Open Opal Dataset Documentation

Tap On and Tap off Opal Data

Contents

Author: Transport for NSW

Date: July 2017

Version: 2

1 Introduction

1.1 What is Opal?

Opal is the smartcard ticketing system which is used on Transport for NSW's public transport. Opal cards are smartcard tickets that are used to pay for travel on public transport. The Opal network covers train, bus, ferry and light rail services in Sydney, the Blue Mountains, Central Coast, Hunter, Illawarra and Southern Highlands.

1.2 Opal Data – Tap on and Tap off

Transport for NSW customers can tap on (depart) and tap off (arrive) at a destination. Tap on and tap off data can be from re-useable stored value cards or from single use tickets.

For the Manly Ferry, customers only tap on (depart), i.e. there is no tap off at the arrival destination.

2 Data overview

The Open Opal dataset contains 6 CSV files covering 28 days across the four (4) public transport modes.

The CSV files (one for each week) are:

- Tap on and off time (single attribute data set)
- Tap on and off location (single attribute data set)
- Tap on and off time and location (double attribute data set)

The 14 days are:

- Monday 25 July to Sunday July 31 2016
  Before paper ticket retirement – note trips made with paper tickets in this week are not included in the data.
- Monday 8 August to Sunday 14 August 2016
  After paper ticket (Magnetic Stripe Tickets/MST) retirement.
- Monday 21 November 2016 – Sunday 27 November 2016
  Post fare reform. More information available at https://www.transport.nsw.gov.au/newsroom-and-events/media-releases/opal-card-fare-freeze-and-weekly-travel-reward-reformed
- Monday 26 December 2016 – Sunday 1 January 2017
  school holidays and New Year's Eve

The four (4) public transport modes are:

- Train
- Bus
- Ferry
- Light rail

3 Privacy protection

Protecting customers' privacy is a priority when releasing open data. The Opal dataset has been treated to ensure no personal information and no information that can be linked to an individual was released in the dataset.

Algorithms with a proof of privacy associated with their output, known as differential privacy, were developed to treat the Opal data to produce a privacy-protected dataset.

The Opal dataset produced by the differential privacy algorithms are protected against all known classes of privacy attacks, such as re-identification. Privacy attacks currently not known also have little chance of success in these datasets.

The algorithms are designed to:

- Remove high risk low count queries – low count is less than 18 customers (excluding zero)
- Include noise in queries that are not removed.

The algorithms modified elements of the Opal data that can be used to identify an individual to produce a privacy-protected dataset. The privacy-protected dataset maintains the statistical accuracy.

The Opal data is modified by algorithms to ensure customers' privacy is protected. This includes the following:

- All information that is or can be linked to an individual Opal card is removed
- Geographic information that could identify travel patterns of individuals on low volume transport services such as bus stops that a small number of passengers use is treated and aggregated.

4 Time Binning

To mitigate privacy risk of re-identification, all times in the dataset have been binned to 15 minutes. For example, a tap on at 3.58pm will be represented in the data as a tap on at 3.45pm.

All times in all tables have been binned to 15 minutes and are outlined in the table below.

Original time	Bin	Example
0 to 14 minutes past the hour	To the hour	**Original time:** 8.05am **Bin:** 8am
15 to 29 minutes past the hour	15 minutes past the hour	**Original time:** 10.20am **Bin:** 10.15am
30 to 44 minutes past the hour	Half past the hour	**Original time:** 11.40pm **Bin:** 11.30pm
45 to 59 minutes past the hour	45 minutes past the hour	**Original time:** 1.49pm

		Bin: 1.45pm

5 Using the data

5.1 Fields

The available fields in the data are mode, date, tap, time, location and count.

Field	Type	Description
mode	text	Transport mode - Either Train, Bus, Ferry or Light Rail indicating the mode of transport the tap on or off is related to
date	number	Date in format of yyyymmdd
tap	text	Tap type - Either on or off indicating departure (on) or arrival (off)
time	time	Binned time in 24 hour format. Time is binned to 15 minutes intervals. -1 indicates “unknown” -- an error with the Opal card reader Time is based on Opal system time
loc	text	Location -1 indicates “unknown” -- an error with the Opal card reader 4 digit number indicates Postcode Named locations are Train Stations or Ferry Wharves or Light Rail stops
count	number	A privacy protected number of tap occurrences

Some data has been hidden for privacy. A return of a zero count is an indication that either there were zero counts in the data or that there were insufficient trips to ensure privacy protection.

5.2 Data variances

In order to preserve privacy, data has been aggregated and treated. Trips have been removed which pose a threat to privacy. Bus locations (bus stops) have been binned by postcode. Light rail stops are not binned but has resulted in some small counts being removed from the released dataset.

A return of a zero count is an indication that there were zero counts or there are insufficient trips to release the information.

The total (sum) of the double attribute data set may not match the details from the single attribute set due to privacy protection measures.

6 Open Opal Data period

6.1 Dates

There are four non- consecutive weeks of data in the Opal dataset. The four (4) 7 day period of data is from

- Monday 25 July to Sunday July 31 2016
- Monday 8 August to Sunday 14 August 2016
- Monday 21 November 2016 to Sunday 27 November 2016
- Monday 26 December 2016 to Sunday 1 January 2017

6.2 Events

The periods were selected to have the fewest incidents across the network. Where possible, comma separated values (csv) files of relevant events can be found on the Open Data hub with the Open Opal dataset.

6.3 Trackwork

Train trackwork for the weeks may skew tap on and tap off data. The usual case during trackwork is that neither tap on or tap off data is collected for the period for the train stations affected. Customers may still tap on, however it is not necessary as the buses which replace train services during trackwork do not require customers to tap on or off (and may not have an enabled Opal card reader).

Comma separated values (csv) files of the trackwork information can be found on the Open Data hub with the Open Opal dataset.

For information about the route details and which stops/stations are affected, please refer to the GTFS specification and https://developers.google.com/transit/gtfs/reference/ and the GTFS trains bundle which can be found on the Open Data Hub the link to the GTFS bundles are stored at https://opendata.transport.nsw.gov.au/dataset/timetables-complete-gtfs/.

6.4 New Year's Eve

One of the week's includes New Year's Eve. New Year's Eve is one of the busiest days for the Transport for NSW network in Sydney.

Often in the busy CBD stations, the Opal gates are left open and passengers do not need to Tap on or Tap off.