Improving the Quality of Event Logs in the Construction Industry for Process Mining

L. Chen1[a], S. Kang2[a], S. Karimidorabati3[b], C. T. Haas4[a]

[a]Department of Civil and Environmental Engineering, University of Waterloo, Canada
[b]Gina Cody School of Engineering & Computer Science, Concordia University, Canada
E-mail: l364chen@uwaterloo.ca, s43kang@uwaterloo.ca, shahin.karimi@concordia.ca, chaas@uwaterloo.ca

Abstract –

Process mining is an emerging tool kit to discover, analyze, and improve workflows. In the construction industry, projects can be large and unique in terms of costs and durations. In such projects, commercial software that support analyses and decision making may be helpful. Process mining software allow construction companies to quickly discover benchmark workflows, conduct conformance checking, and analyze root causes. However, engineers who initially format the event logs, and users who use the event logs for analyses may not share the same knowledge or values. Due to the lack of understanding, knowledge, or experience in these respective domains, companies may not have documented the event logs in the most practical way. With the poorly structured event logs, the software may not be able to provide the most accurate analyses. Therefore, there is a need to pre-process event logs so as to improve their quality. This paper examines a case study on Engineering Change Request (ECR) in the construction industry to explain the importance of pre-processing the event logs before importing them into the commercial process mining software. For large complex projects, the improved quality of event logs will reduce confusion between engineers and analyzers and improve the accuracy of the analysis results produced by the process mining software.

Keywords –

Process Mining; Event Log; Pre-processing

1 Introduction

Process mining, which is a powerful tool in analyzing workflow and improving its efficiency, has been widely used in the academic fields of many different industries, such as healthcare [1] and banking [2]. Because of their standardized processes and the process mining software's ability to deal with big data, process mining has been successful in these industries. Disco, ProM, and Celonis are some example software that have been used in different research [3]–[5]. In the construction industry, the management of workflow makes the processes easier to monitor and control. Collecting and analyzing event data using data mining techniques allow people to track and validate the processes [6]. Process mining uses event logs to perform analyses on workflows. Event logs record information including Case IDs, Activities, Timestamps, and Attributes (such as people, costs, and locations). As process mining aims at utilizing information in event logs to discover, monitor, and improve processes, the quality of the data imported to the software will greatly affect the clarity and effectiveness of process mining [3].

After uploading the event logs, the commercial software creates a discovered model and conducts conformance checking, bottleneck analysis, and root cause analysis. These analyses provide meaningful insights regarding the efficiency of the existing workflow. However, when importing the event logs into the software, the software may have difficulties in analyzing the event logs, if they are poorly structured. Meanwhile, users may also have trouble in understanding the results. For example, the event logs could have ambiguous naming when multiple activities share the same activity name. Hence, there are needs to pre-process the event logs to minimize the ambiguity and enhance the understanding for the users and analyses from the software. This may include the formatting of the event log names, timestamps, and adding or deleting certain logs.

According to Suriadi et al. [3], a "high-quality" event log is defined as the one with minimal information loss, which is also valid in the context of the domain and for the analysis purpose. Resulting from manually recording the data, poor-quality event logs have missing data, incorrect data, imprecise data and irrelevant data. In this context, to clean up event logs means to address these issues by correcting the errors if possible. However, in this paper, the event logs are assumed to have all the information required, including activity names, timestamps, and attributes. A "high-quality" event log is thus defined as a structured data that minimizes the confusion and maximizes the ease and clarity for both software and users to analyze and interpret the workflow.

This paper focuses on pre-processing event logs to improve their quality both on their format and content. An event log including 58 cases of Engineering change requests (ECR) of a mega-construction project was used as a case study. The case study elaborated on the importance of the clarity and accuracy of the activity names and timestamps of event logs. Therefore, companies should be aware of pre-processing the event logs before manipulating them if they want to take full advantage of process mining.

2 Literature Review

Construction projects can happen over a long period of time due to large scope. The planning, design, construction, and maintenance stages can take years to happen and complete. In the meantime, construction projects are prone to delays [7]. High uncertainties due to the weather, resources or changes can result in delays. Recent studies have focused on the automation of change management in construction industry to minimize the delays caused by change management [8]. The processes of construction projects have been the area that the researchers and engineers consider as important and try to standardize [9], [10]. When time is the main concern, process mining may provide suggestions by analyzing the patterns from timestamp records. Despite the fact that process mining has been introduced to other industries to minimize the delay and improve the efficiency of the workflow, few companies have adopted process mining in the construction industry. Process mining has shown potential to help in managing construction projects. Due to the long timespan of capital projects, process mining techniques, such as bottleneck analysis and root cause analysis, could be applied to project workflows to find where delays usually occur and what their causes may be. The information could be used to improve the workflow of the latter project stages. The advantage of applying process mining techniques to the same project is that it will produce more accurate results where the time, location, and human factors are most likely to remain the same.

Process mining takes event logs to automatically obtain process models and check conformance. In this case, event logs are considered as construction information whose quality will be important [11]. Moreover, event logs are the main foundation data to identify bottlenecks and deviations, suggest improvements, and predict processing time. For process mining, event logs are the raw data that need to be dealt with to construct models and analysis. The quality of the data, both form and content, is critical in order for the process mining exercise to be successful [12]. Anomalies and impurities lead to higher cost and less benefit for data processing, potentially making it less applicable to the clients [13]. Therefore, pre-processing event logs to improve their quality before conducting a process mining analysis is a necessary task. However, this pre-stage to clean the data for process mining is often overlooked because the task is considered tedious. The proper handling of potential problems and challenges of event logs should be taken care of before moving on to recording logs [14], [15]. In addition, there are also studies to find patterns after event logs are recorded and diagnose the issues afterwards [16], [17]. This paper focuses on ways to pre-process event logs before conducting further analyses, preventing potential confusion and errors. This includes activity definitions and timestamp alignment.

3 Pre-processing Event Log Framework

Although event logs existed before, how to manage and use event logs were not thoroughly investigated. This paper discusses how to improve the quality of event logs to deliver the meanings and intentions better. While there may be many overlaps, there is no consensus on how to define an event log. Therefore, suggested methodology in this paper may not be the unique path to build the data. However, this methodology will be a stepping stone especially for the construction management domain where there are various participants and activities involved throughout a process. In this section, a general framework to obtain a high quality event log is presented. An event log is a combination of "Case ID", "Activity", and "Timestamp" as the main components. Thus, clarifying these three areas of information can be beneficial.

Especially when it comes to activity names, some information can be added or simplified. In order to pursue clarification, the first task is to differentiate two different activities that share one activity name. If this step is skipped, high confusion is expected once the event log is recorded. Usually, activity names include actions. However, in construction management processes, there are many tasks with the same actions but executed by different position performers. In other words, if only actions without performers are included, overlaps may exist which can create confusion. Therefore, in this study, the form of "Action + Performer" is recommended. This way, clarity can be achieved. Note that "performer" and "user" need to be distinguished. "Performer" refers to a participant who is involved in the process, here, the ECR process. "User" on the other hand, refers to the process analyzer.

Additional information can be provided through attributes. However, in this paper, Case ID, Activity, and Timestamp are the main interests. When naming the activities, the authors recommend to use the form of "Verb" + "by/to Noun" for consistency.

The next task would be regarding timestamps. Every case has a different path from start to end. For some cases, errors exist, and completion time information does not exist. For other cases, the case is not formally closed but ends abruptly. Such cases should be treated differently so that the users can extract as much information as possible. Timestamps that receive "NULL" for the completion time by the workflow software have several reasons. First, the activity happened instantly which means the start time and end time are the same. If an activity is informational (notifications) and is automatically generated by the software, this requires no further action from the participants who get it. Second, when certain activities are not completed by the designated performer, the ECR is sent to another performer or again to the original performer automatically. When this happens, the original performer's previous activity receives a "NULL" by most workflow software such as LANA, Celonis, etc. Third, if an activity is not completed by the user before the predefined due date, the activity receives a "NULL" as the timestamp. Fourth, when the action is performed by a group of performers and there exists at least one "NULL" for the completion time, the performer who got "NULL" will receive a timeout warning. However, performers who completed the action in time may also get a timeout warning. For the timeout warning activities, the performers may again receive "NULL". Fifth, there are abort cases which includes "NULL" for the completion time. A summary on the common problems in event logs and their solutions can be seen in Figure 1.

4 Case Study

A case study was conducted to show some ways to pre-process the event logs before uploading them into the process mining software. The data used in the case study was regarding 58 cases of the Engineering Change Request (ECR) in a mega construction project. Figure 2 summarizes the problems of the raw event data that the user attained.

The first problem observed was the duplicated names. As shown in (1) in Figure 2, the activity "Change Request Participants Verification" happened twice. However, the activities were performed by different performers. This can be inferred from the fact that one was followed by "Review (Engineer)" whereas the other one was followed by "Review (Participants)". Considering the fact that "participants" were also ambiguous, the activity "Change Request Participants Verification" should be further analyzed and re-named properly.

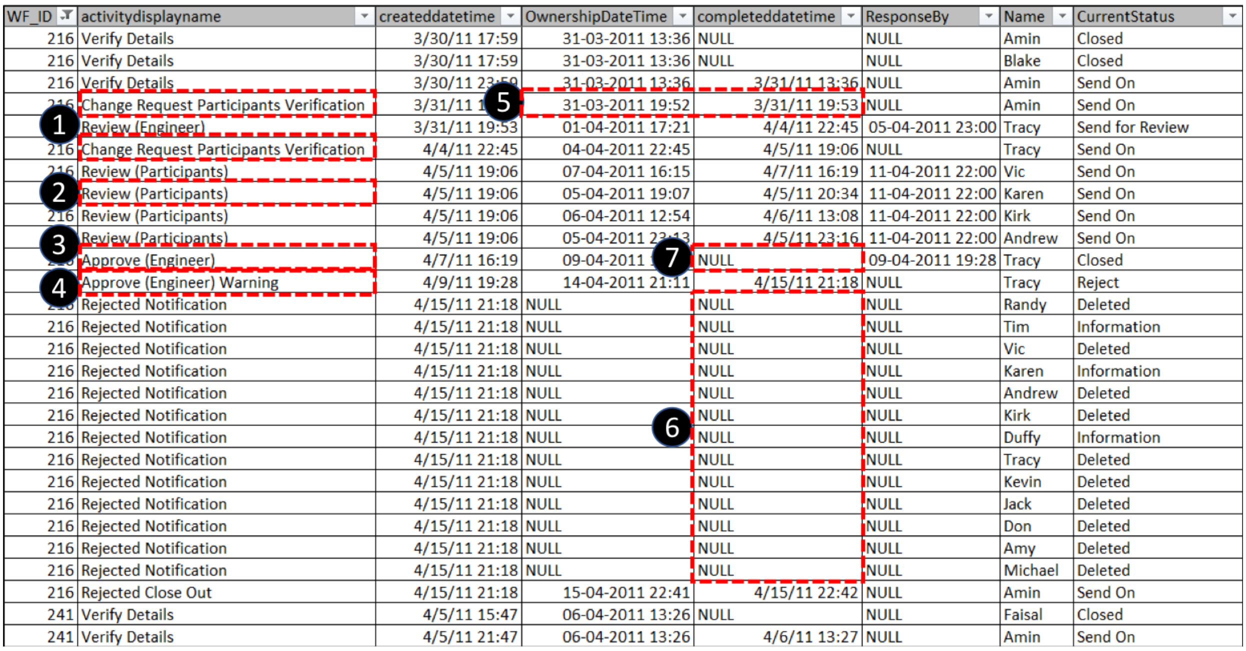
The second problem shown in (2) was regarding the "action (performer)" format. Some activities such as "Review (Participants)" and "Approve (Engineer)" included the action performer in the brackets. However, activities such as "Rejected Notification" and "Rejected Close Out" did not include the action performer nor recipient. This is the inconsistent format mentioned in the framework.

The third problem was ambiguous definition. For example, in the raw event log, there existed confusion in "Approve (Engineer)." It turned out that this can mean both approve and reject based on each case.

The fourth problem was missing critical information, such as reasons for warnings. The fifth problem was the timestamp format. Problems shown in (6) and (7) were both regarding the "NULL" completion time. The "NULL" timestamps happened due to a couple of reasons which will be elaborated in section 4.2.

No	Activity		No	Timestamp	
	Problem	**Solution**		**Problem**	**Solution**
1	- Duplicated names	- Rename into different names	5	- Inconsistent format	- Consistent format
2	- Inconsistent format	- Consistent format (Action + Performer)	6	- Missing values due to instant occasion	- Data imputation (e.g. completion time = start time)
3	- Ambiguous definitions	- Remove ambiguity	7	- Missing values due to overtime or group recipients	- Data imputation
4	- Missing information	- Additional information	8	- Missing values in abandoned activities	- Ignore the missing values

Figure 1. Pre-processing event log framework



WF_ID	activitydisplayname	createddatetime	OwnershipDateTime	completeddatetime	ResponseBy	Name	CurrentStatus
216	Verify Details	3/30/11 17:59	31-03-2011 13:36	NULL	NULL	Amin	Closed
216	Verify Details	3/30/11 17:59	31-03-2011 13:36	NULL	NULL	Blake	Closed
216	Verify Details	3/30/11 23:59	31-03-2011 13:36	3/31/11 13:36	NULL	Amin	Send On
216	Change Request Participants Verification	3/31/11 1	31-03-2011 19:52	3/31/11 19:53	NULL	Amin	Send On
	Review (Engineer)	3/31/11 19:53	01-04-2011 17:21	4/4/11 22:45	05-04-2011 23:00	Tracy	Send for Review
216	Change Request Participants Verification	4/4/11 22:45	04-04-2011 22:45	4/5/11 19:06	NULL	Tracy	Send On
216	Review (Participants)	4/5/11 19:06	07-04-2011 16:15	4/7/11 16:19	11-04-2011 22:00	Vic	Send On
	Review (Participants)	4/5/11 19:06	05-04-2011 19:07	4/5/11 20:34	11-04-2011 22:00	Karen	Send On
216	Review (Participants)	4/5/11 19:06	06-04-2011 12:54	4/6/11 13:08	11-04-2011 22:00	Kirk	Send On
	Review (Participants)	4/5/11 19:06	05-04-2011 23:13	4/5/11 23:16	11-04-2011 22:00	Andrew	Send On
216	Approve (Engineer)	4/7/11 16:19	09-04-2011 1	NULL	09-04-2011 19:28	Tracy	Closed
	Approve (Engineer) Warning	4/9/11 19:28	14-04-2011 21:11	4/15/11 21:18	NULL	Tracy	Reject
216	Rejected Notification	4/15/11 21:18	NULL	NULL	NULL	Randy	Deleted
216	Rejected Notification	4/15/11 21:18	NULL	NULL	NULL	Tim	Information
216	Rejected Notification	4/15/11 21:18	NULL	NULL	NULL	Vic	Deleted
216	Rejected Notification	4/15/11 21:18	NULL	NULL	NULL	Karen	Information
216	Rejected Notification	4/15/11 21:18	NULL	NULL	NULL	Andrew	Deleted
216	Rejected Notification	4/15/11 21:18	NULL	NULL	NULL	Kirk	Deleted
216	Rejected Notification	4/15/11 21:18	NULL	NULL	NULL	Duffy	Information
216	Rejected Notification	4/15/11 21:18	NULL	NULL	NULL	Tracy	Deleted
216	Rejected Notification	4/15/11 21:18	NULL	NULL	NULL	Kevin	Deleted
216	Rejected Notification	4/15/11 21:18	NULL	NULL	NULL	Jack	Deleted
216	Rejected Notification	4/15/11 21:18	NULL	NULL	NULL	Don	Deleted
216	Rejected Notification	4/15/11 21:18	NULL	NULL	NULL	Amy	Deleted
216	Rejected Notification	4/15/11 21:18	NULL	NULL	NULL	Michael	Deleted
216	Rejected Close Out	4/15/11 21:18	15-04-2011 22:41	4/15/11 22:42	NULL	Amin	Send On
241	Verify Details	4/5/11 15:47	06-04-2011 13:26	NULL	NULL	Faisal	Closed
241	Verify Details	4/5/11 21:47	06-04-2011 13:26	4/6/11 13:27	NULL	Amin	Send On

Figure 2. Raw event logs and their problems

4.1 Activity Names

Activity names are one of the most crucial components in event logs as they tell users what happened. Therefore, it is important to name the activity names properly and clearly. This section talks about some naming issues encountered during the process mining analysis of the case study and how they were solved. These problems include duplicated names, ambiguous naming and missing information, which had caused confusion.

4.1.1 Duplicated Names

The overview of the raw data can be seen in Figure 2. From the raw data, it can be seen that in Case 216, the activity "Change Request Participants Verification" was followed by two different activities, "Review (Engineer)" and "Review (Participants)". This proves the fact that the "participants" did refer to different groups of performers. When analyzing the processes using software, the activity "Change Request Participants Verification" caused confusion as they were considered as one activity due to the same activity name.

Based on their following activities, "Change Request Participants Verification" was renamed as "Change Request Participants Verification (Engineer)" and "Change Request Participants Verification (Participants)" respectively. "Change Request Participants Verification (Engineer)" was later renamed as "Verify Senior Engineer by Coordinator" whereas "Change Request Participants Verification (Participants)" was later renamed as "Verify Engineer by Senior Engineer."

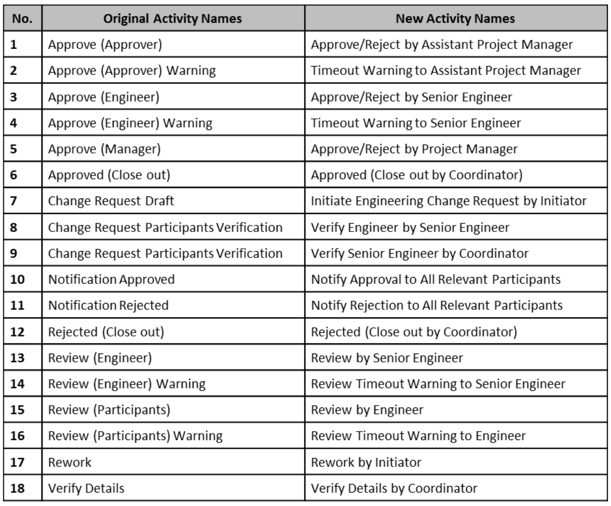
4.1.2 Activity Name Format

The data set has eighteen different activities in total as shown in Figure 3, referred as original activity names. By observation, there are four problems regarding the naming of the event logs. Firstly, the activity names started with different parts of speech. Majority of the activities started with verbs. However, activities such as "Notification Approved" and "Notification Rejected" started with nouns. After consideration, all the activities were decided to start with verbs for they indicated the actions of each activity clearly.

Secondly, some activity names included the performers / recipients involved, but some did not. For example, the activity "Approve (Engineer)" clearly indicated that the performer who approved the change was the engineer. However, the activity "Verify Detail" did not specify who was the performer that had verified the detail.

Thirdly, some activity names were ambiguous on the individuals who performed or received the action. For example, the terms "Approver" and "Participants" are ambiguous pronouns and not proper nouns that identify real roles or positions in the construction project, which also caused confusion.

Ideally, the activity names should be unique, clear and straightforward. By looking at the activity names, users should be able to know what the action was, who performed the action, and whom the action affected. To make the action clear, the activities have been structured to start with verbs. For example, the name "Notification Approved" means to send out the notification of approval to relevant recipients, although it can be interpreted as the notification is approved to be sent out. To avoid the confusion, “Notification Approved” has been renamed as “Notify All Relevant Participants of the Approval”.

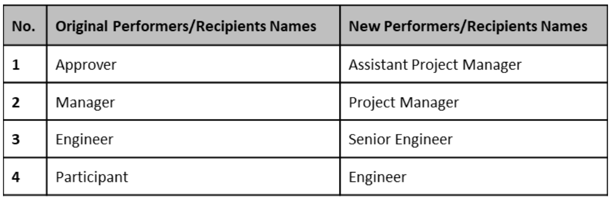


No.	Original Activity Names	New Activity Names
1	Approve (Approver)	Approve/Reject by Assistant Project Manager
2	Approve (Approver) Warning	Timeout Warning to Assistant Project Manager
3	Approve (Engineer)	Approve/Reject by Senior Engineer
4	Approve (Engineer) Warning	Timeout Warning to Senior Engineer
5	Approve (Manager)	Approve/Reject by Project Manager
6	Approved (Close out)	Approved (Close out by Coordinator)
7	Change Request Draft	Initiate Engineering Change Request by Initiator
8	Change Request Participants Verification	Verify Engineer by Senior Engineer
9	Change Request Participants Verification	Verify Senior Engineer by Coordinator
10	Notification Approved	Notify Approval to All Relevant Participants
11	Notification Rejected	Notify Rejection to All Relevant Participants
12	Rejected (Close out)	Rejected (Close out by Coordinator)
13	Review (Engineer)	Review by Senior Engineer
14	Review (Engineer) Warning	Review Timeout Warning to Senior Engineer
15	Review (Participants)	Review by Engineer
16	Review (Participants) Warning	Review Timeout Warning to Engineer
17	Rework	Rework by Initiator
18	Verify Details	Verify Details by Coordinator

Figure 3. Original activity names vs new activity names

To make the action performer and recipient clear, there was an investigation on the hierarchy of the positions and roles involved. Firstly, the “Approver” is not a real position title in the construction company. It is a temporary role, and in this case, it anonymizes the person responsible and results in ambiguous information. As approvers approve and reject the engineering change request, they were identified as the assistant project managers who are at the higher position in the company with the authority to approve or reject the change requests. Note that the names of positions can be flexible, e.g. project director can be added as approver. “Participants” who reviewed the change requests were interpreted as engineers whereas “Engineer” who verified “Participants” were thus inferred as senior engineers. The summary of the original and the new performers’/recipients’ names has been shown in the Figure 4.

Moreover, the activity “Approve (Approver)” could mean that Approver approved the change request, or it could also mean that Approver’s action or decision was approved. To clearly differentiate the action performer from the action recipient, action performers were added to the activity names as “by someone” whereas action recipients were added to the activity names as “to someone”. Realizing that Approver was the one who approved the change request, the activity “Approve (Approver)” was renamed as “Approve by Approver” which was eventually renamed as “Approve/Reject by Assistant Project Manager” as shown in Figure 3.

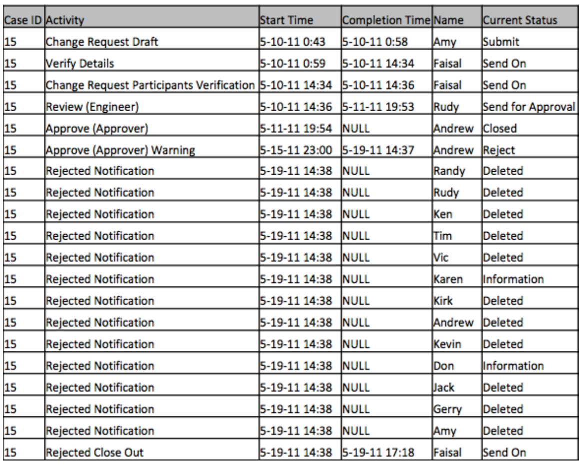


No.	Original Performers/Recipients Names	New Performers/Recipients Names
1	Approver	Assistant Project Manager
2	Manager	Project Manager
3	Engineer	Senior Engineer
4	Participant	Engineer

Figure 4. Original vs. new performers’ / recipients’ names

4.1.3 Ambiguous Activity Names

After having carefully investigated the event logs, it could be seen that some activities were not defined properly. For example, by looking at the activity name “Approve (Approver)”, users might interpret it as such that the approver has approved the engineering change request. However, in the case shown in the Figure 5, “Approve (Approver)” was followed by “Rejected Notification”. Hence, it could be induced that the activity “Approve (Approver)” does not necessarily mean approving. Instead, it refers to the process of making the decision on whether to approve or reject the change request. Therefore, to minimize the confusion, the activity “Approve (Approver)” was renamed as “Approve / Reject by Approver”, which was later renamed as “Approve / Reject by Project Manager”. Similarly, “Approve (Engineer)” was renamed as “Approve / Reject by Senior Engineer” and “Approve (Manager)” was renamed as “Approve / Reject by Assistant Project Manager.” The summary of the original activity names and the new activity names can be seen in the Figure 3.

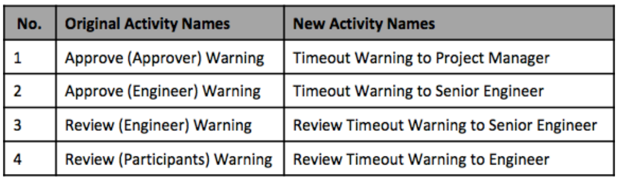


Case ID	Activity	Start Time	Completion Time	Name	Current Status
15	Change Request Draft	5-10-11 0:43	5-10-11 0:58	Amy	Submit
15	Verify Details	5-10-11 0:59	5-10-11 14:34	Faisal	Send On
15	Change Request Participants Verification	5-10-11 14:34	5-10-11 14:36	Faisal	Send On
15	Review (Engineer)	5-10-11 14:36	5-11-11 19:53	Rudy	Send for Approval
15	Approve (Approver)	5-11-11 19:54	NULL	Andrew	Closed
15	Approve (Approver) Warning	5-15-11 23:00	5-19-11 14:37	Andrew	Reject
15	Rejected Notification	5-19-11 14:38	NULL	Randy	Deleted
15	Rejected Notification	5-19-11 14:38	NULL	Rudy	Deleted
15	Rejected Notification	5-19-11 14:38	NULL	Ken	Deleted
15	Rejected Notification	5-19-11 14:38	NULL	Tim	Deleted
15	Rejected Notification	5-19-11 14:38	NULL	Vic	Deleted
15	Rejected Notification	5-19-11 14:38	NULL	Karen	Information
15	Rejected Notification	5-19-11 14:38	NULL	Kirk	Deleted
15	Rejected Notification	5-19-11 14:38	NULL	Andrew	Deleted
15	Rejected Notification	5-19-11 14:38	NULL	Kevin	Deleted
15	Rejected Notification	5-19-11 14:38	NULL	Don	Information
15	Rejected Notification	5-19-11 14:38	NULL	Jack	Deleted
15	Rejected Notification	5-19-11 14:38	NULL	Gerry	Deleted
15	Rejected Notification	5-19-11 14:38	NULL	Amy	Deleted
15	Rejected Close Out	5-19-11 14:38	5-19-11 17:18	Faisal	Send On

Figure 5. Event logs regarding “Approve (Approver)”

4.1.4 Missing Information

When investigating the event logs, there was a set of activities named with warnings. The warning events are missing values as the activity names did not indicate which type of warnings they were. For example, these warnings could mean exceedance of costs, resources or time. The warning activities can be seen in Figure 6. In all the cases, the warning activities were followed by the normal activities without warnings. For example, the activities "Review (Engineer)" were followed by "Review (Engineer) Warning". Keeping such a relationship in mind, it was found that the warning activities happened when the participants failed to respond before the deadline. Therefore, the warnings were identified as the timeout warnings since the performers did not respond in time.



No.	Original Activity Names	New Activity Names
1	Approve (Approver) Warning	Timeout Warning to Project Manager
2	Approve (Engineer) Warning	Timeout Warning to Senior Engineer
3	Review (Engineer) Warning	Review Timeout Warning to Senior Engineer
4	Review (Participants) Warning	Review Timeout Warning to Engineer

Figure 6. Timeout warning activities

4.2 Timestamps

Timestamps are another crucial factor in event logs, which usually indicates the start time and the completion time. In this case study, timestamps included created time, ownership time and completed time as shown in Figure 2. The ownership time refers to the opening time of the tasks by the action performers. However, as there were cases when the performers had already started the work but did not open the task, the ownership time was not the actual start time. The created time was thus used as the start time as that was when the performers are supposed to begin to work on the activities.

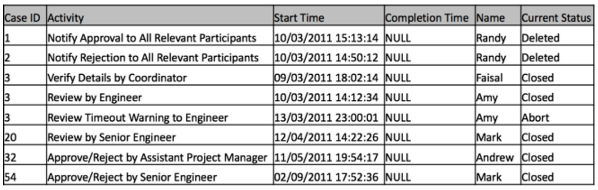
There are two main problems for timestamps. One is the inconsistent formatting, and the other one is the "NULL" timestamps. The "NULL" timestamps happened due to three reasons: the activity happened instantly, the activity was aborted, or the performer did not manage to complete the activity before the due time and data. These cases will be further explained in the following sections.

4.2.1 Inconsistent Timestamp Format

After fixing the activity names, the timestamps were also investigated carefully. In Figure 2, it can be seen that the timestamps were documented in different format. The "createddatetime" (start time) in and the "completeddatetime" (completion time) adopted the format of "MM-dd-yy HH:mm". In contrast, the "OwndershipDateTime" (opening time) and the "ResponseBy" (deadline) adopted the format of "yyyy-MM-dd HH:mm". The inconsistency in the format caused confusion for both users and process mining software. For example, if the date is "April 10th, 2011", it will be written as "04-10-11" which could also possibly mean "October 4th, 2011" without a clear definition. To be consistent and avoid the confusion, all the timestamps were reformatted as "dd/MM/yyyy HH:mm:ss."

4.2.2 "NULL" Cases

In the event logs, some activities were triggered and completed instantly. For example, the activity "Notify Approval to All Relevant Participants" referred to the action of sending notification of approval to performers involved. This activity is merely an automated machine-based action that takes place instantly upon the approval or rejection of the ECR by an authorized performer such as Senior Engineer or Assistant Project Manager. With no further action required from the participants, "Notify Approval to All Relevant Participants" had a completion time of "NULL" as shown in Figure 7. The "NULL" timestamps were assigned values using data imputation. The completion time for "Notify Approval to All Relevant Participants" and "Notify Rejection to All Relevant Participants" were therefore changed to the same time as their start time.

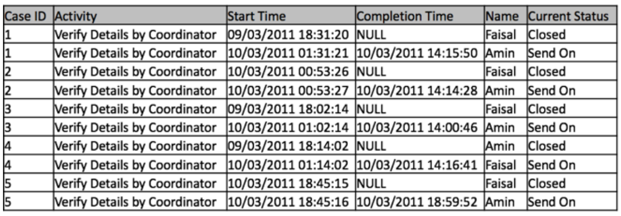


Case ID	Activity	Start Time	Completion Time	Name	Current Status
1	Notify Approval to All Relevant Participants	10/03/2011 15:13:14	NULL	Randy	Deleted
2	Notify Rejection to All Relevant Participants	10/03/2011 14:50:12	NULL	Randy	Deleted
3	Verify Details by Coordinator	09/03/2011 18:02:14	NULL	Faisal	Closed
3	Review by Engineer	10/03/2011 14:12:34	NULL	Amy	Closed
3	Review Timeout Warning to Engineer	13/03/2011 23:00:01	NULL	Amy	Abort
20	Review by Senior Engineer	12/04/2011 14:22:26	NULL	Mark	Closed
32	Approve/Reject by Assistant Project Manager	11/05/2011 19:54:17	NULL	Andrew	Closed
54	Approve/Reject by Senior Engineer	02/09/2011 17:52:36	NULL	Mark	Closed

Figure 7. Activities with "NULL" completion time

"NULL" timestamps were also observed for other activities, which have been listed in Figure 7. By looking into the attributes, it was found that these activities all had the current status of either "Closed" or "Abort". The status of "Abort" indicates that the change request was cancelled and no longer proceeded. As a result, the completion time for those "Abort" activities were left blank, for they did not have a completion time.

"Verify Details by Coordinator" was one of the activities with the greatest number of "NULL" timestamps for their completion time. These cases were usually repeated again with a non-NULL completion time. From Figure 8, it was suggested that the "NULL" timestamp happened when the original performer was not able to complete the task in time. As a result, the original performer appears to have a "NULL" completion time, and another performer, authorized as the coordinator, may complete the task. The current status was thus “Closed” instead of “Send on.”



Case ID	Activity	Start Time	Completion Time	Name	Current Status
1	Verify Details by Coordinator	09/03/2011 18:31:20	NULL	Faisal	Closed
1	Verify Details by Coordinator	10/03/2011 01:31:21	10/03/2011 14:15:50	Amin	Send On
2	Verify Details by Coordinator	10/03/2011 00:53:26	NULL	Faisal	Closed
2	Verify Details by Coordinator	10/03/2011 00:53:27	10/03/2011 14:14:28	Amin	Send On
3	Verify Details by Coordinator	09/03/2011 18:02:14	NULL	Faisal	Closed
3	Verify Details by Coordinator	10/03/2011 01:02:14	10/03/2011 14:00:46	Amin	Send On
4	Verify Details by Coordinator	09/03/2011 18:14:02	NULL	Amin	Closed
4	Verify Details by Coordinator	10/03/2011 01:14:02	10/03/2011 14:16:41	Faisal	Send On
5	Verify Details by Coordinator	10/03/2011 18:45:15	NULL	Faisal	Closed
5	Verify Details by Coordinator	10/03/2011 18:45:16	10/03/2011 18:59:52	Amin	Send On

Figure 8. Activity “Verify Details by Coordinator”

In the case study, there were also timeout warning activities which have been summarized in Figure 6 in Section 4.1.4. The timeout warning activities happened when the performers did not manage to complete the activities by deadline. In this case, the completion time for the normal activity was recorded as “NULL” and the normal activity was followed by a timeout warning. The solution to the timeout activities was to add up the duration of the normal activity and the timeout activity.

5 Conclusion

Process mining has great potential in the construction industry as it allows people to find out the activities causing delays, thus improving the efficiency. To make sure that process mining provides the best result, it is important to pre-process the event logs. Some of the issues with activity names are duplicated naming, ambiguous naming and missing information. These problems should be identified and fixed before the event logs are used for further analysis. Timestamps are also another critical feature of event logs. Some of the issues with timestamps are confusing format, missing values, i.e. “NULL” cases, and activities that are not completed by the deadline. The discovery of the timeout activities had triggered thoughts on conformance checking. In process mining, conformance checking is conducted between the model discovered using event logs and the target model. The difference between the discovered model and the target model, i.e. missing or inserted activities, can be considered as non-conformance. In the real world, projects usually have a deadline by which they are supposed to be completed. Hence, if the performers are not able to finish their tasks before the predefined deadlines, it should be considered as non-conformance too. This paper mainly focused on the first step of process mining, which is to pre-process the event logs extracted from the construction projects. The next step will be to use the pre-processed event logs to conduct analyses such as conformance checking, bottleneck analysis, and root cause analysis. As every industry has workflows and processes, this paper can be further generalized to all industries, not limited to the construction industry, on how to pre-process and improve the quality of event logs. With event log quality improvement, there will be less confusion for both engineers who created the event logs and users, which results in better accuracy in process mining analyses.

References

[1] E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro, “Process mining in healthcare: A literature review,” *J. Biomed. Inform.*, vol. 61, pp. 224–236, Jun. 2016.

[2] C. Moreira, E. Haven, S. Sozzo, and A. Wichert, “Process mining with real world financial loan applications: Improving inference on incomplete event logs,” *PLOS ONE*, vol. 13, no. 12, p. e0207806, Dec. 2018.

[3] S. Suriadi, R. Andrews, A. H. M. ter Hofstede, and M. T. Wynn, “Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs,” *Inf. Syst.*, vol. 64, pp. 132–150, Mar. 2017.

[4] W. van der Aalst, “Spreadsheets for business process management: Using process mining to deal with ‘events’ rather than ‘numbers’?,” *Bus. Process Manag. J.*, vol. 24, no. 1, pp. 105–127, Jan. 2018.

[5] M. Kebede and M. Dumas, “Comparative Evaluation of Process Mining Tools,” 2015.

[6] W. van der Aalst *et al.*, “Process Mining Manifesto,” in *Business Process Management Workshops*, 2012, pp. 169–194.

[7] Lo Tommy Y., Fung Ivan W., and Tung Karen C., “Construction Delays in Hong Kong Civil Engineering Projects,” *J. Constr. Eng. Manag.*, vol. 132, no. 6, pp. 636–649, Jun. 2006.

[8] S. Karimidorabati, C. T. Haas, and J. Gray, “Evaluation of automation levels for construction change management,” *Eng. Constr. Archit. Manag.*, vol. 23, no. 5, pp. 554–570, Sep. 2016.

[9] B. Golzarpoor, C. T. Haas, D. Rayside, S. Kang, and M. Weston, “Improving construction industry process interoperability with Industry Foundation Processes (IFP),” *Adv. Eng. Inform.*, vol. 38, pp. 555–568, Oct. 2018.

[10] B. Golzarpoor, C. T. Haas, and D. Rayside, “Improving process conformance with Industry Foundation Processes (IFP),” *Adv. Eng. Inform.*, vol. 30, no. 2, pp. 143–156, Apr. 2016.

[11] A. J. Antony Chettupuzha and C. T. Haas, “Algorithm for Determining the Criticality of Documents within a Construction Information System,” *J. Comput. Civ. Eng.*, vol. 30, no. 3, p. 04015039, May 2016.

[12] R. P. J. C. Bose, R. S. Mans, and V. D. W. M.P. Aalst, “Wanna improve process mining results? : it’s high time we consider data quality issues seriously,” *2013 IEEE Symp. Comput. Intell. Data Min. CIDM13 Singap. April 16-19 2013*, pp. 127–134, 2013.

[13] H. Mueller and J.-C. Freytag, “Problems , Methods , and Challenges in Comprehensive Data Cleansing,” 2005.

[14] R. Conforti, M. La Rosa, and A. H. M. ter Hofstede, “Noise Filtering of Process Execution Logs based on Outliers Detection,” Report, 2015.

[15] R. Sarno, W. A. Wibowo, K. Kartini, Y. Amelia, and K. Rossa, “Determining Process Model Using Time-Based Process Mining and Control-Flow Pattern,” *TELKOMNIKA Telecommun. Comput. Electron. Control*, vol. 14, no. 1, pp. 349–360, Mar. 2016.

[16] C. Fernandez-Llatas, J. L. Bayo, A. Martinez-Romero, J. M. Benedí, and V. Traver, “Interactive pattern recognition in cardiovascular disease management. A process mining approach,” in *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2016, pp. 348–351.

[17] C. Fernandez-Llatas, A. Lizondo, E. Monton, J.-M. Benedi, and V. Traver, “Process Mining Methodology for Health Process Tracking Using Real-Time Indoor Location Systems,” *Sensors*, vol. 15, no. 12, pp. 29821–29840, Dec. 2015.