Towards Understanding Process Modeling – The Case of the BPM Academic Initiative

Matthias Kunze, Alexander Luebbe, Matthias Weidlich, and Mathias Weske

Hasso Plattner Institute at the University of Potsdam
Prof.-Dr.-Helmert-Str. 2-3
14482 Potsdam, Germany
matthias.kunze@hpi.uni-potsdam.de, alexander.luebbe@hpi.uni-potsdam.de,
matthias.weidlich@hpi.uni-potsdam.de, mathias.weske@hpi.uni-potsdam.de

Summary. Business process models are typically graphs that communicate knowledge about the work performed in organizations. Collections of these models are gathered to analyze and improve the way an organization operates. From a research perspective, these collections tell about modeling styles, the relevance of modeling constructs, and common formal modeling mistakes.
With this paper, we outline a research agenda for investigating the act of process modeling using models of the BPM Academic Initiative. This collection comprises 1903 models, the majority captured in BPMN. The models were created by students from various universities as part of their process modeling education. As such, the collection is particularly suited to investigate modeling practice since it is probably unique in terms of modeling heterogeneity. As a first step, we characterize EPC and BPMN models of the collection using established process model metrics. Further, we investigate the usage of language constructs for these models. Our findings largely confirm the results obtained in prior, smaller studies on modeling in a professional context.

1 Introduction

Business process modeling is at the heart of modern organizations. Process models capture how work is performed in an organization and how business goals are achieved. Large organizations manage literally thousands of process models in process repositories [25]. These models form a knowledge base that is protected since it can be seen as a competitive advantage of an organization. However, analyzing model collections can yield valuable insights for language development and education. Modeling guidelines [21] and best practices for activity labeling [20] have been proposed based on investigations of process model collections. For instance, deriving the most common set of constructs used in a process modeling language can help to distinguish important from unimportant concepts for process model education. Understanding different modeling styles, their advantages and pitfalls, can be used to propose best practices for modelers. Common errors found in process model repositories can be used for education or for developing more easily comprehensible process modeling languages. Existing research on process

model collections has tried to answer these questions partially. However, findings are often limited because the used collections stem from homogeneous groups of modelers. Conclusions drawn have not been validated for a broader public and more research is needed to confirm or revoke existing findings.

In this paper, we outline a research agenda to evaluate a large collection of process models from the BPM Academic Initiative (BPMAI[1]) [13]—a joint venture of academic and industrial partners that aims at providing a mature process modeling platform for researchers and lecturers free of charge. Besides a comprehensive collection of lecture exercises, Signavio[2], industry partner of the BPMAI, offers a set of tools to design and manage business process models online. The modeling languages offered by the BPMAI include, but are not limited to, BPMN [24], EPCs [10, 26], and Petri Nets. The BPMAI is used by more than 4500 people from around 450 universities as part of their curriculum. Our investigations are based on anonymized models of a snapshot of the BPMAI collection from early 2011. It comprises 1903 models created by students, of which a majority of 1210 models was created using a BPMN 2.0 compliant shape set; 135 models are EPCs. Note that, due to the applied anonymization, personal and demographic data such as the level of graduation or the study discipline could not be related to the individual models.

The BPMAI collection is particularly suited to investigate modeling practice. In contrast to process model collections that have been around, e.g., the often cited SAP reference model [2], it shows a high heterogeneity along various dimensions. The models have been created by modelers that originate from universities all over the world. Modelers have different educational backgrounds, e.g., in business administration or computer science, capture processes in different natural and modeling languages, and represent operations from different business domains. Hence, empirical insights that are grounded on the BPMAI collection can be assumed to have a high external validity. Results on modeling styles, the relevance of modeling constructs, and common formal modeling mistakes derived from this collection are likely to be independent of any specific context in which process modeling is conducted. This kind of conclusions can hardly be drawn using homogeneous model collections created within a narrow context.

As a first step of a research agenda for the analysis of modeling practice using the BPMAI collection, this paper focuses on characterizing the collection and investigating the language usage for BPMN and EPC models. First, we derive descriptive statistics for the process models using an established set of process model metrics. This provides us with insights on the characteristics of the BPMAI collection. Second, we take up research on the relevance of modeling constructs. We contribute an analysis of the language usage for BPMN models and EPCs. For BPMN, our findings largely confirm the results of prior studies on professional modeling obtained with rather small sets of process models. Further, we present a comparison of language usage for BPMN and EPC, an aspect that has not been addressed in prior work.

[1] `http://bpt.hpi.uni-potsdam.de/BPMAcademicInitiative`
[2] `http://www.signavio.com`

The remainder of this paper is structured as follows. In Section 2, we review related research on process model collections. Then, we characterize the models of the BPMAI collection with process model metrics in Section 3. Section 4 is devoted to the analysis of language usage. We outline further research questions that relate to the BPMAI collection in Section 5, before we conclude in Section 6.

2 Related Work on Process Model Collections

The SAP reference model [2] is probably the most commonly used model collection in research. Published by SAP in 1997, it contains 604 process diagrams of the reference processes implemented in the SAP R/3 system during the mid nineties. These models have been used, e.g., for formal error analysis [18], for process model metrics development [17], extraction of reusable action patterns [27], and label analysis [16]. Approaches towards meaningful process similarity measures leveraged the reference model in conjunction with human assessment for similarity that has been captured in experiments [5, 4, 11]. Also data structures and algorithms for efficient search in large process model repositories [9, 12, 30] needed to resort to this collection, as virtually no other available collection contains as many models.

In short, these models became the reference for empirical research on model collections in the last ten years. But this set is not sufficient for research questions on modeling practice. First and foremost, it is limited to 604 models that represent one community of practice. The models have been created by a rather small group of modelers that have a similar background and capture only the processes of a single system. Moreover, the SAP reference model is based on EPCs [10, 26] and cannot provide answers to research questions towards other process modeling languages, in particular the Business Process Model and Notation (BPMN) [24].

Existing research on BPMN as a modeling language is limited. The most comprehensive evaluation was performed on a set of 120 models collected from consultants and the web [22]. The focus of this evaluation was on language usage and findings indicated that only a small subset of the BPMN modeling language is used in practice. At that time, it was a significant contribution to the ongoing OMG discussions on a BPMN Core Set. However, the findings resulted from a small collection that was assembled manually by the researchers. The evaluation is limited to the data given, mainly pictures of process models. The authors themselves note that clustering within the set was not possible due to the limited size. To draw conclusions with a high external validity, therefore, requires a large, heterogeneous model collection. In [8], the authors replicated one metric from [22], the syntax complexity graph, based on 166 BPMN models from an online modeling community[3]. The results indicated a quite similar complexity of syntax in both model collections. Unfortunately, further investigation was not presented. In Section 4, we use evaluation mechanisms of [22] to investigate and compare the language usage in the BPMAI collection.

[3] `http://bpmn-community.org`

In summary, existing work on modeling practice either relied on a rather homogeneous model collection, e.g., the SAP reference model, or used only a small set of process models. In this paper, we build upon prior work towards language usage and apply it to a set of 1345 BPMN and EPC models from the BPMAI collection.

3 Analysis with Process Model Metrics

In this section, we explore the models of the BPMAI collection with an existing set of process model metrics. As such, we derive descriptive statistics to characterize the collection. We first recall the metrics used in our analysis in Section 3.1. Then, Section 3.2 presents the obtained results.

3.1 A Set of Process Model Metrics

In essence, a process model is a graph that consists of nodes and edges. The former represent activities and the latter are used to encode a temporal and logical order of their execution [29]. Depending on the applied process definition language, nodes may also represent means to define control flow routing that goes beyond simple sequencing of activities, i.e., gateways in BPMN and connectors in EPCs. We refer to these nodes as routing nodes.

Against this background, it is not surprising that there have been various efforts to adapt generic structural metrics that are defined for graphs for the use case of process models, e.g., [15, 23, 1]. Such efforts are particularly inspired by metrics in software engineering or network analysis. Metrics that have their roots in these domains are conceptually close since they are also applied to graphs that define control flow dependencies. See [19] for a discussion of this relation.

For our analysis, we rely on process model metrics that focus on comprehensibility. The process models of the BPMAI collection have been created by students as part of modeling exercises, so that process documentation and communication can be seen as the primary drivers for model creation. Hence, comprehensibility is the major quality criterion for these models. Although we do not assess comprehensibility explicitly, we leverage single metrics for a descriptive characterization of the collection. To this end, we employ the process model metrics presented by Mendling [19]. They integrate many of the aforementioned metrics and provide a multi-dimensional framework for the analysis of single process models. The metrics have been evaluated, again, using the SAP reference model [2] for their ability to predict EPC modeling errors. Nevertheless, the metrics are more universal because they provide a generic characterization of a process model. We focus on the metrics that cover size, density, routing diversity, cyclicity, and concurrency of a process model. For these metrics, we shortly recall their definitions found in [19]. Note that, albeit commonly referred to as metrics, these measures may not be metrics in the mathematical sense.

Size. First and foremost, size of a process model may be assessed using the number of nodes (NN). This metric does not differentiate between types of nodes, i.e., activities or routing nodes. In addition, we compute the diameter ($Diam$), which is the longest path between any pair of nodes of a process model.

Density. Metrics for density relate the number of nodes and the number of edges of a process model to each other. In particular, we compute the number of edges divided by the (theoretical) maximum number of edges that may be observed for the number of given nodes ($Dens$). Closely related is the coefficient of connectivity (CNC), which is the ratio of edges and nodes. Focusing on the relation of edges and routing nodes, we determine the average and maximum degree of routing ($AvgDR$ and $MaxDR$), which capture the average and maximum number of nodes that a routing node is connected to.

Routing Diversity. To take the diversity of routing nodes into account, we compute the routing heterogeneity (RH) as the entropy over the observed types of routing nodes. In BPMN, activity nodes disclose implicit routing semantics, referred to as uncontrolled flow [24]. Therefore, BPMN activities have to be treated as exclusive routing nodes for incoming edges, and as concurrent routing nodes for outgoing edges.

Cyclicity. Since cyclic structures influence the comprehensibility of a process model, we also consider cyclicity. It is measured by the ratio of nodes that are part of a control flow cycle to all nodes of the process model (CYC).

Concurrency. Comprehensibility is further influenced by the level of concurrency. It is assessed by the token split (TS), which is the sum of the outgoing edges of routing nodes that may create concurrent behavior, i.e., AND or inclusive OR semantics, minus one.

3.2 Evaluation of the BPMAI Collection

Using the metrics introduced in the previous section, we investigated all BPMN and EPC models of the BPMAI collection. We implemented the metrics as part of a Java library that also comprises utility classes to access the models of the BPMAI collection[4]. An overview of the obtained results is presented in Table 1. For each of the metrics, the table depicts the average and the maximal value over all BPMN models or EPCs. To further characterize the obtained values, we also list the median along with the upper and lower quartile. Those indicate which values have been obtained for the 25th percentile (Lower Q), the 50th percentile (Median), or the 75th percentile (Upper Q), respectively. In the following, we discuss the results along the aforementioned dimensions of measurement.

It is worth mentioning that whenever we refer to nodes in the context of a metric, we mean a control flow node, i.e., an activity, event, or routing node. In many process model specifications, edges are captured as a pair of connected nodes. However, in the BPMAI modeling tool, cf. Section 1, edges are components

[4] See `http://code.google.com/p/bpmai/`.

Table 1. Metrics for the BPMN and EPC models of the BPMAI collection

	Size	Density					Rout. Div.	Cyclicity	Concurrency
	NN	*Diam*	*Dens*	*CNC*	*AvgDR*	*MaxDR*	*RH*	*CYC*	*TS*
Results for the BPMN models:									
Avg	15.6	6.52	0.09	0.79	1.15	1.84	0.17	0.03	0.65
Max	156.0	69.0	0.5	1.6	6.0	11.0	0.9	0.75	28.0
Upper Q	19.0	9.0	0.13	1.0	1.4	2.0	0.39	0.0	1.0
Median	11.0	5.0	0.07	0.88	1.23	2.0	0.0	0.0	0.0
Lower Q	7.0	2.0	0.03	0.67	1.0	1.0	0.0	0.0	0.0
Results for the EPC models:									
Avg	19.88	10.84	0.07	0.85	0.9	0.96	0.06	0.03	0.47
Max	123.0	50.0	0.5	1.15	4.0	4.0	1.0	0.57	5.0
Upper Q	27.0	16.25	0.09	1.03	2.0	2.0	0.0	0.0	1.0
Median	15.5	9.0	0.05	0.95	0.0	0.0	0.0	0.0	0.0
Lower Q	8.75	3.0	0.03	0.8	0.0	0.0	0.0	0.0	0.0

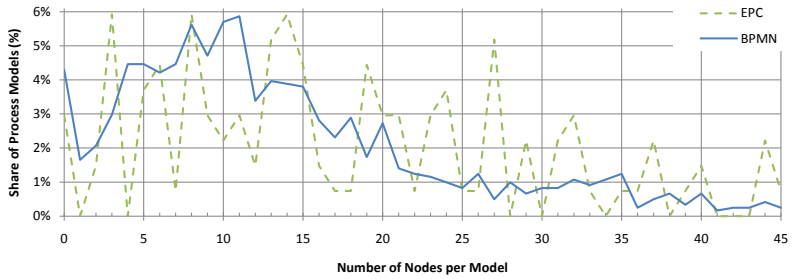


Fig. 1. Share of process models relative to their size, in terms of the number of nodes.

of their own and do not need to be connected to nodes. Thus, we only considered those edges that are connecting nodes with both their ends for computing the metrics.

Size. The average size of the models in the BPMAI collection is around 16 nodes (BPMN) and 20 nodes (EPC), respectively. We consider this to be remarkable because we have not applied any filtering to the collection. Our collection includes *all* models created within a certain timeframe, not only those that are intended to be published. Consequently, we assume the collection to include several model stubs that have not been completed by the modeler. Against this background, the observed sizes hint at a considerable complexity of the models with large models comprising more than hundred nodes. Further, the EPCs contain more nodes on average, which may be explained by the bipartite structure of EPC graphs that requires an alternating order of EPC functions and EPC events. This observation is underpinned by the values for the quartiles. 25% of the EPCs have more than 27 nodes compared to 19

nodes for BPMN. The differences in model size between both languages are also illustrated in Fig. 1. It shows that EPCs show a larger variety in their size compared to the BPMN models.

Our assessment of size in terms of the model diameter indicates that we observe longer paths in EPCs compared to BPMN models. For half of the BPMN models, the longest path comprises at most five edges compared to nine edges for EPCs. This difference, nearly twice the value, is larger than what can be expected from the difference in node size between both modeling languages. Finally, the relation between observed node sizes and diameters allows concluding on the complexity of the model structure. A model that is completely sequential shows a diameter that is the number of nodes minus one. As such, our results indicate that the models are not of such a trivial structure.

Density. Since process models are rarely complete graphs (in which each pair of nodes is directly connected), the observed values for the *Dens* metric are rather small. Here, the maximal values of 0.5 are obtained for minimal models that comprise two nodes that are connected by one edge. With the metrics that leverage all nodes and edges, i.e., *Dens* and *CNC*, we obtain similar results for BPMN models and EPCs. Still, there are differences once density is assessed with a focus on routing nodes. The implementation of control flow routing is notably more elaborated for BPMN models compared to EPCs, as more than half of all BPMN models expose a maximum routing degree of greater or equal to 2, whereas this holds only for 25% of EPCs. Note that, values in the *Average* row for *AvgDR* and *MaxDR* of Table 1 result from computing the average over all models, including those that do not comprise any routing at all and thus have a *MaxDR* of 1. Here, the average maximum routing degree of all BPMN models are twice as high as for EPCs. Further, we observe that a routing node in BPMN models is connected to a maximum of 11 nodes, whereas the maximum number of adjacent nodes is four for routing nodes in EPCs. This indicates that the EPCs do not show nodes that fan out a high number of branches, as it can be observed, e.g., in the EPCs of the SAP reference model.

Routing Diversity. A routing heterogeneity of zero indicates that only one type of routing node is used, a value of one means that all types introduced by the language are used in a model. Our results indicate that BPMN models, on average, rely on a larger share of the possible types of routing nodes. This is remarkable since BPMN defines a lot more different types of routing nodes, i.e., types of gateways or tasks with multiple incoming or outgoing control flow edges. Even though there is an EPC that comprises all kinds of routing nodes, at least 75% of the EPCs comprise only a single type of routing node (connector) or no routing nodes at all.

Cyclicity. The models of the BPMAI turn out to be mostly acyclic. There are 178 BPMN models and 18 EPCs that show a control flow cycle. Apparently, this leads to very low values obtained for the cyclicity metric, which assesses the ratio of nodes that are part of a control flow cycle to all nodes. However, there

are notable exceptions, such as a BPMN model for which 75% of the nodes are part of a cycle (see Table 1). Note that this metric considers only control flow cycles but neglects high-level constructs, e.g., BPMN loop markers, to express repetitive behavior.

Concurrency. The level of concurrency observed in the model collection is rather low, too. There are 355 BPMN models and 41 EPCs that may show concurrent behavior. Again, we observe models with exceptional behavior. For instance, there is a BPMN model with a token split of 28. Note that such a high token split does not mean that there may be 28 concurrent branches. It may also be caused by several routing nodes, each creating a small number of concurrent branches which are synchronized before further concurrent branches are spawned. Comparing BPMN models and EPCs, we observe that EPCs in the BPMAI collection show less concurrency.

4 Analysis of Language Usage

In their joint paper, zur Muehlen and Recker [22] approached the question "How much language is enough?" and evaluated quantitatively, which of the modeling constructs provided by BPMN [24] are used regularly. We applied their investigations to the BPMAI model collection. In contrast to [22], our evaluation investigates EPCs [10, 26] and the more recent version BPMN 2.0 [24]. Again, we implemented the analysis as part of the Java library mentioned in Section 3.2.

4.1 Usage of Process Model Constructs

Process modeling languages usually offer a large set of constructs, each with a unique meaning. While we addressed control flow concepts in Section 3, here we consider the complete spectrum of constructs, BPMN and EPC offer to the modeler, e.g., data, resources, and events. In order to identify, which process model constructs have been used most, we simply counted, for each construct provided by the process modeling language, in how many models it is contained. To further characterize the usage of a modeling language, we also elaborated on the general heterogeneity of process models, i.e., whether process modelers employ the full expressiveness of a language or rather resort to a small share for their models.

In line with [22], we found the BPMN constructs Task, Sequence Flow, Start and End Event to be the most prominent constructs. However, the major share of BPMN models we evaluated also contains Pools, Lanes, and (Databased) Exclusive Gateways, cf. Fig. 2(a). Pools and lanes occur with the same frequency, because, in the BPMAI modeling tool, a pool always contains at least one lane.

A similar usage frequency for corresponding constructs to the above can be observed for EPCs, i.e., Function, Control Flow, and Events occur most frequently, along with the XOR Connector in more than 50% of the models, cf. Fig. 2(b). A Relation is an edge that is supposed to combine a Function or Event

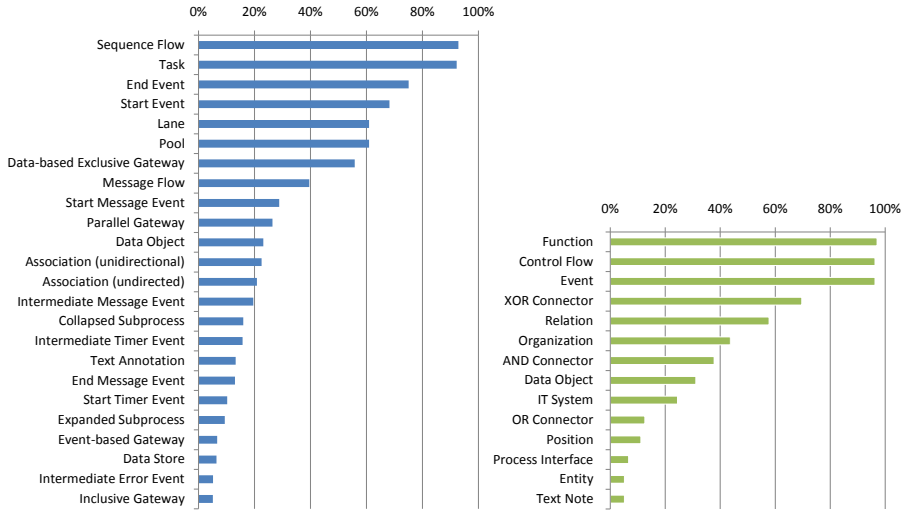


Fig. 2. Usage Frequency of Modeling Constructs from BPMN (a) and EPC (b). (Only constructs with a frequency above 5% are listed.)

with an Organization, Position, Data, or System construct. However, as these constructs expose a significantly lower frequency of occurrence, we discovered that the Relation edge has been used falsely in spite of a control flow edge.

The above observation indicates that most process models are very simply, i.e., consist of Tasks and Sequence Flows; EPCs contain almost equally many Events as Functions, as the modeling tool requires Functions and Events to alternate. In practice, we see many of these models that represent business processes in a coarse grain, where tasks describe phases, rather than individual activities. It is also worth noticing that the Parallel Gateway (BPMN) and the AND Connector (EPC) have been used in less than 40% of all models, which suggests that most of the models describe sequential behavior.

In BPMN, we find Message Flow and Message Start Event at high positions, which indicates that modelers actually use the capability of BPMN process diagrams to describe interacting processes of different parties. This concept is not available in EPCs.

To assess the heterogeneity of process models, we examined the number of unique process modeling constructs, i.e., the types of model elements, that have been used in the models of our collection. For BPMN, we differentiate 63 unique modeling constructs, whereas we did not distinguish Interrupting from Non-Interrupting Events, different types of Data Objects, nor different task types, e.g., Human Task, Service Task. For EPC, we distinguished 14 constructs. These numbers are constituted by the modeling tool, cf. Section 1, that has invariably been used to create the models. Fig. 3 shows an overview of the heterogeneity distribution with regard to the number of used modeling constructs.

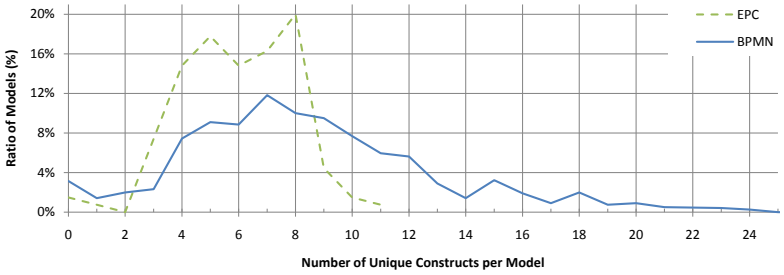


Fig. 3. Frequency of unique constructs per model: BPMN models expose a higher diversity than EPCs.

On average, a BPMN model used 8.46 different modeling constructs, compared to 5.97 unique constructs used in EPC models. The most heterogeneous BPMN model used 27 different constructs, compared to 11 in EPCs. For both modeling languages, 75% of all models employed more than 5 distinct constructs. Nevertheless, BPMN models show a considerable higher diversity than EPCs, cf. Fig. 3.

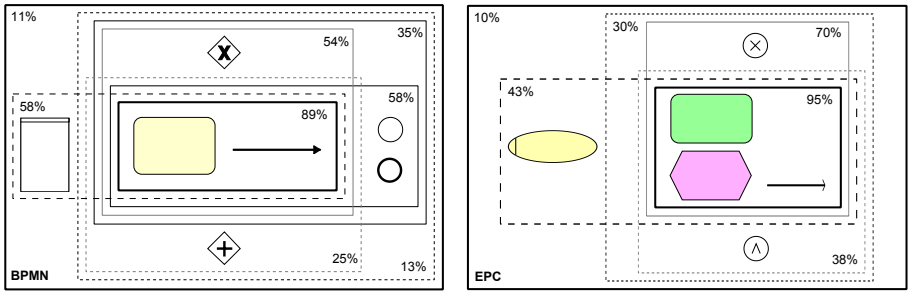
A set of 20 unique constructs of BPMN has been used in less than 10% of all models, but only one has never been used, which is the Event Subprocess. For EPCs there is no single construct that has never been used, and only up to 5 of the 14 unique constructs have been used in less than 10% of all models.

The observations above showed that certain, basic modeling constructs are more prominent than others, which holds true for BPMN and EPC. Also, we discovered that the majority of process models exposes a rather low diversity compared to the expressiveness of the modeling language. This approves the existence of a rather compact vocabulary of modeling constructs used for process modeling.

4.2 Vocabulary of Process Models

In this section, we illustrate subsets of the respective modeling language that are shared among most process models and elaborate on differences of these vocabulary sets between EPC and BPMN. Thus, we iteratively built sets of the most prominent process model constructs, cf. Fig. 2, and counted those models, which contain the particular set of constructs. Starting from the very core of the constructs used together in the most models, we extended this set stepwise by constructs that would exclude the fewest models. The results are visualized in Fig. 4.

Obviously, the most compact subsets comprise the most prominent modeling constructs. The very core of both modeling languages, BPMN and EPC, is simply made of activities and edges between them, i.e., Tasks and Sequence Flow arcs in BPMN, and Functions, Events, and Control Flow arcs for EPCs, respectively. As mentioned earlier, EPCs require a bipartite structure, which yields the combined usage of Functions and Events.



(a) (b)

Fig. 4. Frequency of Used Vocabulary for BPMN (a) and EPC (b) Diagrams.

Direct extensions of the core set of BPMN, cf. Fig. 4(a), include Pools and Lanes on the one hand, and Blank Start Events and Blank End Events on the other hand; almost immediately follows the cluster of Tasks, Sequence Flow and Exclusive Gateways—more than half of all models share either of these subsets. Zur Muehlen and Recker [22] recognized the same aspect and attributed it to the two main application areas for BPMN: The usage of Pools and Lanes without advanced control flow constructs is prevalent to express organizational partitioning of a process according to roles and their responsibilities, whereas routing constructs are used separately for a detailed specification of the control flow of a process.

For EPC models, cf. Fig. 4(b), we see a much higher ratio of models that consist of basic control flow and XOR Connectors (70%) than in BPMN, where only 54% of the models share the basic subset with Databased Exclusive Gateways. This is due to the fact that BPMN allows to model choreographies, while EPCs do not.

For both modeling languages, the Parallel Gateway and AND Connector, respectively, follow way behind: Less than half of the models of the above clusters share constructs for parallel routing. The Inclusive Gateway and OR Connector of BPMN and EPC play only a minor role, with 5% in BPMN and 13% in EPCs, respectively.

5 Further Research Agenda

In this paper, we examined aspects of process models and investigated usage of process modeling languages, whereas conclusions and recommendations for practitioners shall be addressed in future work. Accordingly, we plan to direct further research towards communities of practice and model evolution. In this section, we outline a set of research challenges, for which the model collection might be leveraged. The topics are interrelated and can benefit from each other.

Assessing process similarity measures. Many models in the collection have been created based on a limited set of exercises. Thus, we can expect multiple models to represent the same scenario. The BPMAI model collection is unique in this respect and leads to many interesting opportunities. One of them is to review process model similarity measures, cf. [6, 4, 11], e.g., whether they are able to recover process models sprung from the same scenario.

Identify communities of practice. Based on identification mechanisms, such as similarity measures or the metrics discussed above, we can identify groups of people with similar modeling styles or language sets. A community of practice might result from the education in a university. We might as well be able to identify different communities even within a university. Finally, it might be that a community of practice is established by certain modeling patterns that result from cognitive thinking styles. We can also link that information to people, who have edited the model.

Process model evolution. Based on the versions of each model in the repository, it is possible to trace the process of model creation and get more insights into the act of process modeling. This can also be linked to the number of people that have contributed to the models. On average, models in the BPMAI collection have 4.27 revisions and 20% have 6 or more revisions. One particular BPMN process model even exposed 80 revisions.

Analyze interacting models. The ability to depict processes interacting with other processes is a feature of BPMN but not feasible in EPCs. There exists research on the theoretical aspect of interacting processes, cf. [3]. The BPMAI model collection enables empirical research on the way people use this ability and identify typical mistakes that might result from it. In our snapshot of the BPMAI models, 479 BPMN models contain the Message Flow construct and 458 models consist of more than one Pool, which indicates a reasonably large basis for empirical analysis.

Link modeling errors to process metrics. Analogous to Mendling's [18] evaluation of process model metrics against common EPC modeling errors, this can be done for BPMN and other process modeling languages as well. As a result, one can obtain a weighted set of metrics that predict whether a model might have an error. Comparing this list to the findings by Mendling might result in the identification of a core set of model metrics that are relevant to avoid modeling mistakes. We have to stress, though, that all models where created with a professional tool which excludes many syntactic mistakes by design.

Obtain complex models for understandability tests. Ongoing research on process model understandability [14] and complexity metrics [7] can leverage the models in this sample set. Typically, the models used in those tests have been designed by the researchers because they need to fulfill very specific properties. With the large sample set given and the strong diversity in the collection, it should be possible to identify and reuse models from this collection for empirical research, enabling a higher external validity for the experiments.

Linguistic Analysis. One of the major challenges for recent research topics in BPM, e.g., to compute the similarity of business processes [6, 4, 11] or to manage consistency between processes [28], is the alignment of process models, i.e., the identification of corresponding nodes in two or more business processes. The main obstacle is the heterogeneity of used terms in process model inscriptions, to which many solutions, based on syntactic, semantic, and linguistic approaches, have been proposed. The BPMAI models show over 25 natural languages, English and German being the most commonly used. This offers a source to train algorithms towards aligning process models, as well as to evaluate other means of process modeling, e.g., the usage of a limited vocabulary for labeling.

This list of research opportunities is not a closed set. We hope to collect more ideas as part of the discussion that is triggered with this paper. We do not claim to tackle all aspects in our future research. Instead we want to inspire researchers to leverage the potential given in such a model collection.

6 Conclusion

In this paper, we laid the foundations for research on modeling practice using a particularly suited model collection, the BPMAI collection. This collection allows for investigating modeling practice in a unique setting. The models show a high heterogeneity with respect to the educational backgrounds of the modelers, the used natural and modeling languages, and the considered business domains. Hence, empirical insights that are derived using the BPMAI collection can be assumed to have a high external validity.

While the process models are well suited for many use cases towards understanding process models, they are limited by few aspects. All models have been created by academics, i.e., mostly by students as part of course assignments. While this leads to a high heterogeneity with regards to modeling practice, usage of process model constructs, and terminology, the models may expose certain characteristics, e.g., modeling style, that is attributed to their lecturers. Also, every model was created with the same tool, the Signavio process model editor, which is aware of syntax rules of modeling language specifications, and thus prevents many modeling mistakes. Consequences of this are, e.g., the bipartite occurrence of Functions and Events in EPCs and the co-occurrence of Pools and Lanes in BPMN.

This paper first presented a characterization of the BPMN models and EPCs of the collection, a set of 1345 models, using several established process model metrics. In this data set, BPMN models expose a higher diversity than EPCs in terms of construct heterogeneity, i.e., the number of unique modeling constructs. This is due to the greater and more detailed expressiveness of the BPMN language compared to EPCs, which is also leveraged in process models. At the same time, the overall heterogeneity of process models is rather low among both modeling languages, which suggests that most models are kept concise.

The investigation of language usage in these models showed results in line with [22]. Most modelers resort to a rather limited set of vocabulary, with simple activity sequences at the very core. Remarkably, the vocabulary subsets of BPMN and EPC are fairly similar. It suggests a true core set of relevant modeling concepts.

We concluded our work with an outline of a research agenda that uses the models of the BPMAI collection. These include more topics towards understanding process modeling, for example, through model evolution, communities of practice, assessing modeling mistakes, but also towards other opportunities that rely on a heterogeneous model collection, e.g., similarity and linguistic analysis of process models.

Acknowledgements. We are grateful to Signavio, in particular Gero Decker, for providing access to a snapshot of anonymized models from the BPM Academic Initiative. We also thank Katrin Honauer and Philipp Berger who supported us in the implementation of the experiments.

References

1. Jorge Cardoso. Business Process Control-Flow Complexity: Metric, Evaluation, and Validation. *Int. J. Web Service Res.*, 5(2):49–76, 2008.
2. Thomas Curran, Gerhard Keller, and Andrew Ladd. *SAP R/3 Business Blueprint: Understanding the Business Process Reference Model.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1997.
3. Gero Decker. *Design and Analysis of Process Choreographies.* PhD thesis, Hasso Plattner Institut an der Universität Potsdam, 2009.
4. Remco Dijkman, Marlon Dumas, Boudewijn van Dongen, Reina Käärik, and Jan Mendling. Similarity of Business Process Models: Metrics and Evaluation. *Information Systems*, 36(2):498 – 516, 2011.
5. Remco M. Dijkman, Marlon Dumas, and Luciano García-Bañuelos. Graph Matching Algorithms for Business Process Model Similarity Search. In *BPM '09*, volume 5701 of *LNCS*, pages 48–63. Springer, 2009.
6. Marlon Dumas, Luciano García-Bañuelos, and Remco M. Dijkman. Similarity Search of Business Process Models. *IEEE Data Eng. Bull.*, 32(3):23–28, 2009.
7. K. Figl and R. Laue. Cognitive Complexity in Business Process Modeling. In *CAISE '11*, volume 6741 of *LNCS* 2011, pages 452–466. Springer 2011.
8. A. Grosskopf, J. Brunnert, S. Wehrmeyer, and M. Weske. bpmncommunity.org: A Forum for Process Modeling Practitioners - A Data Repository for Empirical BPM Research. In *ER-BPM '09*, volume 43 of *LNBIP*, pages 101–104. Springer 2009.
9. Tao Jin, Jianmin Wang, Nianhua Wu, Marcello La Rosa, and Arthur H. M. ter Hofstede. Efficient and Accurate Retrieval of Business Process Models through Indexing - (short paper). In *OTM '10*, volume 6426 of *LNCS*. Springer, 2010.
10. G. Keller, M. Nüttgens, and A.-W. Scheer. Semantische Prozessmodellierung auf der Grundlage "Ereignisgesteuerter Prozessketten (EPK)", 1992.
11. Matthias Kunze, Matthias Weidlich, and Mathias Weske. Behavioral Similarity – A Proper Metric. In *BPM '11*, volume 6896 of *LNCS*, pages 166–181. Springer 2011.

12. Matthias Kunze and Mathias Weske. Metric Trees for Efficient Similarity Search in Process Model Repositories. In *IW-PL '10*, volume 66 of *LNBIP*, pages 535–546. Springer, 2010.
13. Matthias Kunze and Mathias Weske. Signavio-Oryx Academic Initiative. In *BPM '10 Demonstration Track*, volume 615 of *CEUR*. 2010.
14. R. Laue and A. Gadatsch. Measuring the Understandability of Business Process Models – Are We Asking the Right Questions. In *BPD 2010*, 2010.
15. Gang Soo Lee and Jung-Mo Yoon. An Empirical Study on the Complexity Metrics of Petri Nets. *Microelectronics and Reliability*, 32(3):323–329, 1992.
16. Henrik Leopold, Jan Mendling, and Hajo A. Reijers. On the Automatic Labeling of Process Models. In *CAISE '11*, volume 6741 of *LNCS*, pages 512–520. Springer, 2011.
17. J. Mendling. Testing Density as a Complexity Metric for EPCs. In *German EPC workshop on density of process models*, 2006.
18. J. Mendling. *Detection and Prediction of Errors in EPC Business Process Models*. PhD thesis, 2007.
19. J. Mendling. *Metrics for Process Models: Empirical Foundations of Verification, Error Prediction, and Guidelines for Correctness*. Springer, 2008.
20. Jan Mendling, Hajo A. Reijers, and Jan Recker. Activity Labeling in Process Modeling: Empirical Insights and Recommendations. *Inf. Syst.*, 35(4):467–482, 2010.
21. Jan Mendling, Hajo A. Reijers, and Wil M. P. van der Aalst. Seven Process Modeling Guidelines (7PMG). *Information & Software Technology*, 52(2):127–136, 2010.
22. Michael Zur Muehlen and Jan Recker. How Much Language is Enough? Theoretical and Practical Use of the Business Process Modeling Notation. In *CAISE '08*, volume 5074 of *LNCS*, pages 465–479, Springer, 2008.
23. Mark E. Nissen. Valuing it Through Virtual Process Measurement. In *ICIS*, pages 309–323, 1994.
24. Object Management Group. Business Process Model and Notation (BPMN) Specification, Version 2.0.
25. Michael Rosemann. Potential Pitfalls of Process Modeling: Part A. *Business Process Management Journal*, 12(2):249–254, 2006.
26. A.W. Scheer, O. Thomas, and O. Adam. Process Modeling Using Event-driven Process Chains. 2005.
27. Sergey Smirnov, Matthias Weidlich, Jan Mendling, and Mathias Weske. Action Patterns in Business Process Models. In *ICSOC '09*, volume 5900 of *LNCS*, pages 115–129. Springer, 2009.
28. Matthias Weidlich, Jan Mendling, and Mathias Weske. Efficient Consistency Measurement based on Behavioural Profiles of Process Models. *IEEE TSE*, 37 (3), 2011, pp. 410-429. IEEE Computer Society.
29. Mathias Weske. *Business Process Management: Concepts, Languages, Architectures*. Springer, 2007.
30. Zhiqiang Yan, Remco M. Dijkman, and Paul Grefen. Fast Business Process Similarity Search with Feature-Based Similarity Estimation. In *OTM '10*, volume 6426 of *LNCS*, pages 60-77. Springer, 2010.