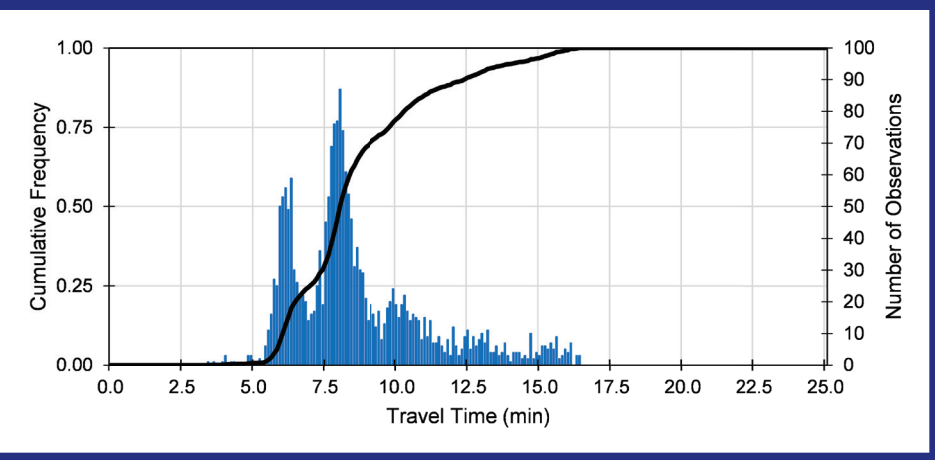
Common Data Formats for Re-Identification and High-Resolution Data



Stanley E. Young, Darcy M. Bullock, Dennis So Ting Fong

Common Data Formats for Re-Identification and High-Resolution Data

Stanley E. Young
Traffax, Inc.

Darcy M. Bullock
Purdue University

Dennis So Ting Fong
Traffax, Inc.

SBIR Phase 3 Joint Transportation Research Project
Traffax, Inc.
Purdue University

June 30, 2015

Deliverable Reference:	D2.2 Common Data Format Report
Project Name:	Sensor Fusion and MOE Development for Off-Line Traffic Analysis of Real Time Data
Contractor:	Traffax, Inc.
Contract Number:	DTFH61-14-C-00035
Contract Term Start	9/4/2014
Contract Term End	9/4/2017
Key Personnel	Stan Young, Darcy Bullock, Dennis So Ting Fong

Recommended Citation
Young, S. E., D. M. Bullock, and D. S. T. Fong. *Common Data Formats for Re-Identification and High-Resolution Data.* Purdue University, West Lafayette, Indiana, 2017. https://doi.org/10.5703/1288284316566

Acknowledgments
This work was supported by Traffax/USDOT SBIR DTFH6114C00035. The contents of this paper reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein, and do not necessarily reflect the official views or policies of the sponsoring organizations. These contents do not constitute a standard, specification, or regulation.

Table of Contents

Introduction

This report targets the delivery of common data formats to facilitate adoption and uniformity in the use of performance measures from re-identification and high-resolution probe data. The traffic data industry is typically segregated by data collection activities performed separately from data analysis activities. A common format for re-identification data allows data collection activities to be performed separate from analysis without having to build custom data interfaces based on the equipment, vendor, or data collection service provider.

Similarly, controller data differs in format and content by vendor, and sometimes by model. A common high-resolution data format minimizes the variability of implementation from vendor to vendor. Much of this work was completed by a team lead by Purdue in 2012, in a document named *INDIANA TRAFFIC SIGNAL HI RESOLUTION DATA LOGGER ENUMERATIONS*. These enumerations, or numbered codes, have been used effectively to combine data from multiple signal control vendors in early implementations.

Common data formats enable use of any performance measures software without the concern and cost of extra integration effort needed to transform or port data. The re-identification format specified below was developed in conjunction with the University of Maryland Center for Advanced Transportation Technology, and tested in software used to evaluate probe data quality, and to calculate performance metrics from probe and re-identification data.

Re-Identification Technology Standards

Standard Name: CATTWORKS STANDARD 5200 RE-IDENTIFICATION DATA SET
Last edited: 2015 June 06 Stan Young, Initial Creation

Purpose

This document establishes standardized terms and data structures to convey traffic data derived from re-identification data. It was first authored to encourage standard performance measure use for signalized arterials based on data collected with Bluetooth traffic monitoring equipment. The traffic data industry is typically segregated by data collection activities performed separately from data analysis activities. Thus a common re-identification data set to support many common data analysis activities is described such that data collection activities can be performed separate from analysis without having to build custom data interfaces based on the equipment, vendor, or data collection service provider. This standard format is intended to support various forms of re-identification technologies such as Bluetooth, Wi-Fi, automated license plates readers, and toll tag readers to name a few. This probe data set standard (CWS5200) is intended to provide a uniform method of conveying observed travel times collected along corridors using some form of re-identification technology.

Definitions

Bluetooth Traffic Monitoring: A form of re-identification technology in which the MAC address of Bluetooth enabled electronic devices in vehicles are recorded at upstream and downstream stations for the purpose of collecting a travel time sample.

Detection Range: A measurement of length, specific to re-identification technology, that describes the detection zone around a sensor. For example, for BTM technology, the detection range is roughly 300 feet before and after the sensor.

Filtering: Any method of identifying data points within a data set that meets criterion for exclusion. For example, outliers that are not reflective of the central tendency of the data may be identified by statistical tests and excluded.

Gapout: As this term relates to re-identification technology, this is a period of time after which an event is determined to have ended during which no additional data is obtained. For example, if a Bluetooth sensor detects an electronic device within a vehicle and then does not detect that same device for more than 30 seconds (the gapout threshold), then it is determined that the vehicle has left the detection zone of the sensor.

Link: See discussion under Node.

MAC ID / Address: A unique identifier programmed into Bluetooth and Wi-Fi enabled electronic communication devices to facilitate electronic data exchange.

Matched pair: A record, including corresponding time information, which signifies that a device at the upstream station was re-identified at the downstream station.

Node: The ‘node’ and ‘link’ terminology used to describe physical networks such as roadway systems is often used in place of ‘station’ and ‘segment’ with respect to re-identification data. As many times sensors are placed at the intersection of roadways, use of the terminology is often apt. Note however that sensors may be located at mid-block and other places that do not reflect junctions on a physical network. For this reason the terms ‘sensor’, ‘station’ and ‘segment’ are

used in this document to avoid confusion, and remain explicit to any re-identification deployment.

Re-identification data: A form of traffic data collection in which a vehicle is observed at an upstream and downstream station. A characteristic of the vehicle such as a license plate number, toll tag identifier, Bluetooth MAC ID, or Wi-Fi MAC ID is used to uniquely identify the vehicle at both the upstream and downstream stations. The difference in time between the upstream and downstream observations provides a travel time sample.

Re-identification technology: Any form of technology used to collect re-identification data. This includes automated license plate recognition, toll tag RFID, Bluetooth traffic monitoring, or Wi-Fi traffic monitoring.

Segment: The route that connects the upstream and downstream stations - sometimes referred to as a *link*.

Sensor: The device used to automatically record re-identification data. One or more sensors are placed at stations to record unique identifiers. Many times the term 'sensor' and 'station' are used interchangeably in casual discussion as it is typical for a single sensor to be used at a single station. However, at times, multiple sensors, of either the same or different technology, may be employed at a station.

Station: The location/s where vehicles are observed, either upstream or downstream, in order to record a unique identifier for the purpose of re-identification data. One or more sensors may be located at a station to record unique identifiers that are detected.

Unique Identifier: An alpha-numeric sequence that uniquely identifies an object (applicable for Bluetooth, WiFi, license plate and toll tag re-identification technologies).

Wi-Fi Traffic Monitoring: A form of re-identification technology in which the MAC address of Wi-Fi enabled electronic devices in vehicles are recorded at upstream and downstream stations for the purpose of collecting a travel time sample.

Acronyms

ALPR: Automated License Plate Reader

BTM: Bluetooth Traffic Monitoring

GMT: Greenwich Mean Time

MAC ID: Media Access Control Identification

TMC: Traffic Message Channel

UID: Unique Identifier

UTC: Coordinated Universal Time

Data Structure

The description of the re-identification data sets is given using a classical structured array format. The format is derived from Matlab ™ syntax (also Octave), but is also similar to the structured formats in many programming languages. The data format is first described using this structured format description, and then various packaging examples, such as XML and CSV, are provided.

The re-identification DATASET contains descriptors and three primary sub-elements: STATIONS, SEGMENTS, and MATCHED_PAIRS, as well as attributes specific only to the dataset. STATION attributes provide information on the upstream and downstream locations where re-identification data is collected. SEGMENT attributes provide information specific to the roadway or path connecting the upstream and downstream stations, and MATCHED_PAIR provides the travel time data specific to a segment. In a structured array format, DATASET elements are prepended by 'ds', STATION elements are prepended by 'station', SEGMENT elements are prepended by 'segment', and MATCHED_PAIR elements are prepended by 'mp'. As STATIONS, SEGMENTS, and MATCHED_PAIR data are sub-elements of a DATASET, each will also be prepended by 'ds'. As an example, a STATION attribute such as its latitude will be designated as 'ds.station.lat'.

Mandatory elements must appear in the definition of any dataset, whereas optional elements may be omitted or left blank.

The dataset definition allows for multiple sensors, stations, segments, and matched pairs data sets within a single structure or data file.

Re-identification Structure Elements

	Element	Mandatory / Optional	Description
Dataset attributes (abbreviated ds)			
	ds.dataformat	Mandatory	'CATTWORKS STANDARD 5200 REIDENTIFICATION DATASET' or 'CWS5200'
	ds.datasetname	Optional	Text field with a descriptive name of the data set
	ds.local_datetime.begin	Mandatory	The beginning date and time, in the local time reference of the entire data set. The begin date-time is preferred to the nearest minute. In the format of yyyy-mm-dd HH:MM:SS For example, if the data set spans a two week period from January 14 to January 28 of 2015, the local_datetime.begin would be '2015-01-14 00:00:00' reflecting the beginning of Jan 14, 2015. Local time implies that any adjustment for Daylight Savings Time has been applied.
	ds.local_datetime.end	Mandatory	The end date and time, in the local time reference of the data set. The end date-time is preferred to the nearest minute. In the format of yyyy-mm-dd HH:MM:SS For example, if the data set spans a two week period from January 14 to January 28 of 2015, the local_datetime.end would be '2015-01-28 23:59:59' reflecting the end of the day of Jan 28, 2015, or alternatively it could be '2015-01-29 00:00:00' The intent is to bracket the timeframe of the dataset. Local time implies that any adjustment for Daylight Savings Time has been applied.
	ds.lengthunits	Mandatory	Text field containing one of the following 'miles' or 'km'
	ds.local_datetime.timezone	Optional	A text field indicating the local time zone. Useful if data may be combined with other data sets that span time zones, or are specified in UTC or GMT date-time formats
	ds.middefinition	Optional	If a mid-point, (or intermediate point) time offset is provided in the matched pair data, this text field describes how the mid-point is defined or derived. Examples include such things as 'highest RSSI reading' or 'median observation'. The method for a mid-point or intermediate point is often times technology dependent.
	ds.datecreated	Optional	The date the data set was created in 'yyyy-mm-dd' format.
	ds.contact.name	Optional	Text field with name of contact person in case of questions.
	ds.contact.number	Optional	Text field with phone number of contact in case of questions.
	ds.contact.email	Optional	Text field with email address of contact in case of questions.
	ds.filename	Optional	Text field with original filename of dataset

Station Elements			
Station elements provide a description of the upstream and downstream stations that comprise the segment. At least two stations must be defined in the dataset.			
	ds.station.name	Mandatory	Name of the station. This is a mandatory text field, and is intended to contain the identifying name of the station as determined or needed by the application, such as a sequential naming scheme. It is a text field. Any combinations of characters are allowed except quotes, brackets, braces, ampersand, or parentheses. Station name must be unique.
	ds.station.uid	Mandatory	Unique identifier of the station. This is a mandatory field. It is intended to be populated with a machine code or other automatically assigned identifier. Station UID must be unique. The station *uid* and *name* may be the same.
	ds.station.lat	Mandatory	Latitude of the station location in decimal degrees
	ds.station.lon	Mandatory	Longitude of the station location in decimal degrees
	ds.station.roadway	Optional	Text field indicating roadway on which station is located. Ex. 'US-40'.
	ds.station.crossroad	Optional	Text field indicating nearest crossroad to the station. Ex. 'Bell Rd.'
	ds.station.notes	Optional	Freeform text field for additional information about the station
Segment Elements			
Segment elements provide a description of the corridor or path connecting the upstream station the downstream station.			
	ds.segment.name	Mandatory	Name of the segment. This is a mandatory field, and is intended to contain the identifying name of the segment as determined or needed by the application. It is a text field. Any combinations of characters are allowed except quotes, brackets, braces, ampersand, or parentheses. Segment name must be unique.
	ds.segment.name2	Optional	Alternate name of the segment. This is an optional field to facilitate a secondary naming scheme. It is a text field. Any combinations of characters are allowed except quotes, brackets, braces or parentheses. Segment secondary name must be unique.
	ds.segment.upstreamstation	Mandatory	Unique name of the upstream station.
	ds.segment.downstreamstation	Mandatory	Unique name of the downstream station.
	ds.segment.length	Mandatory	Length of segment, used to convert travel time to speed. Units are in *ds.lengthunits*
	ds.segment.roadname1	Optional	Primary road designation (such as I-70)
	ds.segment.roadname2	Optional	Secondary road designation (such as PA Tollway)
	ds.segment.direction	Optional	Direction given as a text field. Examples include 'northbound', 'NB', or 'ccw' (for counter-clockwise, as in a beltway).
	ds.segment.description	Optional	Freeform text field for additional information about the segment

Matched Pair Data

The data observed for the period defined. Each data element is a vector of values, each described below. For conciseness, 'Matched pair', is abbreviated to 'mp' in the following definitions.

ds.mp.segment	Mandatory	The unique segment name for which matched pair data is provided
ds.mp.reidentificaiontype	Mandatory	Text field containing one of the following 'BTM', 'WIFI', 'BTMWIFI', 'ALPR', 'TOLLTAG'
ds.mp.notes	Optional	Freeform text field for additional information about the matched pair data
ds.mp.data	Mandatory	The data field contains vectorized matched pair data as defined in the following description. Each row is a matched pair record. The elements of each row are defined in the seven elements below.

uid (Optional)	**upstream_ initial_ datetimeoffset** (Mandatory)	**upstream_ final_ timeoffset** (Mandatory)	**downstream_ initial_ timeoffset** (Mandatory)	**downstream_ final_ timeoffset** (Mandatory)	**upstream_ mid_ timeoffset** (Optional)	**downstream_ mid_ timeoffset** (Optional)
Alphanumeric string unique to the matched pair.*	The offset in decimal days from *ds.local_datetime.begin* Minimum precision is to the nearest second. This may be stored as single precision floating point (see implementation notes).	The offset in seconds from the initial observation at the upstream station to the last observation of the UID at the upstream station.	The offset in seconds from the initial observation at the upstream station to the first observation of the UID at the downstream station.	The offset in seconds from the initial observation at the upstream station to the last observation of the UID at the downstream station.	The offset in seconds from the initial observation at the upstream station to the midpoint of the upstream station.	The offset in seconds from the initial observation at the upstream station to the midpoint of the downstream station.

* For example, a "uid" for ALPR may contain the license plate number, or be based on the actual license plate number, but encrypted so as not to reveal personally identifiable data. Unique identifiers are useful for some applications such as identifying extent of commute vs non-commute traffic.

High Resolution Controller Data Technology Standards

Standard Name: INDIANA TRAFFIC SIGNAL HI RESOLUTION DATA LOGGER ENUMERATIONS
Last edited: 2012 November

Introduction

High resolution traffic signal controller data is used to record the times when certain events occur at a signalized intersection, such as the state changes of signal outputs, vehicle detectors, and other elements relevant to the signal control. The "high resolution" term indicates that the events are recorded as they occur, at a fine time resolution (0.1 seconds in current signal controllers). This contrasts with legacy data formats for volume and occupancy that were reported in aggregate values in fixed, 1-15 minute intervals. A specification for the format of this data was previously established by a working group initially led by the Indiana Department of Transportation and a consortium of controller vendors. A document that fully describes the data format is available at the following link:

- Sturdevant, J. R., T. Overman, E. Raamot, R. Deer, D. Miller, D. M. Bullock, C. M. Day, T. M. Brennan, H. Li, A. Hainen, and S. M. Remias. *Indiana Traffic Signal Hi Resolution Data Logger Enumerations*. Publication . , Indiana Department of Transportation and Purdue University, West Lafayette, Indiana, 2012. Available online at http://docs.lib.purdue.edu/jtrpdata/3/.

In addition, a monograph has been produced that documents a portfolio of performance measures that can be derived from the hjgh resolution controller Data. That document is available at:

- Day, C. M., D. M. Bullock, H. Li, S. M. Remias, A. M. Hainen, R. S. Freije, A. L. Stevens, J. R. Sturdevant, and T. M. Brennan. *Performance Measures for Traffic Signal Systems: An Outcome-Oriented Approach*. Purdue University, West Lafayette, Indiana, 2014. http://dx.doi.org/10.5703/1288284315333

Report Sponsor

The "Small Business Innovation Development Act of 1982" (Pub. L. No. 97-219), along with reauthorizing legislation (Pub. L. No. 99-443 and Pub. L. No. 102-564, the "Small Business Research and Development Enhancement Act of 1992"), seeks to encourage the initiative of the private sector and to use small business effectively to meet federal research and development objectives. To comply with statutory obligations of the Act, the U.S. Department of Transportation established the Small Business Innovation Research (SBIR) Program, which conforms to the guidelines and regulations provided by the Small Business Administration. Annually, small businesses are solicited to submit innovative research proposals that address the high-priority requirements of the U.S. Department of Transportation and that have potential for commercialization.

This report was developed through a partnership between Traffax, Inc., and Purdue University with funding from a Phase III SBIR contract (DTFH6114C00035) with the Federal Highway Administration. The project, entitled "Sensor Fusion and MOE Development for Off-Line Traffic Analysis of Real Time Data," created and refined methods and tools for the characterization of performance along arterial corridors.

Publication

This report is part of a series of reports published in collaboration with USDOT, Traffax, Inc., and Purdue University. The full report series is available for download at http://docs.lib.purdue.edu/apmtp/.

Open Access and Collaboration with Purdue University

The Indiana legislature established the Joint Highway Research Project in 1937. In 1997, this collaborative venture between the Indiana Department of Transportation and Purdue University was renamed as the Joint Transportation Research Program (JTRP) to reflect state and national efforts to integrate the management and operation of various transportation modes. Since 1937, the JTRP program has published over 1,600 technical reports. In 2010, the JTRP partnered with the Purdue University Libraries to incorporate these technical reports in the University's open access digital repository and to develop production processes for rapidly disseminating new research reports via this repository. Affiliated publications have also recently been added to the collection. As of 2017, the JTRP collection had over 1.5 million downloads, with some particularly popular reports having over 20,000 downloads.