 2.1, we explain the reasons for our focus on PDDQM and the role of process modeling. We provide two process modeling languages focusing on data quality to illustrate main differences and why we do not limit our analysis to process modeling languages which are data quality specific. In Sect. 2.2, we outline the relationship between complexity and understandability of process models. With respect to research streams in process model quality, we justify our focus on process model complexity, pragmatic guidelines, and according process model quality metrics.

### 2.1 Data Quality in the Context of Processes, Process Models, and Process Modeling Languages

Techniques to assess and improve data quality can be classified into data-driven and process-driven ones (Batini et al. 2009, p. 5). Data-driven techniques focus on direct modification of data, for example, cleansing, normalization, and integration of data. Therefore, processes creating and updating data are not modified. In contrast, PDDQM techniques focus on optimizing these processes by identifying root causes of errors, eliminating them, and sustaining the improvements (English 1999, pp. 289–301). If defective data are corrected without adjusting the underlying process, the process will continue to produce defective data. Therefore, we focus on process-driven techniques, which outperform data-driven techniques in the long-term.

Processes are logical sequences of tasks in which goods and services are created or where the creation is coordinated using resources (Buhl et al. 2011, p. 163). To emphasize the involvement of business stakeholders as process model users, we focus on "business and manufacturing processes that create, update, and delete data, distribute or disseminate information, and retrieve or present information to information producers and knowledge workers" (English 1999, p. 69). For simplicity, we continue to use the term process.

Process models provide the means to understand and communicate processes and thus are mandatory for conducting process control activities or process redesign (Batini et al. 2009, p. 16). Process models are instantiations of process modeling languages. Process modeling languages provide a vocabulary of model elements and compositional rules which define legal compositions of the vocabulary. A general meaning of the vocabulary's elements is also provided, but should not be confused with the semantics and meaning of the instantiation which relate to a specific (problem) domain (Lindland et al. 1994, pp. 44–45; Moody 2009, p. 757).

Building on Batini et al. (2009), we highlight two different process modeling languages which focus on data quality: First, information chain maps provided within the Cost-Effect of Low Data Quality (COLDQ) methodology (Loshin 2001) to model strategic and operational data flows. Second, information product maps (IP-MAP) to model the production of an IP.

Information chain maps provide generic steps to enable the conversion of raw input data into usable information. Strategic and operational data flows are based on generic steps (e.g., data supply, data processing), which are instantiated through specific processing steps (e.g., data entry, credit card processing). Besides annotations, no further information about the IP or its quality are integrated into the models. Information chain maps show the information flow and, similarly to data flow diagrams (DFDs) (Shankaranarayanan and Wang 2007), no explicit processing sequence can be derived. However, information flows are not depicted between processes but between stakeholders and systems.

Alternatively, processing stages of the data flow, from data supply to data consumption, can be presented in a process sequence. This principle resembles IP-MAPs.

The IP-MAP was introduced as an extension of the information manufacturing system (IMS) (Ballou et al. 1998; Shankaranarayanan et al. 2000), applying concepts from product quality in manufacturing systems. Additionally, the IMS is part of the total data quality management (TDQM) methodology (Wang 1998). The IP attributes and data units can be tracked systematically from the source to the final IP that is delivered to the consumer. Further, the impact of system modifications on the attributes can be analyzed. The design of the IP-MAP is driven by the requirements of the final IP (Shankaranarayanan et al. 2000). Therefore, the final IP provides the basis for the specification of necessary raw or component data. A major change – with respect to the IMS – is the definition of additional modeling elements, namely the decision block, the business boundary block, and the information system boundary block. A comprehensive description of the IP-MAP is provided in Lee (2006).

Both process modeling languages focus on information as a product. However, the illustration of data differs, as do the languages. In contrast to information chain maps, the IP-MAP focuses stringently on the delivery of a specific IP and on the necessary sequential steps to manufacture such an IP. Regarding the sequence, IP-MAPs resemble process flow charts (PFCs) despite their focus on the data flow (Shankaranarayanan and Wang 2007). Additionally, the necessary data and its sources are presented. 'Necessary' means that the presented data flow is limited to the purpose of producing the IP.

Due to the varying applications and customizations of process models, we are aware that data quality aspects might be integrated into other process modeling languages, that is, activity-centric modeling languages (Recker et al. 2009, p. 338). Therefore, we do not exclude instantiations from process modeling languages that do not focus on data quality, such as Petri nets, DFDs, business process modeling language (BPML), and business process modeling notation (BPMN) (Rosemann et al. 2009).

### 2.2 Managing Process Model Understandability and Complexity in the Context of Process Model Quality

When integrating data quality into existing process models, the challenge is to render the process models easy to understand for stakeholders, including novice modelers. Understandability of process models refers to the degree of which information contained in a process model can be easily understood by the reader (Reijers and Mendling 2011, p. 3). A process model is understood if the reader is able to explain the model (Figl and Laue 2011, p. 453). Extant research identifies several factors of model understandability, such as contextual factors (Rosemann et al. 2008), personal factors related to the model reader, and factors related to the model itself (Reijers and Mendling 2011, p. 1). To determine model understandability, complexity metrics, that is, measures concerning the ease or difficulty to understand a model (Laue and Gruhn 2007, p. 13), can be applied. Extant research provides various such measures, however, pragmatic guidelines for improving process models are lacking (e.g., Figl and Laue 2011; Reijers and Mendling 2011). We examine diverse process models, potentially instantiated from different process modeling languages, and analyze how data quality can be integrated. Due to the limited insights into the process modeling languages applied and the context the process models are used in, we focus on model-inherent factors. Therefore, we exclude personal and contextual factors. For the purpose of clarity, we use the term model complexity throughout this article when referring to model-inherent factors.

Regarding the research on process model quality that considers process models instantiations, three research streams can be distinguished: *quality frameworks*, *pragmatic guidelines, and process model quality metrics* (Mendling et al. 2010, p. 128). Current research (Overhage et al. 2012) considers prominent model *quality frameworks* and *process model quality metrics* for assessing process model quality. However, Overhage et al. (2012) mainly rely on the SEQUAL model (Lindland et al. 1994). Hence, a specific process modeling language, providing syntactic rules, and the context of the process model, providing semantic and pragmatic assessment, are necessary to assess the quality of process models. Similarly, the Guidelines of Modeling (GoM) (Becker et al. 2000) allow for an assessment of process model quality in a given context, depending, for instance, on the applied process modeling language and existing conventions. Another quality framework for measuring process model quality builds on software quality characteristics (Gucegliogl u and Demirors 2005). In this framework, however, understandability refers to the completeness of the model whereas complexity is subsumed under syntactic analyzability metrics.

Regarding *pragmatic guidelines*, Overhage et al. (2012) exclude the seven process modeling guidelines (7PMG) (Mendling et al. 2010) from their considerations. Although admitting the modeling style's importance (p. 231), they argue that syntactic and semantic correctness are more important than modeling style. Complementary to process model quality frameworks, pragmatic guidelines may improve existing process models and their complexity and thus their understandability (Mendling et al. 2010, p. 130). Context-independent guidelines and patterns allow improving process models without changing their underlying behavior, and relating guidelines and patterns to metrics allows for a context-independent assessment and control of model complexity. Based on previous research, Gruhn and Laue (2009), La Rosa et al. (2011a; 2011b), and Moody (2009) provide patterns for complexity reduction of process models. La Rosa et al. (2011b) additionally relate their patterns to *process model quality metrics*, facilitating measuring the changes when applying *pragmatic guidelines*.

We apply the aforementioned work to discuss the integration of data quality into existing models and its impact on process model complexity. In consideration of the wide range and varying application of process models, we do not focus on the complexity of specific modeling languages but on characteristics provided by the literature review.

## 3 Research Method

To identify the varying applications of process modeling languages and provide the basis for answering our RQs, we conducted a structured literature review. Focusing on literature dealing with process-driven data quality, we followed

the phases proposed by Webster and Watson (2002). The underlying conceptual framework (Sect. 3.1) and the methodology (Sects. 3.2 and 3.3) are described in the following.

### 3.1 Conceptual Research Framework

We apply the conceptual framework depicted in **Fig. 1** to answer our research questions. We start with RQ1 by examining the *application of process modeling languages* for PDDQM (Sect. 4.1). Following the concept-oriented approach of Webster and Watson (2002), we provide an overview of our primary studies focusing on the applied process models. The next steps are a detailed examination of the process models and the presentation of several data quality-specific and further prominent *process model characteristics* (Sect. 4.2).

As a basis for RQ2, we propose two general *integration approaches* from current literature on process model complexity (Sect. 5), that is, within-model integration (Sect. 5.1) and across-model integration (Sect. 5.2). The *impact on model complexity* is examined in the context of integrating data quality into process models. To counteract increasing complexity, we examine which *patterns for complexity reduction* can be applied, taking into account the existing process model characteristics.

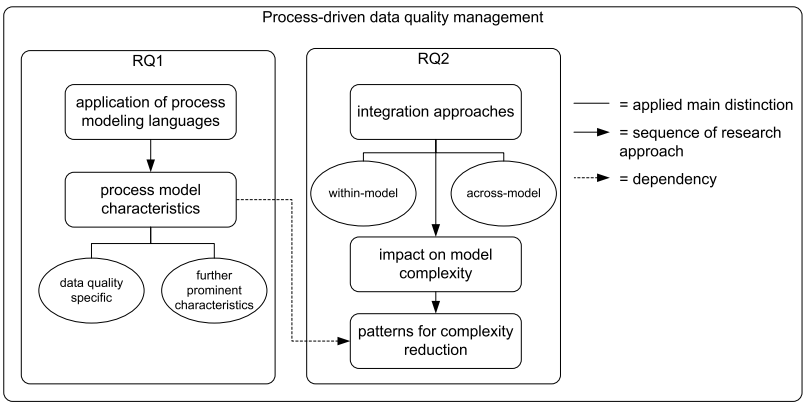


**Fig. 1** Conceptual research framework

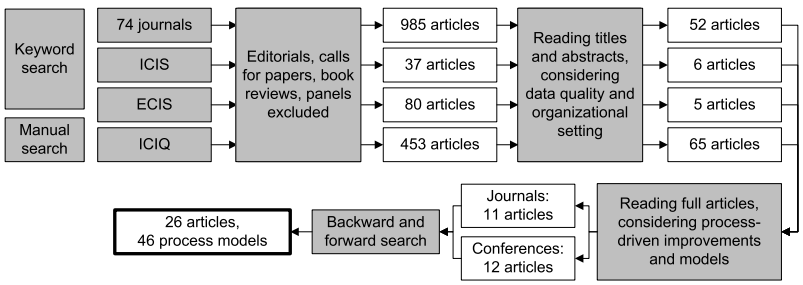


**Fig. 2** Article selection

### 3.2 Keyword and Manual Search

In the first phase we based our keyword search on the Senior Scholars' Basket[1] and the 50 highest ranked journals applying the AIS/MIS journal ranking.[2] Additionally, we included the International Journal of Information Quality and the ACM Journal of Data and Information Quality due to their focus on data quality. Regarding the ACM and IEEE transactions, which contain various journals, we conducted a manual selection. Overall, the approach led to 74 journals, listed in online Appendix A (available online via http://link.springer.com). We considered articles from 1995 onwards, since the already mentioned prominent perspective – to view data as a product – was proposed in that year. To allow a view on the latest developments and broaden the research towards more practice-oriented articles, we also considered three conferences. First, we included the International Conference on Information Systems (ICIS) and the European Conference on Information Systems (ECIS). Second, we included the International Conference on Information Quality (ICIQ) due to its relevance to our topic. Since the ICIQ proceedings are not accessible for a keyword-based search, they were searched manually. Likewise the ECIS proceedings before 2000 had to be searched manually. We derived the keywords based on an initial search, especially considering reviews and overview articles. The keywords were consolidated and supplemented (e.g., information product was supplemented by data product), leading to the following list: *data quality, information quality, data product(s), information product(s), data production, information production, data manufacturing, information manufacturing, data management, information management, data flow(s), information flow(s)*. The keywords were searched in title, abstract, and keywords/subject terms. The conduct of our review is illustrated in **Fig. 2**.

### 3.3 Inclusion and Exclusion Criteria

After having excluded editorials, calls for papers, book reviews, panels, the search yielded 1,555 articles. First, we read the titles and abstracts of the 1,555 articles. If abstracts were unavailable, we read the article in more detail. We considered an article relevant if it focused on data quality or at least on one of its quality dimensions. According to RQ1, the data quality aspect had to be established in an organizational setting, dealing with measures to assess or improve quality of the organization's data. We considered an article within an organizational setting if the measures to assess or improve data quality were conducted in the field and described in the context of the particular organizational setting. That means, we considered, for instance, case studies and case descriptions. We excluded articles in which results were presented isolated from the organizational setting (e.g., the presentation of lessons learned with short examples from conducted case studies or personal experience for corroboration). As to the inclusion criterion when dealing with organizational data, we explic-

[1] http://start.aisnet.org/default.asp?page=SeniorScholarBasket.
[2] http://start.aisnet.org/?JournalRankings.

itly considered data stored and processed by information systems. After this review, 128 articles were selected.

Furthermore, we only selected articles which included process-driven strategies and models. We employed the definition of processes or business processes respectively. Hence, we excluded information systems development processes (English 1999, p. 69) that did not provide business processes, for instance, by focusing on processes that are inherent to IT-systems (e.g., the optimization of data warehouse internal processes). These criteria direct our focus towards the organizational context where models are mainly used for communicating processes, including stakeholders that are non-modeling experts (Rosemann 2006). This review led to 23 articles. Whereas it is not surprising that eleven of the conference articles are from the ICIQ and one from the ECIS, journal articles are from several different journals. Three articles are from the International Journal of Information Quality, which is available since 2007. The remaining eight journal articles are from eight different journals.

We did not identify any further articles in the backward search that constitutes the second phase (Webster and Watson 2002). The forward search, constituting the third phase, led to 3 more articles. We chose 'Google Scholar', since it indexes conference papers in addition to journal papers. In total, the identified articles provide 46 process models.

**Table 1** Primary studies

| Article | IMS, IP-MAP | PFC | DFD |
|---|---|---|---|
| Balka et al. (2012) | – | X | – |
| Ofner et al. (2012) | – | X | – |
| Dejaeger et al. (2010) | – | X | – |
| Xie and Helfert (2010) | – | X | – |
| Gaynor and Shankaranarayanan (2008) | X | – | – |
| Hakim (2008) | – | X | – |
| Laumann and Rosenkranz (2008) | – | X | X |
| Lee et al. (2007) | X | – | – |
| Thi and Helfert (2007) | X | X | X |
| Shankaranarayanan and Cai (2006) | X | – | – |
| Keenan and Simmons (2005) | – | – | X |
| Mielke (2005) | – | X | X |
| Davidson et al. (2004) | – | – | X |
| Klesse et al. (2004) | – | X | – |
| Shankaranarayanan et al. (2003) | X | – | – |
| Katz-Haas and Lee (2002) | – | – | X |
| Kovac and Weickert (2002) | – | X | – |
| Wang et al. (2002) | X | X | – |
| Helfert and von Maur (2001) | – | X | – |
| Kahn et al. (2001) | – | – | X |
| Millard and Lavoie (2000) | – | – | – |
| Ballou et al. (1998) | X | – | – |
| Kovac et al. (1997) | – | X | – |
| Harkness et al. (1996) | – | X | – |
| Meyer and Zack (1996) | – | X | – |
| Zack (1996) | – | X | – |

IMS = Information Manufacturing System; IP-MAP = Information Product Map; PFC = Process Flow Chart; DFD = Data Flow Diagram

## 4 Application of Process Modeling Languages for PDDQM

This section deals with RQ1. We present our primary studies and the process models used for representing data quality aspects. Then we identify process model characteristics from our primary studies and present their application.

### 4.1 Primary Studies

As recommended by Webster and Watson (2002), we structure the articles according to our unit of analysis, that is, process models (**Table 1**), differentiating between PFCs and DFDs (Shankaranarayanan and Wang 2007). Shankaranarayanan and Wang (2007) provide a comparison of IP-MAPs to other modeling languages concerning a possible substitution or complementation of the IP-MAP. In this context, PFCs represent the sequence of process steps without data flows, and DFDs represent data flows without the sequence of process steps. We apply this simplified categorization, since organizations utilize enhanced models and the use of process modeling languages varies. For instance, in practice process models are enhanced (e.g., Katz-Haas and Lee 2002; Lee et al. 2007) or combined (e.g., Davidson et al. 2004; Mielke 2005) to represent data quality within processes. Even differentiating between DFDs and PFCs is not distinct in every case. For example, PFCs may contain further elements such as databases or repositories without presenting data flows (e.g., Meyer and Zack 1996; Helfert and von Maur 2001). Despite providing tasks and data flows or control flows, process models may specifically focus on other aspects (e.g., the viable system model allows the analysis of functions, responsibilities and management requirements within organizational systems (Laumann and Rosenkranz 2008), and the cause-and-effect-diagram is concerned with process improvement, focusing on causes and effects (Ishikawa 1993; Klesse et al. 2004)). Only Thi and Helfert (2007) provide an approach which aims to integrate IMS with other process modeling languages. However, while focusing on representation of dynamic changes to IPs, previously included data quality checks are discarded. Our detailed categorization of each process model from the primary studies is provided in online Appendix B. The respective process model characteristics are examined in the next section with regard to the applied process model.

### 4.2 Process Model Characteristics

The process models used in the primary studies can be classified accord-

**Table 2** Process model characteristics

| Process model | No. of models | Swim lane | Time axis | Sequence | Quality check | Quality dimension | Quality metrics |
|---|---|---|---|---|---|---|---|
| Information Manufacturing System (IMS); Information Product Map (IP-MAP) | 14 | 6 | 3 | 14 | 10 | – | – |
| Process Flow Chart (PFC) | 20 | 10 | 1 | 20 | 5 | 1 | 1 |
| Data Flow Diagram (DFD) | 12 | 1 | – | 8 | – | 2 | 1 |
| Total | 46 | 17 | 4 | 42 | 15 | 3 | 2 |

ing to process model and data quality characteristics as well as to the underlying process modeling language (**Table 2**). Additionally, we consider the sequence of process models, since the sequence is important for differentiating between PFCs and DFDs. The time-axis and swim lanes are interrelated with the sequence and data quality checks and additionally provided as enhancements to the IP-MAP (Lee et al. 2007). Several articles provide more than one process model (**Table 1**), therefore 46 process models are examined.

Swim lanes are applied throughout several process modeling languages. In the IP-MAP, they represent stakeholder groups involved in the IP process (Lee et al. 2007). In PFCs and DFDs swim lanes refer to departments (Kovac and Weickert 2002; Mielke 2005), internal as well as external stakeholders (Dejaeger et al. 2010; Harkness et al. 1996; Kovac et al. 1997; Kovac and Weickert 2002; Ofner et al. 2012), specific roles (Klesse et al. 2004), systems and databases (Helfert and von Maur 2001; Kovac et al. 1997), and tasks (Harkness et al. 1996; Kovac and Weickert 2002). Examples for possible variations within one PFC are provided by Harkness et al. (1996), Kovac et al. (1997), Kovac and Weickert (2002), and Dejaeger et al. (2010). The swim lanes include stakeholders, tasks and products. In one of these cases, swim lanes are applied in rows and columns, constituting a matrix with stakeholders and tasks (Kovac and Weickert 2002). Hence, the process model includes additional information, but it is not possible to add a time axis.

Further characteristics are the time axis and the logical sequence of steps. A time axis shows the time needed to conduct processes or process steps. The logical sequence shows the logical flow of the steps regarding predecessor and successor relations. The time axis is usually represented by the X-axis and shows the flow of the process (in- or excluding data) from left to right (Mielke 2005; Lee et al. 2007). In this context, the sequence of the process steps is also defined. However, most models do not provide a time axis, but the sequence of the process steps.

All PFCs provide a sequence (e.g., Hakim 2008; Kovac and Weickert 2002; Klesse et al. 2004). This observation conforms to the definition of PFCs. From most DFDs, the sequence of the processes can be derived as well (e.g., Davidson et al. 2004; Keenan and Simmons 2005), although this is not inherent in the process modeling language. In several cases DFDs and PFCs seem to have been combined, impeding a clear distinction (e.g., Davidson et al. 2004; Keenan and Simmons 2005; Mielke 2005; Thi and Helfert 2007).

Although the presented models were applied within projects to assess or improve data quality, the integration of data quality directly into the models is rare (**Table 2**). However, it can be observed that data quality is integrated not only with the help of IP-MAPs, but also across PFCs and DFDs.

In the presented process models, employed data quality elements are mostly data quality checks integrated as process steps (e.g., Millard and Lavoie 2000; Thi and Helfert 2007; Zack 1996) or attached to process steps (Helfert and von Maur 2001). In one case, data quality checks are integrated as a swim lane, since the process deliverable is jointly agreed upon between two parties regularly (Harkness et al. 1996).

Within four models, data quality checks are not specific model elements but tasks determining that data quality checks take place (Harkness et al. 1996; Millard and Lavoie 2000; Zack 1996). Without data quality specific information provided beyond the model, such tasks emphasize the importance of the visibility and communication of data quality checks.

Data quality checks within IMS or IP-MAPs are specific model elements which can be referred to the according meta data (Ballou et al. 1998; Shankaranarayanan et al. 2003, 2000). Further sophisticated approaches exist, which contain process models and included data quality checks. Kovac et al. (1997) derive metrics for the data quality dimensions timeliness and accuracy, focusing on process hand-offs between stakeholders. Beside single tasks for checking data quality, defined data quality measures between process hands-offs are indicated by specific arcs. Helfert and von Maur (2001) annotate modeling elements in a data delivery process. The numbered annotations refer to verbalized data flow processes and are used to relate data quality dimensions to data quality indicators, and measuring points to the process elements. Ofner et al. (2012) provide a formalized meta model for assessing data quality within a process across different tasks.

Two other approaches visibly integrate data quality into single process models without using data quality checks. Katz-Haas and Lee (2002) focus on timeliness since a process' cycle time leads to delayed information provision thus causing high costs. To visualize why information does not arrive in a timely manner, they enhance a process model, assigning timestamps to process steps. Without using data quality checks, Mielke (2005) provides quality dimensions and metrics to measure data quality within process models. Moreover, only Mielke (2005) integrates data quality across several models, without applying IP-MAPs. An abstract model provides an overview of the most important data quality dimensions for the main processes and departments. A more detailed level shows the sub-processes and their IP inputs and outputs. Data quality of each sub-process is determined, using weighted key performance indicators, based on the most important data quality dimensions. The performance of the sub-processes adds up to the process performance. The overall degree of data quality performance

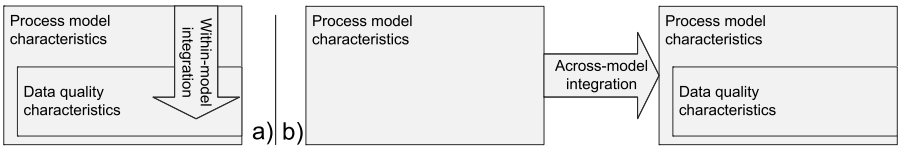


**Fig. 3** Data quality integration into process models

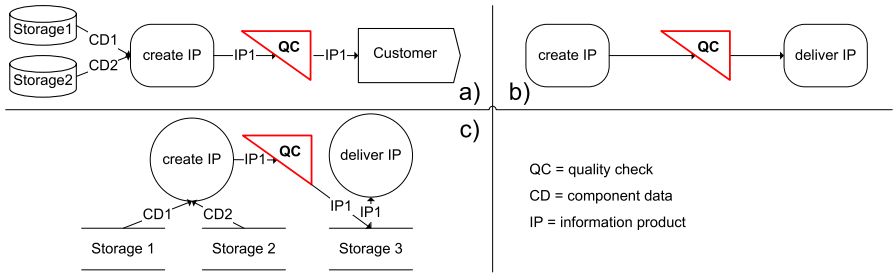


**Fig. 4** Integrating quality checks within (**a**) IP-MAPs, (**b**) PFCs, and (**c**) DFDs

is calculated from data quality across processes.

Regarding our primary studies and the lack of data quality integration, we examine the integration of data quality into process models. Emanating from existing process models, we analyze the impact on model complexity in the next section.

## 5 Data Quality Integration and Process Model Complexity

The primary studies indicate that data quality can be integrated into existing (e.g., Gaynor and Shankaranarayanan 2008; Helfert and von Maur 2001) or new process models (e.g., Lee et al. 2007; Mielke 2005). Current literature also considers complexity within single models and across multiple models and provides respective complexity metrics (La Rosa et al. 2011b). Hence, we adopt this perspective to examine the impact on process model complexity and propose two approaches to represent data quality for PDDQM: within-model and across-model integration (**Fig. 3**).

Within-model integration (Sect. 5.1) is the integration of data quality into existing process models by adding data quality elements. For instance, quality checks may be added as additional tasks. Furthermore, other characteristics, for instance, swim lanes, can be applied with a data quality focus (Harkness et al. 1996). Therefore, the integration of data quality into an existing process model is embedded into the existing characteristics (**Fig. 3a**). Across-model integration (Sect. 5.2 and **Fig. 3b**) considers the representation of data quality in a new process model. In this case, it is necessary to link the process models to maintain their relation.

### 5.1 Within-Model Integration

Data quality characteristics can be integrated into process models with minor changes to the original layout – the visual arrangement of elements – to maintain clarity (Becker et al. 2000, pp. 32–33). Although the enhancement increases the model's complexity, stakeholders familiar with the original process model do not have to cope with a completely new layout or model.

As aforementioned, Harkness et al. (1996) integrate data quality by using a swim lane. We doubt that swim lanes are generally adequate for integrating data quality into existing models due to the impact on the model layout. Replacing existing swim lanes with data quality swim lanes would lead to a new layout and loss of information. Adding swim lanes for reoccurring quality checks (Harkness et al. 1996, p. 360) results in a new layout and additional complexity, since the swim lanes would contain different content types. Alternatively, additional swim lanes might be integrated, leading to a process matrix (Kovac and Weickert 2002, p. 71), which would impede the representation of process sequences. If swim lanes are not provided in the original model, the process model can be enhanced, again with the entailed need to adjust the layout. We argue that similar results can be achieved with data quality elements integrated into existing process models without a loss of information or the need to restructure the model.

Supported by the application in the examined models, we consider integration of data quality by using data quality checks. For facilitating clarity (Becker et al. 2000, pp. 32–33) and thus the understanding and communication of changes within the model, we see the need to use a specific modeling element for data quality checks. Such a specific element allows differentiating between transformation and validation processes in order to produce a flawless IP (Shankaranarayanan et al. 2000, 2003). In contrast to the IP-MAP, the data quality check element is applicable within non-IP-centric models and its purpose therefore differs from extant tasks, which might provide IPs as a by-product. To ensure a differentiation from rectangles or ovals that are used as tasks, we use triangles for the visualization of data quality checks. Additionally, the symbol may be familiar for stakeholders due to its usage in IP-MAPs.

Data quality checks can be applied as process steps within IP-MAPs, PFCs, and DFDs (**Figs. 4a–4c**). The data quality check denotes clearly when a data quality related activity is conducted and implicates that context-dependent data quality should be defined, measured, analyzed and improved (Wang 1998). However, the properties of the data quality check at the meta level and syntactic rules for its usage need to be defined within the given context. The meta data might range from a (1) meta model, defining the content and syntactic use of the element over (2) conceptual models focusing on encapsulated generic steps to define and measure relevant data quality dimensions to (3) textual annotations. Data quality checks can be used after extant tasks to validate the quality of entailed IPs. However, they may also consider further existing data, providing a data quality check before subsequent process steps are conducted. Moreover, data quality checks might even be placed at the beginning of a process model, for instance, if

an IP should be checked by other stakeholders after a process hand-off. Due to these reasons, we do not further limit the potential context-dependent application of the data quality check element.

As shown in **Fig. 4**, syntactic integration of data quality checks can be conducted similarly across process models. However, process models' different foci have to be considered: DFDs and PFCs, especially since we consider already existing models, do not focus on single IPs and their processing. For instance, although focusing on data flows, DFDs are not necessarily able to provide IPs and the component data they are based on. IP-MAPs focus on a specific IP and its manufacturing process (Thi and Helfert 2007). Although more than one IP can be included (Shankaranarayanan et al. 2003), the IP-MAP and its elements, for example, databases, pre- and post-conditions of IPs, belong to specific IPs. To integrate the manufacturing of specific IPs, the process model needs to focus on IP production or a new process model becomes necessary.

### 5.2 Across-Model Integration

As pointed out above, integration of data quality might require a new process model. Due to lack of metrics to measure the complexity of control and data flows within one model, we examine data quality integration into IP-centric process models, that is, PFCs providing the manufacturing of an IP. Extant research considers model complexity only across models of the same process modeling language. In our case, across-model integration refers to the integration of data quality within different process model types, integrating IP-centric process models with PFCs and DFDs. We provide an approach which has low impact on existing non-IP-centric process models.

We consider the IP as the element that links the original model with the new IP-centric model. DFDs represent data and its flow. Therefore, IPs can be related to presented data. In PFCs, even if not explicitly provided, IPs can be referred to within process steps without changing the process model's structure. Although the PFC in **Fig. 5** does not focus on IP production, it is possible to link any process step with an IP-centric process model and utilize the manufacturing process to provide the IP. For instance, the 'stock records' (**Fig. 5**, IP1) can be linked to the IP-centric process model provided in **Fig. 4a**.

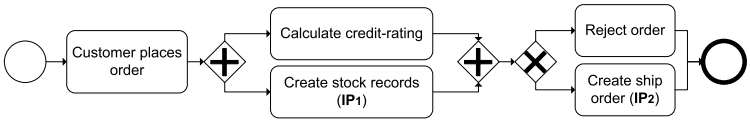


**Fig. 5** Linkage with implicit IP

**Table 3** Integration approaches' impact on complexity metrics of the existing process model

| Metrics | Within-model integration | Across-model integration |
|---|---|---|
| Number of nodes | + | ∘ |
| Number of arcs | + | ∘ |
| Number of tasks | + | ∘ |
| Number of models | ∘ | + |
| Repository size | + | + |
| Average connector degree | ± | ∘ |
| Maximum connector degree | (+) | ∘ |
| Connectivity | ± | ∘ |
| Density | − | ∘ |
| Control flow complexity | (+) | ∘ |
| Cross-connectivity | ± | ∘ |
| Fan-in | ∘ | + |
| Fan-out | ∘ | + |
| Separability | ± | ∘ |
| PST distance | (+) | ∘ |
| Diameter | (+) | ∘ |

∘ = not affected; ± = may increases or decrease; + = increase; (+) = potential increase; − = decrease; (−) = potential decrease

Should existing models already represent IP production and further IP-centric process models be developed, across-model integration may also focus on process models of the same type.

### 5.3 Impact of Integration Approaches on Model Complexity

We examine process model complexity based on related metrics, as accumulated by current research (Figl and Laue 2011; La Rosa et al. 2011b; Reijers and Mendling 2011). We include metrics that change due to data quality integration. For instance, the diameter of a process model may change as a result of the integration of a data quality check. We exclude complexity metrics which concern the nesting of process models of the same type. Our integration approach assumes an existing model into which a data quality check, consisting of a node and two connecting arcs, will be integrated. Accordingly, we further exclude metrics based on connectors. For instance, including data quality checks within a process model will typically not have an impact on the 'token splits' (Reijers and Mendling 2011, p. 5) within a model. However, a complex data quality check with different paths that can be concurrently initiated, could change the metric 'token splits'. Online Appendix C provides the excluded complexity metrics, while online Appendix D provides included complexity metrics, their definitions and according literature.

**Table 3** summarizes the impact of our data quality integration approaches on the process model metrics of existing models. The number of nodes, arcs, and tasks increases when a data quality check is added to a model. If a new model is used, the number of models increases. However, for integrating the new model with respect to the IP, the original model does not require further nodes, arcs, or tasks (**Fig. 5**). The repository size increases in both cases, since it refers to the

summarized size of all models, based on the number of nodes.

Adding new nodes, arcs, or tasks also affects other metrics. Maximum connector degree may increase, depending on where new elements are integrated into the original model. However, average connector degree may decrease, although this does not mean that the model is easier to understand. In contrast, the higher the average connector degree, the more this metric will decrease if new connectors with few incoming and outgoing arcs are integrated. Therefore, a decreasing average connector degree might be an indicator for a rather complex model. Assuming only one arc and node are added, connectivity increases if the minimal number of arcs is given in the existing model ('number of nodes' − 1). Otherwise, connectivity decreases. Density will decrease. Density is the ratio of the number of arcs to the number of maximum of arcs when all nodes are directly connected. With the exception of very small models with less than three nodes, the maximum number of arcs will increase over-proportionally with each node added. The cross-connectivity as "sum of the connectivity between all pairs of nodes in a process model, relative to the theoretical maximum number of paths between all nodes" may change depending on where new elements are included, that is, depending on the weight of the existing and established connections between nodes (Vanderfeesten et al. 2008, p. 482). Control flow complexity may not rise, since no new connectors are added. However, since control flow complexity is the weighted sum of connectors, the weights might increase if incoming and outgoing arcs increase and therefore the complexity of connectors. Separability might rise if new nodes serve as cut-vertices and decrease otherwise. Diameter and PST distance might increase if a new node is added in the longest path (diameter) or in a block with the highest element interactivity. Regarding across-model integration, fan-in and fan-out increase, since references from the original to the new model increase and vice versa.

Only few metrics consider complexity between models at the same level, that is, if the models are not nested. Especially measuring the impact of different modeling languages with different foci is difficult or rather impossible. For instance, references between two models (e.g., fan-in) can be measured. Yet, linking models of different process modeling languages is more complex than linking models instantiated from the same language. Furthermore, a new model type adds new modeling elements that a model reader needs to understand. Complexity measures for the number of different elements within a model (e.g., 'connector heterogeneity') exist (Reijers and Mendling 2011). However, more knowledge regarding the impact of complexity across models would facilitate the decision if, instead of developing a new model, more elements should be integrated into an existing model.

### 5.4 Patterns for Complexity Reduction

La Rosa et al. (2011b) propose twelve patterns for complexity reduction that optimize complexity metrics. However, their applied metrics only constitute a subset of the metrics applied in this study. Thus, we examine the patterns in more detail and assess the impact on all metrics relevant for our data quality integration approaches. Moreover, we evaluate the patterns' interdependencies concerning process model characteristics. In the following, we first exclude the patterns irrelevant for our integration approach. Second, we examine each remaining pattern's impact on the identified process model characteristics.

Referring to connectors, the *block-structuring* pattern (La Rosa et al. 2011b) and *removing unnecessary gateways* (Gruhn and Laue 2009, p. 340–341) are beyond the focus of our proposed data quality integration. Furthermore, the patterns *restriction* and *extension* (La Rosa et al. 2011b) consider syntax restrictions or extensions of specific process modeling languages. The remaining patterns, which deal with removing useless and redundant arcs and nodes, are not relevant since we do not intent to include superfluous arcs or nodes. However, we consider *necessary redundancy* if equal tasks are placed within different swim lanes (Gruhn and Laue 2009, p. 342) or elements are duplicated to decrease process model complexity. The patterns considered in Moody (2009) are included and described in more detail in La Rosa et al. (2011b).

The patterns within the first two pairs (**Table 4**) can be reversed through the application of the other pattern in the respective pair. Consequently, the impact on metrics through the application of patterns can also be reversed. However, depending on the observed metrics, the application of a pattern on an existing model has no clear positive or negative impact, that is, the impact depends on the existing process models. This might also be a reason why La Rosa et al. (2011b) do not provide the impact on all their considered metrics.

Before applying patterns to decrease model complexity, the desired aspects of data quality need to be integrated into existing, new models, or model repositories to capture data quality processes. Integrating new elements generally leads to increased complexity. Depending on the approach chosen, the impact on the metrics (**Table 3**) needs to be monitored. Subsequently the application of the patterns allows to adjust metrics and therefore to manage complexity while integrating data quality.

*Duplication and Compacting* Duplication of model elements aims at simplifying model structure, whereas compacting removes redundant or superfluous model elements (La Rosa et al. 2011b). Compacting reverts the effects of duplication and vice versa. Removing superfluous elements is not relevant within our focus, since superfluous elements need to be considered within a given context.

When including data quality checks into an existing model, congruent data quality checks may be 'compacted' to decrease the number of tasks and model size as depicted in **Fig. 6**. Placing data quality checks into different swim lanes, in order to represent different resources (e.g., systems, stakeholders), limits compacting (Gruhn and Laue 2009, p. 342). Conversely, duplication may be necessary to include swim lanes.

Compacting bears the risk of increasing the model structure's complexity due to the need to reroute arcs within the model to remaining representative elements. Besides potential impacts on connectors and according metrics (e.g., separability, structuredness), the layout of the model tends to become more complex (e.g., due to crossing arcs). Consequently, the changes in structure and layout will have a negative impact on the sequence's understandability as an essential characteristic of process models. We use the term understandability instead of complexity since the changes in the layout go beyond the impact on the considered metrics. At the same time, applying

**Table 4** Dependency of patterns and process model characteristics

| Pattern (pairs) | | Within-/ across-model integration | Data quality (check) integration | Swim lane | Sequence |
|---|---|---|---|---|---|
| Duplication | Compacting | Within | Duplication/merging of conceptually congruent data quality checks | Limited compacting if data quality checks are processed by different resources | Duplication increases understandability of sequence |
| Modularization | Composition | Across | Extract complex or redundant data quality checks into new models | Possibility to keep information from former disjoint models | Modularization increases understandability of sequence |
| Merging | – | Across | – | Possibility to keep information from former disjoint models | Merging decreases understandability of sequence |
| Omission | – | Within | Reverse pattern for data quality integration | Omission and Collapse increase understandability | |
| Collapse | – | Within | Merging of data quality checks with tasks, merging of conceptually non-congruent data quality checks | | |

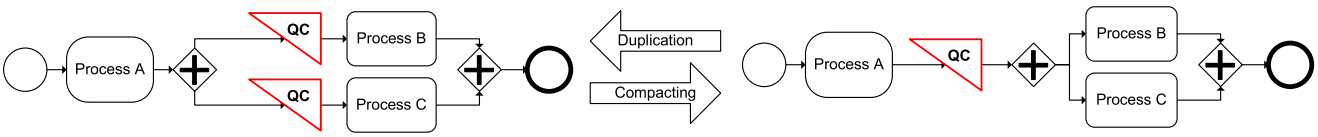


**Fig. 6** Application of the duplication and compacting pattern

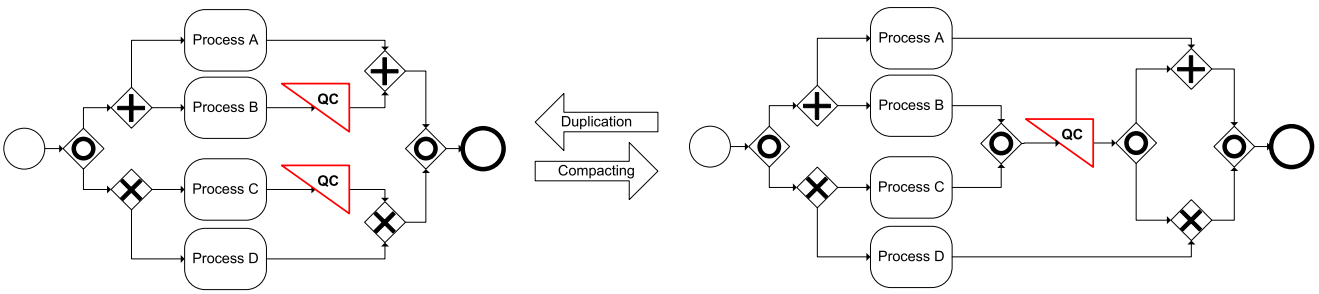


**Fig. 7** Application of the duplication pattern while decreasing the number of nodes and arcs

the compacting pattern, the model size should be reduced (La Rosa et al. 2011b). **Figure 7** refutes this general assumption. The duplication introduces a new data quality check while removing two gateways. Additionally, the duplication decreases the number of arcs, the repository size, and the diameter.

Since the number of nodes and arcs might increase or decrease, the derived metrics may increase or decrease as well (e.g., repository size, diameter, connectivity, density). Additionally, due to structural model changes, further metrics may increase or decrease (e.g., separability).

The impact on the metrics due to the application of this pair of patterns shows two important issues. First, although duplication is applied to improve model structure, related metrics might be impaired and therefore need to be controlled to mitigate undesired effects. Second, the impact of duplication and compacting on complexity is not generally predictable.

*Modularization and Composition* La Rosa et al. (2011b) identify three types of modularization. All these types have the same effect on our relevant metrics since they partition models into smaller parts or subsystems to avoid cognitive overload (Moody 2009, p. 767). In contrast, composition aims at consolidating disjoint models.

Modularization can be applied to extract data quality checks into a new model – whether as smaller parts at the same level or as subsystems. Modularization makes it possible to extract complex structures into a new model, additionally – but not necessarily – eliminating redundancy. Therefore, modularization is an alternative to compacting if compacting is ineffective due to complex data quality checks or complex rerouting of arcs. In such a case, besides the impact on metrics within and across models, the understandability of the sequence is a relevant factor. Although metrics measuring the structure might be impaired (e.g., separability by extracting a series of cut-vertices), the overall impact on

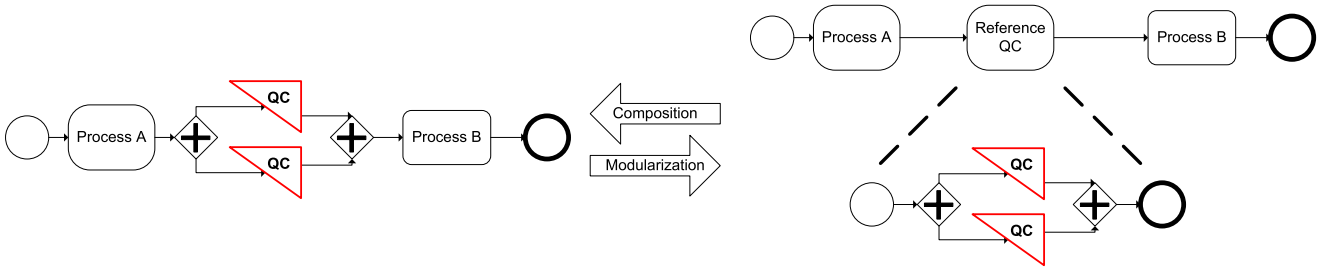


**Fig. 8** Application of the modularization and composition pattern

the sequence will be positive. This effect demonstrates that metrics have to be considered conjointly and that in this case, decreasing the number of nodes – and therefore the model size – might be more important. Extracted parts or subprocesses might be placed in swim lanes to provide related resources and might be referred to by several other models. Swim lanes can also be used in composed models to preserve information from formerly disjoint models.

Modularization decreases several metrics (e.g., number of nodes, arcs, tasks and repository size) within and across models if redundant model parts (within one model or from different models) are extracted. Otherwise, the number of nodes, arcs, and tasks will increase across models since a node and related arcs have to be inserted into the original model (**Fig. 8**). Hence, the impact on metrics might also be an indicator which parts to extract. Besides maximizing the extracted redundant elements, the impact on the structure has to be considered. For instance, the impact on separability should be taken into account in order to improve the ratio of cut-vertices. Whereas some metrics can only decrease through modularization (e.g., PST-distance, maximum connector degree, diameter), other metrics can increase or decrease (e.g., average connector degree, connectivity), depending on the extracted parts. Again, it is necessary to assess interdependency of and impact on the metrics to decide which metrics to improve.

*Merging* Merging is applied to consolidate process model variants (La Rosa et al. 2011b). Therefore, the impact on metrics depends on the merged models' similarity and cannot be compared to just one of the original models. Merging is an alternative to composition if similar process models exist. The possibility to merge similar models fosters modularization allowing extraction of similar process parts while containing the number of new models. Since the information in all former models has to be preserved, the understandability of the sequence will decrease.

The integration of data quality into a model could inhibit merging since data quality integration might be conducted only within one variant or differently across variants. Therefore, difference between variants would increase and thus reduce the benefit of merging models. In contrast, integrating data quality into a model will rather facilitate development of different model variants. However, there is no corresponding pattern related to the complexity metrics. Managing process variants while avoiding redundancy is addressed in Weber et al. (2011) by means of example process models.

Merging will generally lead to a more complex model since all variants have to be represented. However, several metrics, which are based on the ratio of elements, will increase or decrease, depending on the original process model they are compared to. For instance, a simple process model can be merged with a complex one. In such a case, the complex model may hardly change and metrics such as the average connector degree may decrease. In contrast, the complexity increases when compared to the original simple process model.

*Omission and Collapse* Omission of modeling elements implies information loss, whereas collapsing implies information synthesis (La Rosa et al. 2011b). Both patterns are applied to provide models for a specific purpose or audience and they therefore do not contain irrelevant information.

The omission pattern is the reverse pattern to the initial inclusion of the data quality checks. The collapse pattern supports merging data quality checks with preceding or subsequent tasks. In this case, omission and collapse have equal impact on the considered metrics, although the information provided within the model may differ. Furthermore, collapsing can be used to merge non-congruent data quality checks. The impact on the metrics is comparable to the compacting pattern, although different data quality checks are synthesized.

Omission and collapse might be applied to discard irrelevant information. Both patterns have a simplifying impact on complexity since the impact on the metrics of the original model is comparable to the modularization pattern without introducing a new model.

## 6 Discussion

Based on our literature review, we have identified prominent characteristics that are applied throughout the examined process models. Building upon these characteristics, we propose our integration approach focusing on data quality checks, which allows enhancing existing process models. By enhancing or introducing IP-centered process models, the sequence of process steps focuses on IP production. Differentiating between within- and across-model integration has two benefits. First, some metrics do not change if data quality is integrated only into existing models. Second, patterns can be differentiated according to if they have an impact on existing or on potentially new models. We identify the impact of integrating data quality within or across models and show applications of patterns within and across models. On this basis, we facilitate matching integration approaches and patterns to provide guidelines for complexity reduction.

### 6.1 Process Model Complexity Metrics

Based on our proposed integration approach, we see three issues in current

research on process model complexity metrics:

- Lacking complexity metrics
- Ambiguous relationship between complexity metrics and understandability
- A lack of knowledge about interdependency of metrics

The proposed approach shows diverse potential impact on complexity metrics (**Table 3**), which impede prediction of model complexity resulting from data quality integration. Furthermore, the metrics do not allow for measuring following aspects of model complexity and thus additionally impede complexity prediction: (1) Complexity of using control and data flow, that is, different types of sequences, within one model, (2) complexity of using different model types, (3) complexity of process model characteristics (swim lane, time axis).

Process models with a control and data flow are supported by BPMN and therefore used in practice. Generally, the differentiation of arcs allows for the integration of an IP-centric process flow within an existing model. However, we have not been able to detect metrics measuring the impact on complexity. We propose an alternative, that is, developing and linking a new IP-centric model, but the impact on complexity has to be further examined. Regarding our identified process model characteristics, no metrics are provided to measure the impact of a swim lane or time axis on complexity. An alternative is to assess the impact on understandability. An approach which assesses understandability of alternative process models (e.g., La Rosa et al. 2011a) can be used to initially assess modeling alternatives resulting from the above mentioned aspects. However, guidelines for improving model understandability should be linked to complexity metrics to allow for objective improvement of process models.

Research on the development of process model metrics – and especially their impact on process model complexity and understandability – is in its early stages (Vanderfeesten et al. 2008, p. 492; Reijers and Mendling 2011, p. 1). An issue of the evolving state-of-the-art regarding process model complexity and metrics is that no common set of metrics exists. Currently, Reijers and Mendling (2011) are continuing empirical research on process model metrics to identify metrics that significantly impact model understandability. Their research demonstrates the interdependency and ambiguity of complexity metrics. Reijers and Mendling (2011, 2007) find the impact of density on process model understandability significant; however, they kept model size constant to explicitly examine metrics beyond model size to arrive at these results.

Since understandability of process models over-proportionally decreases with their size (Reijers and Mendling 2011, p. 3), we assume that the impact and relevance of metrics will strongly depend on the model size. Within the context of our research, further unexamined interdependencies are important to provide guidelines for the application of the presented patterns. Additional knowledge about complexity metrics interdependencies would support the interpretation of how model complexity and understandability are influenced depending on metric values and their changes. Furthermore, identifying important metrics and relationships might lead to a reduced set of meaningful metrics.

### 6.2 Patterns for Complexity Reduction

Potential improvements of across-model metrics are available (e.g., La Rosa et al. 2011b; Weber et al. 2011). However, despite patterns that show impact across models, state-of-the-art empirical research on process model metrics focuses on the complexity and understandability of single process models (e.g., Figl and Laue 2011; La Rosa et al. 2011a; Laue and Mendling 2010; Reijers and Mendling 2011). Within the plethora of potential changes in process models and entailed impacts on metrics, patterns might provide comprehensible steps to decrease model complexity. Applying duplication might increase or decrease the number of nodes. However, the pattern should be applied to improve model structures despite increasing the number of nodes (La Rosa et al. 2011b). Modularization may be applied to decrease the number of nodes of a complex model or to extract redundancy from several models. However, a trade-off might be necessary to decide if to decrease complexity of single large models or to extract redundancy from several models.

We argue that current research does not provide generic patterns to reliably decrease complexity of process models. First, we showed that extant research examines patterns' impact on a small subset on metrics although several changes to metrics occur. This limited view may be attributed to the intention to foster understandability of patterns. Second, research on potential pitfalls and their impact on complexity is missing. We see the need to control changes in process model complexity to avoid undesired changes since the effect of repeated or even combined applications of patterns is not predictable. In extant research, for instance, structural problems that could arise in the modularization and composition are addressed (Basu and Blanning 2003), and process clones within a process model repository can be identified automatically to reduce repository size (Uba et al. 2011). However, applying patterns and controlling the presented changes in metrics is limited due to missing tool support (La Rosa et al. 2011b).

Referring to our integration approach, compacting and modularization have the highest potential to reduce complexity. Congruent data quality checks can be compacted, reducing the number of nodes. The applicability is limited by the use of swim lanes and a potential decrease in the understandability of the sequence. Regarding modularization, especially with respect to our integration of data quality checks into several models, it is likely that redundant parts are included. Furthermore, complex data quality checks comprising multiple activities can be extracted. To extract data quality specific aspects resembles orthogonal modularization (La Rosa et al. 2011b). Similar to cross-cutting concerns as security and exception handling, data quality checks may be scattered throughout processes. Approaches such as aspect-oriented modeling might be applicable to extract reoccurring data quality aspects, including related data quality dimensions and metrics, into new models or swim lanes (e.g., Cappelli et al. 2009).

### 6.3 Beyond Visible Data Quality Elements in Process Models

Our approach focuses on context-independent measurement of complexity. With data quality checks, which are integrated into different process models, we enable data quality integration into (non-)IP-centric process models without causing fundamental changes. As we seek context-independency, we have not provided a formal process modeling language. However, for visualizing our approach, some design suggestions were inevitable. For visualizing data quality

checks, we use a triangle relying on the IP-MAP. We assume that this choice supports the clarity of the data quality element within process models. However, clarity is extremely subjective (Becker et al. 2000, p. 33) and, as aforementioned, depends on the given context. For instance, the annotations used by Helfert and von Maur (2001) provide an alternative visualization, especially when already applied in existing models. However, when data quality checks are represented by annotating tasks, the sequence of the annotated task and data quality check cannot be derived. At the level of process modeling languages, the potential negative impact on ontological clarity should be considered if a process modeling language (e.g., BPMN) provides high coverage and thus potential use of redundant modeling elements (Recker et al. 2010; Wand and Weber 1993, 1995).

Our primary studies show different approaches to relate data quality requirements and measures to process models. Besides adapting the IP-centric IMS (Ballou et al. 1998) or IP-MAP (Lee 2006; Shankaranarayanan et al. 2000, 2003) and according meta data, further approaches exist which relate data quality meta data to process models, not necessarily using data quality checks (Helfert and von Maur 2001; Kovac et al. 1997; Mielke 2005; Ofner et al. 2012). In the context of our approach, data quality meta data should be related to integrated data quality checks. It depends on the meta data if the data quality check is merely an annotation, a simple task, or a more sophisticated modeling element specifically related to data quality information. If integrated into a specific process modeling language, the meta model additionally needs to define the possible application of the data quality check within the process model instantiations. Ofner et al. (2012) provide a sophisticated enhancement of the BPMN and its meta model for integrating data quality information. While data quality is included for decision-making about process redesign, the visibility of data quality aspects within the process model is neglected. Nevertheless, the extension of the BPMN meta model might be used complementarily to our suggested approach. Despite the high number of applicable modeling elements, BPMN may leverage familiarity due to its broad application. However, further informal approaches are conceivable. For an exploration and definition of data quality requirements, a data quality check might be extended with a meta model providing details about IP production. We currently evaluate the application of the conceptual combined life cycle model by Knight (2011) for an intuitive process-driven approach which allows an in-depth exploration of the production and control of critical IPs (Glowalla et al. 2014). Hence, the application of the approach depends on the context, and if process models based on a sound process modeling language exist at all.

### 6.4 Limitation

Our literature review aims at a detailed analysis of process models within PDDQM and their application in organizational settings. Our rather specific selection led to a small number of articles. Since in some cases our differentiation is rather soft and the inclusion or exclusion of articles had to be discussed, we cannot claim that we captured all relevant articles dealing with process-driven data quality techniques. Our more important aim was a detailed description and the presentation of the broad application of process-driven data quality within process models, based on the identified articles. Considering our keywords, we conducted our literature review from a data quality perspective. Our results have to be reflected in additional research literature, for instance, business process management literature. This is supported by the distribution of our identified articles across several journals and the variety of topics. Data and information quality is a cross-sectional issue that can be examined from several perspectives in different contexts.

### 6.5 Further Research

Empirical research is necessary to address the identified issues in process model complexity and understandability. Further research should enhance guidelines to apply meaningful, practice-relevant patterns for PDDQM modeling while being aware of the current limitations in measuring complexity and assessing understandability. Extant research examines complexity metrics' impact on understandability (Figl and Laue 2011; Reijers and Mendling 2011), complexity reduction patterns (Moody 2009; Gruhn and Laue 2009), and their perceived usefulness (La Rosa et al. 2011a, 2011b). However, we see the need to additionally examine the entailed changes in complexity metrics. Such an approach would facilitate learning from the application of patterns, providing insight into interdependencies of metrics as well as into the relationship between complexity, understandability, and interpreting changes in metrics.

Future research should avoid developing IP-MAPs in an isolated way since existing process models provide useful characteristics and are applied in several organizations. An integrated development would support the application of IP-centric process models since the familiarity with a process modeling language affects its use (Recker 2010, p. 87). An adequate process modeling language has to be identified for an empirical study and the approaches to integration of data quality have to be refined. For instance, when using further modeling elements or different arcs within one model to differentiate control and data flow, further metrics need to be considered or developed. Moreover, extending the proposed approach will also impact the application of the patterns.

Our general modeling approach is derived from extant literature and may substitute varying approaches for integrating data quality into non-IP-centric process models. However, further research is necessary to ground our approach in application scenarios and adequate meta data or meta models.

## 7 Summary and Outlook

We conducted a literature review within 74 IS journals and three conferences, reviewing 1,555 articles from 1995 onwards. We examined 26 articles and 46 process models in detail regarding the varying application of process modeling languages within organizations for PDDQM. Building on this synthesis, we provide two integration approaches, within- and across-model integration, to integrate data quality into existing process models. Furthermore, we examine the impact on the models' complexity with regard to the integration approaches and patterns for complexity reduction.

Regarding RQ1, our literature review provides a synthesis of the varying applications of process modeling languages within organizations for PDDQM. Our categorization of process models into IP-centric models (IMS, IP-MAP), DFDs, and PFCs shows that the process model

characteristics are applied in great variety across process models instantiations. On the one hand, our results show that data quality can be applied across different process models (**Table 2**). On the other hand, the mixed application of other characteristics across the categories impedes a clear-cut categorization of customized process model instantiations. Our background literature and primary studies suggest that formalized IP-centric process models, such as the IP-MAP, are used to visibly integrate data quality. Additional formalized approaches enhancing extant process modeling languages exist; however, they neglect the visible integration of data quality. Beyond that, several other diverse approaches attempt to integrate data quality visibly into process models. Our approach primarily addresses the latter. Following the process models of our primary studies, we consider integration of data quality checks into existing process models a straightforward approach to integrate data quality. Apart from the visual enhancement of the process model, the data quality check allows referring further information outside the model. Since DFDs and PFCs do not focus on IPs, a data quality check is a means to define IPs and their quality requirements. Furthermore, enhancing existing process models creates awareness for data quality aspects without radical changes in the existing model layout. Creating awareness counteracts the problem of not managing data quality at all. Hence, instead of switching to new models, organizations have the option to use well-known process models and enhance them with data quality aspects. To provide an IP-centric data quality perspective, we further considered the integration of an IP-centric model with existing models. Besides linking data quality information to each IP, this allows to represent the sequential steps necessary to manufacture (critical) IPs and identify existing process issues and room for improvement.

Regarding RQ2, we examine within- and across-model integration, their impact on model complexity, and the application of complexity reduction patterns. We find several metrics beyond the ones addressed in current research, which are influenced by the integration of data quality checks. Although we provide an intuitive suggestion of integrating data quality into process models, several entailing issues need to be addressed. To provide a context-independent assessment of process model complexity, we identify, select, and evaluate patterns as well as related changes to metrics. Extant research lacks metrics to measure impact on complexity when a new model is applied and integrated. Especially if the new model is an instantiation of another model type, there is a lack of adequate metrics. The application of complexity-reduction patterns after integrating data quality checks into process models is limited due to several open issues in research on process model complexity and understandability. Thus, current research does not provide generic patterns to reliably decrease complexity of process models.

Our selection and presentation of complexity-reducing patterns supports manual integration of data quality checks into existing process models. However, as indicated above, our approach is also compatible with more formalized approaches. A context-independent approach reducing complexity while maintaining behavior-equivalent process models provides a basis for automated improvement. The diversity of potential changes in complexity metrics additionally shows the need to control especially big process model repositories. By presenting, selecting, and applying current patterns and metrics, we further develop issues on process modeling in general, specifically building on and extending extant approaches (Sect. 5.4) for our purpose. Therefore, our integration approach and the entailed considerations are relevant beyond PDDQM for general process modeling.

## References


Balka E, Whitehouse S, Coates ST, Andrusiek D (2012) Ski hill injuries and ghost charts: socio-technical issues in achieving e-health interoperability across jurisdictions. Information Systems Frontiers 14(1):19–42

Ballou D, Wang R, Pazer H, Kumar Tayi G (1998) Modeling information manufacturing systems to determine information product quality. Management Science 44(4):462–484

Bandara W, Gable GG, Rosemann M (2005) Factors and measures of business process modelling: model building through a multiple case study. European Journal of Information Systems 14(4):347–360

Basu A, Blanning RW (2003) Synthesis and decomposition of processes in organizations. Information Systems Research 14(4):337–355

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

Becker J, Rosemann M, von Uthmann C (2000) Guidelines of business process modeling. In: van der Aalst WMP, Desel J, Oberweis A (eds) Business process management, models, techniques, and empirical studies

Buhl HU, Röglinger M, Stöckl S, Braunwarth KS (2011) Value orientation in process management. Bus Inf Syst Eng 3(3):163–172

Capgemini (2013) IT-Trends-Studie. http://www.de.capgemini.com/it-trends-studie. Accessed 2013-03-25

Cappelli C, Leite JCSP, Batista T, Silva L (2009) An aspect-oriented approach to business process modeling. In: Proc 15th workshop on early aspects, Charlottesville, pp 7–12

Cardoso J (2006) Process control-flow complexity metric: an empirical validation. In: Proc IEEE international conference on services computing, Chicago, pp 167–173

Davidson B, Lee YW, Wang R (2004) Developing data production maps: meeting patient discharge data submission requirements. International Journal of Healthcare Technology and Management 6(2):223–240

Dehnert J, van der Aalst WM (2004) Bridging the gap between business models and workflow specifications. International Journal of Cooperative Information Systems 13(3):289–332

Dejaeger K, Hamers B, Poelmans J, Baesens B (2010) A novel approach to the evaluation and improvement of data quality in the financial sector. In: Proc international conference on information quality, Cambridge

English LP (1999) Improving data warehouse and business information quality: methods for reducing costs and increasing profits. Wiley, New York

Figl K, Laue R (2011) Cognitive complexity in business process modeling. In: Mouratidis H, Rolland C (eds) Advanced information systems engineering. Springer, Heidelberg, pp 452–466

Forrester Research (2011) Trends in data quality and business process alignment. http://www.enterpriseiq.com.au/documents/whitepapers/Trends_in_Data_Quality_and_Business_Process_Alignment.pdf. Accessed 2013-03-25

Gaynor M, Shankaranarayanan G (2008) Implications of sensors and sensor-networks for data quality management. International Journal of Information Quality 2(1):75–93

Glowalla P, Balazy P, Basten D, Sunyaev A (2014) Process-driven data quality management – an application of the combined conceptual life cycle model. In: Proc 47th Hawaii international conference on system sciences, Hawaii

Glowalla P, Sunyaev A (2012) A process management perspective on future ERP system development in the financial service sector. AIS Transactions on Enterprise Systems 3(1):18–27

Glowalla P, Sunyaev A (2013) Managing data quality with ERP systems – insights from the insurance sector. In: Proc European conference on information systems, Utrecht

Gruhn V, Laue R (2006) Complexity metrics for business process models. In: Proc 9th international conference on business information systems, Klagenfurt

Gruhn V, Laue R (2009) Reducing the cognitive complexity of business process models. In: Proc 8th IEEE international conference on cognitive informatics, Los Alamitos, pp 339–345

Gucegliogl AS, Demirors O (2005) Using software quality characteristics to measure

business process quality. In: Aalst W, Benatallah B, Casati F, Curbera F (eds) Business process management. Springer, Heidelberg, pp 374–379
Hakim LA (2008) Modelling information flow for surgery management process. International Journal of Information Quality 2(1):60–74
Harkness WL, Segars AH, Kettinger WJ (1996) Sustaining process improvement and innovation in the information services function: lessons learned at the Bose Corporation. MIS Quarterly 20(3):349–368
Haug A, Zachariassen F, van Liempd D (2011) The costs of poor data quality. Journal of Industrial Engineering and Management 4(2):168–193
Helfert M, von Maur E (2001) A strategy for managing data quality in data warehouse systems. In: Proc conference on information quality, Cambridge
Ishikawa K (1993) Guide to quality control. Asian Productivity Organization, Tokyo
Kahn BK, Katz-Haas R, Strong DM (2001) Organizational realism meets information quality idealism: the challenges of keeping an information quality initiative going. In: Proc conference on information quality, Cambridge
Katz-Haas R, Lee YW (2002) Understanding hidden interdependencies between information and organizational processes in practice. In: Proc international conference on information quality, Cambridge
Keenan SL, Simmons T (2005) CSDQ: a user-centered approach to improving the quality of customer support data. In: Proc international conference on information quality, Cambridge
Klesse M, Herrmann C, Maier D, Mügeli T, Brändli P (2004) Customer investigation process at Credit Suisse: meeting the rising demand of regulators. In: Proc international conference on information quality, Cambridge
Knight S (2011) The combined conceptual life-cycle model of information quality. Part 1. An investigative framework. International Journal of Information Quality 2(3):205–230
Ko RK, Lee SS, Lee EW (2009) Business process management (BPM) standards: a survey. Business Process Management Journal 15(5):744–791
Kovac R, Lee YW, Pipino L (1997) Total data quality management: the case of IRI. In: Proc conference on information quality, Cambridge
Kovac R, Weickert C (2002) Starting with quality: using TDQM in a start-up organization. In: Proc international conference on information quality, Cambridge
Kurzlechner W (2011) Die Top-10-Listen der IT-Trends 2012. http://www.cio.de/strategien/2298020/. Accessed 2012-02-07
La Rosa M, ter Hofstede AH, Wohed P, Reijers HA, Mendling J, van der Aalst WM (2011a) Managing process model complexity via concrete syntax modifications. IEEE Transactions on Industrial Informatics 7(2):255–265
La Rosa M, Wohed P, Mendling J, ter Hofstede AH, Reijers HA, van der Aalst WM (2011b) Managing process model complexity via abstract syntax modifications. IEEE Transactions on Industrial Informatics 7(4):614–629
Laue R, Gruhn V (2007) Approaches for business process model complexity metrics. In: Abramowicz W, Mayr HC (eds) Technologies for business information systems. Springer, Dordrecht, pp 13–24

## Zusammenfassung / Abstract

**Paul Glowalla, Ali Sunyaev**

## Process-Driven Data Quality Management Through Integration of Data Quality into Existing Process Models

### Application of Complexity-Reducing Patterns and the Impact on Complexity Metrics

The importance of high data quality and the need to consider data quality in the context of business processes are well acknowledged. Process modeling is mandatory for process-driven data quality management, which seeks to improve and sustain data quality by redesigning processes that create or modify data. A variety of process modeling languages exist, which organizations heterogeneously apply. The purpose of this article is to present a context-independent approach to integrate data quality into the variety of existing process models. The authors aim to improve communication of data quality issues across stakeholders while considering process model complexity. They build on a keyword-based literature review in 74 IS journals and three conferences, reviewing 1,555 articles from 1995 onwards. 26 articles, including 46 process models, were examined in detail. The literature review reveals the need for a context-independent and visible integration of data quality into process models. First, the authors present the enhancement of existing process models with data quality characteristics. Second, they present the integration of a data-quality-centric process model with existing process models. Since process models are mainly used for communicating processes, they consider the impact of integrating data quality and the application of patterns for complexity reduction on the models' complexity metrics. There is need for further research on complexity metrics to improve the applicability of complexity reduction patterns. Lacking knowledge about interdependency between metrics and missing complexity metrics impede assessment and prediction of process model complexity and thus understandability. Finally, our context-independent approach can be used complementarily for data quality integration with specific process modeling languages.

**Keywords:** Data quality, Information quality, Process modeling, Process model, Model integration, Model complexity, Model understandability

Laue R, Mendling J (2010) Structuredness and its significance for correctness of process models. Information Systems and e-Business Management 8(3):287–307

Laumann M, Rosenkranz C (2008) Analysing information flows for controlling activities within supply chains: an Arvato (Bertelsmann) business case. In: Proc European conference on information systems (ECIS), Galway

Lee YW, Chase S, Fisher J, Leinung A, McDowell D, Paradiso M, Simons J, Yarsawich C (2007) CEIP maps: context-embedded information product maps. In: Proc Americas conference on information systems (AMCIS), Cambridge

Lee YW, Strong DM (2003) Knowing-why about data processes and data quality. Journal of Management Information Systems 20(3):13–39

Lee YW (2006) Journey to data quality. MIT Press, Cambridge

Lindland OI, Sindre G, Solvberg A (1994) Understanding quality in conceptual modeling. IEEE Software 11(2):42–49

Loshin D (2001) Enterprise knowledge management. The data quality approach. Morgan Kaufmann, San Diego

Madnick SE, Wang RY, Lee YW, Zhu H (2009) Overview and framework for data and information quality research. Journal of Data and Information Quality 1(1):1–22

Mendling J (2008) Metrics for process models. Empirical foundations of verification, error prediction, and guidelines for correctness. Springer, Heidelberg

Mendling J, Neumann G, van der Aalst WM (2007) Understanding the occurrence of errors in process models based on metrics. In: Proc OTM conference on cooperative information systems, Vilamoura, pp 113–130

Mendling J, Reijers HA, van der Aalst WMP (2010) Seven process modeling guidelines (7PMG). Journal of Information and Software Technology 52(2):127–136

Mendling J, Strembeck M (2008) Influence factors of understanding business process models. In: Abramowicz W, Fensel D (eds) LNBIP LNCS. Springer, Heidelberg, pp 142–153

Meyer MH, Zack MH (1996) The design and development of information products. Sloan Management Review 37(3):43–59

Mielke M (2005) IQ principles in software development. In: Proc international conference on information quality, Cambridge

Millard FH, Lavoie M (2000) Developing data product maps for total data quality management: the case of Georgia Vital Records. In: Proc conference on information quality, Cambridge

Moody D (2009) The "physics" of notations: toward a scientific basis for constructing visual notations in software engineering. IEEE Transactions on Software Engineering 35(6):756–779

Ofner MH, Otto B, Österle H (2012) Integrating a data quality perspective into business process management. Business Process Management Journal 18(6):1036–1067

Otto B (2011) Data governance. Bus Inf Syst Eng 3(4):241–244

Overhage S, Birkmeier D, Schlauderer S (2012) Quality marks, metrics, and measurement procedures for business process models: the 3QM-framework. Bus Inf Syst Eng 4(5):229–246

Recker J (2010) Continued use of process modeling grammars: the impact of individual difference factors. European Journal of Information Systems 19(1):76–92

Recker J, Indulska M, Rosemann M, Green P (2010) The ontological deficiencies of process modeling in practice. European Journal of Information Systems 19(5):501–525

Recker JC, Rosemann M, Indulska M, Green P (2009) Business process modeling – a comparative analysis. Journal of the Association for Information Systems 10(4):333–363

Redman TC (2004) Data: an unfolding quality disaster. DM Review 14(8):21–23

Reijers HA, Mendling J (2011) A study into the factors that influence the understandability of business process models. IEEE Transactions on Systems, Man, and Cybernetics, Part A 41(3):449–462

Rosemann M (2006) Potential pitfalls of process modeling: part a. Business Process Management Journal 12(2):249–254

Rosemann M, Green P, Indulska M, Recker JC (2009) Using ontology for the representational analysis of process modelling techniques. International Journal of Business Process Integration and Management 4(4):251–265

Rosemann M, Recker JC, Flender C (2008) Contextualisation of business processes. International Journal of Business Process Integration and Management 3(1):47–60

Shankaranarayanan G, Cai Y (2006) Supporting data quality management in decision-making. Decision Support Systems 42(1):302–317

Shankaranarayanan G, Wang R (2007) IPMAP research status and direction. In: Proc international conference on information quality, Cambridge

Shankaranarayanan G, Wang RY, Ziad M (2000) IP-MAP: representing the manufacture of an information product. In: Proc conference on information quality, Cambridge

Shankaranarayanan G, Ziad M, Wang RY (2003) Managing data quality in dynamic decision environments: an information product approach. Journal of Database Management 14(4):14–32

Thi TTP, Helfert M (2007) Modelling information manufacturing systems. International Journal of Information Quality 1(1):5–21

Uba R, Dumas M, García-Bañuelos L, Rosa M (2011) Clone detection in repositories of business process models. In: Rinderle-Ma S, Toumani F, Wolf K (eds) Business process management. Springer, Heidelberg

Vanderfeesten I, Reijers HA, Mendling J, Aalst WM, Cardoso J (2008) On a quest for good process models: the cross-connectivity metric. In: Proc 20th international conference on advanced information systems engineering, Montpellier, pp 480–494

Vanhatalo J, Völzer H, Koehler J (2009) The refined process structure tree. In: Sixth international conference on business process management – five selected and extended papers. Data & Knowledge Engineering, vol 68(9), pp 793–818

Wand Y, Weber R (1993) On the ontological expressiveness of information systems analysis and design grammars. Information Systems Journal 3(4):217–237

Wand Y, Weber R (1995) On the deep structure of information systems. Information Systems Journal 5(3):203–223

Wang RY (1998) A product perspective on total data quality management. Communications of the ACM 41(2):58–65

Wang RY, Allen TJ, Harris W, Madnick S (2002) An information product approach for total information awareness. MIT Sloan working paper no 4407-02; CISL no 2002-15

Weber B, Reichert M, Mendling J, Reijers HA (2011) Refactoring large process model repositories. Computers in Industry 62(5):467–486

Webster J, Watson RT (2002) Analyzing the past to prepare for the future: writing a literature review. MIS Quarterly 26(2):xiii–xxiii

Xie S, Helfert M (2010) Assessing information quality deficiencies in emergency medical service performance. In: Proc international conference on information quality, Cambridge

Zack MH (1996) Electronic publishing: a product architecture perspective. Information & Management 31(2):75–86