**Instructions for GTFS data pre-processing**

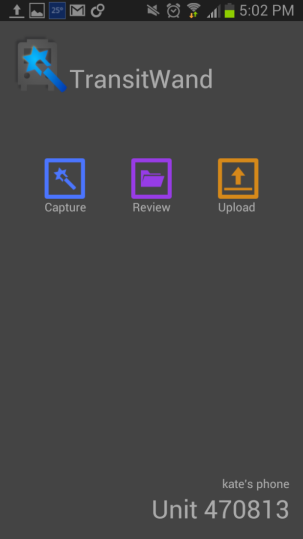
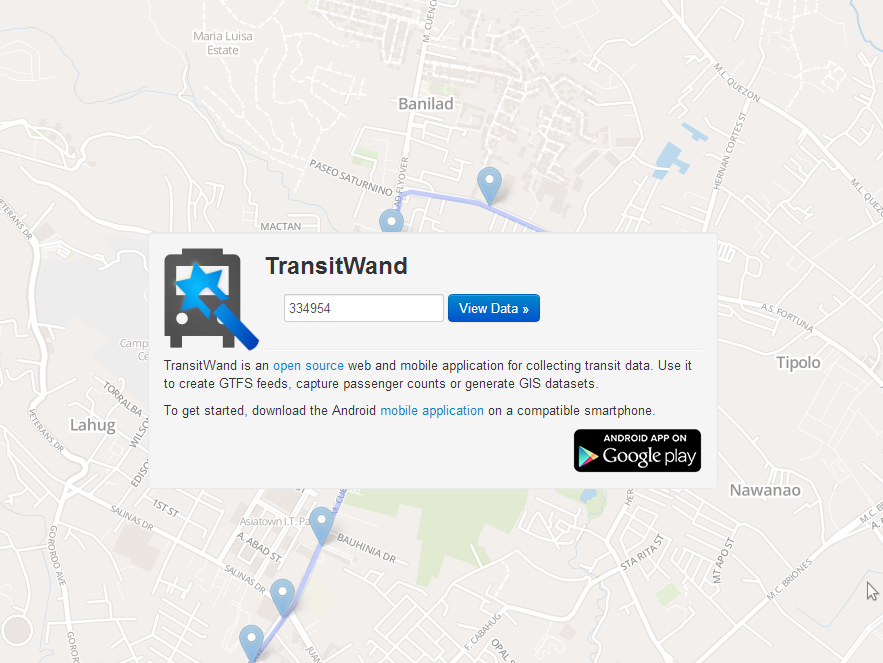
Prepared by World Bank team for Vietnam Maritime University

This instruction document is prepared by the World Bank task team specifically for the GTFS data pre-processing task assigned to the team in Vietnam Maritime University, which illustrates the steps and notes for processing the data collected by riding the buses with mobile app “Transitwand” and additional information from DOT into a preliminary format for GTFS data.

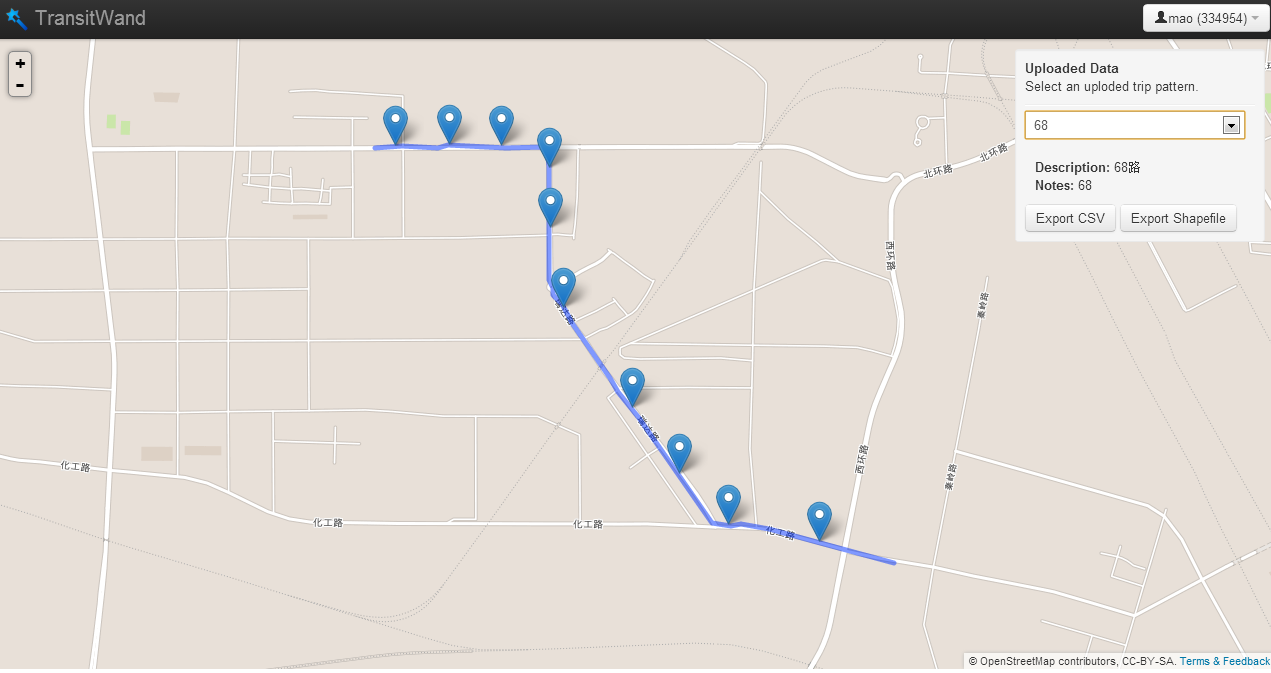
The steps below are assuming that

* data collecting by “Transitwand” has been completed and each group has registered their device and has a unique unit number (device ID) assigned by the app.
* the stop names are recorded in a separate data sheet in line with the data collected by mobile app.

1. **Download the data from website** [**http://transitwand.com**](http://transitwand.com)
   1. Input the unit number showed on the lower-right corner of the mobile app. And click “view data”.

* 1. Download the data collected by clicking “Export CSV” and save the zip file to your computer.



1. **Understanding the data collected**
   1. After extract the files from the downloaded zip file, you will see two separate file in the folder: *routes.txt* and *stops.txt*.
   2. routes.txt records the basic information of the routes that had been collected and the start time of the collection.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| unit\_id | route\_id | route\_name | route\_description | field\_notes | vehicle\_type | vehicle\_capacity | start\_capture |
| 638808 | 4383087 | Route 10011 |  |  |  |  | 1 Jan 2007 01:29:56 GMT |
| 638808 | 4383120 | Route 10010 |  |  |  |  | 1 Jan 2007 00:41:11 GMT |
| 638808 | 4383632 | Route 10009 |  |  |  |  | 1 Jan 2007 00:36:00 GMT |

* 1. stop.txt records the information of stop locations, stop sequence, travel time and dwell time (in seconds)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| route\_id | stop\_sequence | lat | lon | travel\_time | dwell\_time | board | alight |
| 4383120 | 0 | 20.80669 | 106.6241 | 5 | 0 | 0 | 0 |
| 4383120 | 1 | 20.80432 | 106.6243 | 55 | 0 | 0 | 0 |
| 4383120 | 2 | 20.80307 | 106.6265 | 64 | 1 | 40 | 0 |

* 1. Additional information were collected in a separate data sheet which were required to fill in by the data collecting team.

|  |  |  |  |
| --- | --- | --- | --- |
| Route Number | Stop Number | Stop Name | Flag |
| 1 | 1 | Aaa | 1 |
| 1 | 2 | Bbb | 0 |

1. **Processing the data collected into GTFS data format**
   1. GTFS data set consists of a series of files, where six of them are compulsory and the other files are optional. In this task, we will add shapes.txt as a compulsory file to show the alignment of the bus lines.
      * + agency.txt
        + stops.txt
        + routes.txt
        + trips.txt
        + stop\_times.txt
        + calendar.txt
        + shapes.txt (optional but compulsory in this case)

**\* Please see the GTFS reference document for detailed information of the attributes in each of the file.**

* 1. The data collected by “Transitwand” covers basic information for some of the GTFS files as below.
  2. **agency.txt** is the list of all the bus company information. The attributes are straight forward, please see the data template in a separate file.

|  |  |  |  |
| --- | --- | --- | --- |
| agency\_id | agency\_name | agency\_url | agency\_timezone |
| DOT | Demo Transit Authority | http://google.com | Asia/Ho\_Chi\_Minh |

* 1. **stops.txt** is the list of all the stops information in the city area.
     + - stop\_id is a unique number assigned to each stop. The coding rules is flexible, which can be defined by the data collector.
       - stop\_name will be filled by the name collected by the team in the separate data sheet, or information obtained from DOT.
       - stop\_lat/stop\_lon will be filled by the data collected by Transitwand or field survey.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| stop\_id | stop\_code | stop\_name | stop\_desc | stop\_lat | stop\_lon | zone\_id | stop\_url | location\_type | direction | position |
| 2 |  | A Ave & Chandler |  | 45.420595 | -122.676 |  |  |  |  |  |
| 3 |  | A Ave & Second St |  | 45.419386 | -122.665 |  |  |  |  |  |
| 4 |  | A Ave & 10th St |  | 45.420703 | -122.675 |  |  |  |  |  |

* 1. **routes.txt** is the list of basic information of the routes.
     + - route\_id is a unique number assigned to each route. The coding rules is flexible, which can be defined by the data collector.
       - agency\_id should be consistent with the agency.txt file.
       - route\_short\_name will be filled with the route number.
       - route\_long\_name will be filled with the “origin – destination” information.
       - route\_type are all set to 3 as all the routes are bus lines.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| route\_id | agency\_id | route\_short\_name | route\_long\_name | route\_desc | route\_type | route\_url | route\_color | route\_text\_color |
| CD | BC 1 | 1 | CAU RAO - DU NGHIA |  | 3 |  |  |  |
| BT | BC 2 | 2 | BEN BINH - TT VINH BAO |  | 3 |  |  |  |
| BD | BC 3 | 3A | BUU DIEN - DO SON |  | 3 |  |  |  |
| KD | BC 4 | 3B | KHACH SAN DAU KHI - DO SON |  | 3 |  |  |  |

* 1. **trip.txt** is the list of all the trips in all routes. For example, Bus line 1 runs between A to B, then there will be two trips for this route, one is from A to B, and the other one is from B to A. Trip.txt here is basically presenting the information for each direction of each route. Please see the data template in a separate file.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| route\_id | service\_id | trip\_id | trip\_headsign | direction\_id | block\_id | shape\_id |
| CD | FULLW | CD1 | 1 to DU NGHIA | 0 |  | 11 |
| CD | FULLW | CD2 | 1 to CAURAO | 1 |  | 12 |
| BT | FULLW | BT1 | 2 to TT VINH BAO | 0 |  | 21 |
| BT | FULLW | BT2 | 2 to BEN BINH | 1 |  | 22 |

* 1. **stop\_times.txt** is the most important information in GTFS data set. It records the stops sequence in each route and the stop times at each stop throughout the day.
     + - trip\_id is consistent with the one in trip.txt which presents the trip that this stop\_time record belongs to.
       - arrival\_time is the list of the arrival time at this specific stop throughout the day. This will be filled by the time information collected via Transitwand (collecting start time + travel time at each stop) or schedule information obtained from DOT.
       - departure\_time is the list of the departure time at this specific stop throughout the dat. This will be filled by the information collected via Transitwand (collecting start time + travel time + dwell time at each stop) or schedule information obtained from DOT.
       - stop\_sequence is the order of the stops in each trip, the order number can be simply as 1,2,3,4…

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| trip\_id | arrival\_time | departure\_time | stop\_id | stop\_sequence | stop\_headsign | pickup\_type | drop\_off\_type | shape\_dist\_traveled | timepoint |
| 3963496 | 6:47:00 | 6:47:00 | 13170 | 1 |  |  |  |  |  |
| 3963496 | 6:48:18 | 6:48:18 | 7631 | 2 |  |  |  |  |  |
| 3963496 | 6:50:13 | 6:50:13 | 7625 | 3 |  |  |  |  |  |
| 3963496 | 6:52:11 | 6:52:11 | 7612 | 4 |  |  |  |  |  |

* 1. **calendar.txt** is the definition of the service days in a week. It is connected with trips.txt via service\_id. Since the buses are running for all days, there is only one record for calendar.txt as below.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| service\_id | monday | tuesday | wednesday | thursday | friday | saturday | sunday | start\_date | end\_date |
| FULLW | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 20130901 | 20140901 |

* 1. **shapes.txt** is a list of the GPS points of each route. It helps to present the geographic alignment of the route on base maps.
     + - shape\_id is the unique number for each shape. It is linked to trips.txt to identify the alignement for each trip.
       - shape\_pt\_lat and shape\_pt\_lon are the coordinates for each GPS points in the shape. These two fields will be filled with Transitwand shapefile exporting (to be fixed), or adding the alignment in OSM and export the GPS point, or information obtained from DOT.
       - shape\_pt\_sequence is the order of the points in each alignment shapes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| shape\_id | shape\_pt\_lat | shape\_pt\_lon | shape\_pt\_sequence | shape\_dist\_traveled |
| 170517 | 45.52291 | -122.677372 | 1 | 0 |
| 170517 | 45.522921 | -122.67737 | 2 | 3.7 |
| 170517 | 45.522991 | -122.677432 | 3 | 34 |
| 170517 | 45.522992 | -122.677246 | 4 | 81.5 |
| 170517 | 45.523002 | -122.676567 | 5 | 255.7 |
| 170517 | 45.523004 | -122.676486 | 6 | 276.4 |
| 170517 | 45.523007 | -122.676386 | 7 | 302 |
| 170517 | 45.523024 | -122.675386 | 8 | 558.4 |
| 170517 | 45.522962 | -122.67538 | 9 | 581 |
| 170517 | 45.522922 | -122.675383 | 10 | 595.6 |
| 170517 | 45.52288 | -122.675394 | 11 | 611.2 |
| 170517 | 45.522703 | -122.675478 | 12 | 679.2 |
| 170517 | 45.522612 | -122.675527 | 13 | 714.7 |