A Month's Worth of Labelled Active Window Tracking Data

Iris Beerepoot[1]

[1]Utrecht University, Heidelberglaan 8, 3584 CS Utrecht, The Netherlands

Abstract

Recent process mining techniques provide interesting new ways to uncover and comprehend complex work practices within organisations. The efficacy of process mining, however, is contingent upon the accessibility and quality of event logs. This paper introduces and describes a publicly-available dataset containing labelled Active Window Tracking data, capturing my app usage and active screen titles over the course of four weeks. It aims to support diverse studies: classifying activities from window titles, identifying action patterns, and developing new visualisations for detailed process data. By making this resource available, I aim to encourage the development of new process mining techniques that provide detailed insights into work practices.

Keywords

Event data, Active Window Tracking, UI logs, cross-system data

1. Introduction

Process mining has evolved into a means to discover and understand complex work practices of employees in organisations [1]. However, the quality of the results heavily relies on the availability of event logs. Most process mining studies draw on event data from single work systems [2], which leads to incomplete representations of the breadth of work that has taken place. In [1], we proposed the use of so-called Active Window Tracking (AWT) data for mining work practices, and outlined the opportunities of using this type of detailed event data. AWT records the apps that a worker uses and the title of the screen that is active at a certain point in time, providing data that sits in between UI logs and traditional single-system event data.

The window titles and apps provide a very detailed view on how the worker performs certain tasks, which may be interesting in itself for some types of analyses. However, for other analyses there may be a need to abstract window titles into higher-level activities. To support such analyses and to encourage the development of novel abstraction techniques, this resource presents a month's worth of (manually) labelled and pseudonymised AWT data. The data is publicly available on Github (https://github.com/project-pivot/labelled-awt-data). The following sections introduce the details of the resource, provide a preliminary analysis, and outline its possible uses.

Proceedings of the Best BPM Dissertation Award, Doctoral Consortium, and Demonstrations & Resources Forum co-located with 22nd International Conference on Business Process Management (BPM 2024), Krakow, Poland, September 1st to 6th, 2024.

i.m.beerepoot@uu.nl (I. Beerepoot)
http://irisbeerepoot.com/ (I. Beerepoot)
0000-0002-6301-9329 (I. Beerepoot)

2. Description of the Resource

The resource contains a set of files, all of which relate to one month's worth of AWT data. For the recording, I used a tool called Tockler (http://maygo.github.io/tockler/) which has been running on my computer since December 2022. It logs application title, window title, start time and end time. Contrary to many time tracking tools which aggregate the total time spent in a certain app or window, rendering it unusable for process mining, Tockler keeps the timestamps of the individual window titles that were active at a certain point in time.

The repository contains a labelled subset of the resulting data, namely from March 6 to April 2, 2023. The labelling was done locally in Excel, by exporting the data from Tockler and adding two columns to the log, one for the corresponding activity and one for the case. I deductively selected activities from the University Job Classification system used by Dutch universities (https://edu.nl/hkdr4). Examples of such activities are 'Assessing exams and giving marks', and 'Conducting research'. When I could not fit the behaviour within the existing activities, I created a new activity. Examples of added activities include 'Communicating about events', 'Planning teaching activities', and 'Reviewing journal and conference papers'. I selected the cases inductively, e.g., courses that I taught, students that I supervised, research papers that I worked on, events that I organised, etc. This procedure is also described in [1].

Table 1 provides an overview of the datasets in the repository. The first file contains the full pseudonymised version of the labelled data, where I replaced names and other sensitive information with placeholders. The result of this can be found in the file entitled `awt_data_1_pseudonymized`. A snippet is provided in Table 2. The remaining files in the data folder contain versions in which I applied some processing. The Python notebook contains details on the steps that I took. In `awt_data_2_merged_titles`, I merged all sequential events with the same activity label, resulting in an abstracted event log with significantly less events. In `awt_data_3_added_duration`, I added a Duration column to calculate the duration between the start and end time, and in `awt_data_4_added_case_type` I added a case type attribute.

Depending on the use case, you might want to work with the detailed window titles in `awt_data_1_pseudonymized` or the final abstracted data in `awt_data_4_added_case_type`. If you are interested in, e.g., a technique that automatically recognises higher-level tasks performed based on (a set of) window titles, you would want to have a look at `awt_data_1_pseudonymized`. If you want to explore the data with process mining techniques immediately, you better check out `awt_data_4_added_case_type`.

Table 1
Description of the datasets.

File name	Number of events	Description
awt_data_1_pseudonymized.csv	10,066	Full pseudonymised dataset (individual window titles labelled in Excel)
awt_data_2_merged_titles.csv	1,227	Same as previous but with sequential events with same label merged (see preprocessing notebook)
awt_data_3_added_duration.csv	1,227	Same as previous but with duration column (see preprocessing notebook)
awt_data_4_added_case_type.csv	1,227	Same as previous but with case type added (manually)

Table 2
Snippet of the labelled data (corresponding to awt_data_1_pseudonymized.csv).

App	Title	Begin	End	Activity	Case
Windows Explorer	20230309 Gastcollege bachelor NWI	8-03-23 14:14:40	8-03-23 14:14:46	Encouraging and giving lectures	Guest lecture NWI bachelor
Microsoft PowerPoint	20230309 gastcollege NWI - PowerPoint	8-03-23 14:14:46	8-03-23 14:14:49	Encouraging and giving lectures	Guest lecture NWI bachelor
Microsoft PowerPoint	20230309 gastcollege NWI - PowerPoint Presenter View	8-03-23 14:14:49	8-03-23 14:16:01	Encouraging and giving lectures	Guest lecture NWI bachelor
Microsoft PowerPoint	20230309 gastcollege NWI - PowerPoint	8-03-23 14:16:01	8-03-23 14:16:04	Encouraging and giving lectures	Guest lecture NWI bachelor
Windows Explorer	20230309 Gastcollege bachelor NWI	8-03-23 14:16:13	8-03-23 14:16:16	Encouraging and giving lectures	Guest lecture NWI bachelor
Adobe Acrobat	***name10*** - Research proposal.pdf - Adobe Acrobat Pro (32-bit)	8-03-23 14:18:28	8-03-23 14:18:34	Assessing the students' assignments and submitting the assessment to the Board of Examiners	***name10***
Adobe Acrobat	***name10*** - Research proposal.pdf - Adobe Acrobat Pro (32-bit)	8-03-23 14:19:14	8-03-23 14:19:26	Assessing the students' assignments and submitting the assessment to the Board of Examiners	***name10***

3. Preliminary Analysis

As an initial analysis, let us examine `awt_data_4_added_case_type.csv`. It contains 76 hours worth of labelled AWT data over the course of the four weeks. As can be seen from Figure 1, the majority of this time is spent on research projects, but a significant time was also spent on other case types such as courses, conferences, presentations, and more. The time spent on research projects and courses is reflected in the overview of duration per activity (Figure 2). It shows that most of my time was spent on 'conducting research', with 'encouraging and giving lectures' and 'preparing and providing teaching sessions for students' completing the top-3.

In order to demonstrate the readiness of this data for applying process mining, imagine that we are interested in the activities related to the different academic conferences. We load the preprocessed file `awt_data_4_added_case_type.csv` into Fluxicon's Disco (https://fluxicon.com/disco/). We map the columns as follows:

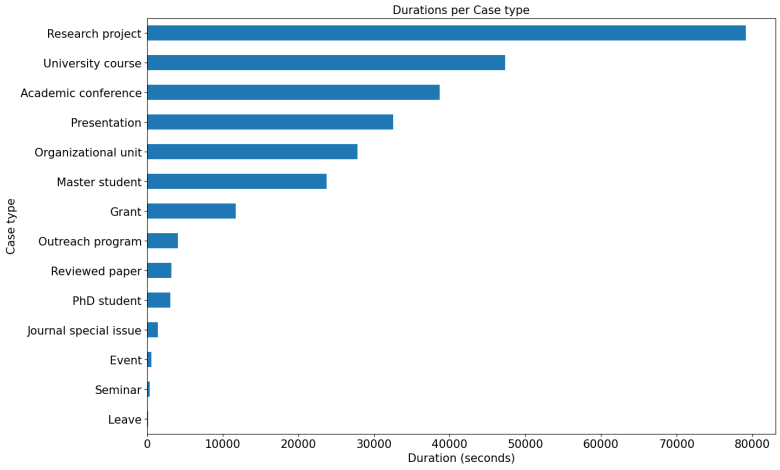


Figure 1: Duration per case type.

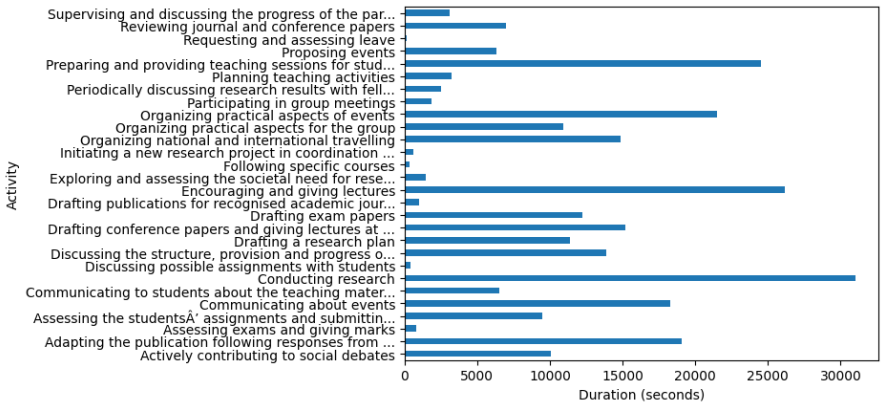
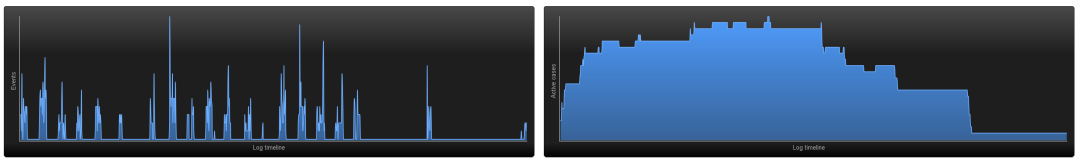


Figure 2: Duration per activity.

- 'Begin' ⟶ Timestamp (Pattern: 'yyyy/MM/dd HH:mm:ss')
- 'End' ⟶ Timestamp (Pattern: 'yyyy/MM/dd HH:mm:ss')
- 'Activity' ⟶ Activity
- 'Case' ⟶ Case ID
- Other attributes: 'Case type'

From the global statistics, the events over time (Figure 3a) and the active cases over time (Figure 3b) are shown. In the former, one can easily spot the working days across the week. In the final week, I was on holiday, doing only some work on the Tuesday, which is reflect in both figures.

Now that we have a general idea of the data, we can discover the process. In this case, we are interested in the work that took place around academic conferences. As such, we filter on 'Case type' ⟶ 'Academic conference'. Figure 4 shows the resulting process map, with activities and paths set to 100%. During the month of March 2023, I was involved in activities related to four cases, i.e., academic conferences: BPM, ECIS, ICIS, and RCIS. The majority of work was done for the BPM conference, where I was part of the organising team and program committee. For ECIS, I was mostly involved in organising a workshop and communicating about the workshop, as well as organising my travelling. For ICIS, there was only a brief activity related to reviewing, and the time spent on the RCIS case revolved around yet more travel organisation. Note that these cases are unfinished; work on these four conferences has taken place before and after these



(a) Events over time.

(b) Active cases over time.

Figure 3: Events and active cases over time.

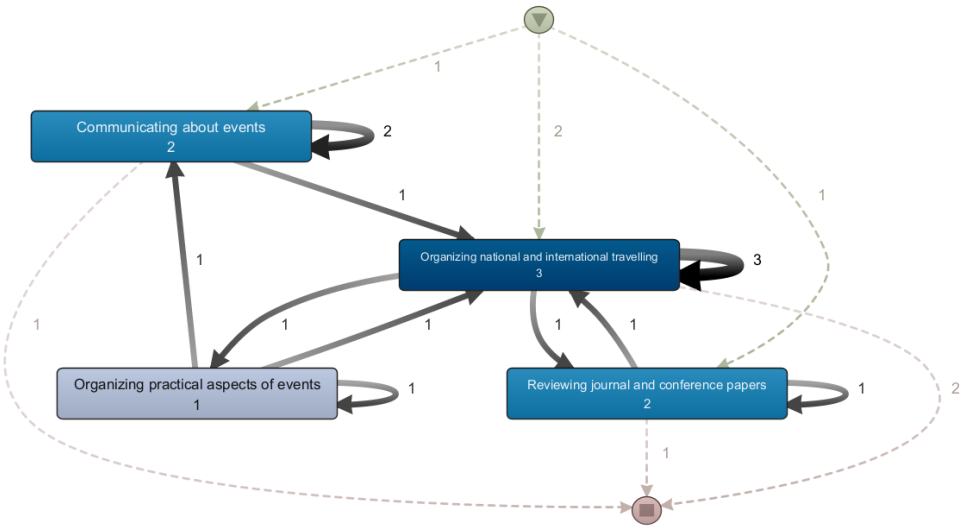


Figure 4: Process map of the activities related to 'Case type' → 'Academic conference'.

four weeks and the same is true for many of the other cases. In order to enable examination of finished cases that span several months without having to manually label all this data, it is vital that we develop and apply techniques to (semi-)automatically recognise higher-level activities and cases from the long-term AWT data, which is ongoing work.

4. Possible Uses

This resource may be used in different ways: (1) to develop techniques that classify (a set of) window titles into higher-level activities, (2) to identify patterns of action in terms of repeated sequences of windows or activities, and (3) to develop new visualisations for detailed process data. Even better, it may spark entirely new ideas about how individuals organise their work, which I hope it does. Feel free to reach out to me with questions or requests to verify insights.

References

[1] I. Beerepoot, D. Barenholz, S. Beekhuis, J. Gulden, S. Lee, X. Lu, S. Overbeek, I. Van De Weerd, J. M. Van Der Werf, H. A. Reijers, A window of opportunity: Active window tracking for mining work practices, in: 2023 5th International Conference on Process Mining (ICPM), IEEE, 2023, pp. 57–64.

[2] M. Thiede, D. Fuerstenau, A. P. Bezerra Barquet, How is process mining technology used by organizations? a systematic literature review of empirical studies, Business Process Management Journal 24 (2018) 900–922.