BPMN in the Wild: A Reprise

Jasmin Türker[1,3], Michael Völske[2], and Thomas S. Heinze[3,4]

[1] Technische Universität Darmstadt
jasmin.tuerker@stud.tu-darmstadt.de
[2] Bauhaus-Universität Weimar
michael.voelske@uni-weimar.de
[3] Institute of Data Science, Jena
German Aerospace Center (DLR)
[4] Cooperative University Gera-Eisenach
thomas.heinze@dhge.de

Abstract. The Business Process Model and Notation (BPMN) language is the de facto standard for business process modeling and automation. While there exists a number of public model collections, there is still a need for ample datasets for empirical analysis on the usage and practice of BPMN. In this paper, we present our repository mining approach for generating a corpus of open source BPMN models by systematically analyzing software projects on GitHub.com. In contrast to our previous work, where we have limited the analysis to a 10% sample of GitHub.com, including 6.1 million projects, we here report the results for our new analysis of the whole 82.8 million projects hosted on GitHub.com at the time of conducting the research. We describe the resulting dataset, containing 79,713 distinct BPMN models from 18,534 open source projects.

1 Introduction

Empirical research on business process modeling can help gaining insight into the usage and practice of modeling languages like BPMN and thus answering questions about, e.g., frequently and rarely used language features or the recurrence of certain modeling styles and preferences. There is however still a lack of publicly available collections of realistic process models, which hinders empirical analysis [11,13]. While, traditionally, researchers have resorted to controlled experiments, surveys or case studies, with often smaller and homogeneous collections of process models, systematically searching for models in open source software projects provides a complimentary approach to tackle the lack of real-world data.

Recent work on mining BPMN model artifacts from open source software repositories hosted on Github.com has highlighted the utility of such data mining efforts to support BPMN tool development [9] and to investigate BPMN usage "in the wild" [1,8,14]. Nevertheless, previous efforts have been rather limited in scale. In this paper, we describe our effort to mine BPMN artifacts from as close to all public Github.com repositories as is reasonably possible, as represented by the repositories included in the most recent GHTorrent dump [5][1]. All told, we

[1] https://ghtorrent.org

J. Manner, D. Lübke, S. Haarmann, S. Kolb, N. Herzberg, O. Kopp (Eds.): 14th ZEUS Workshop, ZEUS 2022, Bamberg, held virtually due to Covid-19 pandemic, Germany, 24–25 February 2022, published at http://ceur-ws.org

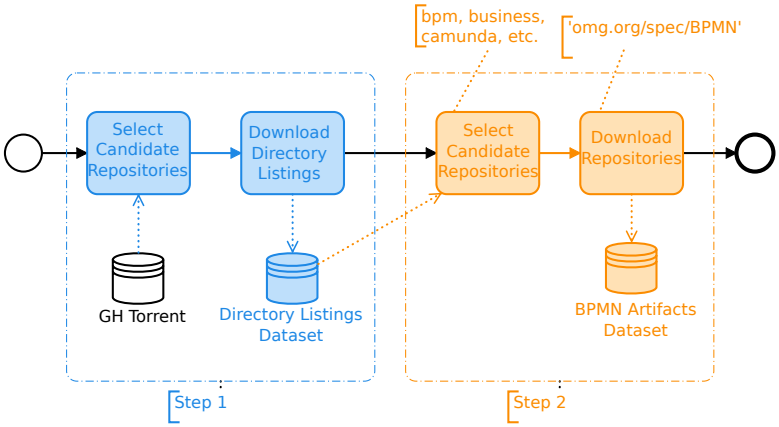

Fig. 1. The data mining pipeline

were able to identify 79,713 distinct BPMN artifacts and thereby can provide the to our knowledge most comprehensive collection of open source BPMN models.

In what follows, we describe our mining pipeline, as well as a preliminary analysis of the collected data. Our data mining pipeline follows a two-step approach, where we first identify interesting candidate repositories based on the names of the files and directories they contain, and inspect the files' contents only after this filtering. Since the initial collection of more than 82 million repositories' directory listings is the most time-consuming step, we make these listings available as a public resource for further research [18]. In the same spirit, we make links to the collected BPMN model artifacts available, as well [17].

2 Mining Approach for `Github.com`

Our data mining pipeline consists of two steps as shown in Fig. 1. In the first step, we retrieve all non-forked, non-deleted, public repositories from the latest *GHTorrent* database dump when conducting our research as of March 6th, 2021, and fetch a shallow and blob-less clone[2] of each repository from `Github.com`. We thus retrieve the file and directory names in the latest revision, and nearly nothing else, in less than 5 % of the time required to gather the same information via the REST API, which is limited to 5,000 requests per hour. The resulting *Directory Listings Dataset*[3] is collected over a span of about four weeks, and contains directory listings for the HEAD revisions of 82.8 million different software repositories. The corpus is available online [18].

To select candidate repositories for the second mining step, we queried the file and directory names in this dataset for the following keywords: *business, bpm,*

[2] `git clone --bare --single-branch --depth=1 --filter=blob:none`
[3] `https://zenodo.org/record/5856129` and `https://zenodo.org/record/5903352`

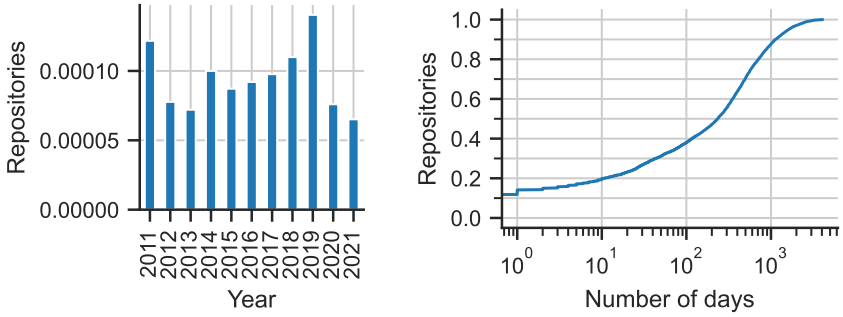

Fig. 2. Left: Fraction of GHTorrent repositories with BPMN artifacts (n=18,126); Right: Days between repository creation and time of latest update (n=12,235).

camunda, activiti, imixs, yaoqiang, modelio, signavio. The first two patterns are strongly connected with BPMN processes in general, and the latter six refer to popular modeling tools frequently used to design BPMN process models. This search matched at least one file or directory name in 1.1 million repositories. Of these, we again fetched a shallow clone, this time also including the blobs of the HEAD revision, i.e., the contents of its files. We filtered these repositories to contain at least one XML-serialized BPMN process model artifact, using the occurrence of the string `‘‘omg.org/sepc/BPMN’’` in at least one file as a heuristic. This left 18,534 repositories for inclusion in our *BPMN Artifacts Dataset*,[3] also available online [17]. Note that we currently only provide links to the identified BPMN process model artifacts and a Python code snippet for retrieving the models, similar to related research [6]. Due to the various licensing, copyright, and related constraints, further work will be required when the models are published on their own. The second mining step was completed in 3 days. A 135-node Kubernetes cluster was used for parallel processing in all steps.

3 Corpus and Preliminary Analysis

Fig. 2 (left) shows the fraction of repositories with BPMN artifacts relative to the total number of repositories in the GHTorrent database [5], broken down by repository creation year (number for 2021 based on the data available up to March). This differs notably from the overall number of repositories collected by GHTorrent: The latter had fewer repos in 2019 than in 2018, but more in 2020 than in either of the former years. Only 30 repositories in the dataset were created before 2011 (not shown in the plot), with the earliest creation date being June 2009. The empirical cumulative distribution for the number of days between repository creation and the most recent update is shown on the right of Fig. 2: While about 10 % of repositories were never updated again after a day past their creation, half of the repositories were still being updated after one year, and 20 %

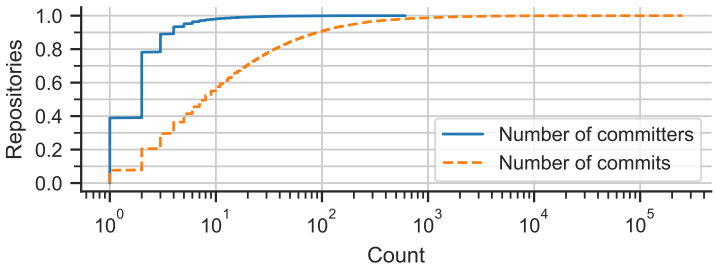

Fig. 3. Cumulative distributions for committers and commits per repository (n=15,773).

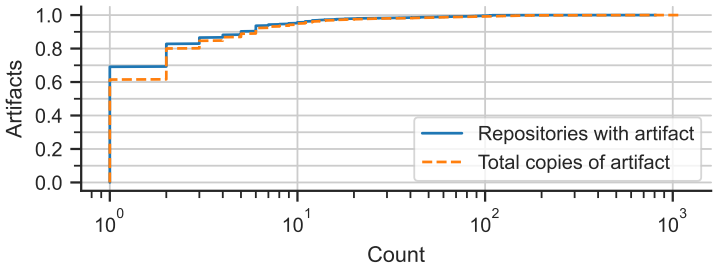

Fig. 4. Distributions for duplication among the 79,713 unique BPMN artifacts.

after two years. In the GHTorrent metadata, the creation date was unknown for 408, and the time of the last update for 6,299 of the repositories we collected; these missing data points are omitted from the above figures.

Fig. 3 shows the cumulative distributions for the number of committers and the number of commits per repository, for the subset of 15,773 repositories in our dataset for which this information is available in the GHTorrent database: Nearly 40 % of the repositories are single-committer projects, and 90 % have fewer than four contributors. Conversely, less than 10 % of the projects have only a single commit, and about 10 % have more than a hundred.

From the 18,534 collected repositories, we have identified 337,436 potential BPMN artifacts—XML files containing a URL for the BPMN 2.0 schema definition. Based on their SHA-1 hashes, we identify 79,713 unique BPMN artifacts; Fig. 4 shows the cumulative distributions for the number of copies of the artifacts and the number of repositories in which they appear. While SHA-1 file hashes provide a straightforward way to identify simple model duplicates, i.e., XML files with the same content, we acknowledge that finding similar though not exactly matching process models would require more laborious effort [7], which is however out of the scope of this paper. Similar to our previous findings in [9], we observe a large number of exact model duplicates. About 60 % of artifacts

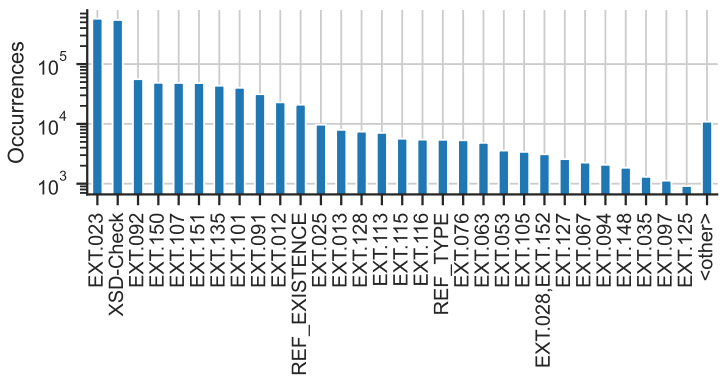

Fig. 5. Most common constraint violations detected by BPMNSpector among 60,779 unique, non-valid BPMN process models.

occur only once, and about 70 % in only one repository (even if multiple times); less than 5 % of all artifacts have ten or more copies across the dataset, meaning most of the copies are from a few thousand, frequently-copied artifacts. Based on several spot-checks, these are typically BPMN artifacts included as test cases with popular software libraries, which end up in many repositories, e.g., through accidentally-committed `node_modules` directories.

For a first look at the quality of the mined BPMN process models, we follow [9] and check all 79,713 distinct BPMN artifacts with the *BPMNSpector* [4] tool, which identifies 18,216 (23%) to be fully standards-compliant and 60,779 (76%) to contain at least one constraint violation (718 artifacts caused the tool to crash or hang); overall, 1,573,635 occurrences of 117 distinct constraint violations were found, the most common of which are shown in the histogram in Fig. 5. In terms of relative frequency, our results are rather similar to those reported by [9]: the most frequently violated constraints are Ext.023 (inconsistent definition of sequence flow; 36.5 % of violations / 38.6 % in [9]) and XsdCheck (violation of BPMN's normative XML schema; 34.8 % / 29.0 %) in both studies, and the next three most common violations, i.e., Ext.107, Ext.092, Ext.101 (inconsistent definition of sequence flow at start/end events, missing/ambiguous sources of data associations), all appear in our top eight, albeit in a slightly different order.

4 Related Work

Empirical research on business process modeling [13] often comes in the form of controlled experiments, surveys or case studies using a limited number of process models, usually with up to hundreds of process models, e.g., [3,12]. In recent years, though, various community initiatives have been started to increase the number of process model collections available for empirical analysis [11,10,2].

The *BPM Academic Initiative*[4], as a notable effort for business process modeling, provided a platform for creating and sharing business process models in academic teaching. In their 2011 publication [11], the initiators reported on 1,903 different process models, including BPMN, created by ca. 4,500 different users and covering various complexities and language features. While the recent number of models is 29,810[5], unfortunately, data collection is discontinued. Various similar platforms and datasets exist, including *GenMyModel*[6] with 12,575 BPMN models, *RePROSitory* [2] with 593 models, and *Camunda BPMN for research*[7] with 3,721 models, respectively. Another process collection has been created in the *BenchFlow* project [16], where companies donated process models for process engine benchmarking. The dataset described in [16] included overall 8,363 models, with a share of 64% of BPMN, but the collection is not publicly available. To the authors knowledge, the corpus presented in our paper with 79,713 distinct process models is the largest publicly available BPMN dataset. Not a dataset by itself, but an open-source business process analytics platform is offered by *Apromore*[8]. The platform has originally been conceived as process model repository [15] and now offers an extensible, service-based framework for a wide range of tools supporting the whole business process lifecycle, including process discovery, analysis, implementation, and monitoring.

In their previous work on repository mining for BPMN, the authors of this paper considered a sample of 10% of software projects on `GitHub.com` [8,9], resulting in a corpus of 8,904 BPMN model artifacts. An analysis proved the diversity of the corpus but also stressed the high number of model duplicates as well as model flaws as detected by BPMNspector [4]. While being a significant sample, the corpus presented and made publicly available in this paper provides ten times more models and thus a more complete and comprehensive picture about the practice of BPMN in open source software projects. The authors' prior dataset has been used in empirical studies on BPMN since then. In a recent study, BPMN models from various public process repositories, including the `GitHub.com` dataset [9] have been combined into a collection of 25,590 models and analyzed according to their complexity and usage frequency of BPMN language elements. Similarly, Lübke and Wutke investigate process layout choices in open source BPMN process models based on the corpus in [14].

5 Conclusion

In this paper, we introduce our approach to extract two hiqh-quality corpora from `Github.com`. Our *Directory Listings Dataset* is publicly available and may be an interesting starting point for research across various fields. Due to the amount of BPMN artifacts it contains, our *BPMN Artifacts Dataset* can contribute to

[4] `http://fundamentals-of-bpm.org/process-model-collections/`
[5] Numbers are reported for the time of writing this paper (25 January 2022).
[6] `https://www.genmymodel.com`
[7] `https://github.com/camunda/bpmn-for-research`
[8] `https://apromore.org`

the investigations on the usage and practice of BPMN process models. Subject of future work will be the exhaustive analysis of the retrieved corpus to further characterize the usage of BPMN in open soource software projects with respect to, e.g., complexity of process models, usage of different BPMN language constructs, utilized modeling tools, correlation of modeling tool with standards compliance, relation of repository metadata and BPMN artifacts, etc.

References

1. Compagnucci, I., Corradini, F., Fornari, F., Re, B.: Trends on the Usage of BPMN 2.0 from Publicly Available Repositories. In: BIR 2021. LNBIP, vol. 430, pp. 84–99. Springer (2021)
2. Corradini, F., Fornari, F., Polini, A., Re, B., Tiezzi, F.: RePROSitory: a Repository Platform for Sharing Business PROcess modelS. In: BPM PhD/Demos 2019. pp. 149–153. CEUR (2019)
3. Fahland, D., Favre, C., Jobstmann, B., Koehler, J., Lohmann, N., Völzer, H., Wolf, K.: Instantaneous Soundness Checking of Industrial Business Process Models. In: BPM 2009. pp. 278–293. Springer (2009)
4. Geiger, M., Neugebauer, P., Vorndran, A.: Automatic Standard Compliance Assessment of BPMN 2.0 Process Models. In: ZEUS 2017. CEUR Workshop Proceedings, vol. 1826, pp. 4–10. CEUR-WS.org (2017)
5. Gousios, G.: The GHTorent dataset and tool suite. In: MSR 2013. pp. 233–236. IEEE (2013)
6. Hebig, R., Quang, T.H., Chaudron, M., Robles, G., Fernandez, M.A.: The Quest for Open Source Projects that use UML: Mining GitHub. In: MODELS 2016. pp. 173–183. ACM (2016)
7. Heinze, T.S., Amme, W., Schäfer, A.: Detecting semantic business process model clones. In: ZEUS. CEUR Workshop Proceedings, vol. 2839, pp. 25–28. CEUR-WS.org (2021)
8. Heinze, T.S., Stefanko, V., Amme, W.: BPMN in the Wild: BPMN on GitHub.com. In: ZEUS 2020. CEUR Workshop Proceedings, vol. 2575, pp. 26–29. CEUR-WS.org (2020)
9. Heinze, T.S., Stefanko, V., Amme, W.: Mining BPMN Processes on GitHub for Tool Validation and Development. In: BPMDS/EMMSAD@CAiSE 2020. LNBIP, vol. 387, pp. 193–208. Springer (2020)
10. Ho-Quang, T., Chaudron, M.R.V., Robles, G., Herwanto, G.B.: Towards an Infrastructure for Empirical Research into Software Architecture: Challenges and Directions. In: ECASE@ICSE 2019. pp. 34–41. IEEE (2019)
11. Kunze, M., Luebbe, A., Weidlich, M., Weske, M.: Towards Understanding Process Modeling – The Case of the BPM Academic Initiative. In: BPMN 2011 Workshops. pp. 44–58. Springer (2011)
12. Leopold, H., Mendling, J., Günther, O.: Learning from Quality Issues of BPMN Models from Industry. IEEE Software **33**(4), 26–33 (2016)
13. Lübke, D., Pautasso, C.: Empirical Research in Executable Process Models. In: Empirical Studies on the Development of Executable Business Processes, pp. 3–12. Springer (2019)
14. Lübke, D., Wutke, D.: Analysis of Prevalent BPMN Layout Choices on GitHub. In: ZEUS 2021. CEUR Workshop Proceedings, vol. 2839, pp. 46–54. CEUR-WS.org (2021)

15. Rosa, M.L., Reijers, H.A., van der Aalst, W.M.P., Dijkman, R.M., Mendling, J., Dumas, M., García-Bañuelos, L.: APROMORE: an advanced process model repository. Expert Syst. Appl. **38**(6), 7029–7040 (2011)
16. Skouradaki, M., Roller, D., Leymann, F., Ferme, V., Pautasso, C.: On the Road to Benchmarking BPMN 2.0 Workflow Engines. In: ICPE 2015. pp. 301–304. ACM (2015)
17. Türker, J., Völske, M., Heinze, T.: Github BPMN Artifacts Dataset 2021 (Jan 2022). https://doi.org/10.5281/zenodo.5903352
18. Türker, J., Völske, M., Heinze, T.: Github Directory Listings Dataset 2021 (Jan 2022). https://doi.org/10.5281/zenodo.5856129