BERMUDA: Participatory Mapping of Domain Activities to Event Data via System Interfaces*

Vlad P. Cosma[1,2][0000-0001-8022-6402], Thomas T. Hildebrandt[1][0000-0002-7435-5563], Christopher H. Gyldenkærne[3][0000-0003-2858-7328], and Tijs Slaats[1][0000-0001-6244-6970]

[1] Copenhagen University, Copenhagen 2200, Denmark
{vco,hilde,slaats}@di.ku.dk
[2] KMD ApS, Ballerup 2750, Denmark
vco@kmd.dk
[3] Roskilde University, Denmark
chrgyl@ruc.dk

Abstract. We present a method and prototype tool supporting participatory mapping of domain activities to event data recorded in information systems via the system interfaces. The aim is to facilitate responsible secondary use of event data recorded in information systems, such as process mining and the construction of predictive AI models. Another identified possible benefit is the support for increasing the quality of data by using the mapping to support educating new users in how to register data, thereby increasing the consistency in how domain activities are recorded. We illustrate the method on two cases, one from a job center in a danish municipality and another from a danish hospital using the healthcare platform from Epic.

Keywords: Data quality · Secondary use · Event extraction · Event matching · Participatory Design

1 Introduction

The abundance of data recorded in information systems and easily accessible technologies for data processing, such as predictive AI models and process mining [1,2], have created huge expectations of how data science can improve the society.

However, there has also been an increasing voicing of concerns [3, 11, 18, 39], pointing out that merely having access to data and technologies is not sufficient to guarantee improvements. In the present paper we focus on data quality and responsible event extraction in the context of secondary use of event data [34] recorded in information systems. That is, data representing events in the domain of use, such as the start and completion of work tasks which has as primary use to support case workers and document the progress of a case, but is intended to be used for secondary purposes, such as building predictive AI models or the discovery of processes using process mining tools.

* A 2-page extended abstract presenting early ideas of the paper was published in [13].

The challenges of event data quality are manifold [9], including handling event granularity, incorrect or missing data and incorrect timestamps of events [17]. A more fundamental problem in the context of secondary use of event-data is that of ensuring a consistent and correct matching of event data to business activities [7].

The lack of research in the area of event log creation has been pointed out in several papers [2,7,9,16,21,26,29,30,36,38]. This task is in general associated with words and expressions like: costly, time consuming, tedious, unstructured, complex, garbage-in garbage-out. Historically, research for data-driven innovation and improving productivity has shown to pay little to no attention to how data is created and by who. Data is often created within a system and its user interface where a given context for capturing and using data has been established through continuous sense-making between people that have local and often individual understanding of why data is generated and for what. Studies claim [22, 41]that data science initiatives are often initiated at high-level and allocated from domain of data creation while the data science product is re-introduced as a model that needs to be adapted by the practice where data is created. While data driven systems can be evaluated with good results on artificial data from the data domain, it is often a struggle to create value for the domain users. This is due to trust of data origin, what it represents and how new intents for its purpose comes through what could be considered a back-door top-down method. A Participatory Design(PD)-study [18] investigated a mismatch between data extraction findings at an administration level of cross-hospital management and how doctors and clinical secretaries represented their ways of submitting data, highlighting a need for re-negotiating data creation and its purpose in a way so data scientists can contribute to better data capture infrastructures as well as giving health-care workers a saying in how such data capture infrastructures are prioritized in their given domains of non-digital work. In PD [8, 23, 32] as a field such presented tensions are not new. Here PD as a design method and practice has sought to create alignment between workers existing understanding of own work and emerging systems through design as a practice for visualising such tensions across actors of an innovation or IT project. PD is from here seeking, in a democratic manner, to find solutions and interests that can match partners across hierarchies.

As a means to facilitate responsible secondary use of event data, we propose in this paper the BERMUDA (*Business Event Relation Map via User-interface to Data for Analysis*) method to capture and maintain the link between domain knowledge and the data in the information system. The method supports involvement of domain experts in the mapping of activities or events in the business domain to user-interface elements, and of system engineers in the mapping of user-interface elements to database records used by data scientists. In other words, the method helps documenting the inter-relationship in the “BERMUDA triangle” between the domain concepts, the user interface and the database, which often disappears. We see that by breaking down the barrier between data-creators and data scientists and building tools for involvement and iterative feedback of data infrastructures and their user front-end, new discussions for data cooperation can occur. The mapping is independent of any specific data analysis, but should of course include the activities and events of relevance for the analysis at hand. In particular, the method contributes to the responsible application of process mining [27] by supporting a collaborative creation of event logs.

The motivation for the method came from research into the responsible engineering of AI-based decision support tools in Danish municipalities within the EcoKnow [19] research project and later the use of the method was also found relevant in a study of a Danish hospital wanting to create an AI-based predictive model for clinical no-shows. The method and prototype were initially evaluated by a consultant employed in a process mining company and a municipal case worker collaborating with the authors in the EcoKnow research project.

The paper is structured as follows. Prior and related work is discussed in Sec. 2. Sec. 3 explains our proposed BERMUDA method, where we also show a prototype tool. Sec. 4 introduces two specific case studies in a job center and a danish hospital. A brief evaluation of the use of the method in the first case along with a discussion on the results is made in Sec. 5. Lastly, in Sec. 6 we conclude and discuss future work.

2 Prior and Related work

Within health-care informatics, problems arising from having a primary use of data (original intend of health-care delivery and services) and different, secondary use of data (emergence of new possibilities through statistics and data science) has been highlighted in several studies [5, 28, 37]. The authors of [5] found that underlying issues for data quality and reuse was attributed to differential incentives for the accuracy of the data; flexibility in system software that allowed multiple routes to documenting the same tasks; variability in documentation practices among different personnel documenting the same task; variability in use of standardized vocabulary, specifically, the internally developed standardized vocabulary of practice names; and changes in project procedures and electronic system configuration over time, as when a paper questionnaire was replaced with an electronic version.

Such underlying socio-technical issues to data capturing can attribute to an overall lower degree of data integrity resulting in little to no secondary usefulness of data representing health-care events. A similar [18] study conducted by this papers co-authors highlighted the need for iteratively aligning data creation and use with domain experts and data creators (i.e. doctors, nurses, secretaries, etc) when conducting data science on operational data from hospitals.

We see event abstraction [40] as a related topic to our paper, however we approach the problem in a top-down manner i.e. from domain knowledge down to the data source. A similar top-down approach exists in database systems [12] where an ontology of domain concepts is used to query the databases. We do not aim to propose techniques for process discovery as there are a plethora of tools already in use for this task, some of which [35] also allow for domain expert interventions. We propose BERMUDA both for pre-processing of data before moving to process discovery or building predictive models, and for training of new users in how to consistently record data suitable for the secondary uses.

The paper [21] provides a procedure for extracting event logs from databases that makes explicit the decisions taken during event log building and demonstrates it through a running example instead of providing tool support. The paper [7] present a semi-automatic approach that maps events to activities by transforming the mapping problem into the a constraint satisfaction problem, but it does not directly handle the event log extraction.

In [29] the authors describe a meta model that separates the extraction and analysis phases and makes it easy to connect event logs with SQL queries. In [30] they associate events from different databases into a single trace and propose an automated event log building algorithm. They point towards the lack of domain knowledge as a driving force for an automated and efficient approach. They discuss that their definition of event log “interestingness” as an objective score ignores aspects of domain level relevance. Both papers bind database scripts and event log concepts in order to build ontologies/meta-models, but do not link to domain knowledge in order to provide traceability to domain experts, such that the limitations of the “interestingness” score may be overcome.

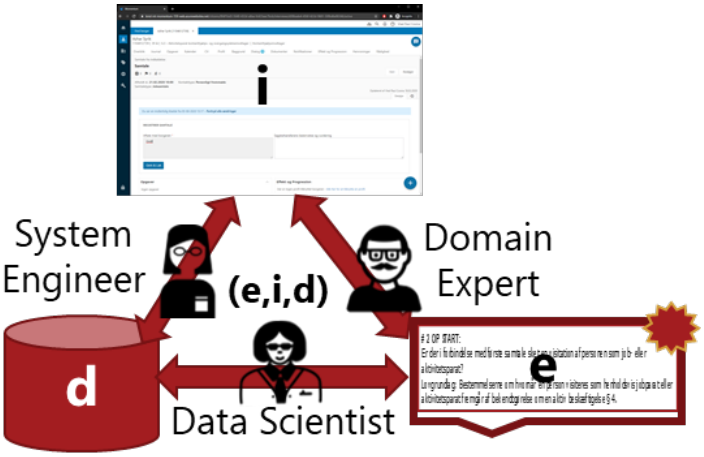
To summarize, most work [6, 9, 10, 16, 17, 24, 25, 33, 38] on event data quality so far has focused on technical means to repair and maintain the quality of event logs [15]. Our approach complements these approaches by focusing on the socio-technical problem of aligning what is done in practice by the users of the information systems, i.e. how is a domain activity registered within the system, and at the other hand, where is this event stored in the database.

3 BERMUDA: Mapping domain events to data

Our method relies on so-called BERMUDA triples **(e,i,d)** as illustrated in Fig.1, recording the relation between respectively a domain event **e**, a user interface element **i** of the information system in which the domain event is registered and the location of the resulting data element **d** in the database. A concrete example from one of our case studies can be seen in Fig.2. Here a domain event ”Register ... during the first interview” is described in a textual audit schema. This is linked by a screen shot to the drop down menu in the user interface, where the case worker performs this concrete registration. And finally, the location of the resulting data element is recorded by an SQL statement that extracts the event.

There are typically three roles involved in the recording such BERMUDA triples: Data scientist (or analyst), domain expert and system engineer. As guidance towards applying our method we recommend following these steps:

1. **Domain to user interface**. For each domain event **e**, the domain experts record an association **(e,i)** between the domain event **e** and an (user or system) interface element **i**.

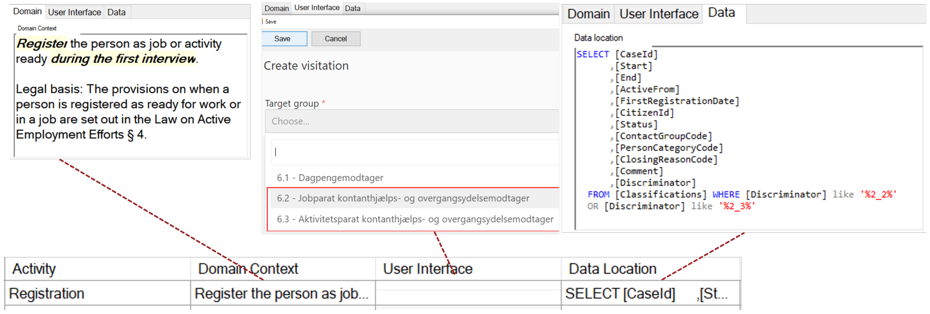


2 OP START:
Er der i forbindelse med første samtale sket en visitation af personen som job- eller aktivitetsparat?
Lovgrundlag: Bestemmelserne om hvornår en person visiteres som henholdsvis jobparat eller aktivitetsparat fremgår af bekendtgørelse om en aktiv beskæftigelse § 4.

Fig. 1. BERMUDA method

2. **User interface to data**. Through code inspection or simulation, system engineers develop the correct database query **d** to extract the data recording the event **e** created via the interface element **i** resulting in a triple **(e,i,d)**.
3. **Triples to event log**. The data scientist merges and refines the database queries and creates the initial version of the event log. The event log entries are enriched with extra attributes that hold a reference to the domain event, the interface element and the data source from where the entry originated.

Prototype tool. To facilitate the adoption of the BERMUDA method we present a prototype tool to illustrate how the triples can be created and an event log extracted. A screenshot from the prototype is shown in Fig. 2. Briefly, the UI consists of 3 input areas in the top for documenting the individual parts of triples (description of domain event, system interface, script for extracting the event from the system), an input area at the bottom for adding and selecting a



Register the person as job or activity ready ***during the first interview***.

Legal basis: The provisions on when a person is registered as ready for work or in a job are set out in the Law on Active Employment Efforts § 4.

Create visitation

Target group *

Choose...

6.1 - Dagpengemodtager

6.2 - Jobparat kontanthjælps- og overgangsydelsemodtager

6.3 - Aktivitetsparat kontanthjælps- og overgangsydelsemodtager

```
SELECT [CaseId]
      ,[Start]
      ,[End]
      ,[ActiveFrom]
      ,[FirstRegistrationDate]
      ,[CitizenId]
      ,[Status]
      ,[ContactGroupCode]
      ,[PersonCategoryCode]
      ,[ClosingReasonCode]
      ,[Comment]
      ,[Discriminator]
  FROM [Classifications] WHERE [Discriminator] like '%2_2%'
  OR [Discriminator] like '%2_3%'
```

Activity	Domain Context	User Interface	Data Location
Registration	Register the person as job...		SELECT [CaseId] ,[St...

Fig. 2. BERMUDA method Prototype

triple to document, and a display area (not shown in the figure) for the resulting event log. [4]

The prototype has a simple role base access control supporting the use of the method in practice. All roles have access to the description of domain events, in order to build trust through a common domain understanding. Domain experts have access to domain events and the user interface input areas. System engineers need access to all areas, but not the production data in the information system. Data scientists are allowed access to all areas except they can not see the data extraction scripts, if they are covered by intellectual propriety rights. They can however run the scripts on the production system, to extract the event data.

4 Cases: Secondary use of Municipal and Health data

We discuss the method in relation to two concrete cases from Denmark where data in respectively a municipality and a hospital were intended to be used for AI-based decision support. Case 1 is elicited at a municipal job center in Denmark and case 2 covers our work with a regional research hospital where a project aiming for producing and using an AI model for no-shows. Both cases unveiled a gap between how data is produced in a local context for its primary purpose of case management and what it represents when extracted and used for decision support. We made an evaluation of our BERMUDA prototype for case one and speculate how it could be used in case two.

Case 1: As part of the EcoKnow research project [19], we had by the software vendor KMD (kmd.dk), been given access to interact with the system engineers that developed the case management system used in danish job centers. Collaborating with colleagues in the EcoKnow research project performing field studies at the job center [4,20,31], we also had the opportunity to gather domain knowledge through workshops, semi-structured interviews and informal methods from job center employees. Finally, we had access to historical data from about 16000 citizens with the purpose of researching the possibilities for improving compliance and the experienced quality of case management in municipalities.

In addition to our case we interviewed a consultant at a process mining company Infoventure (infoventure.dk), doing conformance checking, using the same case management system but a different data source. Their current practice relies on first co-creating a document with employees at the job center, which contained the necessary domain knowledge and screenshots of user interface elements with relevant explanations. Next it was the task of the consultant to build extraction scripts for the identified domain events. During this phase there was ongoing communication with the software vendor and job center employees through meetings, calls or emails, in order to build up the necessary domain and system knowledge. Often he would observe specific data (an exact timestamp or citizen registration number) in the user interface and proceed to search for that exact information in the database. This process was done either offline, with the

[4] The prototype is available at: `https://github.com/paul-cvp/bermuda-method`.

aid of screenshots, or on site by sitting next to a case worker. The links between domain events and the data extracted from the database was recorded in an ad-hoc way and only available to the consultant.

Domain Activities/events: We used a management audit schema comprised of 21 questions. From these questions we define the domain activities/events relevant for the case compliance analysis. For example: From the audit question "Is the first job interview held within one week of the first request? Legal basis: LAB § 31(3)" we can identify several domain event data of interest: first request, first job interview, first week passed.

Graphical User Interface (GUI) Areas for recording domain events A caseworker employed at the job center associated the domain events identified in the audit questions with areas of the user interface where caseworkers record the event. From the 21 questions, 11 domain events could be identified that could be given a user interface association. For 3 of the domain events, the caseworker was unsure where to record it. A data scientist was able to associate 12 of the 21 domain events to a field in the user interface. This relatively low number of associations can be explained by the fact that the audit schema was created by the municipality and not the vendor of the it-system, and thus, some of the domain events relevant for the audit did not have a direct representation in the user interface. Therefore certain events were completely missing or documented in free text fields, while others require access to other systems used by the municipality. In particular, as also observed in [4], the free text field was sometimes used to describe the categorisation of the unemployed citizen (as activity or job ready) or the reason for the choice of categorisation, by selecting the reason "other", instead of using one of the specific predefined values available in the system interface.

Data and database organization. The database contains 133 tables with 1472 columns in total. By having access to source code and the system engineers, we mapped the identified GUI elements to the database. Furthermore this limited our inspection to 8 main tables from which the data was extracted and 4 tables used for mapping table relations, thus ensuring data minimisation as specified in the General Data Protection Regulation (GDPR) [14].

Case 2: In the wake of a grand scale implementation of an EPIC[5] Regional Electronic Health Record-system (EHR-system) purchase and implementation, we have since 2017 been engaged in a longitudinal case-study of facilitating and developing an AI-model for predicting patient no-shows based on clinical event and demographic data. The project was pioneering as the first test of the models developed from local data and appointed a small endoscopy unit at Bispebjerg hospital (a research hospital in the capital region of Denmark). The project have a foundation in participatory design and end-user involvement in pursuit

[5] epic.com

of creating visions for use of data and AI, as well as creating synergy effects for data creation among clinicians, nurses and clinical secretaries as domain experts creating clinical event data used to predict future no-shows.

We extracted 8 different data sets together with the regional data team to learn about implications for applying such data for machine-learning purposes. We here learned, that missing data values and incomplete submissions were largely representing the first data sets and that due to missing guidelines and coordinated workflows each individual health care person had different understanding of the categories used to report clinical appointment statuses.

Domain Events: Interpretations of the events. We conducted 2 follow-up interviews with clinical secretaries to understand the local flow of data submission into the EHR-system. The clinical secretaries demonstrated their data submission practices and their understanding of how to document clinical appointment statuses into the EHR-system. We further conducted four 2-hour workshops involving the clinical secretaries in putting context to their workflow and use of categories to assign meaning to no-show categories. In the same period, we invited Regional data management and extraction teams to learn from practices and iteratively extract data sets with no-show data.

Data and database organization. 8 data sets were extracted in total over a period of 3 months before a machine learning algorithm could be fed with a data set with sufficient domain contexts to remove categories that didn't have meaning for secondary use. The best example of this was again the free text category "other" as a category for assigning reason for no-shows or cancellations of appointments. This category was heavily used by all clinical staff due to its ability to avoid reading through 16 other categories of reason for mentioned outcome. The first data set had 81.000 rows and observations with 2/3 of those past appointments being assigned "other" with text-field inputs sometimes representing the same categories as suggested in the drop-down menu and sometimes left empty or with "other" written in the text-field. A further 11.000 appointments were deemed incomplete or "in process" several months after appointment date. When sorting out unassigned events for appointment status the department only had 2880 observations left for the machine learning algorithm.

5 Initial Evaluation

As an initial qualitative evaluation of the usefulness of the method, we conducted two semi-structured interviews, one with a municipal case worker acting as a domain expert and another with a data scientist working as consultant in the process analysis company Infoventure. Both interview respondents collaborated with the authors in the Ecoknow research project. The municipal case worker was given the task of mapping business activities to user-interface elements of a case and document management system. The consultant was asked about the current practice of documenting event log extraction for process mining, illustrated by a concrete case, and how the Bermuda prototype could support or improve this practice.

Overall, the evaluation indicated, that the BERMUDA method exhibits the following positive proprieties:

- **Transparency, Accountability, and Traceablity.** The BERMUDA triples make it possible to trace the relation between events extracted from a data base, e.g. for the creation of an event log, and domain events. Both interviewees saw the advantage in unambiguously referencing domain events across different roles of a data science project (domain expert, software engineer, data scientist), thereby providing accountability for the data provenance/lineage, while also building trust across different roles.
- **Accuracy.** Through the participatory co-creation of the event log it is possible to observe that the event log correctly captures the relevant domain knowledge. As each of the roles interact with each other, they can observe that the correct steps were taken in the extraction of event data for secondary use. This was already to some extend part of the current practices, but BERMUDA supported the consistent documentation.
- **Maintainability and Training.** The interview participants indicated that the Bermuda method is useful for maintaining event logs over time when changes happen in the domain or system, because the information is documented consistently in one place. They also pointed out, that the method and tool for the same reason could be valuable both in training new data scientists and new case workers.
- **Protection of Intellectual Property.** Since each link in the BERMUDA triangle can be defined independently, the system engineers can provide mappings that can be used to extract events without revealing the code of the system. We observed this in the interaction between the data science consultant and the system engineers developing the job center solution.

Limitations. Firstly, the tool is not mature enough to replace a general SQL scripting environment. Secondly, it does not yet account for data that are not stored in an SQL database, nor for data that is not recorded via user interface, as for instance data recorded automatically by system events.

6 Conclusion and future work

In this paper we presented BERMUDA, a method for facilitating the responsible secondary use of event data in data science projects by supporting the collaboration between domain experts, system engineers and data scientists on associating domain events, via user interfaces to data in the database. This facilitates transparent extraction of event logs for analysis and thereby accountable data lineage. We discussed its use through cases of data science projects at a job center in a Danish municipality and a Danish hospital. In particular, both cases highlight the frequent use of the category “other” in the registration of reasons for domain events, instead of using pre-defined values in drop down menus. We showed through a prototype tool how BERMUDA can facilitate the interactions between domain experts, system engineers and data scientists. Furthermore we conducted interviews in order to lightly evaluate its usefulness and limitations.

In the future we expect to conduct more field trials of the method and interview more practitioners in order to do a thematic analysis for better qualitative feedback. We aim to investigate how the results of applying BERMUDA can be used when training domain experts to use the appropriate categories instead of "other". We also aim to extend the tool with an automatic signaling system to monitor for changes in the user interface and in the database structure to notify the data scientist of possible misalignment in existing processes. We hope to increase the robustness of the tool and its compatibility with existing process mining tools. We also aim to provide the prototype as an online tool in order to facilitate remote cooperative work. Finally we aim to support a broader range of input and output formats by applying the method on diverse data sources from information systems in relevant domains.

Acknowledgements

Thanks to Infoventure, KMD Momentum, Bispebjerg Hospital, The Capital Region of Denmark, Gladsaxe and Syddjurs municipalities, and the reviewers.

References

1. van der Aalst, W.M.P.: Process Mining - Data Science in Action, Second Edition. Springer (2016)
2. van der Aalst, W.M.P., et. al.: Process mining manifesto. In: Daniel, F., Barkaoui, K., Dustdar, S. (eds.) Business Process Management Workshops - BPM 2011 International Workshops, Clermont-Ferrand, France, August 29, 2011, Revised Selected Papers, Part I. Lecture Notes in Business Information Processing, vol. 99, pp. 169–194. Springer (2011)
3. on AI, H.L.E.G.: Ethics guidelines for trustworthy ai, `https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai`
4. Ammitzbøll Flügge, A., Hildebrandt, T., Møller, N.H.: Street-level algorithms and ai in bureaucratic decision-making: A caseworker perspective. Proc. ACM Hum.-Comput. Interact. **5**(CSCW1) (apr 2021). https://doi.org/10.1145/3449114, `https://doi.org/10.1145/3449114`
5. Ancker, J.S., Shih, S., Singh, M.P., Snyder, A., Edwards, A., Kaushal, R., Investigators, H., et al.: Root causes underlying challenges to secondary use of data. In: AMIA Annual Symposium Proceedings. vol. 2011, p. 57. American Medical Informatics Association (2011)
6. Andrews, R., Emamjome, F., ter Hofstede, A.H.M., Reijers, H.A.: An expert lens on data quality in process mining. In: ICPM. pp. 49–56. IEEE (2020)
7. Baier, T., Rogge-Solti, A., Weske, M., Mendling, J.: Matching of events and activities - an approach based on constraint satisfaction. In: Frank, U., Loucopoulos, P., Pastor, Ó., Petrounias, I. (eds.) The Practice of Enterprise Modeling. pp. 58–72. Springer Berlin Heidelberg, Berlin, Heidelberg (2014)

8. Björgvinsson, E., Ehn, P., Hillgren, P.A.: Participatory design and" democratizing innovation". In: Proceedings of the 11th Biennial participatory design conference. pp. 41–50 (2010)
9. Bose, J.C.J.C., Mans, R.S., van der Aalst, W.M.P.: Wanna improve process mining results? In: CIDM. pp. 127–134. IEEE (2013)
10. Bose, R.P.J.C., van der Aalst, W.M.P., Zliobaite, I., Pechenizkiy, M.: Handling concept drift in process mining. In: CAiSE. Lecture Notes in Computer Science, vol. 6741, pp. 391–405. Springer (2011)
11. Cabitza, F., Campagner, A., Balsano, C.: Bridging the "last mile" gap between ai implementation and operation:"data awareness" that matters. Annals of translational medicine **8**(7) (2020)
12. Calvanese, D., Giacomo, G.D., Lembo, D., Lenzerini, M., Rosati, R.: Ontology-Based Data Access and Integration, pp. 2590–2596. Springer New York, New York, NY (2018)
13. Cosma, V.P., Hildebrandt, T.T., Slaats, T.: Bermuda: Towards maintainable traceability of events for trustworthy analysis of non-process-aware information systems. In: EMISA Forum: Vol. 41, No. 1. De Gruyter (2021)
14. Council of European Union: Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. https://publications.europa.eu/s/llVw (May 2016)
15. De Weerdt, J., Wynn, M.T.: Foundations of process event data. Process Mining Handbook. LNBIP **448**, 193–211 (2022)
16. Emamjome, F., Andrews, R., ter Hofstede, A.H.M., Reijers, H.A.: Alohomora: Unlocking data quality causes through event log context. In: ECIS (2020)
17. Fischer, D.A., Goel, K., Andrews, R., van Dun, C.G.J., Wynn, M.T., Röglinger, M.: Enhancing event log quality: Detecting and quantifying timestamp imperfections. In: Fahland, D., Ghidini, C., Becker, J., Dumas, M. (eds.) Business Process Management. pp. 309–326. Springer International Publishing, Cham (2020)
18. H. Gyldenkaerne, C., From, G., Mønsted, T., Simonsen, J.: Pd and the challenge of ai in health-care. In: Proceedings of the 16th Participatory Design Conference 2020-Participation (s) Otherwise-Volume 2. pp. 26–29 (2020)
19. Hildebrandt, T.T., et. al.: EcoKnow: Engineering Effective, Co-Created and Compliant Adaptive Case Management Systems for Knowledge Workers, p. 155–164. Association for Computing Machinery, New York, NY, USA (2020), https://doi.org/10.1145/3379177.3388908
20. Holten Møller, N., Shklovski, I., Hildebrandt, T.T.: Shifting concepts of value: Designing algorithmic decision-support systems for public services. In: Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society. NordiCHI '20, Association for Computing Machinery, New York, NY, USA (2020), https://doi.org/10.1145/3419249.3420149
21. Jans, M., Soffer, P.: From relational database to event log: Decisions with quality impact. In: Business Process Management Workshops. Lecture Notes in Business Information Processing, vol. 308, pp. 588–599. Springer (2017)
22. Jung, J.Y., Steinberger, T., So, C.: Domain experts as owners of data: towards sustainable data science (2022)
23. Kensing, F., Simonsen, J., Bodker, K.: Must: A method for participatory design. Human-computer interaction **13**(2), 167–198 (1998)
24. Leopold, H., van der Aa, H., Pittke, F., Raffel, M., Mendling, J., Reijers, H.A.: Searching textual and model-based process descriptions based on a unified data format. Softw. Syst. Model. **18**(2), 1179–1194 (2019)

25. Li, G., de Murillas, E.G.L., de Carvalho, R.M., van der Aalst, W.M.P.: Extracting object-centric event logs to support process mining on databases. In: CAiSE Forum. Lecture Notes in Business Information Processing, vol. 317, pp. 182–199. Springer (2018)
26. Lux, M., Rinderle-Ma, S.: Problems and challenges when implementing a best practice approach for process mining in a tourist information system. In: BPM (Industry Track). CEUR Workshop Proceedings, vol. 1985, pp. 1–12. CEUR-WS.org (2017)
27. Mannhardt, F.: Responsible process mining. Process Mining Handbook. LNBIP **448**, 373–401 (2022)
28. Meystre, S.M., Lovis, C., Bürkle, T., Tognola, G., Budrionis, A., Lehmann, C.U.: Clinical data reuse or secondary use: current status and potential future progress. Yearbook of medical informatics **26**(01), 38–52 (2017)
29. de Murillas, E.G.L., Reijers, H.A., van der Aalst, W.M.P.: Connecting databases with process mining: a meta model and toolset. Softw. Syst. Model. **18**(2), 1209–1247 (2019)
30. de Murillas, E.G.L., Reijers, H.A., van der Aalst, W.M.P.: Case notion discovery and recommendation: automated event log building on databases. Knowl. Inf. Syst. **62**(7), 2539–2575 (2020)
31. Petersen, A.C.M., Christensen, L.R., Harper, R., Hildebrandt, T.: "we would never write that down": Classifications of unemployed and data challenges for ai. Proc. ACM Hum.-Comput. Interact. **5**(CSCW1) (apr 2021), `https://doi.org/10.1145/3449176`
32. Robertson, T., Simonsen, J.: Participatory design: an introduction. In: Routledge international handbook of participatory design, pp. 1–17. Routledge (2012)
33. Sànchez-Ferreres, J., van der Aa, H., Carmona, J., Padró, L.: Aligning textual and model-based process descriptions. Data Knowl. Eng. **118**, 25–40 (2018)
34. Schrodt, P.A.: The statistical characteristics of event data. International Interactions **20**(1-2), 35–53 (1994)
35. Schuster, D., van Zelst, S.J., van der Aalst, W.M.P.: Cortado—an interactive tool for data-driven process discovery and modeling. In: Buchs, D., Carmona, J. (eds.) Application and Theory of Petri Nets and Concurrency. pp. 465–475. Springer International Publishing, Cham (2021)
36. Slaats, T.: Declarative and hybrid process discovery: Recent advances and open challenges. J. Data Semant. **9**(1), 3–20 (2020). https://doi.org/10.1007/s13740-020-00112-9
37. Smylie, J., Firestone, M.: Back to the basics: Identifying and addressing underlying challenges in achieving high quality and relevant health statistics for indigenous populations in canada. Statistical Journal of the IAOS **31**(1), 67–87 (2015)
38. Suriadi, S., Andrews, R., ter Hofstede, A.H.M., Wynn, M.T.: Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs. Inf. Syst. **64**, 132–150 (2017)
39. Team, R.: Responsible data science, `https://redasci.org/`
40. van Zelst, S., Mannhardt, F., de Leoni, M.: Event abstraction in process mining: literature review and taxonomy, vol. 6, pp. 719–736. Springer New York, New York, NY (2021). https://doi.org/10.1007/s41066-020-00226-2
41. Zhang, A.X., Muller, M., Wang, D.: How do data science workers collaborate? roles, workflows, and tools. Proceedings of the ACM on Human-Computer Interaction **4**(CSCW1), 1–23 (2020)