fluxicon

DATA QUALITY CHECKLIST FOR PROCESS MINING

What exactly do you need to look for to determine whether the quality of your data is good enough for process mining? Here is a checklist with the questions that you can go through to assess the quality of your data.

Yes	Question	Description
☐	**No errors during import?**	The very first check is to make sure that there are no error messages when you import your data set. Error messages can have different root causes, such as Formatting Errors or Missing Timestamps. In some situations, you may want to deliberately induce errors to clean a data quality problem that would otherwise go unnoticed (see Missing Complete Timestamps For Ongoing Activities).
☐	**No gaps in the timeline?**	Verify that there are no unnaturally empty spots along the log timeline in your 'Events over time' chart (see Gaps in the timeline). This would indicate that you are missing a bunch of data.
☐	**Expected amount of data?**	Then, check whether the number of events and the number of cases that were imported correspond to the amount of data that you expected (see Unexpected amount of data).
☐	**Expected distribution of attribute values? No unexpected empty values?**	After checking the volume of the data, take a look at the attributes and their attribute statistics. See if the distribution seems right and if there are unexpected empty values (see Missing Attribute Values). Furthermore, inspect some example cases and verify that the attribute values are correct in their temporal context (see Missing Attribute History).
☐	**No cases with unexpected number of steps?**	If case IDs are overloaded or missing, events that belong to different cases may be grouped into one case. You need to clean such cases from your data set (see Missing Case IDs).
☐	**Expected timeframe? No unexpected long throughput times?**	You have requested the data set for a certain timeframe. Check the earliest and the latest timestamp to see if you have any Zero Timestamps (e.g., 1900, 1970 and 2999).
☐	**No unexpected ordering of sample cases? No unexpected flows in the process map?**	Wrong timestamps mess with the ordering relationships of your process and there can be many different reasons for why they are wrong. Read all of the following examples to know what you should look out for: Missing Activities, Missing Timestamps For Activity Repetitions, Wrong Timestamp Pattern Configuration, Same Timestamp Activities, Different Timestamp Granularities, Unwanted Parallelism, Wrong Timestamp Column Configuration, Recorded Timestamps Do Not Reflect Actual Time of Activities, and Different Clocks.

Yes	Question	Description
☐	**Data validation session with domain expert done?**	You will do all the checks 1-7 on this list for yourself at first. However, before you move to the analysis phase, it is really important that you also check the data quality with a domain expert (see Data Validation Session).
☐	**Documented all quality issues and data questions?**	Throughout the data validation process, write down all the problems, limitations, and questions that emerge as soon as you encounter them. You can download this worksheet as a starting point (see also Keeping an Overview About Your Project).
☐	**If you had to exclude data due to data quality problems, is the remaining data set still representative?**	Keep track which of your original process questions may be affected by the data quality issues that you found. Can certain questions not be answered, because the data is not good enough? Furthermore, be mindful of the amount of data that you remove during the data cleaning steps: After you have fixed all your data quality problems, compare the remaining amount of cases with the initial data set and decide whether this data basis is sufficient for your analysis.

Can you answer “Yes” to all of the points above? Then you can move to the next phase of preparing your analysis: Deal With Incomplete Cases.