Evaluating and Improving Model-Based Assessment of Contextual Data Quality in Smart Grids

Dominik Vereno
Josef Ressel Centre for Dependable System-of-Systems Engineering
Puch/Salzburg, Austria
dominik.vereno@fh-salzburg.ac.at

Katharina Polanec
Josef Ressel Centre for Dependable System-of-Systems Engineering
Puch/Salzburg, Austria
katharina.polanec@fh-salzburg.ac.at

Christian Neureiter
Josef Ressel Centre for Dependable System-of-Systems Engineering
Puch/Salzburg, Austria
christian.neureiter@fh-salzburg.ac.at

***Abstract*—Data science has great potential in smart grids, but applying it must not jeopardize dependability—thus requiring high-quality data. Ideally, quality is ensured in early development stages. To enable this, we propose using architecture models for data quality assessment. The study focuses on a domain-specific language based on the Smart Grid Architecture Model and models created with it. The first goal is to evaluate how suitable such models are for assessing contextual data quality—the data's fitness for use. We discovered the assessability is mainly limited by the modeling language not facilitating standards-independent data definition. Consequently, we set out to improve the language. We arrived at three proposed modifications using a case study–based design-science approach: including an element for generic data definition, separating data and information, and adding attributes to data flows. Our research demonstrates the feasibility of model-based data-quality assessment and takes a step towards integrating data science into smart-grid architecture.**

***Index Terms*—Data science by design, Dependability, Information quality, Model-based systems engineering, SGAM**

I. Introduction

In recent decades, major global trends have affected electricity grids. For example, a worldwide surge in electric vehicles places increasing demand on grids [1]. Additionally, the rising percentage of renewable energy in the energy mix forces power grids to become more flexible. One promising approach for achieving greater flexibility is to transform existing grids to so-called *smart grids*. They are grids intimately linked with information and communications technology, which enables widespread automated control and monitoring [2]. Data science has the potential to improve these tasks, as shown by Zhang et al. [3]. However, power grids are critical infrastructure and hence need to be dependable. Consequently, any data-driven decision making also needs to be dependable and thus requires high-quality data [4]; in particular, *contextual data quality*—the fitness-for-use of data [5]—is necessary to ensure the data is suitable for its intended application.

In addition to smart grids having to be dependable, they are also highly complex. Designing and engineering such complex systems dependably, requires a suitable engineering approach; model-based systems engineering (MBSE) has established itself as an effective method for dealing with such tasks [6]. MBSE is a branch of the interdisciplinary field of systems engineering combined with concepts from model-driven engineering [7]. In the smart-grid domain, the Smart Grid Architecture Model (SGAM) framework is the basis for most MBSE approaches. Notably, Neureiter et al. [8] made the framework usable for concrete engineering task by developing a domain-specific language (DSL) for SGAM.

Thus far, we have established the need for high-quality data in smart grids, making the assessment of data quality necessary. However, most methods for data quality assessment rely on data content, as opposed to just metadata [9]. Thus, at least a prototype or a simulation are required which are available only in later development stages. Unfortunately, significant changes to a complex system so late in the development process are time consuming and costly. MBSE provides a potential solution to this problem: Since models already play a vital role in the engineering of complex systems, they could be used for assessing data quality in early development stages. These assessments would then enable timely countermeasures to data quality issues. This helps ensuring that the completed system facilitates sufficient data quality—thus establishing a basis for *data science by design*. However, we are not aware of any assessment method that only utilizes system models. Therefore, we set out to determine to what extent contextual data quality can be assessed based on a model created using the SGAM DSL. Additionally, we investigate which changes to the language are necessary to improve assessability.

This paragraph outlines the remainder of this paper: The next chapter—*Relevant Background*—provides basic knowledge of MBSE in smart grids and data quality. Subsequently, the research approach is presented in Chapter III. After that, Chapter IV discusses the current data quality assessability of the SGAM DSL. Next, in Chapter V we propose changes to the DSL. And finally, *Conclusions and Future Work* summarizes this paper's findings and prompts further research.

II. Relevant Background

This paper occupies the intersection of data quality research and MBSE in smart grids. The following chapter provides the necessary background in both fields.

The financial support by the Austrian Federal Ministry for Digital and Economic Affairs and the National Foundation for Research, Technology and Development and the Christian Doppler Research Association as well as the Federal State of Salzburg is also gratefully acknowledged.

A. Smart Grid Architecture Model

SGAM is a framework for describing smart-grid architectures in a three-dimensional cuboid which is depicted in Figure 1. It was originally conceived by European standardization bodies with the intent of identifying gaps in standardization [10]. The framework comprises five projections of the SGAM plane, spanned by two axes: the domain and zone axes represent the energy-conversion chain and the automation pyramid, respectively. Since the precise meaning of each domain and zone is not relevant to this paper, they will not be discussed, but a detailed elaboration is available in [11].

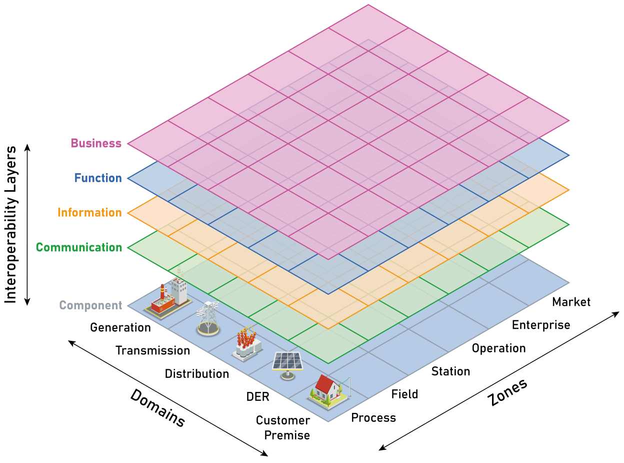

Fig. 1. The Smart Grid Architecture Model (figure based on [11])

Each of the five projections of the SGAM plane deals with different interoperability concerns—from enterprise considerations on the business layer down to technical elements on the component layer. For this research, two layers are especially significant: the function and the information layer. The former contains functions and services which may be derived from use cases and the involved actors. The latter deals with information objects, their flow throughout the system, and the data model standards that define them.

Since its inception, SGAM has outgrown its original purpose; it now serves as the basis for architecture-development approaches. Neureiter et al. [8] have published one such approach: an SGAM-based modeling language. This DSL conveys smart-grid architectures in a way that is understood—and thus accepted—by domain stakeholders. The DSL is implemented using Unified Modeling Language (UML) profiles and is tool independent; it is accompanied by a tool-specific add-in[1] for Enterprise Architect. However, this research only deals with the tool-independent modeling language.

B. Data quality

The following section is a small overview of data quality research relevant to this paper. It establishes data quality as a hard-to-define, multi-dimensional concept. Moreover, it deals with approaches for data quality assessment.

1) Defining data quality: For this research paper, the ISO-derived definition of data quality as being “the degree to which data conform to data specifications” [12, p. 192] is appropriate. However, we must clarify what we mean by data and differentiate it from information. This examination is necessary since the data quality literature does not agree on the distinction between data and information [13] and consequently the distinction between *data quality* and *information quality* [12]. For example, Wang et al. explicitly state that they are using data and information synonymously [14] and Lee et al. implicitly use data quality and information quality interchangeably [15]. In contrast, Zhu et al. mention the tendency to use data quality for technical issue and information quality for non-technical issues [13]. We think this distinction is useful and necessary. Thus, we refer to data as the technical representation and possibly redundant carrier of information. In contrast, information is the technology-neutral, redundancy-free resolution of uncertainty [16].

2) Data quality dimensions: Pipino et al. have labeled data quality a multi-dimensional concept. Each dimension describes ”a general, measurable category for a distinctive characteristic (quality) possessed by data” [17]. Wang and Strong have identified 15 such dimensions relevant to data consumers and split those dimensions into four categories: intrinsic, contextual, representational, and accessibility data quality. The framework is established in the data quality community and provides a suitable taxonomy for this study.

3) Assessing data quality: To adequately define data quality assessment, a distinction has to be made: Batini et al. differentiate between data-driven and process-driven data quality *improvement* [9]. This concept may also be applied to data quality assessment. Data-based data quality assessment refers to evaluating, to what extent data meets the relevant requirements. In contrast, process-based data quality assessment refers to evaluating, to what degree a data-generating or data-modifying process is capable of producing requirements-satisfying data. This distinction is closely related to that made by Aljumaili et al. between content-based and metadata-based assessment [18]. Since model-based data quality assessment does not directly assess the data content, but the metadata, we argue that a model-based approach can be categorized as process- and metadata-based assessment. Furthermore, it should be noted that for Batini et al., assessment includes measurement of data quality followed by the comparison to a reference value. Since many others do not explicitly make the distinction between assessment and measurement— [17], [15] and [18], for instance—this paper will also not do so.

III. Research Approach

We have identified clear value in model-based data quality assessment in smart grids. Therefore, we set out to evaluate how suitable the SGAM DSL is for such assessments, specifically of contextual data quality. Then we intend to improve the assessability of the DSL by modifying its defining metamodel. In this chapter, we briefly describe our research approach and the case study on which we evaluate the language.

[1] available at https://sgam-toolbox.org

A. Methodology

Due to the nature of this research endeavor, we have deemed the design-science research methodology by Peffers et al. [19] to be a suitable approach; it is based on the design-science research paradigm by Hevner et al. [20]. By following the process model by Peffers et al., we first evaluate the current DSL qualitatively to identify problems regarding data quality assessability. Then, requirements for the modification of the language are specified: All five contextual data quality dimensions must be assessable. After the objective-definition step, multiple iterations of three activities are performed: The language is modified, the changes are demonstrated in a case study, and the modified language is evaluated to judge if it meets the previously specified objectives. Since the focus of the evaluation is mainly on feasibility, we classify this study as a proof of concept for model-based data quality assessment.

B. Case Study

To evaluate the modifications to the artifact, we use the following smart-grid use case as a fictitious case study: A distribution system operator (DSO) wants to predict the energy demand of electric-vehicle charging processes to avoid having to buy expensive balancing energy and instead buy it on the intraday market. A charging-station management system sends various data about each charging process to the DSO, who computes the demand estimation. The scenario is modeled with the SGAM DSL following the process proposed by Neureiter et al. [8]. As a result, various artifacts are created. They describe the smart-grid system surrounding the case-study scenario from different viewpoints. All evaluations done in this research paper are based on these artifacts.

IV. Evaluating the Current Assessability

This section describes, to what extent contextual data quality can be assessed in an SGAM-DSL model. For each of the five contextual dimensions identified by Wang and Strong [5], the assessability is evaluated. All evaluation center around information object since they are the only DSL-provided modeling elements representing data. Furthermore, this section aims at identifying, what prevents better assessability.

A. Completeness

Pipino et al. describe completeness as "the extent to which data is not missing and is of sufficient breadth and depth for the task at hand" [17, p. 212]. Breadth is the number of attributes, whereas depth describes the number of rows or samples [21]. Pipino et al. differentiate between three types of completeness: schema, column, and population (i.e. row) completeness. Measuring schema completeness is inherently well suited for metadata-based approaches, since schemata *are* metadata. In an SGAM DSL–based model, such a measurement is not directly possible, since the only way to define an information object is through a data model standard. Therefore, if a suitable standard is specified, the appropriate data model must be identified in that standard; if none exists, the assessment must rely on guesswork based on the information object's name and its context.

Furthermore, model-based assessment of column completeness is not possible, because a missing value is a deviation from the schema, not an inadequacy of the schema. Population completeness, however, can partly be evaluated based on a model: It is possible to evaluate if the modeled number of samples meets the requirements.

B. Appropriate Amount of Data

This quality dimension is a measure for how well the volume of the data fits a given task [5]. Even though this dimension is similar to completeness, there are two major differences: First, completeness is concerned with missing data. Second, completeness refers to the amount of data being *sufficient* for a task, but this dimension refers to the amount of data being *appropriate*. Therefore, the completeness measure would not suffer from there being too much data, whereas the appropriate-amount-of-data measure would. Despite their differences, the model-based assessment of these two dimensions is quite similar. In both cases, one has to rely on a schema for evaluation, since the number of columns and the multiplicities between entities is the only sound basis for an assessment. Thus, the same issue as with completeness arises: If no applicable data model standard exists for an information object, judging if the amount of data is appropriate, is guesswork.

C. Timeliness

According to Kahn et al., data is timely, if it is sufficiently up-to-date [22]. For some tasks, data cannot be older than a few milliseconds, for others, week-old data is sufficient. The SGAM DSL itself does not address timing aspects. However, in UML sequence diagrams, time values can be assigned to all interactions. These diagrams are used to specify use-case behavior in an SGAM-DSL model. Therefore, one can assess the timeliness of an information object by tracing its path through one or more sequence diagrams.

D. Relevancy

Data is to be classified as relevant, if it is applicable and helpful for its intended use [5]. To judge if data is relevant, knowledge about the task as well as the data itself is required. The latter may be provided by the data model standard assigned to the information object. If none is available, one is left with looking at the information object's name and its context. As with completeness and appropriate amount of data, the assessability of relevancy is constrained by the availability of a standardized data model.

E. Value-Added

Data adds value, if it is beneficial for the task at hand and provides advantages from its use [5]. This definition is similar to that of relevancy; in fact, evaluating the relevancy of a piece of data—like an information object—is a prerequisite to determining if it adds value. However, data can be relevant without adding value: If an information object conveys relevant information but that information is already available, it is not valuable. Therefore, assessing the value of data requires an overview over the relevancy of all data available for a task. Furthermore, a more detailed trade-off analysis could be conducted: For example, two information objects could convey the same relevant information, but one can be transmitted in a more timely manner, making it more valuable. In the end, assessing this dimension suffers from the same limitations as relevancy: having no way to specify a non-standardized information object.

V. Modifying the SGAM DSL for Improved Assessability

The previous chapter deals with what is limiting data quality assessability. This chapter proposes changing the DSL to remedy these shortcomings. First, a clear separation of data and information is proposed. Then, a model element for generically describing data is introduced. Finally, we suggest adding attributes to data and information flows.

A. Restructuring the modeling of data and information

The SGAM framework and its DSL provide no clear distinction between information and data. However, distinguishing between them is important: For a logical architecture, the non-technical concept of information is useful. In contrast, a technical architecture requires defining *how* to exchange or store that information; this necessitates the notion of data with a concrete representation, be it digital or not.

To achieve this separation, we propose altering the DSL as illustrated by the current and proposed abstract syntax models in Figure 2 and Figure 3. First, information objects now represent abstract pieces of information without regard to their technical representation. Second, a new model element is introduced: the *data object*. It is bound to a specific representation and can be treated as the technical realization of information. Consequently, the data model standard does not fit this usage of information objects anymore, it rather applies to data objects. However, the data model standard will be replaced with a more generic element, as described in the following section. Furthermore, a *data object flow* is added to represent the transmission of data. Notably, the data object flow connects technical components, whereas the information object flow now connects logical actors. Thus, the modeling of information and its exchange is now limited to the function layer, whereas data is modeled on the information layer.

B. Facilitating generic specification of data representation

We have revealed that the SGAM DSL does not support the definition of non-standardized information objects. This limitation is lifted by swapping the data-model-standard element for the more generic modeling element *data representation*. As argued previously, a data-representation element fits our definition of the data object and is not applicable to information objects. Thus far, it is not clear what means of defining a data-representation element are appropriate. For one, the modified DSL should still support the usage of data model standards. This is easily done by naming the data-representation element accordingly. For describing non-standardized data objects, entity-relationship models are suitable. Since the DSL utilizes the UML profile mechanism, we strongly recommend using UML class diagrams.

C. Adding attributes to data and information object flows

Due to the restructuring of the DSL, newly introduced data objects are used for data quality assessment, not information objects. Therefore, one cannot evaluate the timeliness of data with use case–specifying sequence diagrams since they model the exchange of *information* objects. To ensure the assessability of timeliness, we propose adding time-related attributes to data object flows. The components and their data object flows would thus become a data-flow graph, for which latency can even be assessed symbolically [23]. Furthermore, we argue that timing aspects are also relevant on a functional level and therefore suggest adding similar attributes to information object flows. Finally, we recommend specifying a small set of valid data- and information-flow attributes. This makes formal data-flow analyses easier to realize.

VI. Conclusions and Future Work

We set out to reveal to what extent an SGAM-DSL model can serve as the basis for assessing contextual data quality. We learned that the assessability is severely limited by the DSL not providing a way to describe non-standardized information objects. These findings expose the need to modify the DSL for improved assessability. To lay the groundwork for such modifications, we propose a clear separation between information and data. Then, a new modeling element for generically describing data is introduced; it solves the assessability issues that were uncovered in our research. Additionally, we propose adding time-related attributes—like latency—to data flows to enable analyzing the timeliness of data.

On the one hand, this paper furthers data quality research; it establishes the novel concept of model-based data quality assessment as a close relative to metadata- and process-based approaches. On the other hand, our research contributes to MBSE in smart grids. The proposed changes enable evaluating the fitness-for-use of data in early development stages. Thus, system architects are able to detect issues sooner to avoid late and thus expensive changes; they can ensure that the data created and modified by the implemented system are suitable for their intended task. Therefore, we have taken an important step towards enabling data science by design.

Ultimately, subsequent research needs to be conducted to achieve that goal: First, our findings need to be verified on other, more comprehensive smart-grid case studies. Then it can be tested if the results are transferable to other engineering domains; we assume that frameworks similar to SGAM—like RAMI 4.0 [24]—profit from our findings. Moreover, we have only dealt with contextual data quality so far; other data quality dimensions also have to be considered. And finally, formalizing or even automating data quality assessment could prove valuable for system engineers and architects.

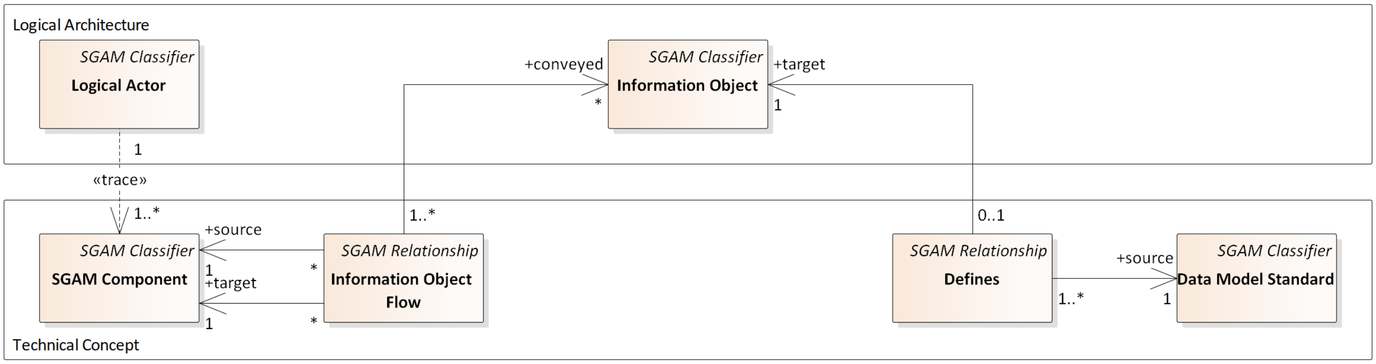

Fig. 2. Current abstract syntax model of the SGAM-DSL metamodel

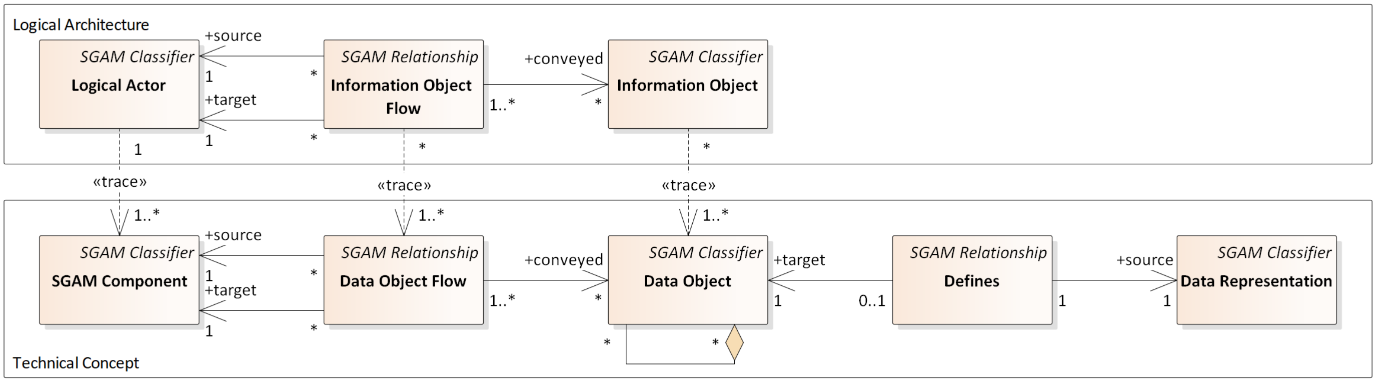

Fig. 3. Proposed abstract syntax model of the SGAM-DSL metamodel

References

[1] H. S. Das, M. M. Rahman, S. Li, and C. W. Tan, "Electric vehicles standards, charging infrastructure, and impact on grid integration: A technological review," *Renewable and Sustainable Energy Reviews*, vol. 120, p. 109618, 2020.

[2] H. Farhangi, "The path of the smart grid," *Power and Energy Magazine, IEEE*, vol. 8, pp. 18–28, 2010.

[3] Y. Zhang, T. Huang, and E. F. Bompard, "Big data analytics in smart grids: a review," *Energy Informatics*, vol. 1, no. 1, 2018.

[4] V. Sessions and M. Valtorta, "The effects of data quality on machine learning algorithms." in *11th International Conference on Information Quality*, 2006, pp. 485–498.

[5] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–33, 1996.

[6] C. Neureiter, C. Binder, and G. Lastro, "Review on domain specific systems engineering," in *2020 IEEE International Symposium on Systems Engineering (ISSE)*. Vienna, Austria: IEEE, 2020, pp. 1–8.

[7] J. A. Estefan, *Survey of Model-Based Systems Engineering (MBSE) Methodologies*, 2008.

[8] C. Neureiter, D. Engel, and M. Uslar, "Domain specific and model based systems engineering in the smart grid as prerequesite for security by design," *electronics*, vol. 5, pp. 1–44, 2016.

[9] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–52, 2009.

[10] M. Uslar, S. Rohjans, C. Neureiter, F. P. Andrén, J. Velasquez, C. Steinbrink, V. Efthymiou, G. Migliavacca, S. Horsmanheimo, H. Brunner, and T. Strasser, "Applying the smart grid architecture model for designing and validating system-of-systems in the power and energy domain: A european perspective," *Energies*, vol. 12, no. 2, p. 258, 2019.

[11] *Smart Grid Reference Architecture*, Smart Grid Coordination Group, Brussels, Belgium, 2012.

[12] J. R. Talburt and Y. Zhou, "ISO data quality standards for master data," in *Entity Information Life Cycle for Big Data*, 2015, pp. 191–205.

[13] H. Zhu, S. Madnick, Y. Lee, and R. Wang, "Data and information quality research: Its evolution and future," in *Computing Handbook, Third Edition*, H. Topi and A. Tucker, Eds., 2014, pp. 16–1–16–20.

[14] R. Y. Wang, M. Ziad, and Y. W. Lee, *Data Quality*. Boston, MA: Springer US, 2002.

[15] Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang, "AIMQ: a methodology for information quality assessment," *Information & Management*, vol. 40, no. 2, pp. 133–146, 2002.

[16] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948, publisher: Institute of Electrical and Electronics Engineers (IEEE).

[17] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Communications of the ACM*, vol. 45, no. 4, pp. 211–218, 2002.

[18] M. Aljumaili, R. Karim, and P. Tretten, "Metadata-based data quality assessment," *VINE Journal of Information and Knowledge Management Systems*, vol. 46, no. 2, pp. 232–250, 2016.

[19] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, Dec. 2007.

[20] Hevner, March, Park, and Ram, "Design science in information systems research," *MIS Quarterly*, vol. 28, no. 1, p. 75, 2004.

[21] L. Sebastian-Coleman, *Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework*. Waltham, MA: Morgan Kaufmann, 2013.

[22] B. K. Kahn, D. M. Strong, and R. Y. Wang, "Information quality benchmarks: Product and service performance," *Communications of the ACM*, vol. 45, no. 4, 2002.

[23] A. Bouakaz, P. Fradet, and A. Girault, "Symbolic Analyses of Dataflow Graphs," *ACM Transactions on Design Automation of Electronic Systems*, vol. 22, no. 2, pp. 1–25, Mar. 2017.

[24] *Umsetzungsstrategie Industrie 4.0*, BITKOM e.V., VDMA e.V., and ZVEI e.V., 2015.