

DA4RDM: Data Analysis for Research Data Management Systems

M. Amin Yazdi^a, David Schimmel^b, Marcel Nellesen^c, Marius Politze^d and Matthias Müller^e
IT Center, RWTH Aachen University, Aachen, Germany

Keywords: Pre-processing Pipeline, Web Application, Research Data Management, Data Analysis, Process Mining, Requirement Engineering.

Abstract: Research Data Management (RDM) systems are becoming an essential part of every researcher's academic career. Often, researchers use various resources and web applications to handle their research data, causing complications for maintaining data and assessing research projects against FAIR principles. Consequently, RDM platforms help researchers with data administration tasks while providing the necessary tools for managing research projects. Furthermore, user engagement with such RDM platforms leaves traces of user interaction with research data; thus, studying user behaviors over research data becomes an exciting territory. However, running periodic data analysis studies proves to be a time-consuming and challenging task and requires the help of scientific staff to run pre-and post-processing pipelines per use case in order to be able to produce results that are usable by domain experts. This paper introduces Data Analysis for Research Data Management systems (DA4RDM) as a scalable web application that supports reusing pre-defined pre-and post-processing pipelines to enable domain experts to utilize the system without the need for scientific expertise. We use real data acquired from an RDM system, explain the tool's applicability, and present the preliminary findings, demonstrating its use cases and capabilities.


1 INTRODUCTION


In the era of digitalization, researchers frequently face the challenges of managing their research data to present their findings to other scholars and ultimately enable the reusing of research data. The Research Data Life Cycle (RDLC) in figure 1 represents the stages that every researcher faces during his/her academic career (Yazdi, 2019). Every RDLC stage includes several sub-activities, and each University provides various IT solutions per activity. On the one hand, RDLC navigates researchs throughout their research processes, and on the other hand, RDM systems encourage researchers to follow the FAIR principles. The FAIR principle is a set of guidelines and standards to increase transparency and impact of research results by making the research data Findable, Accessible, Interoperable, and Reusable (FAIR) (Wilkinson et al., 2016). Thus, data FAIRness enables humans or machines understand the semantics


and purpose of the data (Gargiulo et al., 2021).


Accordingly, there is a need for a system to support researchers throughout the RDLC and promote FAIR guidelines for research projects. Collaborative Scientific Integration Environment (Coscine) is a software platform at RWTH Aachen University to help researchers throughout RDLC and guide them toward FAIRness principles (Politze et al., 2020). Currently, Coscine enables the integration of multiple data resources and handles metadata management for research projects. The users of this platform can define research projects, invite other researchers to projects, integrate multiple storage resources, and define specialized metadata schema to describe research data while storing and archiving data. Coscine aims to provide a centralized web platform for RDLC activities while encouraging users to adhere to FAIR principles.


RDM systems provide the means for managing metadata and studying the RDLC process via discovering user interactions with research data. Although metadata management assists scholars in understanding the semantics of research data, investigating user interactions assists Principal Investigators (PI) and domain experts to discover bottlenecks in the respective processes and to capture implicit user

^a  <https://orcid.org/0000-0002-0628-4644>

^b  <https://orcid.org/0000-0002-1719-8928>

^c  <https://orcid.org/0000-0002-1830-5780>

^d  <https://orcid.org/0000-0003-3175-0659>

^e  <https://orcid.org/0000-0003-2545-5258>

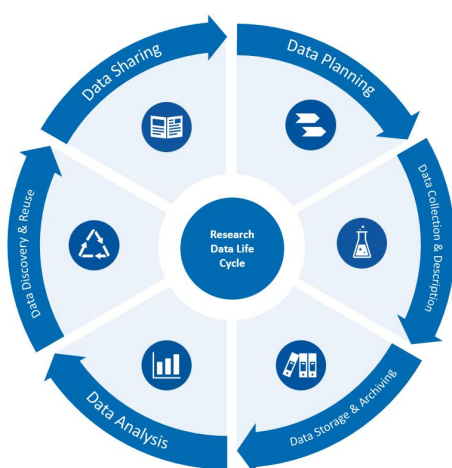


Figure 1: Presumed research data life cycle.

requirements (Yazdi and Politze, 2020). By collecting and capturing the footprints of user engagement, we can discover non-trivial changes to the research data and reveal insightful information, which facilitates further analytical investigations. Process and data mining techniques can effectively discover non-trivial process models and empower data-centric studies (van der Aalst, 2016). However, despite available process and data mining tools, most solutions are either offline tools or developed for specialized use cases and are too advanced for users without technical expertise (Kebede and Dumas, 2015; Malkawi et al., 2020; Celik and Akçetin, 2018). Thus, we need a software solution that can support the full spectrum of data preparation, pre-and post-processing pipelines, modeling, and even action suggestions for RDM systems. DA4RDM, on the contrary, enables us to connect to live data sources, run project-specific data pre-processing pipelines, discover process models, and allows for further scalability and adaptability of our system.

In the remainder of this paper, in section 2, we elaborate on the characteristics and functionalities of the DA4RDM, then in section 3, we review a system under study and demonstrate the findings. In section 4, we address the challenges and possible future work, and lastly, in section 5 we give a brief summary of our work.

2 DA4RDM WEB APPLICATION

DA4RDM is based on the Flask web framework, and it consists of 4 modules. As illustrated in figure 2, module 1 configures a data source, module 2 provides data cleansing tools, module 3 enables trans-

forming data according to projects' scope, and module 4 presents the findings with customizable post-processing and data modelings. The actual implementation for the DA4RDM web application is also publicly available on GitLab ¹.

2.1 Characteristics

DA4RDM is developed with the following characteristics and properties:

2.1.1 Scalability

DA4RDM provides web services for connecting data sources, customizing and executing project-specific pre-processing pipelines, and allows for integrating python packages for process and data mining projects. The current implementation of DA4RDM has embedded the PM4PY python package (Berti et al., 2019) into the existing service infrastructure for process discovery tasks.

2.1.2 Extensibility

As shown in figure 2, the modular construction of the DA4RDM allows for the integration or extension of third-party Python packages to satisfy the needs of a data analysis study. For instance, by extending the appropriate modules for data emulation, outlier detection, or data balancing and normalization, we can define and perform various data modeling projects.

2.1.3 Accessibility

We split the accessibility of DA4RDM into two parts. Firstly, the pipeline needs to be configured by a data scientist. It includes connecting to a data source, specifying a suitable data query, and setting the pre-processing pipeline per project. Once a pre-processing pipeline is defined, the steps are stored for later re-use as predefined projects. Secondly, a non-technical user can reuse a predefined project and apply additional web-based filters, attributes, and algorithms on top of data, supporting post-processing the data models without programming knowledge. Thus, the DA4RDM client application enables principal investigators to analyze the RDM system without technical background knowledge.

2.2 Functionalities

In the following section, we elaborate on the functionalities of the DA4RDM web application. We have divided the functionality of DA4RDM into three main

¹<https://git.rwth-aachen.de/AminYazdi/da4rdm>

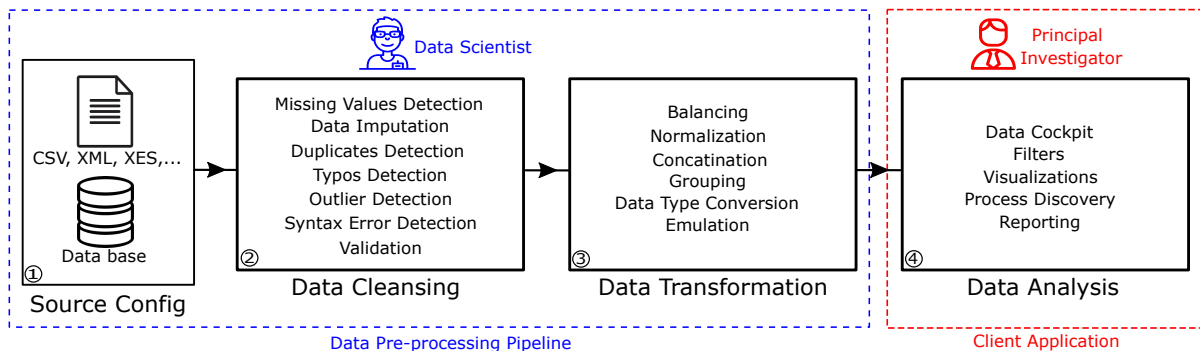


Figure 2: The DA4RDM overarching infrastructure and its modules.

parts, namely data source, data pre-processing, and process analysis.

2.2.1 Data Source

The data source module allows for the supply of unprocessed data for the DA4RDM. As shown in step 1 of figure 2, the data input module currently supports uploading local files in CSV, XES formats, or a direct connection to a relational database table. In the data source user interface (see figure 3a), users can provide additional parameters for each data source to further specify a file configuration or a database query command. Despite local data storage, every fresh execution of pre-processing pipeline fetches the latest instance of a specified data source. Moreover, this module is responsible for serializing data structure into Panda data-frame, allowing for data modifications required for later steps in the process. Note that XES file formats are standardized XML data-types used by Process Mining algorithms; thus, DA4RDM handles converting input data into XES format and preparing for process mining algorithms.

2.2.2 Data Pre-processing

The data pre-processing interface shown in figure 3b is responsible for data cleansing and transformation. The user interface allows for selecting a pre-defined data source and a pre-processing pipeline to initialize a data-driven study. Although the pre-processing pipeline is heavily dependent on the input data properties and objectives of a data analysis project, this tool allows for reusing a pipeline on top of operational data. Consequently, a data scientist specifies a data source, consequently develops or modifies a pre-processing pipeline according to his/her needs by utilizing methods mentioned in steps 2 and 3 of figure 2 to adhere to a data structure or a use-case in hand. Moreover, DA4RDM creates a user session maintaining the designated data source and pre-processing

pipeline outcomes to ensure a continuous workflow throughout the client application.

2.2.3 Process Analysis

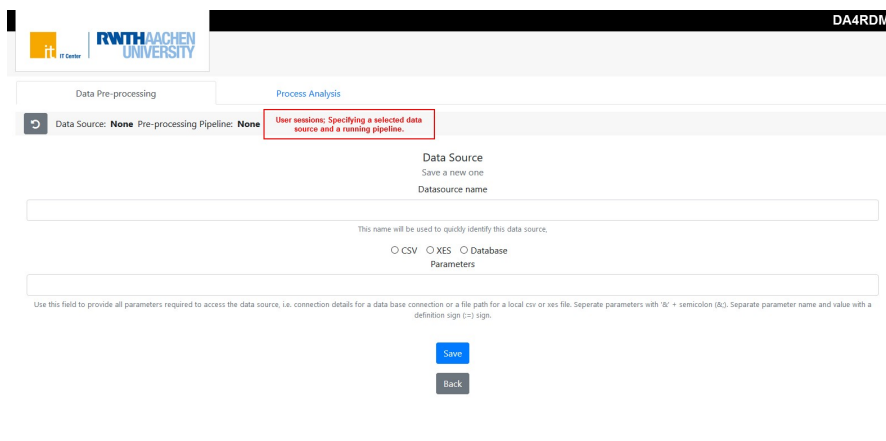
The process analysis shown in figure 3c provides the toolset and user interface required by principal investigators enabling them to further investigate data without technical knowledge. At this stage, users can utilize a few process mining algorithms provided by the Pm4py library to discover the process models and study particular user journey scenarios based on either frequency or performance analysis. Users of DA4RDM can conveniently specify the main event attributes (timestamp, case id, and activity) necessary for a process discovery based on the existing data features. Moreover, the filter module allows for further narrowing the analysis over a specified dataset. Additionally, the information module provides essential statistics of the selected dataset for specified criteria and filters. The results section provides a visualization of the discovered model according to a chosen algorithms such as Alpha miner, Heuristic miner, Inductive miner, or Directly Follows Graph (DFG) (van der Aalst, 2016).

3 SYSTEM UNDER STUDY

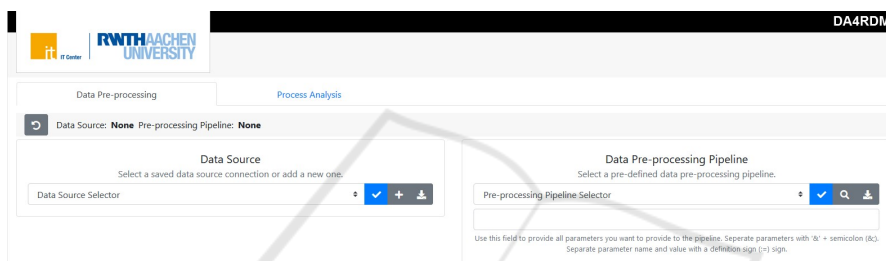
In the following section, we demonstrate the benefits of DA4RDM for an RDM system. We begin by elaborating on the data collection method and its properties over the Coscine platform as an RDM-system under study and then demonstrate the procedures involved to enable principal investigators to utilize data and deduce valuable insights.

3.1 Coscine Platform

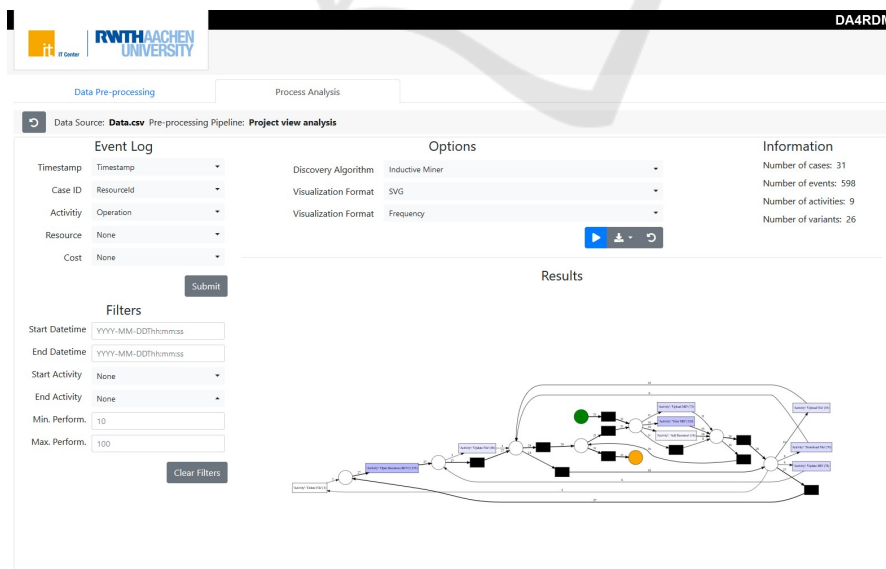
As mentioned earlier, Coscine is a platform to assist researchers with data management tasks and guide a



(a) UI for configuration of a data source using either a local file or a database.



(b) UI for selection of a data source and a pre-defined pre-processing pipeline.



(c) UI for Process analysis over a selected data source and pre-processing pipeline.

Figure 3: User interfaces for DA4RDM web application.

Table 1: Sample of transformed data objects into features and labels, enabling further data analysis using DA4RDM.

Id	Operation	Timestamp	UserId	RoleId	SessionId	ProjectId	PID	MetadataSchema	MetadataComp.	License	Discipline	Organizations
...
1021	View Project	1579108918	29613-d8...	be29c-4e...	4b15f...	4e9f-97...	NULL	NULL	NULL	MIT	Mechanical Eng.	ETH, Darmstadt
1022	View Resource	1579109840	29613-d8...	be29c-4e...	4b15f...	4e9f-97...	ef9175...	EngMeta	NULL	MIT	Mechanical Eng.	ETH, Darmstadt
1023	Update Metadata	1579109897	29613-d8...	be29c-4e...	4b15f...	4e9f-97...	Xn3on4...	EngMeta	73%	MIT	Mechanical Eng.	ETH, Darmstadt
...

research project toward the FAIR maturity paradigm. In addition, researchers may use Coscine to store their research data, provide specialized metadata, and collaborate within research projects. Coscine is under active development and benefits from the microservices software architecture model. Various APIs and applications are employed to interact with the system such that each API is distinguishable by topic. For instance, all requests handling file modifications are processed by the same API.

3.2 Data Model

Coscine allows us to collect user-based RDM activities while formalizing appropriate data objects and relationships between attributes. The obtained information is a set of user requests received by the server-side and processed according to particular application domains. Coscine generates and captures this data in a serialized JSON object format due to its ease of scalability to incorporate supplementary attributes and entities without a need for extending database tables. The data fields include detailed insights into the sequence of actions and respective RDM-relevant entities. For instance, the listing 1 represents a JSON object triggered by a user, attempting to update metadata of a research data entries; respectively, we can deduce meta information from a single user action such as its PID, selected metadata schema, percentage of metadata provided, discipline, and additional associated properties.

```
{
  "Operation": "Update Metadata",
  "Timestamp": "1579109897",
  "UserId": "29613-d8...",
  "RoleId": "be29c-4e...",
  "SessionId": "4b15f...",
  "ProjectId": "4e9f-97...",
  "PID": "ef9175...",
  "MetadataSchema": "EngMeta",
  "MetadataCompleteness": "70%",
  "License": "MIT",
  "Discipline": ["Mechanical Engineering"],
  "Organizations": ["ETH", "Darmstadt"]
}
```

Listing 1: Sample JSON object captured on Coscine upon user action.

Subsequently, DA4RDM provides the means to utilize a produced dataset by Coscine or any other RDM system via the user interface shown in figure 3b. It allows selecting a preprocessing pipeline to evaluate and transform data samples into a new data format, suitable for a data modeling algorithm. Accordingly all JSON objects are collected and stored in a relational database ready for importing to DA4RDM. Table 1 is an example of a dataset converted into columns of features and labels ready for running in-depth process mining analysis using the DA4RDM interface displayed in figure 3c.

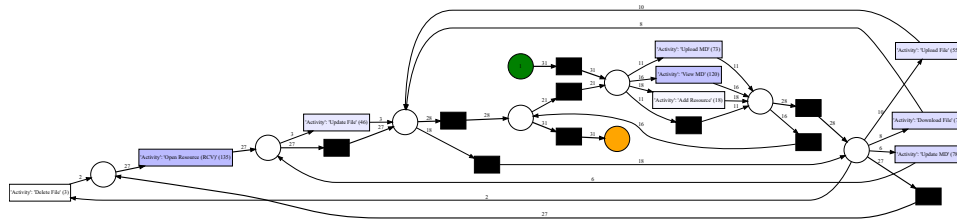
3.3 Privacy

Concerning responsible data mining principles, it is essential to maintain privacy within user-based datasets (Rafiei et al., 2018). Accordingly, Coscine, as a system under study, generates and contains no human-readable names. Instead, GUIDs are used as unique identifiers to enable analyzing a user's actions and user behavior study without compromising users' identity. Moreover, by relying on server-side event logs, we eliminate the need for a client-side logger, which may cause unnecessary complications or include undesired sensitive information.

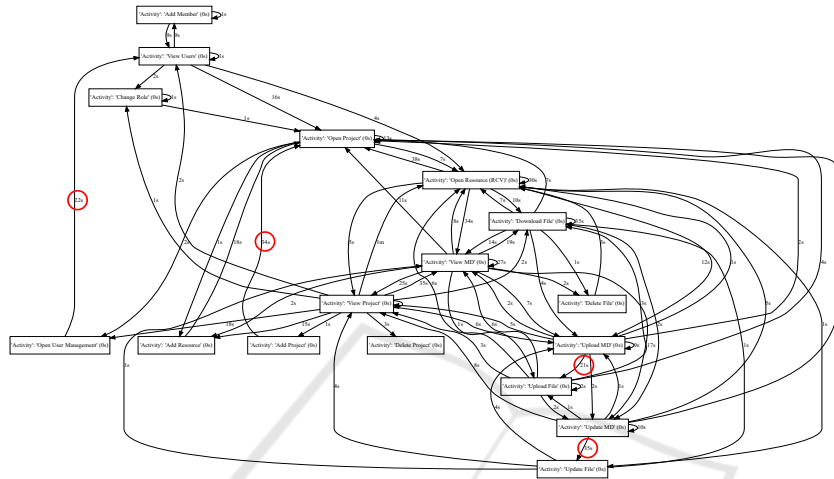
3.4 Preliminary Findings

The following section elaborates on the qualitative and quantitative findings from two case studies using DA4RDM that are also shown in figure 4. Initially, we investigated user activities over resource PIDs and then analyzed overall system performance for research projects on Coscine. The goal is to demonstrate the preliminary applications of DA4RDM in its initial stages by discovering process models for non-trivial user interaction paths and identifying non-functional requirements in the system.

In the first use case, we ran frequency analysis using inductive miner over resource PIDs and discovered a Petri-net model for the most typical user journey paths over interacting with different resources and research data. We refer our readers to (Kindler et al., 2006) for further explanation of Petri-net modeling. On the top right section of figure 3c, DA4RDM reports brief statistics over the number of cases, events, activities, and variations of activities. With the help of



(a) Frequency analysis using Petri-net model for all involving activities over resource PIDs.



(b) Performance analysis using Data-Follow-Graph for all research projects. Red circles indicate violated KPIs in the system.

Figure 4: Discovered process models using the Inductive Miner via DA4RDM web application.

DA4RDM illustrated in figure 4a, we discovered that the number of execution for the *Open Resource(RCV)* activity is oddly excessive with respect to the number of incoming and outgoing transitions, indicating the existence of loops in the code and the necessity to refactor code in order to avoid unnecessary server-side requests.

In the second use case, we investigated the overall performance of the RDM system shown in figure 4b using DFG process model and evaluated the findings (shown in red circles) with the help of a domain expert against key performance indicators. Despite the expected complex and unstructured process model due to the freedom of user interaction within the system, DA4RDM successfully assisted us in identifying bottlenecks in the system that were previously unknown. For instance, the transition from *Open User Management* to *View Users* actions should not take longer than a few seconds, or the process of creating new projects (transition from *Add Project* to *Open Project*) on Coscine was discovered to take over 70 seconds. Accordingly, discovering the evidence of software execution bottlenecks resulted in gaining the developer teams’ attention, and also, as shown in figure 4b, the

performance for this process is greatly improved.

The empirical study over real data, implies that the cyclic order of operations within the RDLC shown in figure 1 may not reflect the reality of how researchers interact with research data. Furthermore, despite the limited data sample size, both use cases manifest DA4RDM capabilities and how the web application could be extended to serve non-technical users to investigate their research projects against certain criteria by extending the data modeling technique or visualization for a target use case.

4 CHALLENGES AND FUTURE WORK

In order to evaluate the functionality of DA4RDM with a real dataset while developing DA4RDM, we spent a considerable amount of time finding a suitable method for obtaining and extracting event logs from Coscine as an evolving RDM system. Nevertheless, DA4RDM is designed to be independent of its input data model and allows for data processing according to foreseeable use cases. Additionally, the current im-

plementation only allows for a single source of truth for data input which is a limitation for RDM systems where the datasets are scattered in multiple locations. The seen complex process model shown in figure 4b results from actual user behavior in an environment where users are free to interact with a system without a specific order. Thus, there is a need to extend DA4RDM with a pre-processing pipeline suggested by the authors in (Yazdi et al., 2021) to abstract event logs to achieve structured and simple process models. We have to acknowledge that Coscine is a developing RDM platform and is currently in its pilot phase; hence the sample dataset available is limited to its beta users, and the current preliminary findings may not entirely reflect the actual user behavior in a mature RDM system.

The future work includes adding additional interfaces for conformance checking of the user process, identifying unfinished user journeys and triggering automatic actions, and extending the collected dataset to produce a FAIR maturity dashboard for every research project and suggest user actions to increase research data FAIRness. Although the current DA4RDM UI allows for a PI to interact with the system, the post-processing of data models proved to be too advanced; therefore, we may need to pre-scope the available features of our web application for the target audience.

5 CONCLUSION

In this paper, we discussed the benefits of DA4RDM for an RDM system. We began with a technical review of the developed web application, its characteristics, and functionalities. Furthermore, we elaborated on an RDM platform (Coscine) as a candidate system under study and the approach used to obtain a real dataset for the goal of data modeling and analysis. Preliminary findings demonstrated the benefits of such a system toward user behavior studies and discovering non-functional requirements. Although the extracted data from Coscine is entirely based on a service layer, DA4RDM is showing to be adaptable to any log format according to the needs of a data modeling algorithm. Therefore, the contributions of DA4RDM are to allow for a scalable web application that enables non-technical staff to reuse pre-defined pre-and post-processing pipelines to execute data-driven studies without technical or scientific expertise.

REFERENCES

- Berti, A., van Zelst, S. J., and van der Aalst, W. (2019). Process mining for python (pm4py): bridging the gap between process-and data science. *arXiv preprint arXiv:1905.06169*.
- Celik, U. and Akçetin, E. (2018). Process mining tools comparison. *Online Academic Journal of Information Technology*, 9:97–104.
- Gargiulo, P., Galimberti, P., Tammaro, A. M., and Zane, A. (2021). Fair rdm (research data management): Italian initiatives towards eosdc implementation. In *IRCDL*, pages 42–52.
- Kebede, M. and Dumas, M. (2015). Comparative evaluation of process mining tools. *University of Tartu*.
- Kindler, E., Rubin, V., and Schäfer, W. (2006). Process mining and petri net synthesis. In *International Conference on Business Process Management*, pages 105–116. Springer.
- Malkawi, R., Saifan, A. A., Alhendawi, N., and Bani-Ismaeel, A. (2020). Data mining tools evaluation based on their quality attributes. *International Journal of Advanced Science and Technology*, 29(3):13867–13890.
- Politze, M., Claus, F., Brenger, B., Yazdi, M. A., Heinrichs, B., and Schwarz, A. (2020). How to manage it resources in research projects? towards a collaborative scientific integration environment. *European Journal of Higher Education IT*, 2.
- Rafiei, M., von Waldthausen, L., and van der Aalst, W. M. (2018). Ensuring confidentiality in process mining. *Proceedings of the 8th International Symposium on Data-driven Process Discovery and Analysis-SIMPDA*, 18:3–17.
- van der Aalst, W. (2016). *Process mining: data science in action*. Springer.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.
- Yazdi, M. A. (2019). Enabling operational support in the research data life cycle. In *Proceedings of the First International Conference on Process Mining*, pages 1–10.
- Yazdi, M. A., Farhadi, P., and Heinrichs, B. (2021). Event log abstraction in client-server applications. In *IC3K 2021: Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management: KDIR*. SciTePress.
- Yazdi, M. A. and Politze, M. (2020). Reverse engineering: The university distributed services. In *Proceedings of the Future Technologies Conference*, pages 223–238. Springer.