# Privacy, Security and Ethics in Process Mining

When I moved to the Netherlands 12 years ago and started grocery shopping at one of the local supermarket chains, Albert Heijn, I initially resisted getting their Bonus card (a loyalty card for discounts), because I did not want the company to track my purchases. I felt that using this information would help them to manipulate me by arranging or advertising products in a way that would make me buy more than I wanted to. It simply felt wrong.

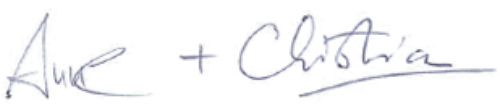
The truth is that no data analysis technique is intrinsically good or bad. It is always in the hands of the people using the technology to make it productive and constructive. For example, while supermarkets could use the information tracked through the loyalty cards of their customers to make sure that we have to take the longest route through the store to get our typical items (passing by as many other products as possible), they can also use this information to make the shopping experience more pleasant, and to offer more products that we like.[1]

Most companies have started to use data analysis techniques to analyze their data in one way or the other. These data analyses can bring enormous opportunities for the companies and for their customers, but with the increased use of data science the question of ethics and responsible use also grows more dominant. Initiatives like the Responsible Data Science seminar series [1] take on this topic by raising awareness and encouraging researchers to develop algorithms that have concepts like fairness, accuracy, confidentiality, and transparency built in [2].

Process Mining can provide you with amazing insights about your processes, and fuel your improvement initiatives with inspiration and enthusiasm, if you approach it in the right way. But how can you ensure that you use process mining responsibly? What should you pay attention to when you introduce process mining in your own organization?

In this article, we provide you four guidelines that you can follow to prepare your process mining analysis in a responsible way.

Your friends from Fluxicon,



[1] Listen to this Planet Money podcast episode about why the milk is in the back of the store if you want to dive deeper into the world of supermarket theories.

## 1. Clarify Goal of the Analysis

The good news is that in most situations Process Mining does not need to evaluate personal information, because it usually focuses on the internal organizational processes rather than, for example, on customer profiles. Furthermore, you are investigating the overall process patterns. For example, a process miner is typically looking for ways to organize the process in a smarter way to avoid unnecessary idle times rather than trying to make people work faster.

However, as soon as you would like to better understand the performance of a particular process, you often need to know more about other case attributes that could explain variations in process behaviours or performance. And people might become worried about where this will lead them.

Therefore, already at the very beginning of the process mining project, you should think about the goal of the analysis. Be clear about how the results will be used. Think about what problem are you trying to solve and what data you need to solve this problem.

### Do:

- *Check whether there are legal restrictions regarding the data*. For example, in Germany employee-related data cannot be used and typically simply would not be extracted in the first place. If your project relates to analyzing customer data, make sure you understand the restrictions and consider anonymization options (see guideline No. 3).
- *Consider establishing an ethical charter* (see Appendix A for an example charter) that states the goal of the project, including what will and what will not be done based on the analysis. For example, you can clearly state that the goal is not to evaluate the performance of the employees. Communicate to the people who are responsible for extracting the data what these goals are and ask for their assistance to prepare the data accordingly.

### Don't:

- *Start out with a fuzzy idea and simply extract all the data you can get*. Instead, think about what problem are you trying to solve? And what data do you actually need to solve this problem? Your project should focus on business goals that can get the support of the process managers you work with (see guideline No. 4).
- *Make your first project too big*. Instead, focus on one process with a clear goal. If you make the scope of your project too big, people might block it or work against you while they do not yet even understand what process mining can do.

## 2. Responsible Handling of Data

Like in any other data analysis technique, you must be careful with the data once you have obtained it. In many projects, nobody thinks about the data handling until it is brought up by the security department. Be that person who thinks about the appropriate level of protection and has a clear plan already prior to the collection of the data.

### Do:

- *Have external parties sign a Non Disclosure Agreement (NDA) to ensure the confidentiality of the data*. This holds, for example, for consultants you have hired to perform the process mining analysis for you, or for researchers who are participating in your project. Contact your legal department for this. They will have standard NDAs that you can use.
- *Make sure that the hard drive of your laptop, external hard drives, and USB sticks that you use to transfer the data and your analysis results are encrypted*.

### Don't:

- *Give the data set to your co-workers before you have checked what is actually in the data*. For example, it could be that the data set contains more information than you requested, or that it contains sensitive data that you did not think about. For example, the names of doctors and nurses might be mentioned in a free-text medical notes attribute. Make sure you remove or anonymize (see guideline No. 3) all sensitive data before you pass it on.
- *Upload your data to a cloud-based process mining tool* without checking that your organization allows you to upload this kind of data. Instead, use a desktop-based process mining tool (like Disco or ProM) to analyze your data locally or get the cloud-based process mining vendor to set-up an on-premise version of their software within your organization. This is also true for cloud-based storage services like Dropbox: Don't just store data or analysis results in the cloud even if it is convenient.

## 3. Consider Anonymization

If you have sensitive information in your data set, instead of removing it you can also consider the use of anonymization. When you anonymize a set of values, then the actual values (for example, the employee names “Mary Jones”, “Fred Smith”, etc.) will be replaced by another value (for example, “Resource 1”, “Resource 2”, etc.).

If the same original value appears multiple times in the data set, then it will be replaced with the same replacement value (“Mary Jones” will always be replaced by “Resource 1”). This way, anonymization allows you to obfuscate the original data but it preserves the patterns in the data set for your analysis. For example, you will still be able to analyze the workload distribution across all employees without seeing the actual names.

Some process mining tools (Disco and ProM) include anonymization functionality. This means that you can import your data into the process mining tool and select which data fields should be anonymized. For example, you can choose to anonymize just the Case IDs, the resource name, attribute values, or the timestamps. Then you export the anonymized data set and you can distribute it among your team for further analysis.

### Do:

- *Determine which data fields are sensitive and need to be anonymized* (see also the list of common process mining attributes and how they are impacted if anonymized in Appendix B).
- *Keep in mind that despite the anonymization certain information may still be identifiable*. For example, there may be just one patient having a very rare disease, or the birthday information of your customer combined with their place of birth may narrow down the set of possible people so much that the data is not anonymous anymore.

### Don’t:

- *Anonymize the data before you have cleaned your data*, because after the anonymization the data cleaning may not be possible anymore. For example, imagine that slightly different customer category names are used in different regions but they actually mean the same. You would like to merge these different names in a data cleaning step. However, after you have anonymized the names as “Category 1”, “Category 2”, etc. the data cleaning cannot be done anymore.
- *Anonymize fields that do not need to be anonymized*. While anonymization can help to preserve patterns in your data, you can easily lose relevant information. For example, if you anonymize the Case ID in your incident management process, then you cannot look up the ticket number of the incident in the service desk system anymore. By establishing a collaborative culture around your process mining initiative (see guideline No. 4) and by working in a responsible, goal-oriented way, you can often work openly with the original data that you have within your team.

## 4. Establish a Collaborative Culture

Perhaps the most important ingredient in creating a responsible process mining environment is to establish a collaborative culture within your organization. Process mining can make the flaws in your processes very transparent, much more transparent than some people may be comfortable with. Therefore, you should include change management professionals, for example, Lean practitioners who know how to encourage people to tell each other “the truth”, in your team [3].

Furthermore, be careful how you communicate the goals of your process mining project and involve relevant stakeholders in a way that ensures their perspective is heard. The goal is to create an atmosphere, where people are not blamed for their mistakes (which only leads to them hiding what they do and working against you) but where everyone is on board with the goals of the project and where the analysis and process improvement is a joint effort.

### Do:

- *Make sure that you verify the data quality before going into the data analysis*, ideally by involving a domain expert already in the data validation step [4]. This way, you can build trust among the process managers that the data reflects what is actually happening and ensure that you have the right understanding of what the data represents.
- *Work in an iterative way* and present your findings as a starting point for discussion in each iteration. Give people the chance to explain why certain things are happening and let them ask additional questions (to be picked up in the next iteration). This will help to improve the quality and relevance of your analysis as well as increase the buy-in of the process stakeholders in the final results of the project.

### Don’t:

- *Jump to conclusions*. You can never assume that you know everything about the process. For example, slower teams may be handling the difficult cases, people may deviate from the process for good reasons, and you may not see everything in the data (for example, there might be steps that are performed outside of the system). By consistently using your observations as a starting point for discussion, and by allowing people to join in the interpretation, you can start building trust and the collaborative culture that process mining needs to thrive.
- *Force any conclusions that you expect*, or would like to have, by misrepresenting the data (or by stating things that are not actually supported by the data). Instead, keep track of the steps that you have taken in the data preparation and in your process mining analysis. If there are any doubts about the validity or questions about the basis of your analysis, you can always go back and show, for example, which filters have been applied to the data to come to the particular process view that you are presenting.

## *Appendix A: Example Project Charter*

*Project Charter 'Process Study using Process Mining'*

Process Mining is a new process analysis technique. From the records (log files) available in software packages that are used, Process Mining can map and analyze the sequences of activities that make up the process.

This technique is a complement to the more traditional approach to the study of processes which is based on interviews with stakeholders. Compared to the conventional approach, Process Mining delivers more comprehensive and more accurate process maps. The quality of the process analysis is thereby enhanced.

Like other work analysis methods, badly framed, malicious and reckless use of Process Mining could lead to unfair and unlawful supervision or control in monitoring the behavior of employees at their workplace.

In order to manage both the interests of the company and the protection of privacy of employees, this Charter states that:

- The first aim of this study by Process Mining is to look for process improvement in using the GECO software;
- This study is not intended to monitor or control the behavior of employees at their workplace;
- Results of the study will not be used for individual evaluation;
- The results of this study will be presented such that we can not draw conclusions on the (individual) behavior of employees;
- Employees will be regularly informed about the progress of the study;
- In order to better understand and clarify some discoveries, employees may be invited to attend meetings with the person responsible for this study. Their managers will not have to know the exact content of these discussions;
- Employees have full access to the final report of this study.

Lausanne, dd/mm/yyyy

Signed by the head of the department, the responsible for the study and the collaborator

*This example project charter has been contributed by Léonard Studer from the City of Lausanne.*

## *Appendix B: Anonymization of Common Process Mining Fields*

*Process mining attributes and why you might want (or might not want) to anonymize them:*

**Resource name**: Removing the names of the employees working in the process is one of the more common anonymization steps. It can help to decrease friction and put employees more at ease when you involve them in a joint analysis workshop. Anonymizing employee names certainly is a must if you make your data publicly available in some form.

Be aware that it may still be possible to trace back individual employees. For example, if you look up a concrete case based on the case ID in the operational system, you will see the actual resource names there.

Finally, keep in mind that anonymizing employee names for an internal process mining analysis also removes valuable information. For example, if you identify process deviations or an interesting process pattern, normally the first step is to speak with the employees who were involved in this case to understand what happened and learn from them.

**Case ID**: Anonymizing the case ID is a must if it contains sensitive information. For example, if you analyze the income tax return process at the tax office, then the case ID will be a combination of the social security number of the citizen and the year of the tax declaration. You will have to replace the social security information for obvious reasons.

However, for data sets where the case ID is less sensitive it is a good idea to keep it in place as it is. The benefit will be that you can look up individual cases in the operational system to verify your analysis or obtain additional information. Losing this link will limit your ability to perform root cause analyses and take action on the process problems that you discover.

**Activity name**: Normally, you would not anonymize the activity name itself. The activities are the process steps that appear in the process map and in the variant sequences in the Process Mining tool. The reason why you do not want to replace the activity names by, for example, “Activity 1”, “Activity 2”, “Activity 3”, etc., is that most processes become very complex very quickly and without the activity names you have no chance to understand the process flows you are analyzing. Your analysis becomes useless.

Keeping the activity names in full is usually not a problem, because they describe a generic process step (like “Email sent”). However, especially if you have many different activity names in your data, you should review them to ensure they contain no confidential information (e.g., “Email sent by lawyer X”).

## *Appendix B: Anonymization of Common Process Mining Fields (cont'd)*

*Process mining attributes and why you might want (or might not want) to anonymize them:*

**Other Attributes**: Sensitive information is often contained in additional attribute columns. For example, even if you are analyzing an internal ordering process, there might be additional data fields revealing information about the customer.

You can either completely remove data columns that you don't need, or you can anonymize their values. Keep the attribute columns that are not sensitive in their original form, because they can contain important context information when you inspect individual cases during your Process Mining analysis.

Finally, be aware that sensitive information can also be hidden in a 'Notes' attribute or some other kind of free-text field, where the employees write down additional information about the case or the process step. Simply anonymizing such a free-text field would be useless, because the whole text would be replaced by "Value 1", "Value 2", etc. To preserve the usefulness of the free-text field while removing sensitive information requires more work in the data pre-processing step and is not something that process mining tools can do for you automatically.

**Timestamps**: Sometimes, the time at which a particular activity happened already reveals too much information and would make it possible to identify one of your business entities in an unwanted way. In such situations, you can anonymize the timestamps by applying an offset. This means that a certain number of days, hours, and minutes will be added to the actual timestamps to create new (now anonymized) timestamps.

Keep in mind that some of the process patterns may change when you analyze data sets with anonymized timestamps. For example, you might see activities appear on other times of the day than you would see in the original data set. For this reason, timestamp anonymization is mostly used if data sets are prepared for public release and not if you analyze a process within your company.

### *Was this guide useful to you?*

This guide was created as part of the Process Mining News initiative (you can **register here for free** to make sure you do not miss any future editions).

Please get in touch via **support@fluxicon.com** to let us know how it worked for you and which other issues you encountered. We would love to hear from you!

The Process Mining News is brought to you by **Fluxicon**. Founded in 2009 by Dr. Anne Rozinat and Dr. Christian W. Günther, Fluxicon has been at the forefront of the process mining movement ever since. Our process mining software **Disco** is based on proven scientific research, and loved by professionals worldwide for setting the gold standard in performance and user experience.

As the most experienced process mining team in industry, Fluxicon supplies a whole range of companies on all continents, many of them in the Global Fortune 100 and Fortune 500 ranks.

Fluxicon also organizes the annual process mining conference, Process Mining Camp (**www.processminingcamp.com**), helps to raise the visibility of process mining as a new data analysis method through numerous invited talks and articles, and supports more than 400 universities through the Fluxicon Academic Initiative (**http://fluxicon.com/academic/**).

## Acknowledgements

We would like to thank **Frank van Geffen** and **Léonard Studer**, who initiated the first discussions in the workgroup around responsible process mining in 2015. Furthermore, we would like to thank **Moe Wynn**, **Felix Mannhardt** and **Wil van der Aalst** for their feedback on earlier versions of this document. Moe's input, based on her experience with performing ethical process mining projects at the Queensland University of Technology, has been particularly helpful and most of her suggestions were included in the final article. Finally, we are very grateful to **Léonard Studer** for allowing us to share his example ethical charter with you in this guide.

## References

[1] Responsible Data Science (RDS) initiative: http://www.responsibledatascience.org.

[2] Wil van der Aalst's presentation on *Responsible Data Science* at Process Mining Camp 2016: http://coda.fluxicon.com/assets/downloads/Camp/2016/8-Wil-van-der-Aalst.pdf

[3] Anne Rozinat & Frank van Geffen. Success Criteria for Process Mining: http://www.kdnuggets.com/2016/07/success-criteria-process-mining.html

[4] Anne Rozinat. Data Quality Problems in Process Mining and What To Do About Them — Part 11: Data Validation Session with Domain Expert: http://fluxicon.com/blog/2016/10/data-quality-problems-in-process-mining-and-what-to-do-about-them-part-11-data-validation-session-with-domain-expert/