SAP Signavio Academic Models: A Large Process Model Dataset

Diana Sola[1,2] , Christian Warmuth[1] , Bernhard Schäfer[1,2] ,
Peyman Badakhshan[1] , Jana-Rebecca Rehse[2] ,
and Timotheus Kampik[1(✉)]

[1] SAP Signavio, Berlin, Germany
{diana.sola,christian.warmuth,bernhard.schaefer,peyman.badakhshan,
timotheus.kampik}@sap.com
[2] University of Mannheim, Mannheim, Germany
rehse@uni-mannheim.de

Abstract. In this paper, we introduce the *SAP Signavio Academic Models* (SAP-SAM) dataset, a collection of hundreds of thousands of business models, mainly process models in BPMN notation. The model collection is a subset of the models that were created over the course of roughly a decade on academic.signavio.com, a free-of-charge software-as-a-service platform that researchers, teachers, and students can use to create business (process) models. We provide a preliminary analysis of the model collection, as well as recommendations on how to work with it. In addition, we discuss potential use cases and limitations of the model collection from academic and industry perspectives.

Keywords: Process models · Data set · Model collection

1 Introduction

Process models depict how organizations conduct their operations. They represent the basis for understanding, analyzing, redesigning, and automating processes along the business process management (BPM) lifecycle [9]. As such, many organizations posses large repositories of process models [11]. Having access to such repositories would be tremendously beneficial for developing and testing algorithms in the area of BPM, e.g., for process model querying [19] or reference model mining [20]. Also, the growing interest in applying machine learning in the BPM field, e.g., for process model matching [1], process model abstraction [27] or process modeling assistance [24], underlines the relevance for large model collections that can, for example, serve as training datasets.

However, researchers rarely have access to large collections of models from practice. Such models can contain sensitive information about the organization's internal operations. Legal aspects and the fear of losing competitive advantage thus discourage companies from publishing their business (process) models [25]. This inherent dilemma has so far largely prevented the publication of large-scale model collections for research, as they are common in related research fields [25].

In this paper, we introduce *SAP Signavio Academic Models* (SAP-SAM), a model collection that consists of hundreds of thousands of process and business models in different notations. We provide a basic overview of datasets related to SAP-SAM, as well as the origin and structure of it. Subsequently, we present selected properties and use cases of SAP-SAM. Finally, we discuss limitations of the dataset along with recommendations on how to work with it.

2   Related Datasets

Compared to SAP-SAM, existing process model collections are rather small. The hdBPMN [21] dataset, for example, contains 704 BPMN 2.0 models. This collection has the special feature that the models are handwritten and can be parsed as BPMN 2.0 XML. Another example is RePROSitory [5] (Repository of open PROcess models and logS) which is an open collection of business process models and logs, meaning users can contribute to the repository by uploading their own data. At the time of writing, RePROSitory also contains around 700 models. Some models included in SAP-SAM have already been published [28]. However, the previously published dataset contains only 29,810 models that were collected over a shorter period of time.

In the process mining community, the BPI challenge datasets, e.g., the BPI challenge 2020 [8], have become important benchmarks. Unlike SAP-SAM, these datasets consist of *event logs* from practice. Therefore, the applications of the BPI challenge datasets only partially overlap with those of SAP-SAM.

3   Origins and Structure of SAP-SAM

SAP-SAM contains 1,021,471 process and business models that were created using the software-as-a-service platform of the SAP Signavio Academic Initiative[1] (SAP-SAI), roughly from 2011 to 2021[2]. Most models are in Business Process Model and Notation (BPMN 2.0[3]). SAP-SAI allows academic researchers, teachers, and students to create, execute, and analyze process models, as well as related business models, e.g., of business decisions. The usage of SAP-SAI is restricted to non-commercial research and education. Upon registration, users consent that the models they create can be made available for research purposes, either *anonymized* or *non-anonymized.* SAP-SAM contains those models for which users have

---

[1] See: signavio.com/bpm-academic-initiative/ (accessed at 2022-07-25).

[2] The total number includes vendor-provided example models, which are automatically added to newly created workspaces (process repositories that users register). About 470,000 models in the dataset bear the name of an example model, but this can only be a rough estimate of the number of example models in the dataset.

[3] Technically, the latest version of BPMN is, at the time of writing, BPMN 2.0.2. However, little has changed between 2.0 and 2.0.2. We assume that the informal cross-vendor alignment efforts of the OMG BPMN Model Interchange Working Group are more substantial than formal progress between minor versions. In the following, we therefore use *BPMN 2.0* to refer to any version among 2.0 and 2.0.2.

consented to non-anonymized sharing. Still, anonymization scripts were run to post-process the models, in particular to remove email addresses, student registration numbers, and—to the extent possible—names.

The models in SAP-SAM were created between July 2011 and (incl.) September 2021 by a total of 72,996 users, based on a count of distinct user IDs that are associated with the creation or revision of a model. The models were extracted from the MySQL database of SAP-SAI and are in SAP Signavio's proprietary JSON-based data format. The BPMN models are conceptually BPMN-2.0-standard-compliant, i.e., individual models can be converted to BPMN 2.0 XML using the built-in functionality of SAP-SAI. Decision Model and Notation (DMN) models can be exported analogously. The dataset contains models in the following notations:

- Business Process Model and Notation (BPMN): BPMN is a standardized notation for modeling business processes [15]. SAP-SAM distinguishes between BPMN process models, collaboration models, and choreography models, and among BPMN process models between BPMN 1.1 and BPMN 2.0 models.
- Decision Model and Notation (DMN): DMN is a standardized notation for modeling business decisions, complementing BPMN [17].
- Case Management Model and Notation (CMMN): CMMN is an attempt to supplement BPMN and DMN with a notation that focuses on agility and autonomy [16].
- Event-driven Process Chain (EPC): EPC [22] is a process modeling notation that enjoyed substantial popularity before the advent of BPMN.
- Unified Modeling Language (UML): UML is a modeling language used to describe software (and other) systems. It is subdivided into class and use case diagrams.
- Value Chain: A value chain is an informal notation for sketching high-level end-to-end processes and process frameworks.
- ArchiMate: ArchiMate is a notation for the integrated modeling of information technology and business perspectives on large organizations [13].
- Organization Chart: Organization charts are tree-like models of organizational hierarchies.
- Fundamental Modeling Concepts (FMC) Block Diagram: FMC block diagrams support the modeling of software and IT system architectures.
- (Colored) Petri Net: Petri nets [18] are a popular mathematical modeling language for distributed systems and a crucial preliminary for many formal foundations of BPM. In SAP-SAM, colored Petri nets [12] are considered a separate notation.
- Journey Map: Journey maps model the customer's perspective on an organization's business processes.
- Yet Another Workflow Language (YAWL): YAWL is a language for modeling the control flow logic of workflows [26].
- jBPM: jBPM models allowed for the visualization of business process models that could be executed by the jBPM business process execution engine before the BPMN 2.0 XML serialization format existed. However, recent versions of jBPM rely on BPMN 2.0-based models.

- Process Documentation Template: Process documentation templates support the generation of comprehensive PDF-based process documentation reports. These templates are technically a model notation, although they may practically be considered a reporting tool instead.
- XForms: XForms is a (dated) standard for modeling form-based graphical user interfaces [2].
- Chen Notation: Chen notation diagrams [3] allow for the creation of entity-relationship models.

SAP-SAM is available at https://zenodo.org/record/7012043. Its license supports non-commercial use for research purposes, e.g., usage for the evaluation of academic research artifacts, such as algorithms and related software artifacts.

4   Properties of SAP-SAM

SAP-SAM comprises models in different modeling notations and languages, as well as of varying complexity. In this section, we provide an overview of selected properties of SAP-SAM. The source code that we used to examine the properties is available at https://github.com/signavio/sap-sam.

Modeling Notations. Figure 1 depicts the number of models in different notations in the dataset, as well as the according percentages (in brackets). We aggregate notations which are used for less than 100 models respectively into *Other*: Process Documentation Template (86 models), jBPM 4 (76 models), XForms (20 models), and Chen Notation (3 models). The primarily used modeling notation is BPMN 2.0, which confirms that it is the de-facto standard for modeling business processes [4]. Therefore, we will focus on BPMN 2.0 models as we examine further properties.

Languages. Since SAP-SAI can be used by academic researchers, teachers and students all over the world, the models in SAP-SAM are created using different languages. For example, SAP-SAM includes BPMN 2.0 models in 41 different languages. Figure 2 shows the ten most frequently used languages for BPMN 2.0 models. Note that the vendor-provided example models, which are added to newly created workspaces, exist in English, German, and French. When a SAP-SAI workspace is created, the example models added to it are in German or French if the language configured upon creation is German or French, respectively; otherwise, the example models are in English. This contributes to the fact that more than half of the BPMN 2.0 models (57.43 %) are in English.

Elements. Figure 3 illustrates the occurrence frequency of different element types in the BPMN 2.0 models of SAP-SAM. It can be recognized that the element types are not equally distributed, which confirms the findings of prior research [14]. The number of models that contain at least one instance of a particular element type is much higher for some types, e.g., sequence flow (98.88 %) or task (98.11 %), than for others, e.g., collapsed subprocess (25.23 %) or start message events (25.42 %). Note that Fig. 3 only includes element types that are
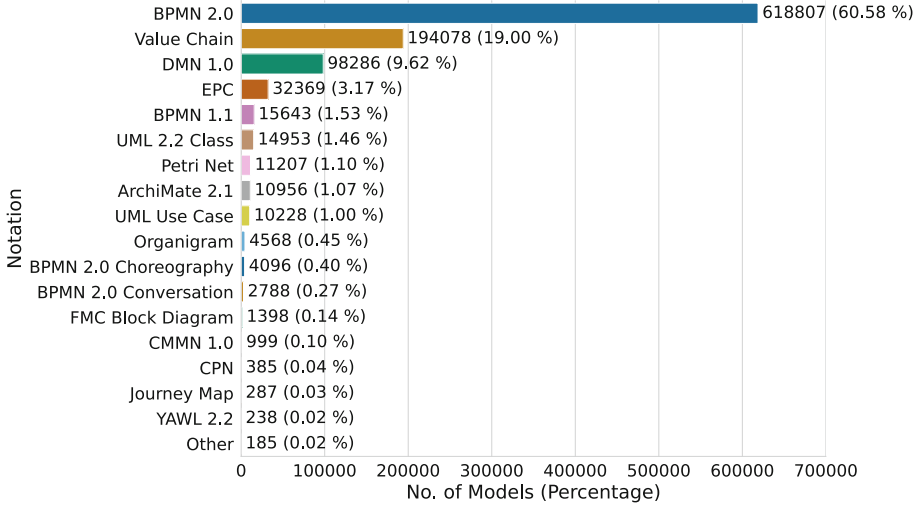
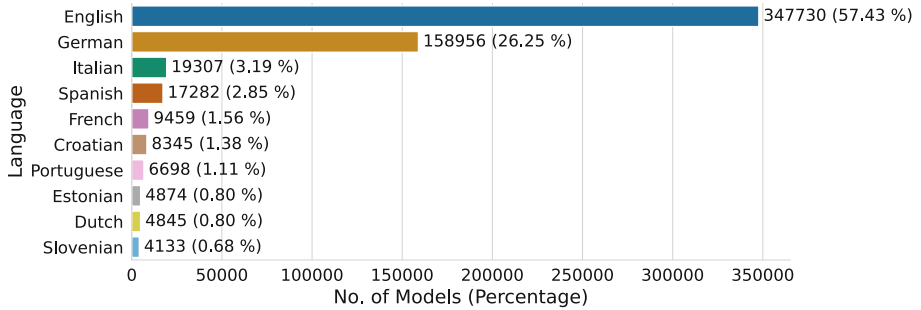**Fig. 1.** Usage of different modeling notations.

Fig. 2. Usage of different languages for BPMN 2.0 models.

used in at least 10 % of the BPMN 2.0 models. More than 30 element types are used by less than 1 % of the models. On average, a BPMN 2.0 model in SAP-SAM contains 11.3 different element types (median: 11) and 46.7 different elements, i.e., instances of element types (median: 40).

Table 1 shows the number of elements per model by type. For a compact representation, we aggregate similar element types by arranging them into groups. On average, connecting objects, which include associations and flows, make up the largest proportion of the elements in a model (mean: 23.1, median: 20).

Labels. All elements of a BPMN 2.0 model can be labeled by the modeler, which results in a total of 2,820,531 distinct labels for the 28,293,762 elements of all BPMN 2.0 models in SAP-SAM. Figure 4 depicts the distribution of label usage frequencies. We sorted the labels based on their absolute usage frequency in descending order and aggregated them in bins of size 10,000 to visualize the
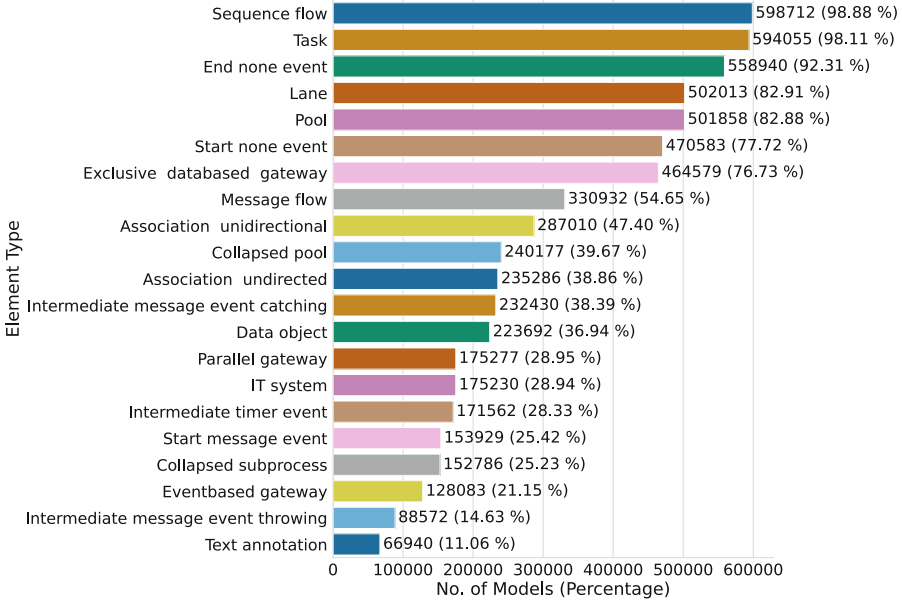
Fig. 3. Occurrence frequency of different BPMN 2.0 element types.

unevenness of the distribution. The first bin (leftmost bar in the chart) therefore contains the 10,000 most frequently used labels for the elements in the BPMN 2.0 models. Overall, 53.9 % of all elements in the BPMN 2.0 models are labeled with these first 10,000 labels. On the other hand, the long-tail distribution indicates that many of the labels are used for only one element of all BPMN 2.0 models. More precisely, 1,829,891 (64.9 %) of the labels are used only one time. The unevenness of the label usage distribution can again partly be explained by the vendor-provided examples in the dataset: The labels of the example processes appear very frequently in the dataset.

5   SAP Signavio Academic Models Applications

As pointed out above, large process model collections like SAP-SAM are a valuable and critical resource for BPM research. Process models from practice codify organizational knowledge about business processes and methodical knowledge about modeling practices. Both types of knowledge can be used by research, for example, for deriving recommendations for the design of future models. In addition, large process model collections are required for evaluating newly developed BPM algorithms and techniques regarding their applicability in practice.

To illustrate the potential value of SAP-SAM for the BPM community, the following list describes some application scenarios that we consider to be particularly relevant. It is neither prescriptive nor comprehensive; researchers can use SAP-SAM for many other purposes.

Table 1. Statistics of the number of elements per BPMN 2.0 model by type (grouped).

Element type groups	Mean	Std	Min	25%	50%	75%	Max
Activities	8.6	8.4	0	4	7	10	1543
Events	5.2	5.1	0	2	5	6	157
Gateways	3.7	4.4	0	2	3	4	303
Connecting objects	23.1	21.8	0	14	20	25	2066
Swimlanes	3.8	2.6	0	3	4	5	227
Data elements	1.3	3.4	0	0	0	2	266
Artifacts	0.9	4.0	0	0	0	1	529
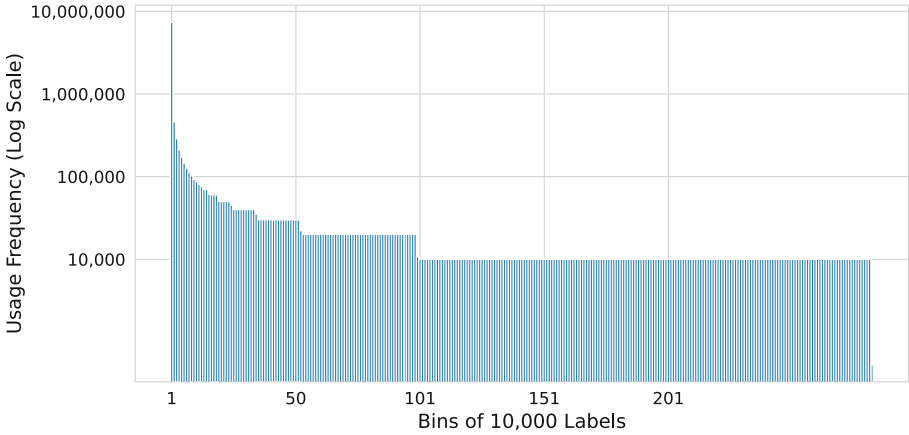
Fig. 4. Distribution of the label usage frequency in BPMN 2.0 models. Each bar represents a bin of 10,000 distinct labels.

Knowledge Generation. Process models depict business processes, codifying knowledge about the operations within organizations. This knowledge can be extracted and generalized to a broader context. Hence, SAP-SAM can be considered as a knowledge base to generate new insights into the contents and the practice of organizational modeling. Example applications include:

– Reference model mining [20]: Reference models provide a generic template for the design of new processes in a certain industry. They can be mined by merging commonalities between existing processes from different contexts into a new model that abstracts from their specific features. By applying this technique to subsets of similar models from SAP-SAM, we can mine new reference models for process landscapes or individual processes, including, e.g., the organizational perspective. Similarly, we could identify, analyze, and compare different variants of the same process.
– Identifying modeling patterns [10]: Process model patterns provide proven solutions to recurring problems in process modeling. They can help in

streamlining the modeling process and standardizing the use of modeling concepts. A dataset like SAP-SAM which contains process models from many different modelers, provides an empirical foundation both for finding new modeling patterns and for validating existing ones. This also extends to process model antipatterns, i.e., patterns that should be avoided, as well as modeling guidelines and conventions.

Modeling Assistance. The modeling knowledge that is codified in SAP-SAM can also be used for automated assistance functions in modeling tools. Such assistance functions support modelers in creating or updating process models, accelerating and facilitating the modeling process. However, many assistance functions are based on machine learning techniques and therefore require a large set of training data to generate useful results. With its large amounts of contained modeling structures and labels, SAP-SAM offers a substantial training set, for example, for the following applications:

– Process model auto-completion [23]. By providing recommendations on possible next modeling steps, process model auto-completion can speed up modeling and facilitate consistency of the terms and modeling patterns that are used by an organization. Besides structural next element type recommendations, text label suggestions or even recommendations of entire process segments are possible. SAP-SAM can be used to train machine learning models for these purposes.
– Automated abstraction techniques [27]: One important function of BPM is process model abstraction, i.e., the aggregation of model elements into less complex, higher-level structures to enable a better understanding of the overall process. Such an aggregation entails the identification and assignment of higher-level categories to groups of process elements. SAP-SAM can provide the necessary training data for an NLP-based automated abstraction.

Evaluation. Managing large repositories of process models is a key application of BPM [7]. Researchers have developed many different approaches to assist organizations with this task. To make these approaches as productive as possible, they need to be tested on datasets that are comparable to those within organizations. Since SAP-SAM goes well beyond the size of related datasets, it can be used for large-scale evaluations of existing process management approaches on data from practice. Examples for these approaches include process model querying [19], process model matching [1], and process model similarity [6].

6   Limitations and Recommendations for Usage

As explained in the previous section, SAP-SAM can be used by the academic community to test and evaluate a plethora of tools and algorithms that address a wide range of process querying and business process analytics use cases. However, in the context of any evaluation, the limitations of the dataset need to be taken

into account. Considering the nature of SAP-SAM as a model collection that has been generated by academic researchers, teachers, and students, the following limitations must be considered:

– Many models in SAP-SAM exist multiple times, either as direct duplicates (copies) or as very similar versions. This includes vendor-provided example models or standard academic examples that are frequently used in academic teaching and research. The existence of these models can be used to evaluate variant identification and fuzzy matching approaches in process querying, but it negatively affects the diversity, i.e., the breadth of the dataset.
– Many models may be of low technical quality, in particular the models that are created by "process modeling beginners", i.e., early-stage students, for learning purposes. Although it can be interesting to analyze the mistakes or antipatterns in such models, flawed models can, for example, be problematic when using the dataset for generating modeling recommendations based on machine learning. Also, the mistakes that students make are most likely not representative of mistakes made by process modeling practitioners.
– Because many of the models have most likely been created for either teaching, learning, or demonstrating purposes, they presumably present a simplistic perspective on business processes. Even when assuming that all researchers, teachers, and students are skilled process modelers[4] and have a precise understanding of the underlying processes when modeling, the purpose of their models is typically fundamentally different from the purpose of industry process models. Whereas academic models often emphasize technical precision and correctness, industry models usually focus on a particular business goal, such as the facilitation of stakeholder alignment.

Let us note that this list may not be exhaustive; in particular, limitations that depend on a particular use case or evaluation scenario need to be identified by researchers who will use this dataset. Still, it is also worth highlighting that the rather "messy" nature of the model collection reflects the reality of industrial data science challenges, in which a sufficiently large amount of high-quality data (or models) is typically not straightforwardly available [11]; instead, substantial efforts need to be made to separate the wheat from the chaff, or to isolate usecases in which the flaws in the data do not have an adverse effect on business value, or any other undesirable organizational or societal implications. However, most process models go beyond *A-B-C* toy examples from exercises and the overall SAP-SAM dataset is of sufficient relevance and quality for facilitating research, for example, in the directions that we have outlined in the previous section.

When using SAP-SAM for academic research purposes, it typically makes sense to filter it, i.e., to reduce it to a subset of models that satisfy desirable properties. Here, we provide some recommendations to help with this step.

---

[4] Considering the previous point, that means even when focusing on the subset of the model collection that only entails models carefully created by skilled advanced students, teachers, and researchers.
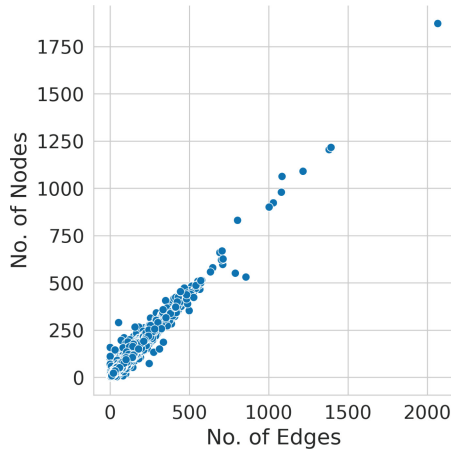
Fig. 5. Correlation of the number of nodes and edges in BPMN 2.0 models.

- It typically makes sense to filter out the vendor-provided example models that are created by the SAP-SAI system upon workspace creation.
- For many use cases, researchers may want to sort out process models that contain a very small or a very large number of elements. As can be expected for BPMN 2.0 models and is shown in Fig. 5, the number of nodes and the number of edges in a model are highly correlated. Hence, it is sufficient to filter according to the number of nodes. There is no need to additionally filter according to the number of edges.
- Similarly, researchers may want to sort out process models where the element labels have an average length of less than, for example, three characters to ensure that only models with useful labels are included.

Let us again highlight that example code that demonstrates how the dataset can be queried, as well as the code for the analysis in this paper is available at https://github.com/signavio/sap-sam.

7   Conclusion

In this paper, we have presented the SAP-SAM dataset of process and other business models. We are confident in our assumption that SAP-SAM is, by far, the largest publicly available collection of business process models. Hence, it can—despite the limitation that it entails "academic" models created by researchers, teachers, and students and not by process management professionals—serve as an excellent basis for developing and evaluating tools and algorithms for process model querying and analysis.

In the future, SAP-SAM can potentially be augmented by including the following additional data objects:

– Business objects/dictionary entries: In addition to models, SAP-SAI supports the creation of business objects, so-called *dictionary entries*. These objects represent, for example, organizational roles, documents, or IT systems and can be linked to models to then be re-used across a process landscape that entails many models. Dictionary entries facilitate process landscape maintenance, as well as reporting.
– Standard-conform XML serializations: The models in the SAP-SAM dataset are serialized using a non-standardized JSON format that *i)* supports a generalization of modeling notations and *ii)* is more convenient to use than XML-based serializations within the JavaScript-based front-ends of modern web applications. However, proprietary components exist that can—in the case of BPMN, DMN, and CMMN models—generate XML serializations which are compliant with the corresponding Object Management Group standards. Adding these XML serializations to the dataset can facilitate academic use, as many open-source and prototypical software tools support the open standards.
– PNG or SVG image representations: Similarly, to allow for a more straightforward visualization of models, PNG and SVG representations of the SAP-SAM models can be generated and included.

References

1. Antunes, G., Bakhshandelh, M., Borbinha, J., Cardoso, J., Dadashnia, S., et al.: The process model matching contest 2015. In: Enterprise Modelling and Information Systems Architectures, pp. 127–155. Köllen (2015)
2. Bruchez, E., Couthures, A., Pemberton, S., Van den Bleeken, N.: XForms 2.0. W3C Working Draft, World Wide Web Consortium (W3C) (2012). https://www.w3.org/TR/xforms20/
3. Chen, P.P.S.: The entity-relationship model-toward a unified view of data. ACM Trans. Database Syst. **1**(1), 9–36 (1976)
4. Chinosi, M., Trombetta, A.: BPMN: an introduction to the standard. Comput. Stan. Interfaces **34**(1), 124–134 (2012)
5. Corradini, F., Fornari, F., Polini, A., Re, B., Tiezzi, F.: Reprository: a repository platform for sharing business process models. In: BPM (PhD/Demos), pp. 149–153 (2019)
6. Dijkman, R., Dumas, M., Van Dongen, B., Käärik, R., Mendling, J.: Similarity of business process models: metrics and evaluation. Inf. Syst. **36**(2), 498–516 (2011)
7. Dijkman, R., La Rosa, M., Reijers, H.: Managing large collections of business process models-current techniques and challenges. Comput. Ind. **63**(2), 91–97 (2012)
8. van Dongen, B.: Bpi challenge 2020: domestic declarations (2020). https://doi.org/10.4121/UUID:3F422315-ED9D-4882-891F-E180B5B4FEB5
9. Dumas, M., La Rosa, M., Mendling, J., Reijers, H.A.: Fundamentals of Business Process Management. Springer, Berlin (2013)
10. Fellmann, M., Koschmider, A., Laue, R., Schoknecht, A., Vetter, A.: Business process model patterns: state-of-the-art, research classification and taxonomy. Bus. Process Manag. J. **25**(5), 972–994 (2018)
11. Houy, C., Fettke, P., Loos, P., van der Aalst, W.M., Krogstie, J.: Business process management in the large. Bus. Inf. Syst. Eng. **3**(6), 385–388 (2011)

12. Jensen, K.: Coloured petri nets. In: Petri Nets: Central Models and their Properties, pp. 248–299. Springer, Heidelberg (1987)
13. Lankhorst, M.M., Proper, H.A., Jonkers, H.: The architecture of the archimate language. In: Halpin, T., et al. (eds.) BPMDS/EMMSAD -2009. LNBIP, vol. 29, pp. 367–380. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-01862-6_30
14. Muehlen, M., Recker, J.: How much language is enough? theoretical and practical use of the business process modeling notation. In: Seminal Contributions to Information Systems Engineering, pp. 429–443. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36926-1_35
15. OMG: Business Process Model and Notation (BPMN), Version 2.0.2 (2013). http://www.omg.org/spec/BPMN/2.0.2
16. OMG: Case Management Model and Notation (CMMN), Version 1.1 (2016). http://www.omg.org/spec/CMMN/1.1
17. OMG: Decision Model and Notation (DMN), Version 1.3 (2021). http://www.omg.org/spec/DMN/1.3
18. Petri, C.A.: Kommunikation mit automaten. Westfäl. Inst. f. Instrumentelle Mathematik an der Univ, Bonn (1962)
19. Polyvyanyy, A.: Process querying: methods, techniques, and applications. In: Polyvyanyy, A. (ed.) Process Querying Methods, pp. 511–524. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-92875-9_18
20. Rehse, J.R., Fettke, P., Loos, P.: A graph-theoretic method for the inductive development of reference process models. Softw. Syst. Model. **16**(3), 833–873 (2017)
21. Schäfer, B., van der Aa, H., Leopold, H., Stuckenschmidt, H.: Sketch2bpmn: automatic recognition of hand-drawn BPMN models. In: La Rosa, M., Sadiq, S., Teniente, E. (eds.) Advanced Information Systems Engineering. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-79382-1_21
22. Scheer, A.W., Thomas, O., Adam, O.: Process Modeling using Event-Driven Process Chains, Chapter 6, pp. 119–145. John Wiley & Sons Ltd., Hoboken (2005)
23. Sola, D., Aa, H.V.D., Meilicke, C., Stuckenschmidt, H.: Exploiting label semantics for rule-based activity recommendation in business process modeling. Inf. Syst. **108**, 102049 (2022)
24. Sola, D., Meilicke, C., van der Aa, H., Stuckenschmidt, H.: On the use of knowledge graph completion methods for activity recommendation in business process modeling. In: Marrella, A., Weber, B. (eds.) BPM 2021. LNBIP, vol. 436, pp. 5–17. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-94343-1_1
25. Thaler, T., Walter, J., Ardalani, P., Fettke, P., Loos, P.: The need for process model corpora. In: FMI 2014, p. 14 (2014)
26. van der Aalst, W., ter Hofstede, A.: Yawl: yet another workflow language. Inf. Syst. **30**(4), 245–275 (2005)
27. Wang, N., Sun, S., OuYang, D.: Business process modeling abstraction based on semi-supervised clustering analysis. Bus. Inf. Syst. Eng. **60**(6), 525–542 (2018)
28. Weske, M., Decker, G., Dumas, M., La Rosa, M., Mendling, J., Reijers, H.A.: Model collection of the business process management academic initiative (2020). https://doi.org/10.5281/zenodo.3758705