Enabling Operational Support in the Research Data Life Cycle

M. Amin Yazdi[0000−0002−0628−4644]

IT Center, RWTH Aachen University, 52074 Aachen, Germany
yazdi@itc.rwth-aachen.de

Abstract. Since 2015 a set of preliminary design studies were started on how to promote the stewardship of research data at RWTH Aachen University. This has resulted in a bottom-up software architecture approach that has created fundamentals for interconnection of Research Data Management (RDM) services. It has facilitated the development of essential services for the collection of structured data-sets with unique persistent identifiers. However, this service-oriented architecture has to be complemented by a set of web technologies to support the exploration and discovery of relevant data or, track and trace data within the research life cycle. With respect to lessons learned from the RDM project and literature reviews, besides technical improvements, investigation on scientists' research process and providing means for operational support (data detection, prediction, and recommendations) are essential. Thus, this research project plans to enable operational support for RDM services across the research data life cycle while at the same time, keeping an eye on data privacy concerns. The goal is to build control-flow models, predict deviations and recommend personalized solutions by analyzing and discovering users' process model with the help of process intelligence techniques.

Keywords: Process mining · Operational support · Process discovery · Research data management.

1 Introduction

Scientific research is one of the core university processes. Therefore, RWTH Aachen University is investing in research and technical development of eScience to support access to research data and its processes. eScience and, more specifically, RDM offers opportunities to improve research processes such as reproducibility and reliability of scientific experiments and as well as of further secondary data analysis. Despite researchers' rising demand for IT infrastructures to deliver business value, still supporting technology for scientific research processes lacks necessary tools for Research Data Life Cycle (RDLC) services. Thus, further investigations are required to identify and implement flexible services that researchers wish to use within their research processes. Through the course of the RDM project at RWTH University, a RDLC model has been presumed as it is shown in Figure 1. This RDLC model consists of six sections:

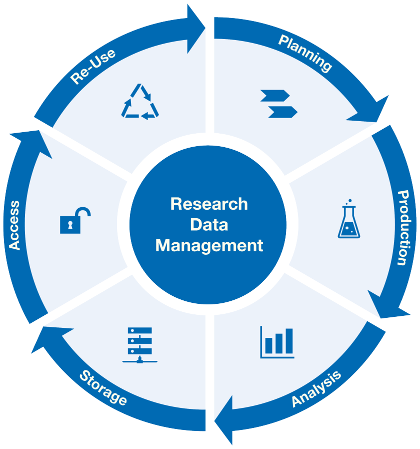


Fig. 1. Presumed research data life cycle.

(a) Data *planning* to determine data type, format and standards.
(b) *Production* of data along side metadata management.
(c) Data *analysis* to identify patterns and explore data through data mining.
(d) *Storage* and archiving, provide means to facilitate data backup.
(e) *Accessing* Data and data sharing to ensure availability for further publications.
(f) Discovery and *reuse* of data to prepare it for future use and identifying relevant data.

Currently, there are number of services that support researchers within *planning*, *Production*, *Storage* and *Accessing* phases of research, but this RDLC lacks support for data *analysis* and *reuse*. Therefore, this research project is planning to employ data science and process mining techniques to deal with the shortcomings.

2 Research Questions

RDM involves a large number of researchers with diverse expertise and roles from many disciplines. This diversity has resulted in the development of a infrastructure for RDM with number of services for researchers. Currently, users lack a system that allows them the dynamic exploration and discovery of relevant data that originated from across scientific disciplines. Moreover, there is no clear process model of researchers to identify most common obstacles and potentials for improving their research process. To fill into the gaps, the following research questions are proposed.

RQ1: How can process intelligence provide means in RDLC to incorporate existing heterogeneous and distributed services?

Currently, distributed services for data management are trending and come with advantages such as autonomy, ease of local data management, privacy or data protection. However, despite the advantages of distributed services, some degree of centralization is essential. At IT Center of the University, there are a number of tools and services provided to handle research data based on users' requirements. In order to offer transparency in current research practices, scholars can be supported with a Computer-Supported Cooperative Work (CSCW) platform to integrate decentralized and centralized services alike, that are involved in RDLC. Through such an integration platform, user interactions can be recorded and collected in the format of event logs (Case ID, timestamp, Activity name). Event logs are then prepared for further analytical investigations. The process mining can efficiently discover non-trivial process models, locate and extract patterns of use. By incorporating and structuring event logs and running process mining algorithms (such as Fuzzy miner, Multi-phase miner, Clustering, Decision rule mining, α-algorithm, etc.) Across available services, a control-flow model for users' RDLC can be generated. These techniques provide means to monitor data circulations, analyze research habits, discover bottlenecks and improve key performance indicators.

RQ2: How to develop a self-steering tool for researchers in a complex research network to explore relevant data and opportunities for research collaborations?

Across the RDLC, it is essential for researchers to monitor and steer their research throughout the period of their investigations. Lack of research environment awareness and difficulty of networking within the academic world, increases the potentials for research redundancies or missing alluring research opportunities. Therefore, a tool has to be implemented to trace and collect users' scientific contributions and allows users to configure their own metrics for self-evaluations or key performance indicators. By extending this system with in-depth social network analysis based on previous cooperations, users should be able to explore their network for relevant data and to identify potentials for research collaborations. Further, user experience and usability evaluations should be the pillars to development of such a system to guarantee technology acceptance and success of this tool.

RQ3: How to offer operational support in the RDLC while predicting research models and carrying out personalized recommendations?

In the academic world, scientists are involved with many heterogeneous services that often discourage researchers from active participation within RDM. Hence, a successful RDM system should motivate researchers to get involved with the required tools through a seamless integration of the RDM services in an integration platform. After identification of the RDLC model using current data, it would be possible to offer on-the-fly operational support for researchers. Thus, a web service has to collect and analyze researchers current event logs and check the compliance of the users' process model iteratively, correspondingly detecting and predicting its deviations. Additionally, besides empowering an efficient data discovery, this tool would allow monitoring of researchers performance and contributions. Furthermore, the impact and validity of the developed service has to be evaluated by experimental analysis to ascertain the usability of the system in interdisciplinary and disciplinary research environments.

RQ4: How to address technical concerns toward research data security and privacy protection throughout the RDLC?

Digitalization poses a continuous risk of breach of private data or secret information. In the context of research data management, researchers are constantly concerned with losing their control over their data that being used by different services or shared between colleagues. Therefore, very strict privacy policies have been enforced to avoid a data breach, but often this comes with the price of losing atomization and knowledge discovery. Of course, with every digitalized service, there is a tradeoff between anonymity versus publicity. However, protection of data and privacy is considered as one of the main challenges to the involvement of users in any system. Hence, it is essential to infer which information is going to be utilized and for what purpose. To protect the rights and ethics of generated or shared data, it is particularly important to address these issues by initially identifying the concerns throughout the RDLC and later discover potentials for improvements. Further, the findings and possible guidelines should be prescribed as requirements to implement any kind of tool within RDLC to gain users' consent and hence, increase trust and usage.

3 Related Work

To tackle the lack of sufficient services for RDLC, above all, it is required to have a brief overview of the current state of the art and related research in this field. Moreover, the set of challenges in this project is formed by the incorporation of multiple computer science disciplines. Namely those are: Human-computer interaction, data and process science, and data privacy. Therefore, there are three main areas that contribute to the complex. Privacy and data protection concerns on RDLC, the role of users in recommender systems and business process intelligence.

3.1 Privacy and Data Protection in RDM

Obviously, privacy and data protection concerns are two main challenges that hinder any tool from gaining users' attention and influence on users participation. In this regard, one needs to understand the concerns in the field and additionally, utilize the findings on the design of the system.

Kokolakis [13] and Adler et al. [12] have reported on a phenomenon of "privacy paradox" and users behavioral inconsistencies toward sharing information. They found that, despite major privacy concerns in social networks, people are willing to reveal personal information when perceived benefits surpass observed risks. Further, by pointing out to privacy calculus theory, authors postulated that individuals perform a calculus between the expected loss of privacy and potential gain of the disclosure. Thus, as is described by Sayogo et al. [16] users look for potential gains such as a need for social relationship, social validation, and self-representation to outweigh expected privacy concerns. Authors have suggested that studies on privacy to increase the users' participation should examine evidence of actual behavior rather than only self-reported behavior.

3.2 Recommender Systems and Role of Users

Principally, a recommender system "is a subclass of information filtering system that seeks to predict the rating a user would give to an item" [17]. By recording users interactions within a system and running analysis, it is possible to derive usage patterns that assists computer to understand users' needs and develop personalized recommendations. Moreover, there are many algorithms and techniques in the field of recommender systems to assist us in the personalization of suggested data. Nonetheless, it is not trivial, the extent that user interaction can be used to improve information personalization. For instance, Chen et al. [7][19] have emphasized on the role of users during information seeking process and their impact on the recommender systems. In particular, they have achieved a better result by proposing an optimized process recommendation model. They suggested that recommending a suitable seeking process rather than recommending a final result has optimized the users trust toward recommender systems. Furthermore, a case study has proved the importance of users behavior in improving recommender systems' accuracy. By incorporating additional contextual information, it is viable to cluster users based on behavioral patterns and then improve the KPIs via personalized recommendations [18].

3.3 Business Process Intelligence

User-orientated design includes an understanding of how the system is being used and how our users would like to extract information from a large set of data. Business Process Intelligence (BPI) assists with the general modeling of processes and software architecture. By comparing user interaction models and expected process models, many potentials for improvement of our systems can be identified.

As is described by Donoho [8] and Press [15], data science includes but not limited to data extraction, preparation, transformation, presentation, predictions and visualization of a huge amount of structured or unstructured data that are static or streaming. Van der Aalst [2] has described process mining as a mean to bridge and brings together traditional model-based process analysis and data-centric analysis techniques. Moreover, BPI is described as a dispute between the process models that were expected and event data that is observed in the real world, and is used to extract knowledge and identify deviations [3] [1]. Van der Aalst has suggested three main types of process intelligence techniques that can participate in a software (re)design phases [2]. These techniques are namely, *Discovery* techniques such as α-algorithm is usually used as a starting point for analysis to extract a process model from raw event logs. *Conformance checking* compares an existing process model with an event log of the same process to check if the extracted model complies to that of reality and vice versa. *Enhancement* aims to improve the existing process by extending the former model.

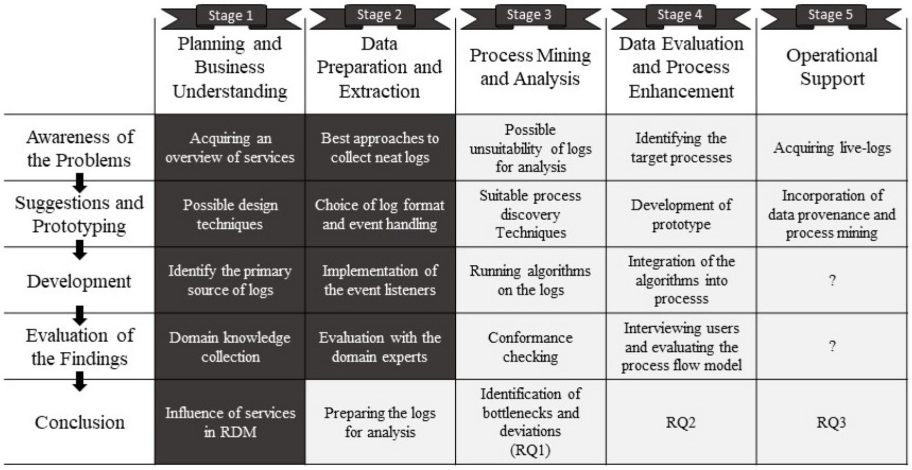
Researchers in the field have also emphasized on the suitability of offering operational support only for structured processes [14], [4] and [5]. However, despite the high ambitions and difficulties of offering operational support for an unstructured process, such projects often result in interesting findings and allow for various significant improvements. Therefore, techniques such as "combination of abstraction" and "clustering" are proposed to simplify the unstructured processes and to prepare it for operational support and process intelligence analysis. Moreover, using holistic data and building a user trust model for predicting the relative data files have successfully enhanced the business intelligence processes performance where users required to discover relevant data [11] [10].

4 Methodology

In order to answer the research questions, this research project is inspired from PMPL [6], PM^2 [9] and L*life-cycle[2] methodologies. Taking into account that applying process mining projects in practice is not a trivial task, it might be necessary to run each stage iteratively to steer to an optimal conclusion. Figure 2 presents the research methodology and its stages. There are five steps that are repeated within every stage. In particular: 1$^{\text{st}}$) *Awareness of the Problems*, 2$^{\text{nd}}$) *Suggestions and Prototyping*, 3$^{\text{rd}}$) *Development*, 4$^{\text{th}}$) *Evaluation of Findings* and 5$^{\text{th}}$) *Conclusion.* The conclusions and results from every stage are going to be used as input for the next stage.

The stages within this research methodology are planned to gradually shape the research project and answer the research problems as it gets mature and evolves.

Stage 1) *Planning and business understanding:* Dedicated to the understanding of the domain knowledge, available services, and its respective processes. The results of this stage is a set of research questions and an awareness of software architecture limitations in place. Respectively, a set of "question-driven" and "goal-driven" research questions have been determined to fulfill the aforementioned research objectives. **Stage 2)** *Data preparation and extraction:* To proceed with the determined research questions, this stage should locate and enable exploration of data. Also, select and prepare data to create event logs, then, extract and remove the noise in the data. Additionally, it has to be verified with respect to the research goals. Further, if necessary, renew or refine the main research questions to fit the real-world problems. **Stage 3)** *Process mining and analysis:* To acquire a suitable event log structure and produce a control flow model, it is required to aggregate events, create views, enrich and filter logs.



	Stage 1 Planning and Business Understanding	Stage 2 Data Preparation and Extraction	Stage 3 Process Mining and Analysis	Stage 4 Data Evaluation and Process Enhancement	Stage 5 Operational Support
Awareness of the Problems	Acquiring an overview of services	Best approaches to collect neat logs	Possible unsuitability of logs for analysis	Identifying the target processes	Acquiring live-logs
Suggestions and Prototyping	Possible design techniques	Choice of log format and event handling	Suitable process discovery Techniques	Development of prototype	Incorporation of data provenance and process mining
Development	Identify the primary source of logs	Implementation of the event listeners	Running algorithms on the logs	Integration of the algorithms into processs	?
Evaluation of the Findings	Domain knowledge collection	Evaluation with the domain experts	Conformance checking	Interviewing users and evaluating the process flow model	?
Conclusion	Influence of services in RDM	Preparing the logs for analysis	Identification of bottlenecks and deviations (RQ1)	RQ2	RQ3

Fig. 2. The research methodology. The dark-gray squares indicate the accomplished steps, while the rest are ideas for future research approach.

Prior to running any process mining analysis, structured event data has to be obtained in order to enable process discovery techniques, conformance checking and process enhancement. **Stage 4)** *Data evaluation and process enhancement:* The findings from the previous step within a control flow model have to get enriched and integrated with additional perspectives to better help understand the as-is process. This stage supports us to verify the conformance of the process and identify bottlenecks and further suggest for actions for enhancement. **Stage 5)** *Operational support:* Operational support is the state when a system is capable of detecting deviations, predict resulting events and recommend actions on the fly using pre-mortem data(live/current event data). However, active operational support requires structured processes.

Presently, at the current research phase, the feasibility of an active operational support for RDM is unknown and is debatable.

5 Conclusion and Expected Contribution

The resulting research should generate a control-flow model of researchers across RDLC and facilitate a personalized operational support to cope with the high demand for RDM and knowledge exchange. During the course of this research, use case studies should be carried out to ensure the validity and orientation of the research. Moreover, apart from continuous development, the findings from research questions will be reflected in iterative design cycles and will be utilized as a proof of concept.

Despite the unpredictability nature of the process mining, it is expected to extract insightful information and build interaction models from event logs within RDLC. Further, this model has to get extended to integrate other relevant perspectives that typical researchers utilize with their research life cycle. Furthermore, this research project should empower the data provenance and exploration of research data. By running the analysis on involved software components, it is expected to discover bottlenecks in semi-distributed systems and enhance the productivity of researchers by identifying key performance indicators. Using this methodology, we should be able to extract usability issues and further identify requirements for further development for the scope users.

Additionally, by tracking and tracing event logs within the RDLC, we should manage to obtain pre-mortem event data and enable on-the-fly events' prediction, and data/process recommendations. Finally, as the user interaction and his data are the centers of the system, it is important to have a realistic understanding of the privacy issues in the field. Throughout the evolution of this project, we should obtain and provide technical solutions that elaborate privacy concerns and allow a user to achieve complete control over his data.

References

1. van der Aalst, W.: Process Mining Discovery, Conformance and Enhancement of Business Processes, vol. 2. Springer (2011)
2. van der Aalst, W.: Process mining: data science in action. Springer (2016)
3. van der Aalst, W., Adriansyah, A., van Dongen, B.: Replaying history on process models for conformance checking and performance analysis. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **2**(2), 182–192 (2012)
4. van der Aalst, W., Gunther, C.W.: Finding structure in unstructured processes: The case for process mining. In: Application of concurrency to system design, 2007. ACSD 2007. Seventh international conference on. pp. 3–12. IEEE (2007)
5. Agrawal, R., Gunopulos, D., Leymann, F.: Mining process models from workflow logs. In: International Conference on Extending Database Technology. pp. 467–483. Springer (1998)
6. Belfiore, J.C., Rekaya, G., Viterbo, E.: The golden code: A 2 x 2 full-rate space-time code with non-vanishing determinants. In: International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings. pp. 310–310. IEEE (2004)
7. Chen, J., Zhou, X., Jin, Q.: Recommendation of optimized information seeking process based on the similarity of user access behavior patterns. Personal and ubiquitous computing **17**(8), 1671–1681 (2013)
8. Donoho, D.: 50 years of data science. Journal of Computational and Graphical Statistics **26**(4), 745–766 (2017)
9. van Eck, M.L., Lu, X., Leemans, S.J., van der Aalst, W.M.: Pm^2: A process mining project methodology. In: International Conference on Advanced Information Systems Engineering. pp. 297–313. Springer (2015)
10. Groeger, C., Schwarz, H., Mitschang, B.: Prescriptive analytics for recommendation-based business process optimization. In: International Conference on Business Information Systems. pp. 25–37. Springer (2014)
11. Huang, X., Lu, T., Ding, X., Gu, N.: Enabling data recommendation in scientific workflow based on provenance. In: 2013 8th ChinaGrid Annual Conference. pp. 117–122. IEEE (2013)

12. Kim, Y., Adler, M.: Social scientists data sharing behaviors: Investigating the roles of individual motivations, institutional pressures, and data repositories. International Journal of Information Management **35**(4), 408–418 (2015)
13. Kokolakis, S.: Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. Computers & Security **64**, 122–134 (2017)
14. Kumar, M., Thomas, L., Annappa, B.: Distilling lasagna from spaghetti processes. In: Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence. pp. 157–161. ACM (2017)
15. Press, G.: A very short history of data science. Forbes. com (2013)
16. Sayogo, D.S., Pardo, T.A.: Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. Government Information Quarterly **30**, S19–S31 (2013)
17. Skulimowski, A.M., Kacprzyk, J.: Knowledge, information and creativity support systems: Recent trends, advances and solutions. In: KICSS2013-8th International Conference on Knowledge, Information, and Creativity Support Systems. vol. 364. Springer (2016)
18. Terragni, A., Hassani, M.: Analyzing customer journey with process mining: from discovery to recommendations. In: The IEEE International Conference on future IOT and cloud, 2018. FiCloud 2018. 6th international conference on. IEEE (2018)
19. Yazdi, M., Valdez, A.C., Lichtschlag, L., Ziefle, M., Borchers, J.: Visualizing opportunities of collaboration in large research organizations. In: International Conference on HCI in Business, Government and Organizations. pp. 350–361. Springer (2016)