BPMN in the Wild: BPMN on GitHub.com

Thomas S. Heinze[1], Viktor Stefanko[2], and Wolfram Amme[2]

[1] German Aerospace Center (DLR)
thomas.heinze@dlr.de
[2] Friedrich Schiller University Jena
[wolfram.amme,viktor.stefanko]@uni-jena.de

Abstract. We present our efforts in creating and analyzing a corpus of BPMN process models by mining software repositories. Systematically searching for BPMN process artifacts in 6,163,217 repositories or 10% of all repositories hosted on GitHub.com, at the time of conducting our research, resulted in a diverse corpus of 8,904 BPMN 2.0 process models.

1 Introduction

Within the last years, an increasing number of software projects have shifted towards using platforms such as GitHub.com for their software development. Using these platforms as a source of data for empirical research allows for addressing a wide range of questions on the practice of software development and receives more and more attention, as indicated by the popularity of the flagship conference on the topic: *International Conference on Mining Software Repositories (MSR)*[1].

Research in the domain of business process modeling can as well benefit from such a data-driven approach. Due to characteristics of the domain, i.e., "process equals product", there is a lack of larger and commonly available datasets with real-world process models, which hinders empirical research in this area [2,11,13]. *Mining software repositories*, i.e., systematically retrieving, processing and analyzing process models from software repositories hosted on platforms such as GitHub.com, can help to overcome this lack and provides a complimentary approach to empirical research besides existing methods like case studies, experiments, and surveys. For example, research questions on how a language such as the *Business Process Model and Notation (BPMN)* [1] is used in practice can be addressed, in order to differentiate between the frequently and the rarely used parts of the language, thus advancing language and tool development. Analyzing modeling styles furthermore allows for investigating best practices and guidelines to help process designers. Eventually, best practices and tools as proposed by academic research or industry can be evaluated more realistically [12].

In this paper, we present our approach for mining software repositories on GitHub.com to create and analyze a corpus of BPMN process models. Due to the sheer number of repositories on GitHub.com and time constraints, we limited our approach to a randomly selected subset of 6,163,217 repositories or 10% of

[1] http://www.msrconf.org

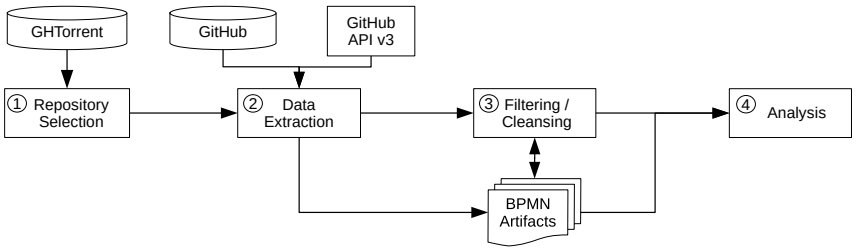

Fig. 1. Schematic illustration of the mining pipeline.

all software repositories on `GitHub.com` at the time of conducting the research. As a result, we were able to identify and analyze 8,904 distinct process models which are defined using BPMN 2.0's XML-based serialization format.

2 Related Work

The *Lindholmen dataset* has been an inspiration for this paper [5,12]. In Hebig et al. [5], the authors describe their approach to mine `GitHub.com` for UML models and report on gained insights. The dataset is considerably larger than our corpus, counting 93,596 models [12]. UML though is a family of general-purpose modeling languages while BPMN is one domain-specific modeling language. We are not aware of other work, which mines software repositories for BPMN models.

There have also been community efforts to create model collections [9]. The *BPM Academic Initiative* provides a platform to create and share business process models for academic teaching [11]. According to Ho-Quang et al. [9], the recent number of models is 29,285, but data collection has discontinued and the focus is on conceptual models as most models originate from students. A similar platform has been introduced last year under the name *RePROSitory* [2], including 174 business process models in its current database. Another initiative is the *BenchFlow project*, where business process models were collected from industrial partners. The authors claim to have collected 8,363 models, with a share of 64% of BPMN [13]. Unfortunately, the collection is not publicly available.

3 Mining BPMN on `GitHub.com`

Mining software repositories is a data mining task, consisting of steps of defining a research objective, selecting and extracting appropriate data, preprocessing and data cleansing, data analysis, and finally interpreting the analysis results.

In the first step of our implemented data mining pipeline, compare with Fig. 1, we got a list of all software repositories on `GitHub.com` by querying a local instance of the *GHTorrent*[2] database. We then randomly selected a subset

[2] `http://ghtorrent.org/`

of 6,163,217 non-forked repositories. All 6,163,217 repositories were examined for potential BPMN process model artifacts using the *GitHub API*[3] in the second step. To this end, the default branch and its file structure were queried for each repository. Potential BPMN process model artifacts were then identified by searching for the term `"bpmn"` in their file name and file extension. Among the analyzed repositories, we found 1,251 repositories, with at least one potential BPMN process model artifact and overall 21,306 artifacts. We downloaded the identified repositories and artifacts. In the third step, since the artifacts included a wide range of file formats, we filtered for BPMN 2.0's XML-based serialization format, which lowered the number of artifacts to 16,907. Additionally removing duplicates[4], yielded the corpus of 8,904 distinct BPMN 2.0 models. All the BPMN artifacts were finally subject to a preliminary analysis in the fourth step. Information on the corpus and analysis outcomes are available online [7][5].

4 Preliminary Analysis

In our preliminary analysis, we were mainly interested in the diversity of the found BPMN process model artifacts. We here sketch some of the results. Looking at the artifacts' age, more than each third was modified in the last year at the time of conducting our research. We though also found artifacts older than 8 years. Using the locations of repository contributors allowed us to reason on the artifacts' geographical origin, where China, USA, and Germany played prominent roles. The corpus spans a range of different model sizes. While half of the process models are smaller than 20 nodes, we also identified 57 models with more than 1,000 nodes. We were also able to confirm the finding reported in [5], that models play a rather static role in software repositories. Up to three quarter of all the BPMN process model artifacts were thus never updated at all.

Since the design of BPMN process models is known to be error-prone, we were also interested in the number of errors found in the models and the need for analysis tools to help process designers in avoiding those. Various analysis tools have been developed in recent years, ranging from simple linters [4], over tools based on data flow analysis [6,8], to full-fledged model checkers [3]. Note that most of the tools are evaluated using case studies or artifical process models. Therefore, evaluating analysis tools using our corpus of 8,904 BPMN process models allows to verify existing tool evaluations based upon a complimentary empirical means. We have chosen the linting tool *BPMNspector*[6] [4] for checking process models with respect to their compliance with the BPMN 2.0 standard [1]. Running the linter revealed violations of the standard's rules for almost all of the process models in the corpus. Only 1,471 models were identified as valid BPMN process models, thus confirming the results for the case study used to evaluate *BPMNspector* in [4], which found 42 invalid among overall 66 BPMN models.

[3] `https://developer.github.com/v3`
[4] `http://doubles.sourceforge.net`
[5] `https://github.com/ViktorStefanko/BPMN_Crawler`
[6] `https://github.com/uniba-dsg/BPMNspector`

5 Conclusion

In this paper, we introduced our approach of systematically extracting a corpus of BPMN business process models from software repositories hosted on `GitHub.com`. Mining a fraction of 10% of all software repositories, at the time of conducting our research, resulted in 8,904 distinct serialized BPMN 2.0 process models. We believe that our corpus of BPMN models provides a starting point for understanding more about the practice of BPMN. Note though the general limitations of the idea of repository mining [10]. In future work, besides increasing the coverage of analyzed software repositories, we want to research on questions about BPMN's use on `GitHub.com`, e.g., what are frequently and rarely used constructs or are there certain characteristics that can be used to predict modeling errors [11].

References

1. Business Process Model and Notation (BPMN), Version 2.0. Object Management Group (OMG) Standard (2011), `https://www.omg.org/spec/BPMN/2.0/PDF`
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. pp. 4–10. CEUR (2017)
5. 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)
6. Heinze, T.S., Amme, W., Moser, S.: Static analysis and process model transformation for an advanced business process to Petri net mapping. Softw.: Pract. & Exp. **48**(1), 161–195 (2018)
7. Heinze, T.S., Stefanko, V., Amme, W.: Mining von BPMN-Prozessartefakten auf GitHub. In: KPS 2019. pp. 111–120. DHBW Stuttgart (2019)
8. Heinze, T.S., Türker, J.: Certified Information Flow Analysis of Service Implementations. In: SOCA 2018. pp. 177–184. IEEE (2018)
9. 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)
10. Kalliamvakou, E., Gousios, G., Blincoe, K., Singer, L., German, D.M., Damian, D.E.: The Promises and Perils of Mining GitHub. In: MSR 2014. pp. 92–101. ACM (2014)
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. Robles, G., Ho-Quang, T., Hebig, R., Chaudron, M., Fernandez, M.A.: An extensive dataset of UML models in GitHub. In: MSR 2017. pp. 519–522. IEEE (2017)
13. Skouradaki, M., Roller, D., Leymann, F., Ferme, V., Pautasso, C.: On the Road to Benchmarking BPMN Workflow Engines. In: ICPE 2015. pp. 301–304. ACM (2015)