

DALG: The Data Aware Event Log Generator

David Jilg², Joscha Grüger^{1,2}, Tobias Geyer² and Ralph Bergmann^{1,2}

¹Artificial Intelligence and Intelligent Information Systems, University of Trier, 54296 Trier, Germany

²German Research Center for Artificial Intelligence (DFKI), Branch University of Trier, 54296 Trier, Germany

Abstract

Data and process mining techniques can be applied in many areas to gain valuable insights, but accessibility to real-world process data is severely limited. However, research, but especially the development of new methods, depends on a sufficient basis of realistic data. With adequate quality, synthetic data can be a solution to this problem. The SAMPLE [1] approach aims to mitigate this problem by generating multi-perspective synthetic event logs that make sense on a semantic level. In this paper, we present the tool DALG: The Data Aware Event Log Generator, which allows users to generate synthetic event logs using the SAMPLE approach.

Keywords

Event Log Generation, Synthetic Data, Process Mining

1. Introduction

The development and application of event log-based data analysis methods opens up great potential in many domains. However, there is a lack of data to develop and evaluate new approaches. This is especially true in domains where personal data is processed (e.g., medicine) or domains where business secrets are protected (e.g., industry or business).

One way to meet this challenge is to use high-quality synthetic data. An option to generate synthetic data in the procedural domain is to use process models and apply techniques such as token-based simulation, automata simulation, abduction, constraint satisfactory problem, or the boolean satisfiability problem. Tools such as Renew [2], RT-PLG [3], or MuDePS [4] use these approaches to generate synthetic event logs. However, all approaches and tools largely focus on the control flow perspective. If they can also generate synthetic data for the data perspective, they generate it based on conditions defined in the process model. In complex domains, this leads to synthetic event logs with semantically unrealistic data and thus to low quality.

The SAMPLE approach [1] addresses this challenge by providing a semantic description of the attributes, which is used as a basis for generating the data perspective during the event log generation. The semantic description supports the process that the generated data is more realistic. In this paper, we present the Data Aware Event Log Generator (DALG), which is an implementation of the SAMPLE approach. DALG was developed as a stand-alone desktop


BPM 2023: Demos and Resources

✉ david.jilg@dfki.de (D. Jilg); grueger@uni-trier.de (J. Grüger); tobias.geyer@dfki.de (T. Geyer); bergmann@uni-trier.de (R. Bergmann)

🆔 0000-0002-4841-3512 (D. Jilg); 0000-0001-7538-1248 (J. Grüger); 0000-0003-3072-1709 (T. Geyer); 0000-0002-5515-7158 (R. Bergmann)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

application and allows to define semantic attribute descriptions for a given data Petri net (DPN) and generate event logs based on them.

In the following, Sect. 2 describes how DALG differs from already existing tools for synthetic data generation by highlighting its innovative characteristics. Subsequently, Sect. 4 evaluates the maturity of the tool before we draw a conclusion in Sect. 5.

2. Innovation and characteristics

DALG permits users to generate synthetic event logs from available data Petri nets, which also support control-flow-only Petri nets. The initial step involves loading a data Petri net modeled in the *Petri Net Markup Language* format [5]. From this point on, the user can configure a wide variety of simulation parameters and can provide additional information about the provided model. Due to the variety of the configuration options and the resulting complexity for the user, the tool provides multiple features to assist the user during configuration. After the user starts the simulation, information about the simulation status is provided and the event logs are exported in the *eXtensible Event Stream* format [6] once the simulation has finished.

Semantic information. In developing the SAMPLE approach, it was determined that a data Petri net simply does not describe a process accurately enough and, therefore, does not contain enough semantic information about the process to produce realistic synthetic data in the data perspective without using additional information [1]. Therefore, SAMPLE was introduced as an approach for complementary semantic description of the data perspective in the generation of synthetic event logs. DALG implements and extends the SAMPLE approach and enables the semantic description of variables and transitions in a data Petri net. Via a user interface or a configuration file, the tool allows, among other things, to define intervals, distribution functions, value ranges and dependencies for variables. In contrast to existing tools, DALG uses this additional semantic information about the properties of variables and transitions to generate more realistic values by, for example, only generating values inside the given value ranges of the numerical variables. This enables the generation of semantically meaningful data in the data perspective.

- **Dependencies:** To achieve the generation of realistic values for the variables in a data Petri net, the dependencies between variables have to be considered. For example, consider a process model describing the treatment process for diagnosing and treating cancer patients. If this model has two variables describing a patient's gender and cancer type, then these two variables can be interdependent, as certain cancer types are exclusive to male patients (e.g., prostate cancer) or female patients (e.g., ovarian cancer). DALG lets the user express these dependencies using pairs of logical expressions and value range restrictions. The previously presented example could be partially described by adding the following dependencies to the variable describing the cancer type.

```
'gender == "female" => (!=,'prostate cancer')  
'gender == "male" => (!=,'ovarian cancer')
```

These two dependencies specify that the variable describing the cancer type cannot be "prostate cancer" if the patient is female, and "ovarian cancer" if the patient is male.

- **Distribution:** For numeric variables, DALG lets users define distribution functions for generated values. In the cancer treatment example, if age is included in the process model, its distribution can be specified. This ensures realistic value distributions in synthetic event logs.
- **Values:** A data Petri net modelled in the PNML does not specify what values should be written when a transition writes to a variable. Therefore, DALG allows the user to specify a list of values for each variable that is used when generating values for that variable. Additionally, each value can be specified with a weight that affects how likely the value is picked when a value for the variable is needed. For numeric variables, it is also possible to specify a value range instead of individual values.

Simulation Configuration. DALG offers comprehensive event log generation control. The simulation setup allows choosing generation modes like randomized traces or experimental complete play-outs for the control flow perspective. Configuration options encompass trace quantity, length range, loop handling, duplicates, non-conforming traces, and timestamps.

Transition Configuration. The transition configuration allows you to set individual weights for the transitions and mark them as invisible. In addition, the tool extends the SAMPLE approach by the configurability of individual time constraints for transitions and thus addresses a limitation identified in the evaluation of the SAMPLE approach.

Usability. DALG prioritizes user-friendliness for researchers across domains, ensuring ease of use without programming expertise in generating synthetic data for their studies. For this purpose, a graphical user interface based on the modern QT6 framework¹ was developed, which provides the user with convenient access to all configuration options. Tooltips aid user configuration, while automated analysis of the process model generates a preliminary configuration, laying the foundation for semantic definition.

Furthermore, the system checks and alerts users about invalid models or configurations, like when minimum values exceed maximum values in variable intervals, before each simulation. In addition, users can export their configurations to a JSON² file for future use or sharing, enhancing reproducibility. Moreover, result reproducibility is secured as users can define a seed for all simulation decisions, relying on pseudo-random number generation.

3. Evaluation

The SAMPLE approach implemented in DALG is described and evaluated in [1]. During the approach's evaluation, the need for semantic description of temporal information was uncovered and then implemented in DALG. During interviews with process mining experts, the application was presented and tested. In the process, the need for additional features, such as the specification of semantic information regarding the time aspect of transitions, came up. These features have also been added. DALG itself was tested using functional testing and synthetic DPNs specially prepared to test all of DALG features. Additionally, synthetic event logs based on real-world process models were generated with DALG and evaluated by

¹<https://www.qt.io/product/qt6>

²<https://www.json.org/json-en.html>

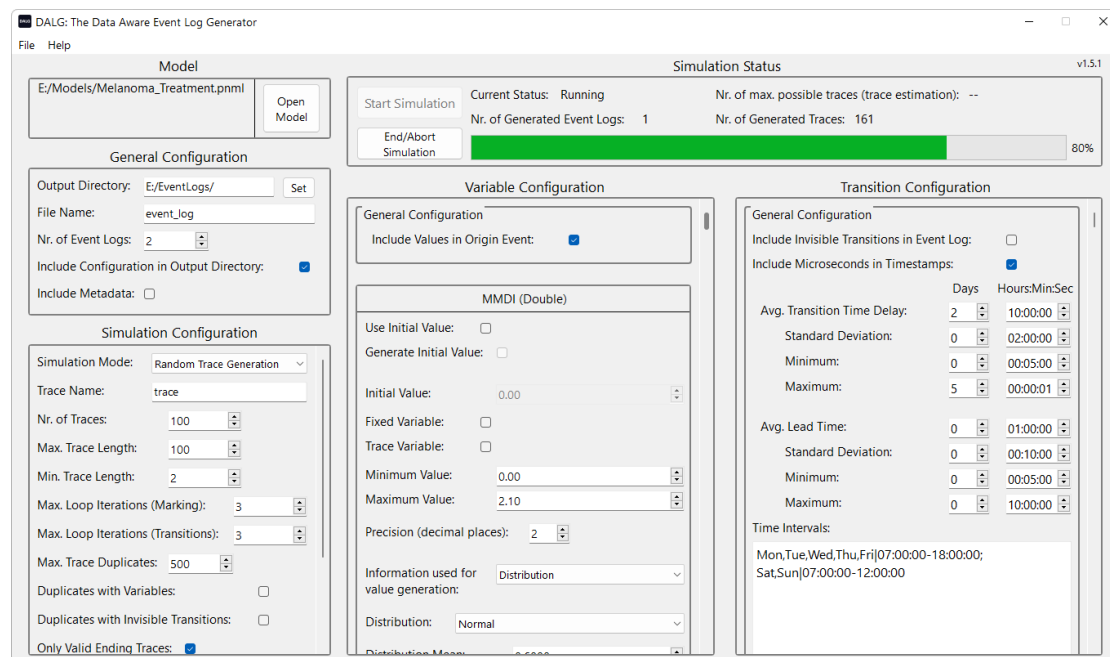


Figure 1: Graphical user interface of the DALG-tool

experts from the models’ domains. For example, DALG was applied to a model describing the treatment of melanoma, and doctors from the University Hospital Muenster evaluated the synthetic treatment traces. They found that the synthetic data was mostly realistic, but some unrealistic properties were identified. However, the unrealistic aspects were all found to be caused by inaccurate semantic information supplied to DALG since the medical professionals were not available to configure these. Petri nets defined using the PNML come in many different shapes and sizes, DALG was also tested with publicly available data Petri nets. Scientific data repositories such as 4Tu.ResearchData³ were searched to acquire DPNs. Subsequently, the features of these DPNs were identified, and it was ensured that they are supported by DALG to ensure a broad compatibility with existing DPNs. The usability of DALG was evaluated with user studies. One limitation found during the evaluation is the great effort required to configure the semantic information. This step is very time-consuming and error-prone. However, it was also found that the amount of semantic information cannot be reduced if the goal is to generate realistic data. A way of mitigating this problem could be to at least partially source the semantic information automatically from external structures, such as ontologies.

4. Tool maturity

DALG a fully functional standalone tool, whose development has been completed. Additionally, the correctness of the implementation of the SAMPLE approach has been evaluated. The tool

³<https://data.4tu.nl/>

and its source code are available for free use and development under the GNU General Public License 3 on GitHub⁴. The GitHub repository also includes a tutorial document⁵ and a video showcase⁶. DALG can be easily installed with the installer provided on GitHub and supports Windows and Linux based operating systems. Additionally, a user manual is provided to guide users through configuring the semantic information and running DALG.

In summary, DALG is a fully functional stand-alone tool that is ready to be used by researchers to generate synthetic multi-perspective event logs across many domains.

5. Conclusions and Future Work

This paper presents DALG, a stand-alone implementation of the SAMPLE approach, that enables the generation of multi-perspective event logs with a realistic data perspective. The tool aims at experts in the field of Business Process Management who do not have access to suitable event logs or are struggling to acquire relevant event logs. With DALG, users can continue in their process with reliable data. Because of the process model conformity, another use is the generation of event logs for evaluations.

The presented tool is available for download and can be extended. In future research, the tool will be extended to include the functionality of an auto configurator. This allows learning the configuration based on event logs. This should especially address privacy aspects in event log availability and significantly simplify the configuration. In addition, the user interface is to be expanded and revised to make it easier to use. Since it is currently only possible to generate compliant traces, an extension for the controlled generation of non-compliant traces should also be developed and integrated.

References

- [1] J. Grüger, T. Geyer, D. Jilg, R. Bergmann, Sample: A semantic approach for multi-perspective event log generation, in: M. Montali, A. Senderovich, M. Weidlich (Eds.), *Process Mining Workshops*, Springer Nature Switzerland, Cham, 2023, pp. 328–340.
- [2] O. Kummer, F. Wienberg, M. Duvigneau, J. Schumacher, M. Köhler-Bußmeier, D. Moldt, H. Rölke, R. Valk, *An extensible editor and simulation engine for petri nets: Renew*, volume 3099, 2004, pp. 484–493.
- [3] B. N. Yahya, Y. Khosiawan, W. Choi, N. A. D. Do, *Rt-plg: Real time process log generator*, volume 431, 2016.
- [4] L. Ackermann, S. Schönig, *Mudeps: Multi-perspective declarative process simulation*, in: *International Conference on Business Process Management*, 2016.
- [5] M. Weber, E. Kindler, *The Petri Net Markup Language*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2003, pp. 124–144.
- [6] *Ieee standard for extensible event stream (xes) for achieving interoperability in event logs and event streams*, IEEE Std 1849-2016 (2016) 1–50.

⁴<https://github.com/DavidJilg/DALG>

⁵<https://github.com/DavidJilg/DALG/blob/main/src/documentation/step%20by%20step%20guide.pdf>

⁶<https://github.com/DavidJilg/DALG/blob/main/src/documentation/DALG%20Showcase.mp4>