STATE OF CALIFORNIA • DEPARTMENT OF TRANSPORTATION

TECHNICAL REPORT DOCUMENTATION PAGE

TR0003 (REV 10/98)

ADA Notice

For individuals with sensory disabilities, this document is available in alternate formats. For information call (916) 654-6410 or TDD (916) 654-3880 or write Records and Forms Management, 1120 N Street, MS-89, Sacramento, CA 95814.

1. REPORT NUMBER	2. GOVERNMENT ASSOCIATION NUMBER	3. RECIPIENT'S CATALOG NUMBER
CA 19-2906		

4. TITLE AND SUBTITLE	5. REPORT DATE
TASAS (Traffic Accident Surveillance and Analysis System) and Injury Data Base Development	6/19/2019
	6. PERFORMING ORGANIZATION CODE: AHMCT Research Center, UC Davis

7. AUTHOR	8. PERFORMING ORGANIZATION REPORT NO.
Patricia Fyhrie, Grachya Hovhannisyan, Travis Swanston and Bahram Ravani: Principal Investigator	UCD-ARR- 19-06-30-0

9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. WORK UNIT NUMBER
AHMCT Research Center UCD Dept. of Mechanical & Aerospace Engineering Davis, California 95616-5294	
	11. CONTRACT OR GRANT NUMBER: 65A0560, Task ID 2906
	13. TYPE OF REPORT AND PERIOD COVERED: Final Report 10/01/15 to 06/30/19
12. SPONSORING AGENCY AND ADDRESS California Department of Transportation P.O. Box 942873, MS #83 Sacramento, CA 94273-0001	14. SPONSORING AGENCY CODE: Caltrans

15. SUPPLEMENTARY NOTES

16. ABSTRACT

The California Highway Patrol reports on motor vehicle accidents on California Highways in Traffic Collison Reports (TCRs). Caltrans processes approximately 200,000 of such reports on a yearly basis to extract critical data that is used for safety evaluation and planning purposes. Caltrans Traffic Accident Surveillance and Analysis System (TASAS) branch codifies and organizes such critical data into a computerized database (TASAS accident database). The extraction and input of the data into the TASAS accident database is performed by trained personnel and can be resource intensive. The objective of this research study was to determine the extent to which TASAS accident data could be processed automatically and to develop a software system that could potentially assist in the processing of TCRs. The emphasis was on coding the accident location (referred to as Card 8a) and coding of the sequence of events in an accident for simple two party collisions (referred to as Card 8b). The software system developed is termed “TCRPRO” and was found that in terms of Card 8a only a small percentage of TCRs (approximately 13.8%) can be coded within the TASAS branch accuracy requirements. In terms of coding for Card 8b, the system was able to handle much of the coding for simple two party collisions with the exception of the lane of travel in which the collision occurred. TCRPRO is successful, however, in other important areas. It can extract the “Summary” portion from the Narrative section in the TCRs. This automated collection of the Summary text will be able to replace the existing labor intensive methods of manual extraction. TCRPRO also provides data such as the GPS coordinates of the collision location. Interface programs can also be developed to flag certain kinds of reports with specific collision attributes that reside solely in the narrative section.

17. KEY WORDS	18. DISTRIBUTION STATEMENT
TASAS, Location Coding, Traffic Collision Reports, Automatic Coding, Collision Summaries, Latitude and Longitude, Sequence of Events	No restrictions. This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

19. SECURITY CLASSIFICATION *(of this report)*	20. NUMBER OF PAGES	21. COST OF REPORT CHARGED
Unclassified	127	

DISCLAIMER

The research reported herein was performed by the Advanced Highway Maintenance and Construction Technology (AHMCT) Research Center, within the Department of Mechanical and Aerospace Engineering at the University of California – Davis, for the Division of Research, Innovation and System Information (DRISI) at the California Department of Transportation. AHMCT and DRISI work collaboratively to complete valuable research for the California Department of Transportation.

The contents of this report reflect the views of the author(s) who is (are) responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the STATE OF CALIFORNIA or the FEDERAL HIGHWAY ADMINISTRATION. This report does not constitute a standard, specification, or regulation. This report does not constitute an endorsement by the Department of any products or services described herein.

The contents of this report do not necessarily reflect the official views or policies of the University of California. This report does not constitute an endorsement by the University of California of any products or services described herein.

For individuals with sensory disabilities, this document is available in alternate formats. For information, call (916) 654-8899, TTY 711, or write to the California Department of Transportation, Division of Research, Innovation and System Information, MS-83, P.O. Box 942873, Sacramento, CA 94273-0001.

Advanced Highway Maintenance and Construction Technology Research Center

Department of Mechanical and Aerospace Engineering
University of California at Davis

TASAS (Traffic Accident Surveillance and Analysis System) and Injury Data Base Development

Patricia Fyhrie, Grachya Hovhannisyan, Travis Swanston and Bahram Ravani: Principal Investigator

Report Number: CA19-2906
AHMCT Research Report: UCD-ARR- 19-06-30-02
Final Report of Contract: 65A0560, Task ID 2906

6-20-2019

California Department of Transportation

Division of Research, Innovation and System Information

EXECUTIVE SUMMARY

Problem, Need, and Purpose of Research

Currently Caltrans must process approximately 200,000 Traffic Collision Reports (TCRs) annually. These reports are provided by California Highway Patrol (CHP) and contain information that can be used for safety and other planning purposes. California Department of Transportation (Caltrans) has developed TSN (Transportation System Network) database, maintained by the Traffic Accident Surveillance and Analysis System (TASAS) branch. The TSN database contains 1) accident data (for collisions related to state highways), and 2) highway inventory data (e.g. district, county, route, postmile, number of lanes, geometric attributes of lanes, shoulders, medians, intersections and ramps). The TSN accident database (TASAS accident database) is populated with critical information extracted from these TCRs. The extraction of critical data that Caltrans stores in the TASAS accident database is performed by trained personnel and can be resource intensive. The purpose of this research study was to determine the extent to which TASAS accident data could be processed automatically and developing a system that could potentially assist in the processing of these reports.

The specific objectives of this research study on processing data for populating TASAS were as follows:

- Investigate the extent that computer software can accurately determine the collision locations in terms of postmile marker values as well as information on " Sequence of Events" in a collision.
- How can use of electronic digitization technology be implemented into Caltrans workflow for coding and populating the TASAS accident database?

Background

- Until 2016, TCRs were "paper based" with data processed by both CHP and Caltrans.
- In 2016, CHP completed a roll out of electronic Traffic Collision Reports (eTCR). This roll out facilitated the work performed in this research study.
- This research study helped making sure that all the information needed for the TASAS database are included in the eTCRs.
- This research study was initially focused on location coding which is referred to as "Card 8a" processing and later was expanded to include coding for Sequence of Events for 'simple cases" which are two party collisions.
- This research study developed a software system referred to as "TCRPRO" standing for TCR Processing with the following functionality:

1. Reads in eTCRs individually or in batch form and results are directly returned.
2. Makes use of snapshot of Clean Roads File
3. Outputs Summary for Location Coding and partial information on Sequence of Events for simple cases.
4. Extracts collision summaries from the report narrative.
5. If given Location, it can return Latitude, Longitude.
6. If given the location and clean roads file, then we can return TASAS variables.

The functionalities described in items 4-6 were developed beyond the original scope of this research study since they do enhance and provide potential cost savings in existing Caltrans operations.

Major Results

The key contribution of this research is developing TCRPRO that can successfully extract certain data contained within the eTCR provided by CHP. This functionality has the following positive benefits:

- TCRPRO extracts the “Summary” portion from the Narrative section in the eTCR and relays it to Caltrans by means of a web service. This automated collection of the Summary text will be able to replace labor intensive methods of manually extracting the Summary for subsequent utilization.
- TCRPRO determines the GPS coordinates (“Lat” and “Long”) using its internal LRS system based on the Location code of the eTCR. TCRPRO has the capability of performing this task either by calculating the location code internally or have it passed in via the Caltrans based web service.
- TCRPRO is able to return the corresponding highway inventory information at a given postmile value on a given highway and county. Using either the internally calculated Location Code or that provided via Caltrans web service, TCRPRO uses the “Clean Roads File” to extract TASAS variables such as the median type, barrier type, access controls and number of lanes (both left and right).
- The capability of extracting any data field from the eTCR sets up a framework where the content of all eTCRs can be made accessible to any application outside of the Location Coding and Sequence of Events coding functionality.
- In its present form, TCRPRO can handle much of the Sequence of Events coding for simple two party collisions with the exception of the lane of travel in which the collision occurred in cases where the lane numbering is not standard such as when there are High Occupancy Vehicle (HOV) lanes.
- TCRPRO can handle location coding within certain degree of accuracy. The testing of 651,094 cases 294,617 (45.25%) were automatically coded by TCRPRO with the following accuracy:

  - From the 294,617 Collisions which were Location Coded compared to the TASAS coding, 281,217 (95.45%) had everything except postmile differences greater than zero..
  - 77,278 had postmiles within 0.01 miles of TASAS coding.
  - 82,565 had postmiles between 0.01 to 0.1 miles of TASAS coding.
  - 107,622 had postmiles between 0.1 to 1.0 mile of TASAS coding.
  - 13,526 had postmile mismatch of 1 mile or more from TASAS coding.
- The accuracy in location coding and generation of partial data in Sequence of Events coding can potentially benefit the Coding Group but may require changes in their workflow and or the need for additional graphical interfaces that would need to be developed by Caltrans Information Technology (IT) group.
- TCRPRO can be used outside of the Caltrans Coding Unit to provide data such as "Summary" and the GPS coordinates of the collision location. Interface programs could also be developed to flag certain kinds of reports with specific collision attributes that reside solely in the narrative section.

Recommendations

1. Changing the workflow protocol for the Coding Group so that they would be able to handle partially completed TASAS data without the need to redo everything over again.
2. Developing a graphical user interface for the Coding Group's use of TCRPRO on Caltrans system can enhance the productivity of the Coding Group and their potential utilization of some of the features of TCRPRO.
3. Having the user interface display the collision summary and the diagram that are generated by TCRPRO in one screen to facilitate the coding process and efficiency for the Coding Group.
4. Working with CHP to modify Form 555 (TCR form) to remove ambiguity in traffic collision descriptions. This can lead to more automation of the coding process.
5. Incorporating some form of an "eDiagram" where known entities (e.g. vehicles, direction of travel, barrier, etc.) are placed on a parametrically drawn roadway where the collision occurred in eTCRs. Based on the parameters and relative placement of the vehicles, the location of the impact events could be reliably detected and TCRPRO would be able to accurately provide the Sequence of Events code.
6. To address the difficulty with location coding, the method of recording the GPS location of a collision needs to be standardized within police operations. This can eliminate some of the variations. Furthermore, methods should be considered that would allow converting a GPS location to a postmile value.

7. In the Clean Roads Files, it is recommended that the landmark fields be made more consistent with respect to naming conventions. Such changes will facilitate matching landmark values from the eTCR with the Clean Roads File data. Furthermore, expanding the landmark description can eliminate possible ambiguities with other landmarks.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS AND ABBREVIATIONS

Acronym	Definition
AHMCT	Advanced Highway Maintenance and Construction Technology Research Center
AOI	Area of Impact
Apache HTTP Server	An open-source web server developed and maintained by the Apache Software Foundation
Apache Tomcat	An open-source implementation of several Java EE server specifications, including the Java Servlet API
API	Application Programming Interface
Caltrans	California Department of Transportation
CHP	California Highway Patrol
COTS	Commercial Off–The-Shelf
CPython	The reference implementation of the Python runtime environment
DOT	Department of Transportation
DRISI	Caltrans Division of Research, Innovation and System Information
eTCR	High quality digital or Electronic Traffic Collision Report
GIS	Geographic Information System
HTTP	Hypertext Transfer Protocol
IT	Information Technology
iTCR	Image Traffic Collision Report
Java Servlet	An API used to forward web service requests to Java applications; may also refer to an object that implements this API
LRS	Linear Referencing System
mod_wsgi	A module for Apache HTTP Server that implements the WSGI specification
OCR	Optical Character Recognition
PDF	Portable Document Format

Acronym	Definition
PMM	Postmile Marker
Python	A high-level programming language created by Guido van Rossum; may also refer to a runtime environment for executing Python, such as CPython
R&D	Research & Development
SOE	Sequence of Events
SR	State Route
SWITRs	Statewide Integrated Traffic Records System
TAG	Technical Advisory Group
TASAS	Traffic Accident Surveillance and Analysis System
TCR	Traffic Collision Report
TCRPRO	Traffic Collision Report Processing (Software System)
TMA	Truck Mounted Attenuator
TSN	Transportation System Network
WSGI	Web Server Gateway Interface: a standardized calling convention used to forward web service requests to Python applications

CHAPTER 1: INTRODUCTION

Problem

Approximately 200,000 collisions occur annually in California on state highways which are maintained by the California Department of Transportation (Caltrans). In order to assess and monitor the safety performance of these highways, Caltrans has put in place a database referred to as TSN (Transportation System Network) database, maintained by the Traffic Accident Surveillance and Analysis System (TASAS) branch. The TSN database contains 1) accident data (for collisions related to state highways), and 2) highway inventory data (e.g. district, county, route, postmile, number of lanes, geometric attributes of lanes, shoulders, medians, intersections and ramps). The TSN accident database (TASAS accident database) is populated with data extracted from the California Highway Patrol (CHP) generated Traffic Collision Reports (TCRs). These reports provide most of the data for the collision, but they do not provide the location of the collision in terms of a postmile value. Caltrans need the postmile location as well as the Sequence of Events (SOE) in a collision for their safety and other evaluations. Determining and coding this additional information must be done by Caltrans personnel, which can be a very time consuming practice.

This research was aimed at utilizing some of the advancement in software and computer technology to assist the process of data coding and data extraction for populating the TASAS database. This research has benefited from the recently implemented electronic version of CHP's traffic collision reports, which was rolled out in October of 2015. The launch of the electronic TCR has streamlined the processing of the TASAS database and the report coding functions performed by Caltrans.

Objectives

The main objective of the proposed research is to develop methods that would facilitate data extraction from TCRs and subsequent processing for Location Coding and the coding of the Sequence of Events. More specifically, this research investigated the extent to which computer software can be used to automate the process of coding the collision locations in terms of postmile marker (PMM) values and Sequence of Events in a collision. In addition work was done in automatically generating the collision summaries as well as latitude and longitude of the accident location.

Scope

The scope of the proposed research is to develop methods that would facilitate data extraction and processing for both coding portions (location coding and coding SOE). Beyond its scope, this research also investigated automatic population of other items in the TASAS accident database as well as generating accident summaries and latitude/longitude information for the collision location. This work was to expand upon the previous research, provided in Appendix C, on the possible usefulness of software

manipulations of the electronic version of the TCR. The results of this research provides data and answers to the following research questions:

1. The extent at which data coding in terms of determination of District, County, Route, Post Mile, Travel Direction and Post Mile Markers can be digitized and streamlined.
2. The extent at which data from electronic TCRs can be digitally extracted and automatically put into the TASAS data base.
3. The extent at which data from the narrative portion of the Police TCRs can be automatically extracted and codified and used to digitally populate the TASAS data base.
4. How can use of electronic digitization technology be implemented into Caltrans workflow for coding and populating the TASAS data base?
5. The extent at which collision summaries can be automatically extracted from TCRs.

Background

The TASAS accident database is an electronic database and data processing system that contains data for collisions that are state highway related. Each collision record in the accident database is referenced to a postmile address that ties to the highway inventory database. The highway inventory database contains data on 15,200 miles of highway, 20,000 intersections, and 16,000 ramps in California [1].

Caltrans Collision Postmile Coding unit processes the police TCRs and assigns specific location values. The collision detail information is then transferred to Statewide Integrated Traffic Records System (SWITRS) and TASAS databases. There are variations in the police TCRs since these are provided by approximately 100 CHP area offices plus over 400 different local police departments making the coding and data extraction process based on somehow non-uniformly prepared TCRs.

CHP rolled out an electronic system in October of 2015 so that all the TCRs from the 100 CHP area offices will be prepared electronically. It was, therefore, feasible to investigate developing techniques and algorithms for automatically extracting data from such digitized system into the TASAS accident database. Initial research into the feasibility of this type of processing was performed and determined that the implementation of a system would show promise (see Appendix C) In addition, there was a need for methods to automatically, or at least semi-automatically, extract similar data from non-digitally generated TCRs from other local police agencies who may not have digitized their process. This research study addressed these issues.

Research Methodology

The methodology used in this research involves building upon the previous work of the Advanced Highway Maintenance and Construction Technology Research Center (AHMCT) in developing an Optical Character Recognition (OCR) system for the California work zone accident injury data base that can semi-automatically extract a limited amount of data from Traffic Collision Reports. Along with using this established knowledge base,

advantage was also taken of CHP's new digital format of the electronic Traffic Collision Report (eTCR). This research developed a general software package (named "TCRPRO" standing for TCR PROcessing) aimed at automating at least part of the manual coding process for the TASAS accident database. The software developed to perform these tasks was deployed for testing at Caltrans with the help of Caltrans Information Technology (IT) applications group.

Research Approach

The research approach involved first working with CHP in making sure that the electronic version of TCRs (eTCRs) capture similar data to that of traditional paper TCRs that would be needed by Caltrans. Second, the approach involved working with the Caltrans Coding Group to understand their process and the coding requirements. The approach also incorporated a Technical Advisory Panel from Caltrans that guided the work of the researchers. The third step involved developing the computerized software system TCRPRO. Finally, the last step working with Caltrans IT applications group to develop methods for Caltrans to be able to securely interface with TCRPRO. Caltrans IT applications group is developing the user interface for Caltrans Coding Group to use TCRPRO. This interface development is to be done by Caltrans IT, since it requires access to Caltrans computing system which is not made available to AHMCT researchers. TCRPRO, which is the main outcome of this research study, is designed to have the following capabilities:

- Read eTCRs individually or in batch forms and directly return results.
- Make use of snapshot of Clean Roads File.
- Output Summary.
- If given location, it can return Latitude, Longitude.
- If given the location and clean roads file, then it can return TASAS variables.

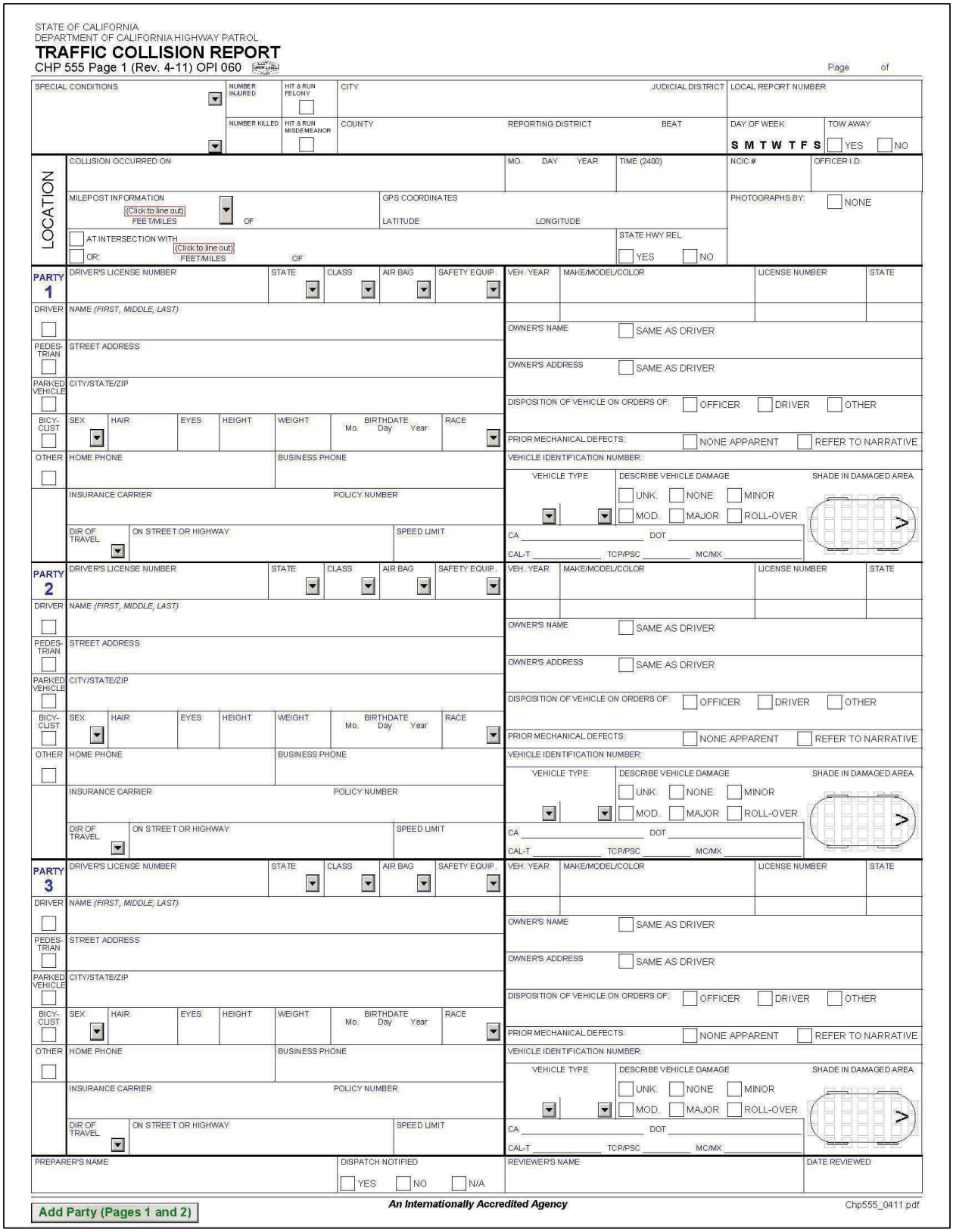
TCRPRO is described in more detail in Chapter 3.

CHAPTER 2: TRAFFIC COLLISION REPORT CODING PROCESS

Traffic Collision Reports (TCR) Overview

The purpose of a Traffic Collision Report is to document all the details related to a collision on the California roadway system. The TCR Form 555 is used in California for both state and regional roadways. These forms contain a vast amount of information and much is packed into a small space.

TCRs have been in existence for many decades in California. They have been developed over the years to facilitate the reporting and the description of a collision as objective as possible. Due to the nature of the language and the complication of communication, the fields within the report can contain many variations making it challenging for automated processing by a computer software. For the purposes of this report, only the segments of this form that are relevant to TASAS accident database are described. These relevant sections are pages one and two of the TCR that contain the collision facts and the narrative section that can be in a subsequent page and has more details about the accident. The form for the first page of all collision reports is shown in Figure 2.1. It is titled the “Traffic Collision Report” page and contains the date, time, and location of the collision as well as information about all involved parties. The second page (Figure 2.2) is titled, “Traffic Collision Coding” and provides additional data on both the collision and the parties. Figure 2.3 shows an example of the “Narrative Supplemental” pages. This example shows typical sections within the Narrative such as Notification, Statements, Cause, Summary and Area of Impact.



STATE OF CALIFORNIA
DEPARTMENT OF CALIFORNIA HIGHWAY PATROL
TRAFFIC COLLISION REPORT
CHP 555 Page 1 (Rev. 4-11) OPI 060

Page of

SPECIAL CONDITIONS | NUMBER INJURED | HIT & RUN FELONY | CITY | JUDICIAL DISTRICT | LOCAL REPORT NUMBER
NUMBER KILLED | HIT & RUN MISDEMEANOR | COUNTY | REPORTING DISTRICT | BEAT | DAY OF WEEK S M T W T F S | TOW AWAY YES NO

LOCATION
COLLISION OCCURRED ON | MO. DAY YEAR | TIME (2400) | NCIC # | OFFICER I.D.
MILEPOST INFORMATION (Click to line out) FEET/MILES OF | GPS COORDINATES LATITUDE LONGITUDE | PHOTOGRAPHS BY: NONE
AT INTERSECTION WITH OR: (Click to line out) FEET/MILES OF | STATE HWY REL. YES NO

PARTY 1
DRIVER'S LICENSE NUMBER | STATE | CLASS | AIR BAG | SAFETY EQUIP. | VEH. YEAR | MAKE/MODEL/COLOR | LICENSE NUMBER | STATE
DRIVER | NAME (FIRST, MIDDLE, LAST) | OWNER'S NAME SAME AS DRIVER
PEDES-TRIAN | STREET ADDRESS | OWNER'S ADDRESS SAME AS DRIVER
PARKED VEHICLE | CITY/STATE/ZIP | DISPOSITION OF VEHICLE ON ORDERS OF: OFFICER DRIVER OTHER
BICY-CLIST | SEX | HAIR | EYES | HEIGHT | WEIGHT | BIRTHDATE Mo. Day Year | RACE | PRIOR MECHANICAL DEFECTS: NONE APPARENT REFER TO NARRATIVE
OTHER | HOME PHONE | BUSINESS PHONE | VEHICLE IDENTIFICATION NUMBER:
INSURANCE CARRIER | POLICY NUMBER | VEHICLE TYPE | DESCRIBE VEHICLE DAMAGE UNK. NONE MINOR MOD. MAJOR ROLL-OVER | SHADE IN DAMAGED AREA
DIR OF TRAVEL | ON STREET OR HIGHWAY | SPEED LIMIT | CA DOT CAL-T TCP/PSC MC/MX

PARTY 2
DRIVER'S LICENSE NUMBER | STATE | CLASS | AIR BAG | SAFETY EQUIP. | VEH. YEAR | MAKE/MODEL/COLOR | LICENSE NUMBER | STATE
DRIVER | NAME (FIRST, MIDDLE, LAST) | OWNER'S NAME SAME AS DRIVER
PEDES-TRIAN | STREET ADDRESS | OWNER'S ADDRESS SAME AS DRIVER
PARKED VEHICLE | CITY/STATE/ZIP | DISPOSITION OF VEHICLE ON ORDERS OF: OFFICER DRIVER OTHER
BICY-CLIST | SEX | HAIR | EYES | HEIGHT | WEIGHT | BIRTHDATE Mo. Day Year | RACE | PRIOR MECHANICAL DEFECTS: NONE APPARENT REFER TO NARRATIVE
OTHER | HOME PHONE | BUSINESS PHONE | VEHICLE IDENTIFICATION NUMBER:
INSURANCE CARRIER | POLICY NUMBER | VEHICLE TYPE | DESCRIBE VEHICLE DAMAGE UNK. NONE MINOR MOD. MAJOR ROLL-OVER | SHADE IN DAMAGED AREA
DIR OF TRAVEL | ON STREET OR HIGHWAY | SPEED LIMIT | CA DOT CAL-T TCP/PSC MC/MX

PARTY 3
DRIVER'S LICENSE NUMBER | STATE | CLASS | AIR BAG | SAFETY EQUIP. | VEH. YEAR | MAKE/MODEL/COLOR | LICENSE NUMBER | STATE
DRIVER | NAME (FIRST, MIDDLE, LAST) | OWNER'S NAME SAME AS DRIVER
PEDES-TRIAN | STREET ADDRESS | OWNER'S ADDRESS SAME AS DRIVER
PARKED VEHICLE | CITY/STATE/ZIP | DISPOSITION OF VEHICLE ON ORDERS OF: OFFICER DRIVER OTHER
BICY-CLIST | SEX | HAIR | EYES | HEIGHT | WEIGHT | BIRTHDATE Mo. Day Year | RACE | PRIOR MECHANICAL DEFECTS: NONE APPARENT REFER TO NARRATIVE
OTHER | HOME PHONE | BUSINESS PHONE | VEHICLE IDENTIFICATION NUMBER:
INSURANCE CARRIER | POLICY NUMBER | VEHICLE TYPE | DESCRIBE VEHICLE DAMAGE UNK. NONE MINOR MOD. MAJOR ROLL-OVER | SHADE IN DAMAGED AREA
DIR OF TRAVEL | ON STREET OR HIGHWAY | SPEED LIMIT | CA DOT CAL-T TCP/PSC MC/MX

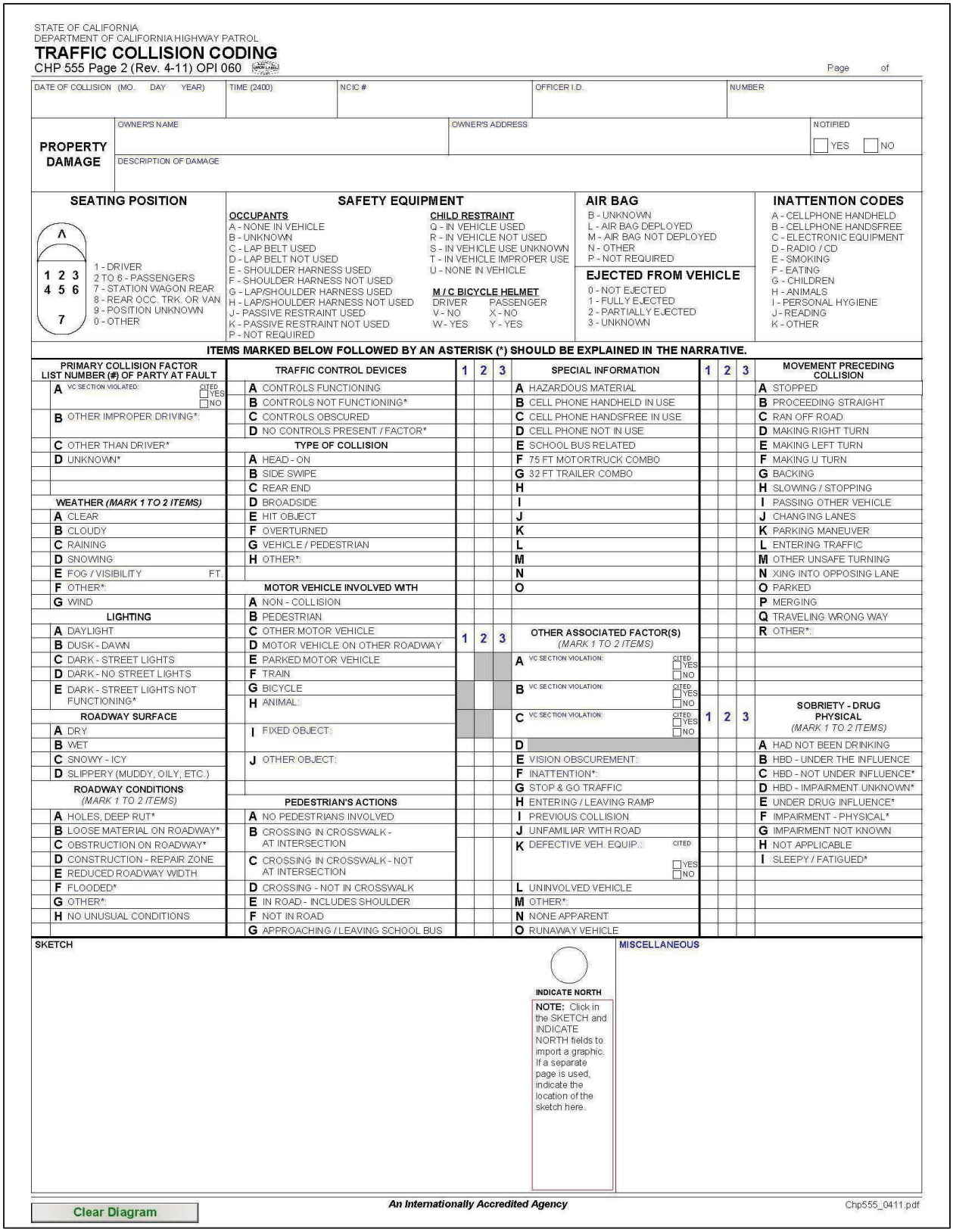
PREPARER'S NAME | DISPATCH NOTIFIED YES NO N/A | REVIEWER'S NAME | DATE REVIEWED

Add Party (Pages 1 and 2)

An Internationally Accredited Agency

Chp555_0411.pdf

Figure 2.1: Page 1 of a TCR



STATE OF CALIFORNIA
DEPARTMENT OF CALIFORNIA HIGHWAY PATROL

TRAFFIC COLLISION CODING

CHP 555 Page 2 (Rev. 4-11) OPI 060

Page of

DATE OF COLLISION (MO. DAY YEAR)	TIME (2400)	NCIC #	OFFICER I.D.	NUMBER

PROPERTY DAMAGE

OWNER'S NAME	OWNER'S ADDRESS	NOTIFIED
		☐ YES ☐ NO

DESCRIPTION OF DAMAGE

SEATING POSITION

A
1 2 3
4 5 6
7

1 - DRIVER
2 TO 6 - PASSENGERS
7 - STATION WAGON REAR
8 - REAR OCC. TRK. OR VAN
9 - POSITION UNKNOWN
0 - OTHER

SAFETY EQUIPMENT

OCCUPANTS
A - NONE IN VEHICLE
B - UNKNOWN
C - LAP BELT USED
D - LAP BELT NOT USED
E - SHOULDER HARNESS USED
F - SHOULDER HARNESS NOT USED
G - LAP/SHOULDER HARNESS USED
H - LAP/SHOULDER HARNESS NOT USED
J - PASSIVE RESTRAINT USED
K - PASSIVE RESTRAINT NOT USED
P - NOT REQUIRED

CHILD RESTRAINT
Q - IN VEHICLE USED
R - IN VEHICLE NOT USED
S - IN VEHICLE USE UNKNOWN
T - IN VEHICLE IMPROPER USE
U - NONE IN VEHICLE

M / C BICYCLE HELMET
DRIVER: V - NO, W - YES
PASSENGER: X - NO, Y - YES

AIR BAG
B - UNKNOWN
L - AIR BAG DEPLOYED
M - AIR BAG NOT DEPLOYED
N - OTHER
P - NOT REQUIRED

EJECTED FROM VEHICLE
0 - NOT EJECTED
1 - FULLY EJECTED
2 - PARTIALLY EJECTED
3 - UNKNOWN

INATTENTION CODES
A - CELLPHONE HANDHELD
B - CELLPHONE HANDSFREE
C - ELECTRONIC EQUIPMENT
D - RADIO / CD
E - SMOKING
F - EATING
G - CHILDREN
H - ANIMALS
I - PERSONAL HYGIENE
J - READING
K - OTHER

ITEMS MARKED BELOW FOLLOWED BY AN ASTERISK (*) SHOULD BE EXPLAINED IN THE NARRATIVE.

PRIMARY COLLISION FACTOR LIST NUMBER (#) OF PARTY AT FAULT

A VC SECTION VIOLATED: CITED ☐ YES ☐ NO
B OTHER IMPROPER DRIVING*:
C OTHER THAN DRIVER*
D UNKNOWN*

WEATHER (MARK 1 TO 2 ITEMS)
A CLEAR
B CLOUDY
C RAINING
D SNOWING
E FOG / VISIBILITY FT.
F OTHER*:
G WIND

LIGHTING
A DAYLIGHT
B DUSK - DAWN
C DARK - STREET LIGHTS
D DARK - NO STREET LIGHTS
E DARK - STREET LIGHTS NOT FUNCTIONING*

ROADWAY SURFACE
A DRY
B WET
C SNOWY - ICY
D SLIPPERY (MUDDY, OILY, ETC.)

ROADWAY CONDITIONS (MARK 1 TO 2 ITEMS)
A HOLES, DEEP RUT*
B LOOSE MATERIAL ON ROADWAY*
C OBSTRUCTION ON ROADWAY*
D CONSTRUCTION - REPAIR ZONE
E REDUCED ROADWAY WIDTH
F FLOODED*
G OTHER*:
H NO UNUSUAL CONDITIONS

TRAFFIC CONTROL DEVICES
A CONTROLS FUNCTIONING
B CONTROLS NOT FUNCTIONING*
C CONTROLS OBSCURED
D NO CONTROLS PRESENT / FACTOR*

TYPE OF COLLISION
A HEAD - ON
B SIDE SWIPE
C REAR END
D BROADSIDE
E HIT OBJECT
F OVERTURNED
G VEHICLE / PEDESTRIAN
H OTHER*:

MOTOR VEHICLE INVOLVED WITH
A NON - COLLISION
B PEDESTRIAN
C OTHER MOTOR VEHICLE
D MOTOR VEHICLE ON OTHER ROADWAY
E PARKED MOTOR VEHICLE
F TRAIN
G BICYCLE
H ANIMAL:
I FIXED OBJECT:
J OTHER OBJECT:

PEDESTRIAN'S ACTIONS
A NO PEDESTRIANS INVOLVED
B CROSSING IN CROSSWALK - AT INTERSECTION
C CROSSING IN CROSSWALK - NOT AT INTERSECTION
D CROSSING - NOT IN CROSSWALK
E IN ROAD - INCLUDES SHOULDER
F NOT IN ROAD
G APPROACHING / LEAVING SCHOOL BUS

1	2	3	SPECIAL INFORMATION
			A HAZARDOUS MATERIAL
			B CELL PHONE HANDHELD IN USE
			C CELL PHONE HANDSFREE IN USE
			D CELL PHONE NOT IN USE
			E SCHOOL BUS RELATED
			F 75 FT MOTORTRUCK COMBO
			G 32 FT TRAILER COMBO
			H
			I
			J
			K
			L
			M
			N
			O

1	2	3	OTHER ASSOCIATED FACTOR(S) (MARK 1 TO 2 ITEMS)
			A VC SECTION VIOLATION: CITED ☐ YES ☐ NO
			B VC SECTION VIOLATION: CITED ☐ YES ☐ NO
			C VC SECTION VIOLATION: CITED ☐ YES ☐ NO
			D
			E VISION OBSCUREMENT:
			F INATTENTION*:
			G STOP & GO TRAFFIC
			H ENTERING / LEAVING RAMP
			I PREVIOUS COLLISION
			J UNFAMILIAR WITH ROAD
			K DEFECTIVE VEH. EQUIP.: CITED ☐ YES ☐ NO
			L UNINVOLVED VEHICLE
			M OTHER*:
			N NONE APPARENT
			O RUNAWAY VEHICLE

1	2	3	MOVEMENT PRECEDING COLLISION
			A STOPPED
			B PROCEEDING STRAIGHT
			C RAN OFF ROAD
			D MAKING RIGHT TURN
			E MAKING LEFT TURN
			F MAKING U TURN
			G BACKING
			H SLOWING / STOPPING
			I PASSING OTHER VEHICLE
			J CHANGING LANES
			K PARKING MANEUVER
			L ENTERING TRAFFIC
			M OTHER UNSAFE TURNING
			N XING INTO OPPOSING LANE
			O PARKED
			P MERGING
			Q TRAVELING WRONG WAY
			R OTHER*:

1	2	3	SOBRIETY - DRUG PHYSICAL (MARK 1 TO 2 ITEMS)
			A HAD NOT BEEN DRINKING
			B HBD - UNDER THE INFLUENCE
			C HBD - NOT UNDER INFLUENCE*
			D HBD - IMPAIRMENT UNKNOWN*
			E UNDER DRUG INFLUENCE*
			F IMPAIRMENT - PHYSICAL*
			G IMPAIRMENT NOT KNOWN
			H NOT APPLICABLE
			I SLEEPY / FATIGUED*

SKETCH

INDICATE NORTH

NOTE: Click in the SKETCH and INDICATE NORTH fields to import a graphic. If a separate page is used, indicate the location of the sketch here.

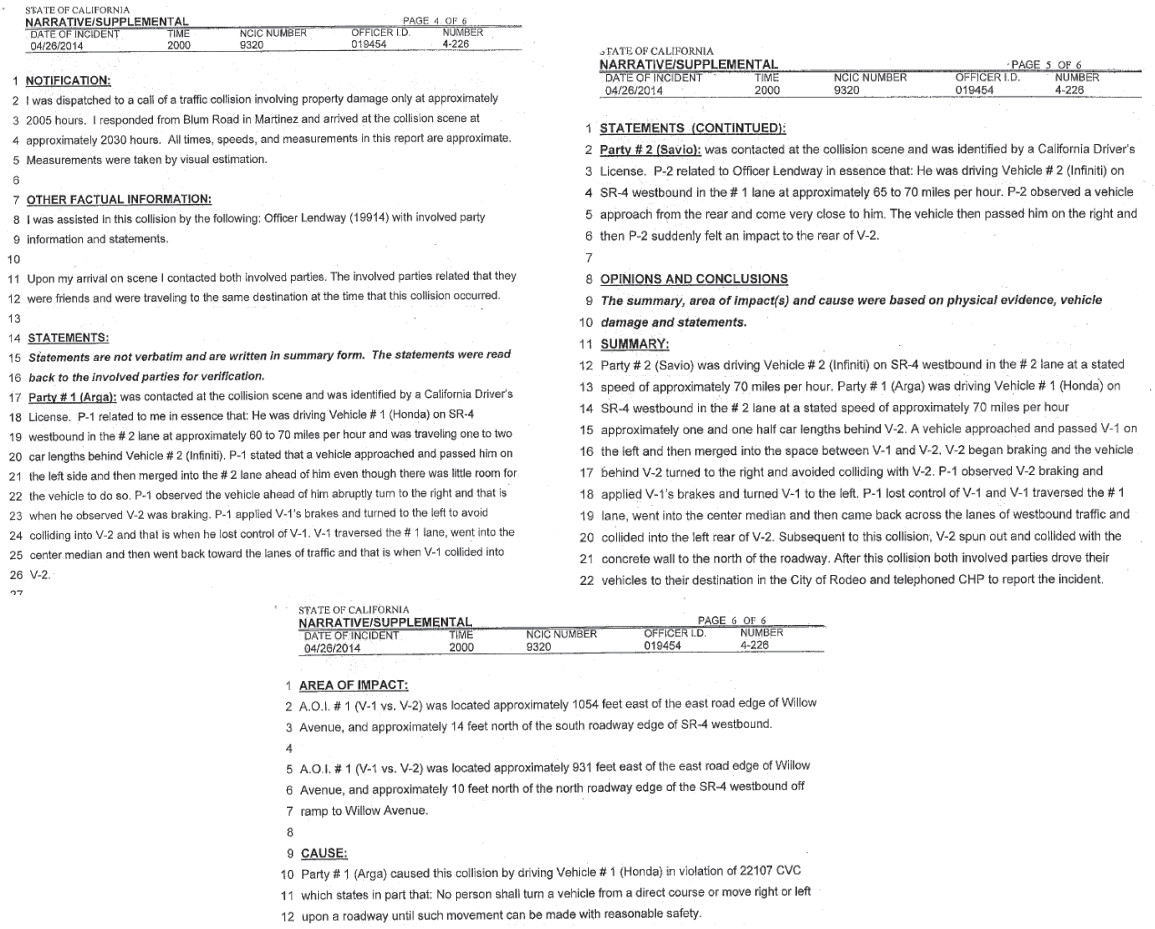
MISCELLANEOUS

Clear Diagram

An Internationally Accredited Agency

Chp555_0411.pdf

Figure 2.2: Page 2 of a TCR



STATE OF CALIFORNIA
NARRATIVE/SUPPLEMENTAL PAGE 4 OF 6

DATE OF INCIDENT	TIME	NCIC NUMBER	OFFICER I.D.	NUMBER
04/26/2014	2000	9320	019454	4-226

NOTIFICATION:
I was dispatched to a call of a traffic collision involving property damage only at approximately 2005 hours. I responded from Blum Road in Martinez and arrived at the collision scene at approximately 2030 hours. All times, speeds, and measurements in this report are approximate. Measurements were taken by visual estimation.

OTHER FACTUAL INFORMATION:
I was assisted in this collision by the following: Officer Lendway (19914) with involved party information and statements.

Upon my arrival on scene I contacted both involved parties. The involved parties related that they were friends and were traveling to the same destination at the time that this collision occurred.

STATEMENTS:
Statements are not verbatim and are written in summary form. The statements were read back to the involved parties for verification.
Party # 1 (Arga): was contacted at the collision scene and was identified by a California Driver's License. P-1 related to me in essence that: He was driving Vehicle # 1 (Honda) on SR-4 westbound in the # 2 lane at approximately 60 to 70 miles per hour and was traveling one to two car lengths behind Vehicle # 2 (Infiniti). P-1 stated that a vehicle approached and passed him on the left side and then merged into the # 2 lane ahead of him even though there was little room for the vehicle to do so. P-1 observed the vehicle ahead of him abruptly turn to the right and that is when he observed V-2 was braking. P-1 applied V-1's brakes and turned to the left to avoid colliding into V-2 and that is when he lost control of V-1. V-1 traversed the # 1 lane, went into the center median and then went back toward the lanes of traffic and that is when V-1 collided into V-2.

STATE OF CALIFORNIA
NARRATIVE/SUPPLEMENTAL PAGE 5 OF 6

DATE OF INCIDENT	TIME	NCIC NUMBER	OFFICER I.D.	NUMBER
04/26/2014	2000	9320	019454	4-226

STATEMENTS (CONTINTUED):
Party # 2 (Savio): was contacted at the collision scene and was identified by a California Driver's License. P-2 related to Officer Lendway in essence that: He was driving Vehicle # 2 (Infiniti) on SR-4 westbound in the # 1 lane at approximately 65 to 70 miles per hour. P-2 observed a vehicle approach from the rear and come very close to him. The vehicle then passed him on the right and then P-2 suddenly felt an impact to the rear of V-2.

OPINIONS AND CONCLUSIONS
The summary, area of impact(s) and cause were based on physical evidence, vehicle damage and statements.
SUMMARY:
Party # 2 (Savio) was driving Vehicle # 2 (Infiniti) on SR-4 westbound in the # 2 lane at a stated speed of approximately 70 miles per hour. Party # 1 (Arga) was driving Vehicle # 1 (Honda) on SR-4 westbound in the # 2 lane at a stated speed of approximately 70 miles per hour approximately one and one half car lengths behind V-2. A vehicle approached and passed V-1 on the left and then merged into the space between V-1 and V-2. V-2 began braking and the vehicle behind V-2 turned to the right and avoided colliding with V-2. P-1 observed V-2 braking and applied V-1's brakes and turned V-1 to the left. P-1 lost control of V-1 and V-1 traversed the # 1 lane, went into the center median and then came back across the lanes of westbound traffic and collided into the left rear of V-2. Subsequent to this collision, V-2 spun out and collided with the concrete wall to the north of the roadway. After this collision both involved parties drove their vehicles to their destination in the City of Rodeo and telephoned CHP to report the incident.

STATE OF CALIFORNIA
NARRATIVE/SUPPLEMENTAL PAGE 6 OF 6

DATE OF INCIDENT	TIME	NCIC NUMBER	OFFICER I.D.	NUMBER
04/26/2014	2000	9320	019454	4-226

AREA OF IMPACT:
A.O.I. # 1 (V-1 vs. V-2) was located approximately 1054 feet east of the east road edge of Willow Avenue, and approximately 14 feet north of the south roadway edge of SR-4 westbound.

A.O.I. # 1 (V-1 vs. V-2) was located approximately 931 feet east of the east road edge of Willow Avenue, and approximately 10 feet north of the north roadway edge of the SR-4 westbound off ramp to Willow Avenue.

CAUSE:
Party # 1 (Arga) caused this collision by driving Vehicle # 1 (Honda) in violation of 22107 CVC which states in part that: No person shall turn a vehicle from a direct course or move right or left upon a roadway until such movement can be made with reasonable safety.

Figure 2.3: An Example of the Narrative Section of a TCR

The Workflow in eTCR Processing for TASAS

In an effort to illustrate the functionality of the TASAS processing method considered in this research study, it is important to describe the workflow in processing eTCRs for coding the TASAS accident database.

Once the reporting Police Officer submits an eTCR, CHP processes the data and the majority of fields are put into the SWITRS database where the field data content (narratives and diagrams) are entered. This data, along with the eTCR in pdf format, are then transferred to Caltrans headquarters where they are further processed by the Caltrans Coding Group. Card 8 is a reference to a past form used by Caltrans when it was desired to have more detailed information in a database for monitoring traffic and safety performance. The TASAS accident database is the culmination of all these efforts. The TASAS accident database will be discussed in more detail later in this report.

Collision coding is a process in which the exact location (postmile) of the collision is identified using the information stated in the TCR. The area of impact (AOI) on the

roadway is identified using several resources including use of Highway Sequence Listings, maps, etc. The goal of the coding personnel in extracting the AOI is to be as accurate as possible, because this location data is the building block for safety data used by transportation and safety engineers in implementing improvements on the state highways.

Postmiles are the values assigned to the highway and they typically increase in the direction of the highway from west to east and from south to north (there are seven highways which are an exception for which the directions are reversed). Postmiles start at the point where a route begins, or when highway enters a new county and ends with the end of the route or the end of the county. For example: when Interstate 5 (I-5), which is a south to north route, enters Sacramento County, the postmile begins at 0.000. These postmiles increase as the highway proceeds north. All structures; i.e. under-crossings, overcrossings, and intersections, etc., have valid postmiles.

The source of data determined by Caltrans for the TASAS accident database contains specific collision location information, such as the presence of roadway features and geometry of the highway.

In a TCR, the reporting officer describes in words and diagrams where the collision occurred. Caltrans coders convert the information on the location into postmile values. A specific protocol is followed to translate the TCR information into precise location on a highway to ensure repeatability. The TCRPRO program attempts to emulate this set of protocols. The information extracted to describe the location is referred to “Card 8a” by the Coding Group. This name is based on how the reports were processed in past years.

The second set of data required by Caltrans is referred to as “Card 8b” or “Sequence of Events”. This set of data is party based and essentially describes what vehicle hit what and where. It also includes some sort of descriptors on the type of parties involved in the collision itself. An example of the Card 8a and Card 8b is shown in Figure 2.4. The top half of the form shown in Figure 2.4 has designated spaces for the employee to write down the postmile value, route and direction. The bottom half also has designated spaces for the sequence of events information. The form is designed for accurate data entry procedures that follow the coding process.

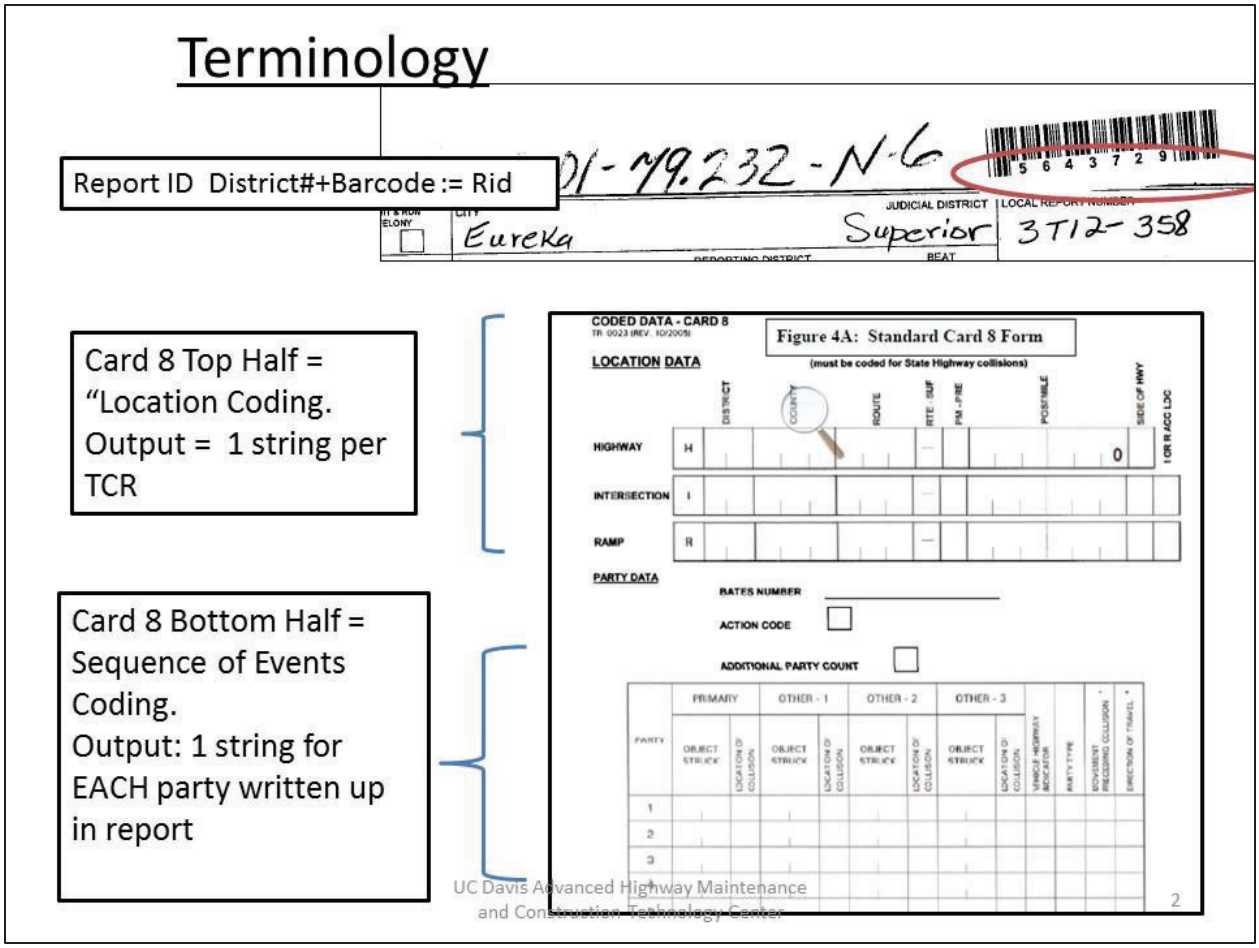


Figure 2.4: Traffic Collision Report Coding Card 8a and Card 8b

Input Obtained from the TCR

The following areas of a TCR have information that are extracted as input to the TASAS accident database:

The Location Box

- This is where the officer describes in words the location of the collision. Direction of Travel is sometimes taken from the location box or it can be found in Part 1 section of a TCR.

- Please note, that although there is a checkbox noting whether the collision occurred at an intersection or not; this box is not regularly checked even when an intersection is involved.

- Furthermore, the Global Positioning System (GPS) values denoted by the officer in a TCR are not always accurate and not necessarily indicative of the actual collision site.

Diagrams

- There can be any number of diagrams provided in the TCR ranging from zero to any needed amount. Straightforward collisions typically only have one diagram whereas complicated collisions involving multiple vehicles will require many diagrams.

In addition, the following information are also extracted from other areas of a TCR:

Area of Impact (AOI)

- This is described in terms of the distance and landmarks by the CHP officer and typically corresponds with the collision diagrams in the TCR. If there is more than one AOI then only the first one is used for the location coding.

Object Struck

- The object struck codes identify any type of impact or event which contributed to damage to a vehicle or property or injury to a person or domestic animal.
- The card 8b allows for four or less consecutive object struck events per party. If there are more objects struck than the coding sheet allows, the first object struck is coded and then the most major damage creating events are coded.
- In multi-car collisions, only the first nine parties are coded for highways on the Card 8.

Sequence Listings

The Coding Group use a set of documents referred to as “Sequence Listing” which provide the relationship between any location on a California state highway with a corresponding postmile marker value. Each postmile marker contains the county, route number and suffix, PMM prefix (if any) and the linear distance along the highway in terms of miles in a 000.000 format. There is a great deal of information in Sequence Listings and more detailed information can be found in the training manual for the postmile system Geographic Information System (GIS) training guide.

The Sequence Listing is related to the highway database which contains all the “descriptors” that pertain to the freeway at each point along all California highways. The same landmark denotations are present in both the sequence listings and highway database stored in Caltrans Transportation System Network (TSN). An example of a Sequence Listing is shown in Figure 2.5.It can be seen in Figure 2.5 that the columns of data are: County, City, postmile value, highway group, file type, and landmark description. Each page of the Sequence Listing references a designated route and a District, which is listed at the top of each Sequence Listing page.

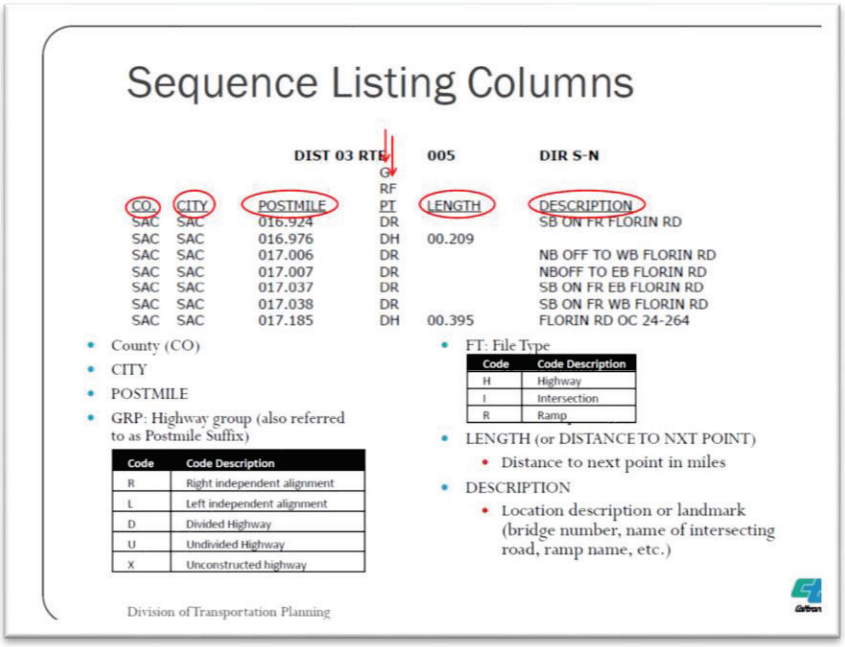


Figure 2.5: Sequence Listings.

TASAS Accident Database

The final objective for coding the Traffic Collision Report for Caltrans purposes is to put this coded data into the TASAS accident database. For each collision report, there is a set of "Collision Data" and a set of "Party Data". The collision data contains the following fields:

Year of the collision, Caltrans District, county, postmile info, highway, accident controls, median type, barrier type, num. lanes left/right, population code, file-type, intersection-ramp, side of highway, day of the week, date, time, accident number, primary collision factor, weather, lighting, road surface, rd. condition, right of way controls, type of collision, and the number vehicles involved.

For the Party Data, the fields that are included are:

Party: Party type, direction of travel, vehicle highway indicator, special information, num. persons killed, number of persons injured, primary object struck, location of primary object struck, other-a object struck, other-a-location, object/location for other-b, object/location for object-c, other attributing factor(1 and 2), movement preceding collision, sobriety/drug/

A snapshot of a TASAS printout can be seen in Figure 2.6 below. The column headings are very short abbreviations for the data fields but follow a specific protocol. The left-hand portion of the printout contains the collision data where the right-hand portion contains the party data. Typically, the TASAS data is analyzed and viewed using computer applications such as Excel and Access where the data fields are easily identified.

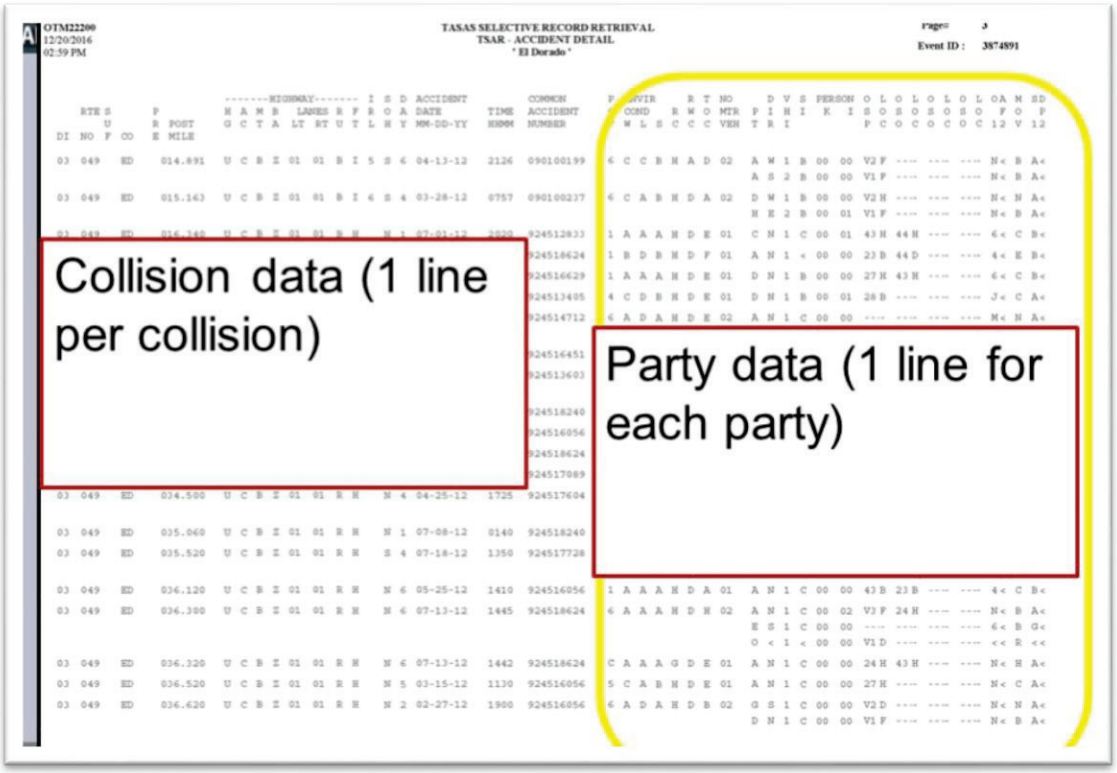
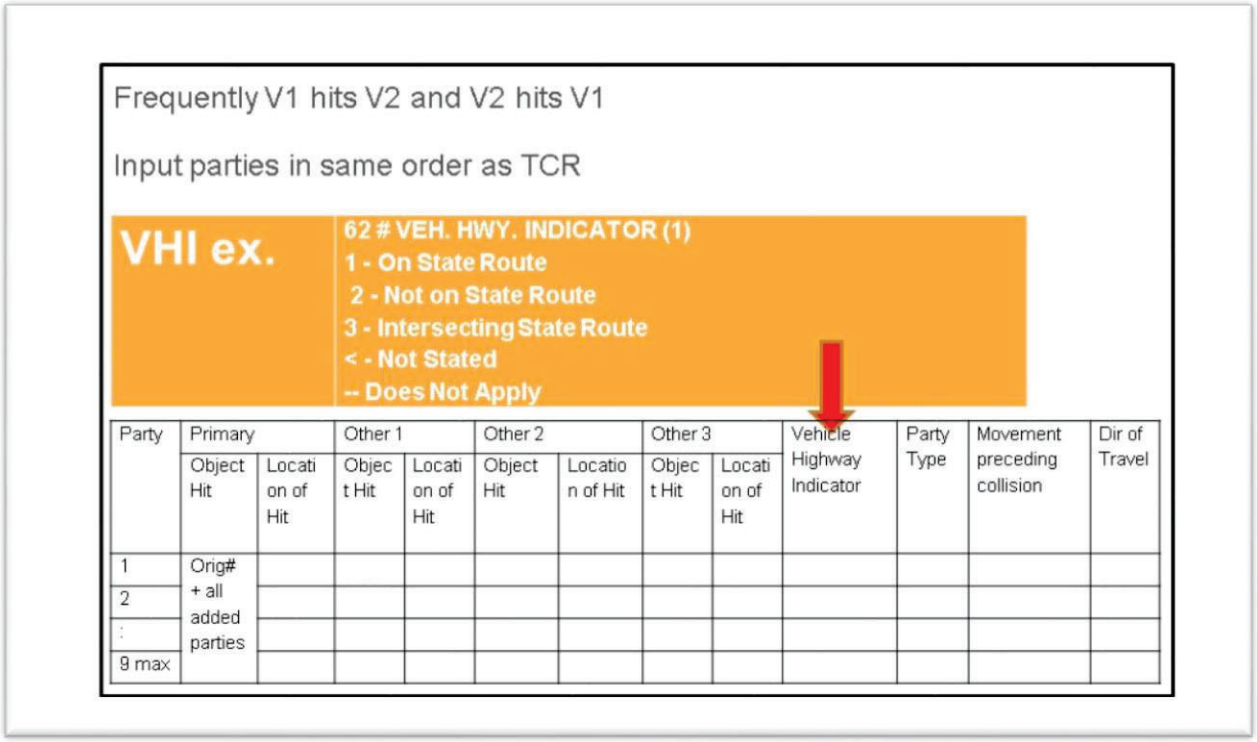


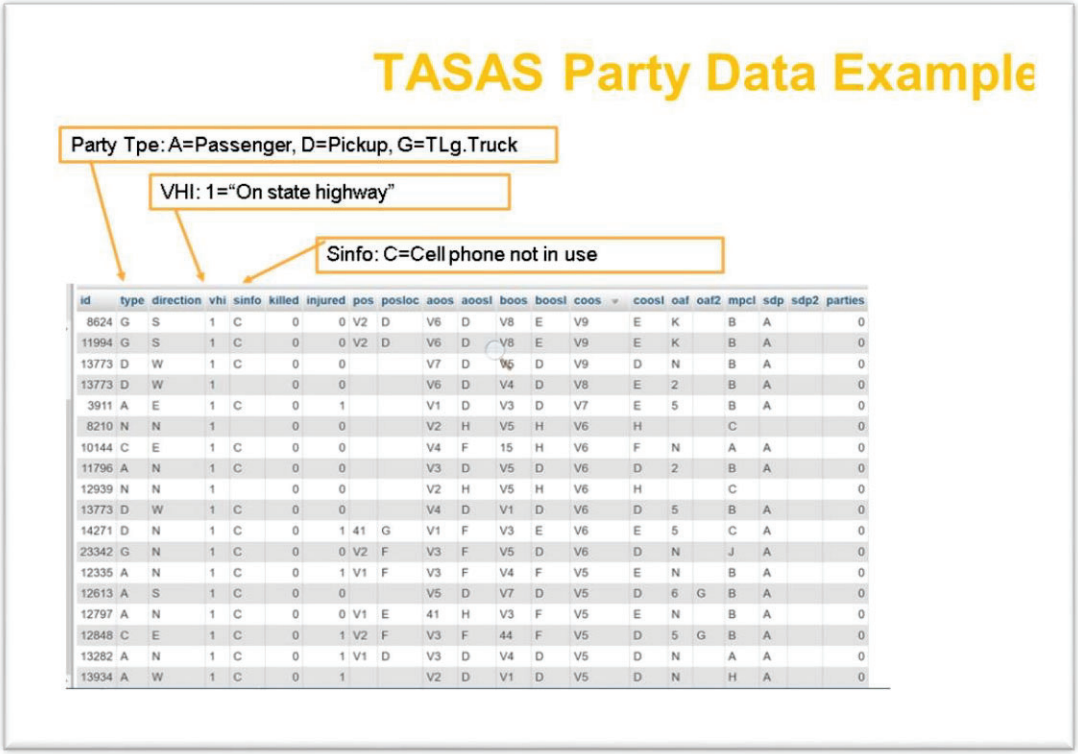
Figure 2.6: Sample printout of the TASAS Database with the collision data on the left-hand side and the party data on the right-hand side.

A close up of the Sequence of Events information in the TASAS table from Figure 2.6 is shown in Figure 2.7. Within this form, the coder must enter the Party Number, and the "Primary" object that was hit and where on the highway the impact occurred. If "Other" collisions occurred that the party was involved with, these are entered in the other 3 segments capturing, the "Sequence of Events". Additional party data must also be entered for the Vehicle Highway Indicator, the movement preceding the collision, the Party Type and the direction of travel for that particular party. An alternative format for the party data can be seen in Figure 2.8. Here the party data is listed with a single line for each party and the "ID" represents the collision identification method. It can be seen in Figure 2.8 the listed fields represent from left to right as, the collision identification, party type, direction of travel, vehicle highway indicator, special information, number people killed, number of people injured, primary object struck, location of the primary collision, then three sets of object struck and location, followed by other attributing factors to the collision such as under the influence of alcohol or drugs. Figure 2.9 shows the legend for the codes used in the party data such as "Other Associated Factors", how the "location" of a collision is coded and how drug/alcohol involvement is coded.



Party	Primary		Other 1		Other 2		Other 3		Vehicle Highway Indicator	Party Type	Movement preceding collision	Dir of Travel
	Object Hit	Location of Hit	Object Hit	Location of Hit	Object Hit	Location of Hit	Object Hit	Location of Hit				
1	Orig# + all added parties											
2												
:												
9 max												

Figure 2.7: Close up of the Sequence of Events information in the Party data in the TASAS database.



id	type	direction	vhi	sinfo	killed	injured	pos	posloc	aoos	aoosl	boos	boosl	coos	coosl	oaf	oaf2	mpcl	sdp	sdp2	parties
8624	G	S	1	C	0	0	V2	D	V6	D	V8	E	V9	E	K		B	A		0
11994	G	S	1	C	0	0	V2	D	V6	D	V8	E	V9	E	K		B	A		0
13773	D	W	1	C	0	0			V7	D	V5	D	V9	D	N		B	A		0
13773	D	W	1		0	0			V6	D	V4	D	V8	E	2		B	A		0
3911	A	E	1	C	0	1			V1	D	V3	D	V7	E	5		B	A		0
8210	N	N	1		0	0			V2	H	V5	H	V6	H			C			0
10144	C	E	1	C	0	0			V4	F	15	H	V6	F	N		A	A		0
11796	A	N	1	C	0	0			V3	D	V5	D	V6	D	2		B	A		0
12939	N	N	1		0	0			V2	H	V5	H	V6	H			C			0
13773	D	W	1	C	0	0			V4	D	V1	D	V6	D	5		B	A		0
14271	D	N	1	C	0	1	41	G	V1	F	V3	E	V6	E	5		C	A		0
23342	G	N	1	C	0	0	V2	F	V3	F	V5	D	V6	D	N		J	A		0
12335	A	N	1	C	0	1	V1	F	V3	F	V4	F	V5	E	N		B	A		0
12613	A	S	1	C	0	0			V5	D	V7	D	V5	D	6	G	B	A		0
12797	A	N	1	C	0	0	V1	E	41	H	V3	F	V5	E	N		B	A		0
12848	C	E	1	C	0	1	V2	F	V3	F	44	F	V5	D	5	G	B	A		0
13282	A	N	1	C	0	1	V1	D	V3	D	V4	D	V5	D	N		A	A		0
13934	A	W	1	C	0	1			V2	D	V1	D	V5	D	N		H	A		0

Figure 2.8: Example TASAS party data

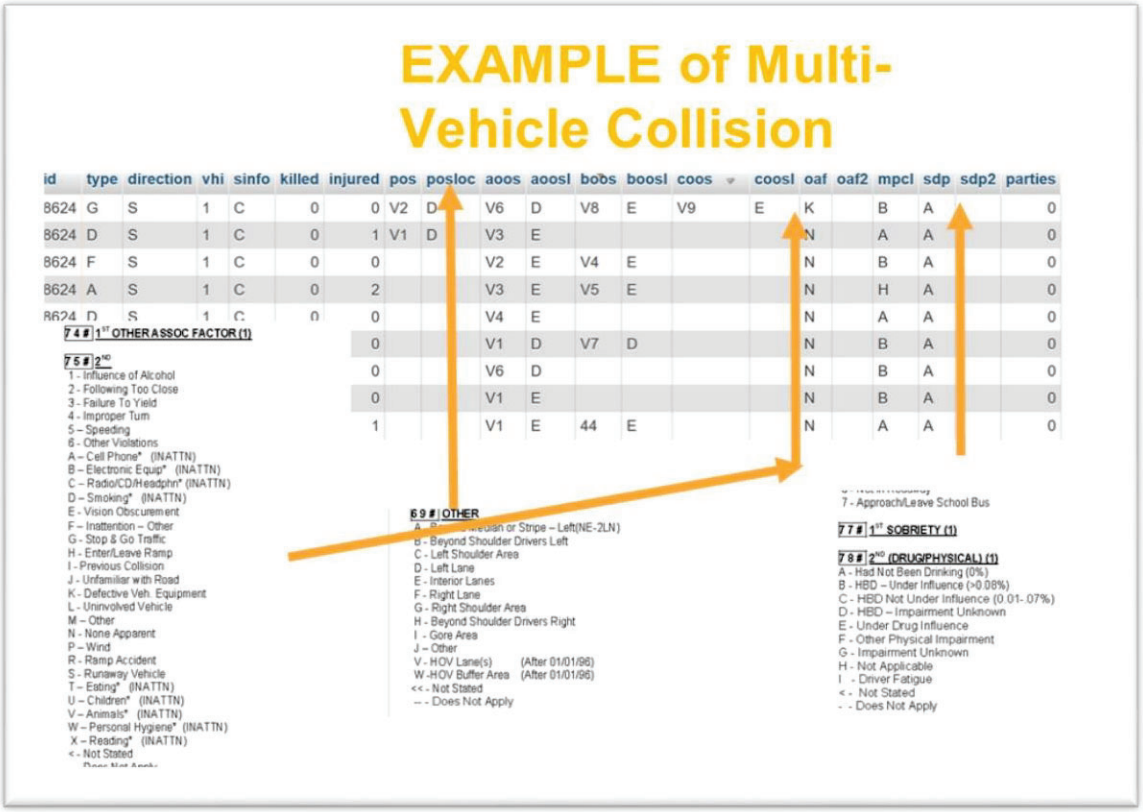


Figure 2.9: Example TASAS party data where two vehicles are involved

CHAPTER 3: TRAFFIC COLLISION REPORT PROCESSING SOFTWARE

The TCRPRO Application

Overview

TCRPRO developed in this research study is a self-contained software application that provides various TCR-processing services. It is designed to be easily deployed into the Caltrans IT infrastructure to perform TCR processing tasks involved in populating the TASAS database. The integration into Caltrans IT infrastructure is done primarily by means of a web-services Application Programming Interface (API). Services provided include location coding, sequence-of-events coding, narrative summary extraction, and more. System requirements are summarized in Table 3.1.

Table 3.1: TCRPRO System Requirements

Component	Requirement
System Architecture	x86-64
Operating System	Red Hat 7 or CentOS 7
Java	Server JRE 8
Python	3
Apache HTTP	2.4
Apache Tomcat	7
PostgreSQL	9.6

The usage model envisioned during development is one in which TCR PDF files, upon receipt by Caltrans from CHP systems, are transmitted to TCRPRO by a Caltrans TSN web-services client. The results of the TCR processing would then be received from TCRPRO by this client and stored in the TSN database. This model is illustrated in Figure 3.1. This figure shows that CHP web service will communicate with TSN web service client. The TSN web service client then works with TCRPRO and TSN collision database and performs the appropriate coding functions.

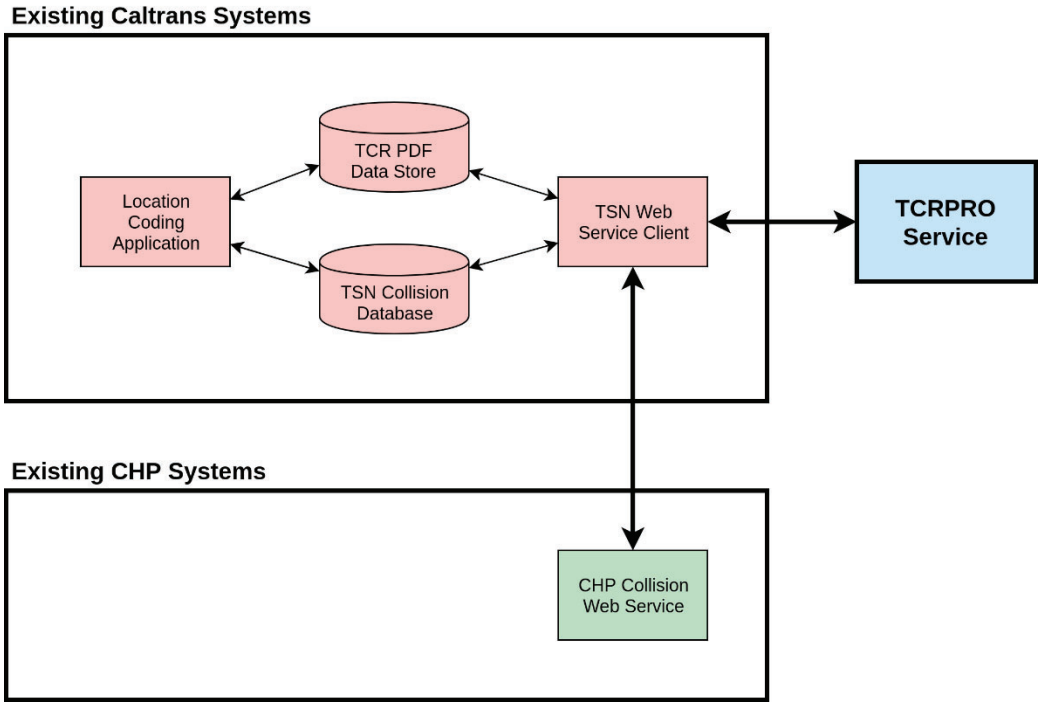


Figure 3.1: TCRPRO Service Integration

TCRPRO Architecture

TCRPRO is designed to be a web-service-centric application, and relies on Apache HyperText Transfer Protocol (HTTP) and Apache Tomcat for HTTP request handling. An Apache HTTP instance handles all web-service requests to TCRPRO components. The outward-facing TCRPRO API service, as well as the postmile referencing service, are implemented as Java servlets, hosted by Apache Tomcat, proxied by Apache HTTP. Python components of the system are implemented as internal Web Server Gateway Interface (WSGI) services hosted via mod_wsgi. A Post-Geographic Information System (GIS) database houses a roadway information database used by the system's postmile referencing service. This configuration is illustrated in Figure 3.2.

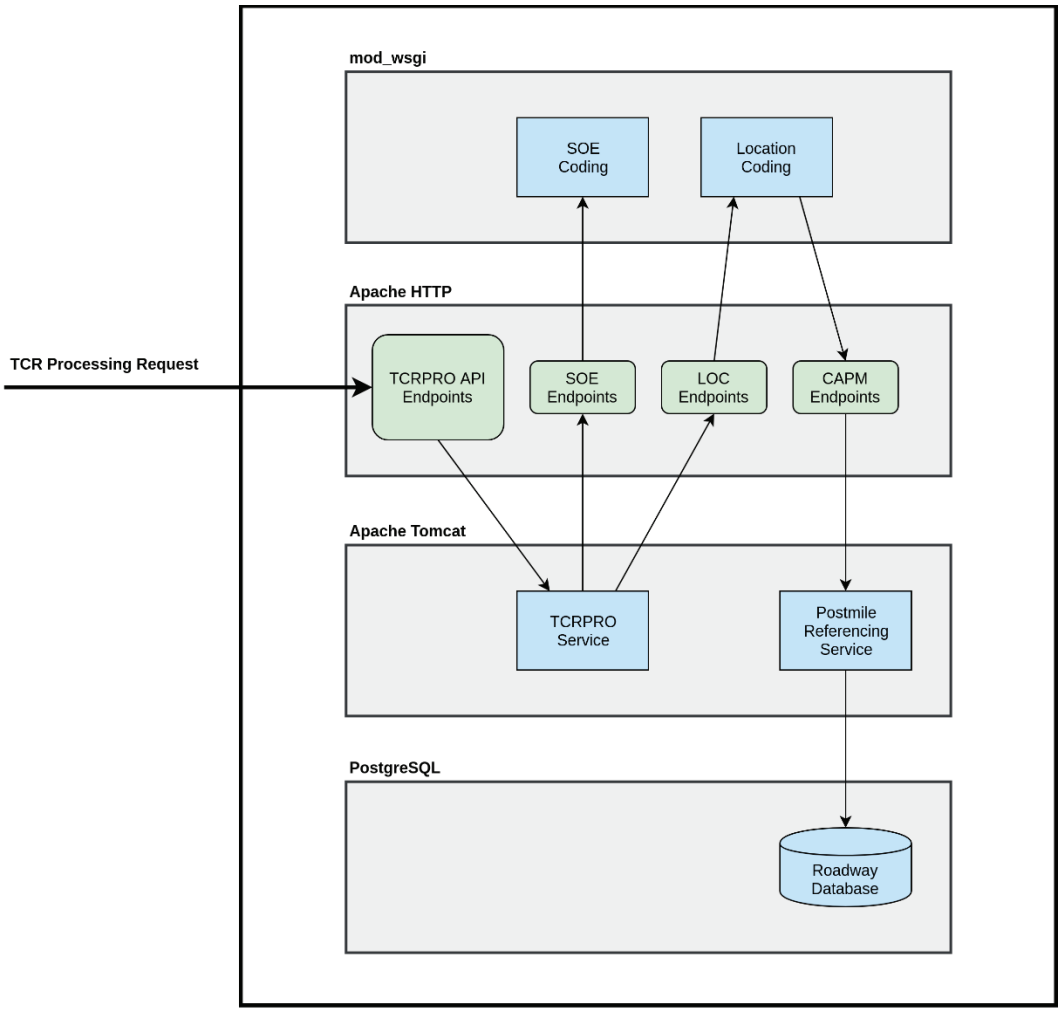


Figure 3.2: TCRPRO Architecture

TCRPRO Source Code

The TCRPRO source code is comprised of several discrete packages, summarized in Table 3.2.

Table 3.2: TCRPRO Source Packages

Package	Description	Language	Lines of Code
`formutil`	General-purpose PDF form processing library	Java 8	1,801
`formconfig`	formutil configuration tool	Java 8	856
`tcrpro`	TCR processing library	Java 8	16,180
`wsgigw`	WSGI gateway	Python 3	225
`libcoding`	location/SOE coding library	Python 3	4,214
`libcapm`	postmile referencing client library	Python 3	1,902
`tcpproapi`	TCRPRO web services API	Java 8	1,315
`tcrextract`	TCR extraction utility	Java 8	1,416
`capmsvc`	postmile referencing service	Java 8	5,217
`assets`	static assets, configuration, etc.	n/a	0
`bdutil`	build and deployment utilities	Bash	1,659

TCR PDF Extraction

TCRPRO accepts TCR form documents that are encoded in PDF format. TCR field data is extracted from the documents, then interpreted and validated by TCRPRO before being used for Location and Sequence of Events coding purposes.

The primary challenge in successfully processing TCRs lies in handling the many variations of these documents that are encountered. While many TCR pages may at first appear to originate from standardized forms with identical layout and geometry, in reality they originate from a family of forms, each similar to the others but possessing significant, yet often subtle, differences. This situation complicates the task of determining the precise location of the form fields, entailing a more sophisticated approach that might be expected.

At the highest level, the TCR extraction process is composed of four stages as follows:

1. PDF Page Extraction
2. Page Type/Variant Detection
3. Field Extraction
4. Semantic Validation

These four stages are illustrated in Figure 3.3 where it is shown how pre-designed page masks are used to extract information from a PDF document.

As input, the module consumes a PDF TCR document. It produces, as output, an object containing the data extracted from the TCR and a set of functions to query that data. A log of any warnings that occurred during the extraction is also produced.

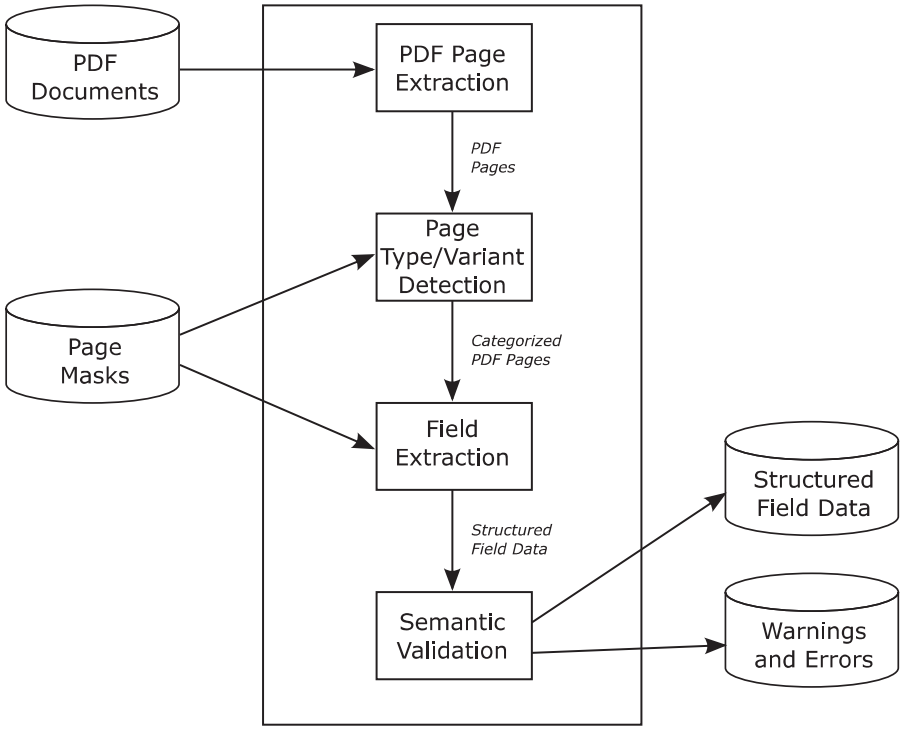


Figure 3.3: TCRPRO PDF Extraction Data Flow

PDF Page Extraction

During the PDF page extraction stage, the PDF object is read and interpreted into a PDF document object model. References to each page object are then passed down to the next stage, which is the Page Type/Variant Detection stage.

Page Type/Variant Detection

During the Page Type/Variant Detection stage, the contents of each page are compared to the set of known TCR form page variants in order to determine the page's type (e.g., CHP 555 Page 2), as well as the particular variant of that type. Most TCR page types have multiple variants in circulation and accurate TCR data extraction cannot take place unless TCRPRO has been preconfigured with a form template for the particular form page variant used by the submitted TCR.

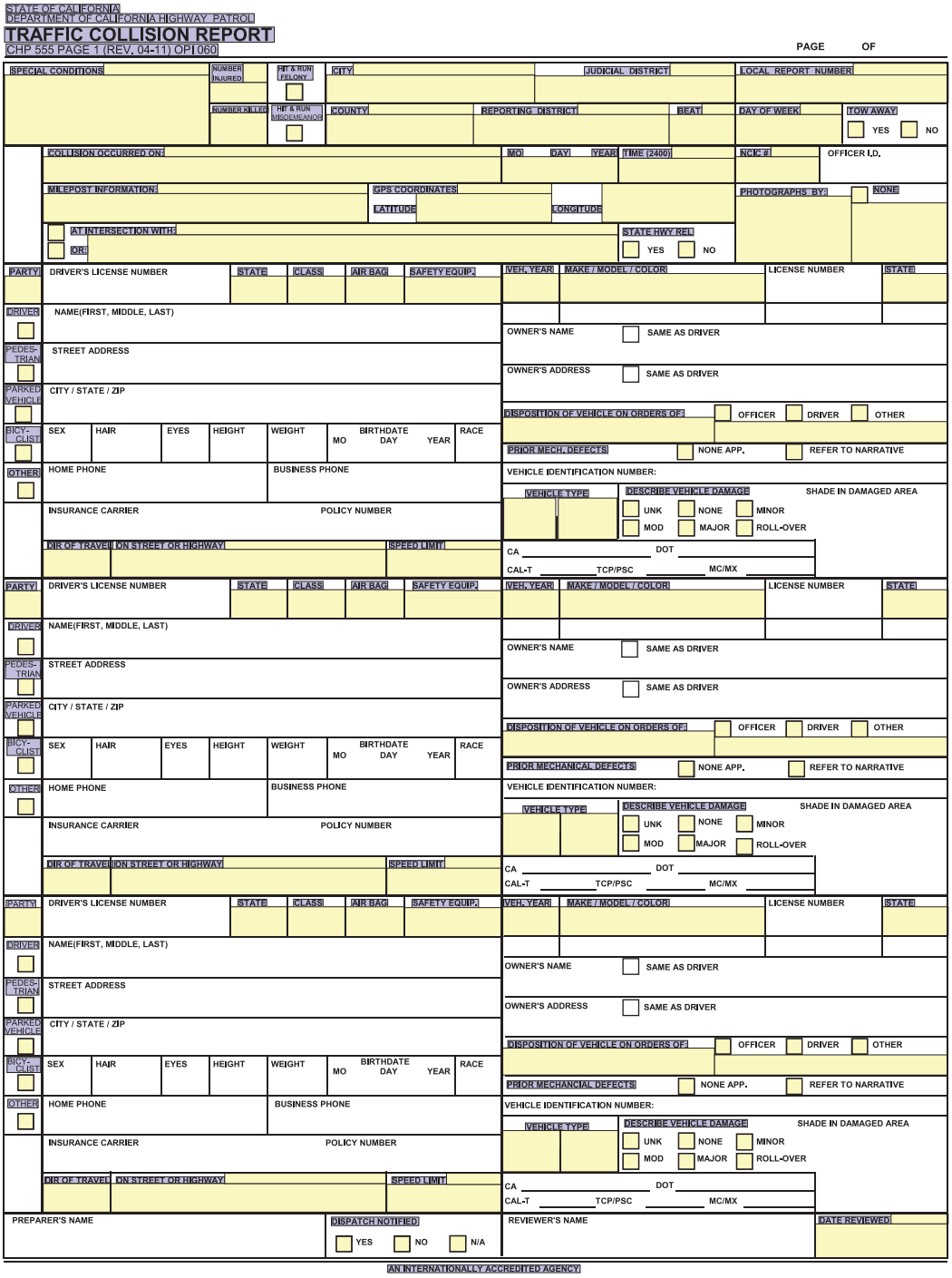
Once each page type and variant is identified (or deemed unidentifiable), the pages are categorized, grouped, and passed down to the next stage, which is the Field Extraction stage.

Field Extraction

The purpose of the Field Extraction stage is to extract the dynamic textual content from each page. For the TCR forms typically handled by Caltrans, this dynamic textual content is represented as a simple text layer superimposed onto a static form template. Because of this relatively unstructured representation, we must make use of the geometry of the textual elements to determine the form element to which each element should be associated. The textual geometry is also used to accurately infer word boundaries.

To accomplish this, AHMCT has developed form templates for the most common variants of TCR pages that are handled by Caltrans. These form templates are, in essence, annotated page masks consisting of a set of bounding boxes, each of which corresponds to an area of the page where text that is relevant to our purposes is expected to be found (see Figure 3.4 and Figure 3.5 for examples). Each bounding box in Figures 3.4 and 3.5 is annotated with an ID which maps it to its respective form component. In this way, the text found within any particular bounding box may be correlated with the correct form component, enabling page/variant detection and accurate field text extraction.

At the end of the Field Extraction stage, after all text content found within the page mask's bounding boxes is extracted and associated with its respective form component, the extracted information is then organized into a data structure and passed down to the next stage, which is the *Semantic Validation stage*.


STATE OF CALIFORNIA
DEPARTMENT OF CALIFORNIA HIGHWAY PATROL
TRAFFIC COLLISION REPORT
CHP 555 PAGE 1 (REV. 04-11) OPI 060

PAGE OF

SPECIAL CONDITIONS | NUMBER INJURED | HIT & RUN FELONY | CITY | JUDICIAL DISTRICT | LOCAL REPORT NUMBER
NUMBER KILLED | HIT & RUN MISDEMEANOR | COUNTY | REPORTING DISTRICT | BEAT | DAY OF WEEK | TOW AWAY YES NO
COLLISION OCCURRED ON: | MO DAY YEAR | TIME (2400) | NCIC # | OFFICER I.D.
MILEPOST INFORMATION: | GPS COORDINATES LATITUDE LONGITUDE | PHOTOGRAPHS BY: NONE
AT INTERSECTION WITH: OR: | STATE HWY REL YES NO

PARTY | DRIVER'S LICENSE NUMBER | STATE | CLASS | AIR BAG | SAFETY EQUIP. | VEH. YEAR | MAKE / MODEL / COLOR | LICENSE NUMBER | STATE
DRIVER | NAME(FIRST, MIDDLE, LAST) | OWNER'S NAME SAME AS DRIVER
PEDES-TRIAN | STREET ADDRESS | OWNER'S ADDRESS SAME AS DRIVER
PARKED VEHICLE | CITY / STATE / ZIP | DISPOSITION OF VEHICLE ON ORDERS OF: OFFICER DRIVER OTHER
BICY-CLIST | SEX | HAIR | EYES | HEIGHT | WEIGHT | BIRTHDATE MO DAY YEAR | RACE | PRIOR MECH. DEFECTS NONE APP. REFER TO NARRATIVE
OTHER | HOME PHONE | BUSINESS PHONE | VEHICLE IDENTIFICATION NUMBER:
INSURANCE CARRIER | POLICY NUMBER | VEHICLE TYPE | DESCRIBE VEHICLE DAMAGE UNK NONE MINOR MOD MAJOR ROLL-OVER | SHADE IN DAMAGED AREA
DIR OF TRAVEL | ON STREET OR HIGHWAY | SPEED LIMIT | CA DOT | CAL-T TCP/PSC MC/MX

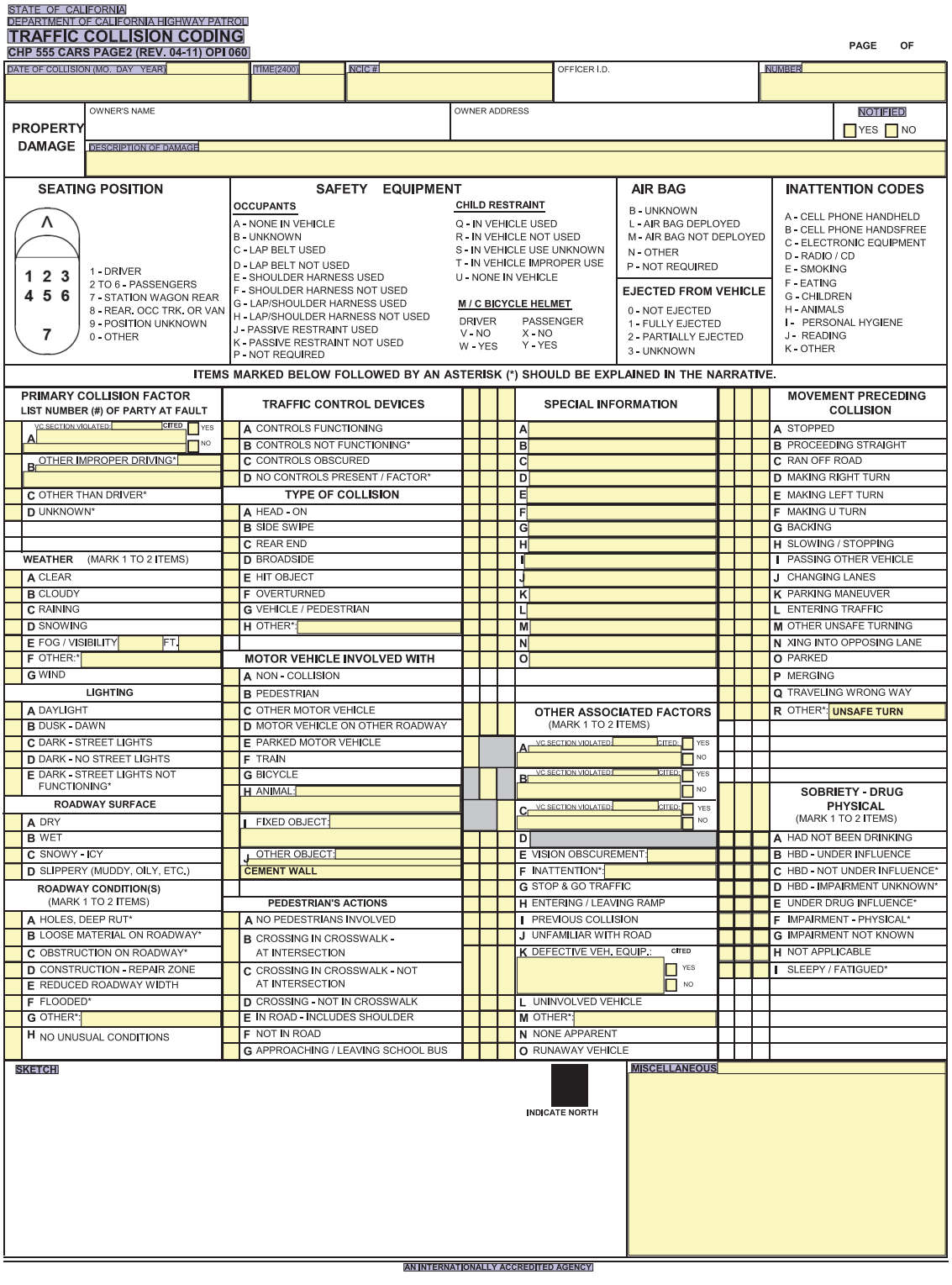
PARTY | DRIVER'S LICENSE NUMBER | STATE | CLASS | AIR BAG | SAFETY EQUIP. | VEH. YEAR | MAKE / MODEL / COLOR | LICENSE NUMBER | STATE
DRIVER | NAME(FIRST, MIDDLE, LAST) | OWNER'S NAME SAME AS DRIVER
PEDES-TRIAN | STREET ADDRESS | OWNER'S ADDRESS SAME AS DRIVER
PARKED VEHICLE | CITY / STATE / ZIP | DISPOSITION OF VEHICLE ON ORDERS OF: OFFICER DRIVER OTHER
BICY-CLIST | SEX | HAIR | EYES | HEIGHT | WEIGHT | BIRTHDATE MO DAY YEAR | RACE | PRIOR MECHANICAL DEFECTS NONE APP. REFER TO NARRATIVE
OTHER | HOME PHONE | BUSINESS PHONE | VEHICLE IDENTIFICATION NUMBER:
INSURANCE CARRIER | POLICY NUMBER | VEHICLE TYPE | DESCRIBE VEHICLE DAMAGE UNK NONE MINOR MOD MAJOR ROLL-OVER | SHADE IN DAMAGED AREA
DIR OF TRAVEL | ON STREET OR HIGHWAY | SPEED LIMIT | CA DOT | CAL-T TCP/PSC MC/MX

PARTY | DRIVER'S LICENSE NUMBER | STATE | CLASS | AIR BAG | SAFETY EQUIP. | VEH. YEAR | MAKE / MODEL / COLOR | LICENSE NUMBER | STATE
DRIVER | NAME(FIRST, MIDDLE, LAST) | OWNER'S NAME SAME AS DRIVER
PEDES-TRIAN | STREET ADDRESS | OWNER'S ADDRESS SAME AS DRIVER
PARKED VEHICLE | CITY / STATE / ZIP | DISPOSITION OF VEHICLE ON ORDERS OF: OFFICER DRIVER OTHER
BICY-CLIST | SEX | HAIR | EYES | HEIGHT | WEIGHT | BIRTHDATE MO DAY YEAR | RACE | PRIOR MECHANICAL DEFECTS NONE APP. REFER TO NARRATIVE
OTHER | HOME PHONE | BUSINESS PHONE | VEHICLE IDENTIFICATION NUMBER:
INSURANCE CARRIER | POLICY NUMBER | VEHICLE TYPE | DESCRIBE VEHICLE DAMAGE UNK NONE MINOR MOD MAJOR ROLL-OVER | SHADE IN DAMAGED AREA
DIR OF TRAVEL | ON STREET OR HIGHWAY | SPEED LIMIT | CA DOT | CAL-T TCP/PSC MC/MX

PREPARER'S NAME | DISPATCH NOTIFIED YES NO N/A | REVIEWER'S NAME | DATE REVIEWED

AN INTERNATIONALLY ACCREDITED AGENCY

Figure 3.4: An example CHP 555 Page 1 page mask



STATE OF CALIFORNIA
DEPARTMENT OF CALIFORNIA HIGHWAY PATROL
TRAFFIC COLLISION CODING
CHP 555 CARS PAGE2 (REV. 04-11) OPI 060

PAGE OF

DATE OF COLLISION (MO. DAY YEAR) | TIME(2400) | NCIC # | OFFICER I.D. | NUMBER

PROPERTY DAMAGE — OWNER'S NAME | OWNER ADDRESS | NOTIFIED ☐ YES ☐ NO
DESCRIPTION OF DAMAGE

SEATING POSITION
1 2 3
4 5 6
7
1 - DRIVER
2 TO 6 - PASSENGERS
7 - STATION WAGON REAR
8 - REAR, OCC TRK, OR VAN
9 - POSITION UNKNOWN
0 - OTHER

SAFETY EQUIPMENT
OCCUPANTS
A - NONE IN VEHICLE
B - UNKNOWN
C - LAP BELT USED
D - LAP BELT NOT USED
E - SHOULDER HARNESS USED
F - SHOULDER HARNESS NOT USED
G - LAP/SHOULDER HARNESS USED
H - LAP/SHOULDER HARNESS NOT USED
J - PASSIVE RESTRAINT USED
K - PASSIVE RESTRAINT NOT USED
P - NOT REQUIRED

CHILD RESTRAINT
Q - IN VEHICLE USED
R - IN VEHICLE NOT USED
S - IN VEHICLE USE UNKNOWN
T - IN VEHICLE IMPROPER USE
U - NONE IN VEHICLE

M / C BICYCLE HELMET
DRIVER: V - NO, W - YES
PASSENGER: X - NO, Y - YES

AIR BAG
B - UNKNOWN
L - AIR BAG DEPLOYED
M - AIR BAG NOT DEPLOYED
N - OTHER
P - NOT REQUIRED

EJECTED FROM VEHICLE
0 - NOT EJECTED
1 - FULLY EJECTED
2 - PARTIALLY EJECTED
3 - UNKNOWN

INATTENTION CODES
A - CELL PHONE HANDHELD
B - CELL PHONE HANDSFREE
C - ELECTRONIC EQUIPMENT
D - RADIO / CD
E - SMOKING
F - EATING
G - CHILDREN
H - ANIMALS
I - PERSONAL HYGIENE
J - READING
K - OTHER

ITEMS MARKED BELOW FOLLOWED BY AN ASTERISK (*) SHOULD BE EXPLAINED IN THE NARRATIVE.

PRIMARY COLLISION FACTOR
LIST NUMBER (#) OF PARTY AT FAULT
A VC SECTION VIOLATED: CITED ☐ YES ☐ NO
B OTHER IMPROPER DRIVING*
C OTHER THAN DRIVER*
D UNKNOWN*

WEATHER (MARK 1 TO 2 ITEMS)
A CLEAR
B CLOUDY
C RAINING
D SNOWING
E FOG / VISIBILITY FT.
F OTHER:*
G WIND

LIGHTING
A DAYLIGHT
B DUSK - DAWN
C DARK - STREET LIGHTS
D DARK - NO STREET LIGHTS
E DARK - STREET LIGHTS NOT FUNCTIONING*

ROADWAY SURFACE
A DRY
B WET
C SNOWY - ICY
D SLIPPERY (MUDDY, OILY, ETC.)

ROADWAY CONDITION(S) (MARK 1 TO 2 ITEMS)
A HOLES, DEEP RUT*
B LOOSE MATERIAL ON ROADWAY*
C OBSTRUCTION ON ROADWAY*
D CONSTRUCTION - REPAIR ZONE
E REDUCED ROADWAY WIDTH
F FLOODED*
G OTHER*:
H NO UNUSUAL CONDITIONS

TRAFFIC CONTROL DEVICES
A CONTROLS FUNCTIONING
B CONTROLS NOT FUNCTIONING*
C CONTROLS OBSCURED
D NO CONTROLS PRESENT / FACTOR*

TYPE OF COLLISION
A HEAD - ON
B SIDE SWIPE
C REAR END
D BROADSIDE
E HIT OBJECT
F OVERTURNED
G VEHICLE / PEDESTRIAN
H OTHER*:

MOTOR VEHICLE INVOLVED WITH
A NON - COLLISION
B PEDESTRIAN
C OTHER MOTOR VEHICLE
D MOTOR VEHICLE ON OTHER ROADWAY
E PARKED MOTOR VEHICLE
F TRAIN
G BICYCLE
H ANIMAL:
I FIXED OBJECT:
J OTHER OBJECT:
CEMENT WALL

PEDESTRIAN'S ACTIONS
A NO PEDESTRIANS INVOLVED
B CROSSING IN CROSSWALK - AT INTERSECTION
C CROSSING IN CROSSWALK - NOT AT INTERSECTION
D CROSSING - NOT IN CROSSWALK
E IN ROAD - INCLUDES SHOULDER
F NOT IN ROAD
G APPROACHING / LEAVING SCHOOL BUS

SPECIAL INFORMATION
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O

OTHER ASSOCIATED FACTORS (MARK 1 TO 2 ITEMS)
A VC SECTION VIOLATED: CITED: ☐ YES ☐ NO
B VC SECTION VIOLATED: CITED: ☐ YES ☐ NO
C VC SECTION VIOLATED: CITED: ☐ YES ☐ NO
D
E VISION OBSCUREMENT:
F INATTENTION*:
G STOP & GO TRAFFIC
H ENTERING / LEAVING RAMP
I PREVIOUS COLLISION
J UNFAMILIAR WITH ROAD
K DEFECTIVE VEH. EQUIP.: CITED ☐ YES ☐ NO
L UNINVOLVED VEHICLE
M OTHER*:
N NONE APPARENT
O RUNAWAY VEHICLE

MOVEMENT PRECEDING COLLISION
A STOPPED
B PROCEEDING STRAIGHT
C RAN OFF ROAD
D MAKING RIGHT TURN
E MAKING LEFT TURN
F MAKING U TURN
G BACKING
H SLOWING / STOPPING
I PASSING OTHER VEHICLE
J CHANGING LANES
K PARKING MANEUVER
L ENTERING TRAFFIC
M OTHER UNSAFE TURNING
N XING INTO OPPOSING LANE
O PARKED
P MERGING
Q TRAVELING WRONG WAY
R OTHER*: UNSAFE TURN

SOBRIETY - DRUG PHYSICAL (MARK 1 TO 2 ITEMS)
A HAD NOT BEEN DRINKING
B HBD - UNDER INFLUENCE
C HBD - NOT UNDER INFLUENCE*
D HBD - IMPAIRMENT UNKNOWN*
E UNDER DRUG INFLUENCE*
F IMPAIRMENT - PHYSICAL*
G IMPAIRMENT NOT KNOWN
H NOT APPLICABLE
I SLEEPY / FATIGUED*

SKETCH
INDICATE NORTH

MISCELLANEOUS

AN INTERNATIONALLY ACCREDITED AGENCY

Figure 3.5: An example CHP 555 Page 2 page mask

Semantic Validation

The Semantic Validation stage serves as a final consistency check for the extracted data. Due to the unstructured nature of the TCR form content, it's possible for data to be improperly extracted, even with the page mask approach described above. For example, dynamic textual elements or checkmarks may have been improperly placed near (but not within) their designated field areas, or textual elements may overflow the bounds of their designated field areas. The Semantic Validation stage allow for a final "sanity check" on the extracted data, increasing the probability of detecting an erroneous extraction.

Postmile Referencing Service

TCRPRO's postmile referencing service is a machine-to-machine HTTP service that is able to handle the two primary types of postmile-related GIS queries:

- **Arbitrary postmile geolocation:** This class of query allows the precise geographic coordinates of any postmile to be determined. The postmile value, along with various roadway and alignment parameters are specified in a query, and the corresponding results are returned to the caller.
- **Postmile proximity search:** This class of query allows the nearest postmile(s) to a particular set of geographic coordinates to be determined. A latitude, a longitude, and a search range, along with various roadway and alignment parameters (to filter the results) are specified in a query, and the corresponding results are returned to the caller.

The service is implemented as a set of Java HTTP servlets which make queries to a Post-GIS database which houses a roadway information database. The data in this database is derived from publicly-available Caltrans datasets[1].

[1] http://www.dot.ca.gov/hq/tsip/gis/datalibrary/

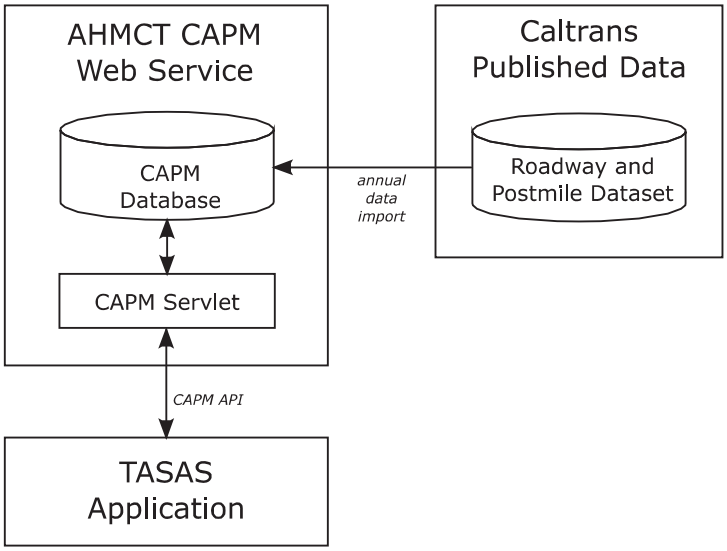


Figure 3.6: Postmile Referencing Service Data Flow

Iterative Development and Productionization

AHMCT researchers worked closely with Caltrans to design the functionality of TCRPRO and its interface to maximize ease of integration with the existing TCR processing infrastructure. It was decided relatively early in the research study, that TCRPRO would be designed as a self-contained, standalone system in order to facilitate initial integration and to simplify future maintenance. A virtual machine was allocated for TCRPRO on the Caltrans VMware infrastructure, and TCRPRO was deployed onto that machine. The interface and functionality of TCRPRO was then refined iteratively based on feedback from Caltrans.

Results

TCRPRO's TCR processing system was tested with a set of 504,290 TCRs. Of these, TCRPRO successfully identified and extracted data from the form pages for 503,184 (99.8%), and failing to do so for 1106 (0.2%). Overall a total of 651,094 TCRs were processed by Caltrans IT using TCRPRO. The results are as follows:

- 600,101 (92.1%) were partially SOE coded by TCRPRO.
- 294,617 (45%) were Location Coded by TCRPRO.

It is important to note that in terms of SOE coding TSRPRO was only designed to handle simple cases and as it was pointed out earlier the lane location could not be automated. Therefore, TCRPRO in its current version can only partially code the Sequence of Events.

In terms of location coding, TCRPRO was not able to code approximately 55% of the TCRs that were processed. This is because automatic coding requires sufficient and consistent information about the accident location within the TCR. When there is insufficient and/or inconsistent data on the accident location, a human coder can make an arbitrary decision to either select between two or more alternatives or utilize other data

beyond the information within a TCR. A computer program, however, follows precise instructions and cannot make arbitrary choices. If specific algorithm on making such choices can be agreed upon by the Coding Group then such an algorithm can be implemented into TCRPRO increasing its ability to perform location coding. Table A.6 shows a summary of the errors that were experienced and caused the 258,130 eTCRs to be returned without the location code. The largest percentage of errors (74.3%) were caused when the information from the “Location Box” portion of the report along with the “Area of Impact” portion extracted from the narrative, provided insufficient information.

The primary challenge in successfully processing TCRs lies in handling the many variations of these documents that are encountered. Early versions of TCRPRO were unable to successfully extract data from a significant percentage of TCRs tested. Through an iterative process of testing, analyzing results, and creating page templates, this number was lowered to a much more tolerable level. More page templates could be created and, while this may be worthwhile to a limited extent, the creation of page templates is a time-consuming process and the benefits of such an effort should be weighed against the cost. An alternative approach is to work with CHP to better standardize the TCR forms in a manner to eliminate variations.

At this time, TCRPRO supports a total of 30 page templates, covering the vast majority of TCRs we have studied. A summary of these can be found in Table 3.3.

Table 3.3: Number of TCRPRO templates by page type

TCR page type	# TCRPRO templates
Page 1	2
Page 2	2
Page 3	2
Narrative/Supplemental	9
Factual Diagram	7
Sketch Diagram	8

CHAPTER 4: CONCLUSIONS AND FUTURE RESEARCH

Key contributions of this research project:

The key contributions from this research are as follows:

- A computerized system for automated data extraction from digital traffic collision reports (eTCRs) has been developed. This system is referred to as "TCRPRO" for TCR Processing.
- TCRPRO extracts the "Summary" portion from the Narrative section in the eTCR and relays it to Caltrans by means of a web service. This automated collection of the Summary text will be able to replace labor intensive methods of manually extracting the Summary for subsequent utilization.
- TCRPRO determines the GPS coordinates ("Lat" and "Long") using its internal LRS system based on the Location code of the eTCR. TCRPRO has the capability of performing this task either by calculating the location code internally or have it passed in via the Caltrans based web service.
- TCRPRO is able to return the corresponding highway inventory information at a given postmile value on a given highway and county. Using either the internally calculated Location Code or that provided via Caltrans web service, TCRPRO uses the "Clean Roads File" to extract TASAS variables such as the median type, barrier type, access controls and number of lanes (both left and right).
- The capability of extracting any data field from the eTCR sets up a framework where the content of all eTCRs can be made accessible to any application outside of the Location Coding and Sequence of Events coding functionality.
- In its present form, TCRPRO can handle much of the Sequence of Events coding for simple two party collisions with the exception of the lane of travel in which the collision occurred in cases where the lane numbering is not standard such as when there are High Occupancy Vehicle (HOV) lanes.
- TCRPRO can handle location coding within certain degree of accuracy. The testing of 651,094 cases 294,617 (45.25%) were automatically coded by TCRPRO with the following accuracy:
    - From the 294,617 Collisions which were Location Coded compared to the TASAS coding, 281,217 (95.45%) had everything except postmile differences greater than zero..
    - 77,278 had postmiles within 0.01 miles of TASAS coding.
    - 82,565 had postmiles between 0.01 to 0.1 miles of TASAS coding.

- 107,622 had postmiles between 0.1 to 1.0 mile of TASAS coding.
- 13,526 had postmile mismatch of 1 mile or more from TASAS coding.

Limitations

- It has been determined that automation of the Sequence of Events portion of the process is not feasible for identifying the lane location without use of the diagrams. Current technology at the time of this report was not sufficiently advanced to "interpret" diagrams to the level needed to perform this task.
- The variations in the way information important for TASAS are currently entered in the CHP Form 555 limits the ability to fully automate TCR processing with full accuracy.

TCRPRO Additional Value Added Results

- TCRPRO can process the narrative and search for key words such as Truck Mounted Attenuator (TMA), distraction, etc. This is because TCRPRO can extract all data fields in a TCR.

- TCRPRO is able to also read the Injury page so that injury details such as the description and "transported to" fields can be saved for future analysis.
- Returning summary and diagrams with each eTCR can be used to potentially help

  the Location Coding Group in the future to enable a "single screen" coding system. Values extracted directly from the PDF can be delivered to the Coding Screen along with the summary and collision diagrams to facilitate the coding process. TCRPRO can also provide the geometry of the collision location since due to its ability to access the Clean Roads Files.

- If given an accurate GPS value from the CHP generated eTCR, then TCRPRO could return an accurate post mile marker value.
- Ability to work with static version of Clean Roads Files can also lead to automation

  of other fields populated in the TASAS database such as Access controls, NOLR, median type.

Recommendations

1. Changing the workflow protocol for the Coding Group so that they would be able to handle partially completed TASAS data without the need to redo everything over again.
2. Developing a graphical user interface for the Coding Group use of TCRPRO on Caltrans system can enhance the productivity of the Coding Group and their potential utilization of some of the features of TCRPRO.
3. Having the user interface display the collision summary and the diagram that are generated by TCRPRO in one screen to facilitate the coding process and efficiency for the Coding Group.
4. Working with CHP to modify Form 555 (TCR form) to remove ambiguity in traffic collision descriptions. This can lead to more automation of the coding process.
5. Incorporating some form of an “eDiagram” where known entities (e.g. vehicles, direction of travel, barrier, etc.) are placed on a parametrically drawn roadway where the collision occurred in eTCRs. Based on the parameters and relative placement of the vehicles, the location of the impact events could be reliably detected and TCRPRO would be able to accurately provide the Sequence of Events code.
6. To address the difficulty with location coding, the method of recording the GPS location of a collision needs to be standardized within police operations. This can eliminate some of the variations. Furthermore, methods should be considered that would allow converting a GPS location to a postmile value.
7. In the Clean Roads Files, it is recommended that the landmark fields be made more consistent with respect to naming conventions. Such changes will facilitate matching landmark values from the eTCR with the Clean Roads File data. Furthermore, expanding the landmark description can eliminate possible ambiguities with other landmarks.

   A large percentage of eTCRs could not be location coded by TCRPRO due to the ambiguity or the inability to find matching landmark references between the eTCR and the Clean Roads File. To improve the reliability and accuracy of the internal procedures within TCRPRO, it is critical to find a link between the CHP description of the collision location with the tools available to the Location Coders and TCRPRO. One avenue would be to improve the landmark matching algorithm. Some examples are as follows:

   - Misspellings either in the Clean Roads File or the TCR
   - Incorrect decorator
   - Lack of a decorator
   - Lack of a distinct decorator
   - In case of a landmark with multiple words, there can be multiple variations of writing it

Here are some modifications that would make the Clean Roads File more useful for this use case:

- Separate the landmark(s) and the decorator(s) into separate columns
- Move unnecessary information, such as (10-124), into a separate column(s)
- Be consistent with the use of acronyms and abbreviations
- Be consistent in the way Junctions are written

APPENDIX A: EMULATION OF LOCATION CODING

The process of coding a collision report is complicated and includes prescribed protocols to ensure data integrity. The purpose of this appendix is to describe how the TCRPRO software emulates this process via computer programming and digital processing.

Extracting Data

To begin the process of coding a report, whether it is the Location Code or the Sequence of Events, the first step is to extract the needed pieces of information from the report before the rest of the processes can be executed.

The benefit of processing eTCRs is that a methodology can be implemented where text and other needed information can be accurately extracted without human intervention. The extraction software within TCRPRO is able to discern the text unlike that of a non-digital image TCR originating from a scanned version of a hard copy report. In the case of the scanned file, “Optical Character Recognition (OCR)” techniques must be employed to “estimate” what each character is and can be inaccurate depending on the quality of the scan.

To emulate the coding operations, the text contained in the narrative portions of the TCR is further processed to extract particular segments. Specifically, the segment of text with the “Summary” heading is extracted as well as the segment with the heading, “Area of Impact”.

Determining the Viability of a TCR for TCRPRO Processing

There are two major reasons why an eTCR is not able to be processed by TCRPRO. The first condition is when the form structure is unfamiliar to the software. Although TCRPRO is able to recognize over 99% of the forms in the 2016-2017 eTCR dataset, there are numerous additional variations of any given form, presenting anomalies, which can prohibit data extraction from an eTCR. Often, each of these particular anomalies can exist in only a small handful of individual eTCRs. In its present version, TCRPRO has two templates for pages 1-3 of a standard traffic collision report. The pages which contain the narrative and diagrams however, hove more variation in their formats and so a higher number of templates are needed for these pages. Table A.1 contains the number of templates that were developed to accommodate the variations in report formats. It can be seen in this table that the first two pages and the Injury page needed only 2 templates whereas the narrative page required 9 templates.

Table A.1: Current number of TCRPRO form templates by page type

Form of eTCR Page Type	Number of Templates Currently Supported
Traffic Collision Report (Page 1)	2
Traffic Collision Report (Page 2)	2
Injured/Witness/Passengers (Page 3)	2
Narrative/Supplemental	9
Factual Diagram	7
Sketch Diagram	8

The second reason an eTCR would not be viable for processing by TCRPRO is if the data in the “Location Box” and in the “Area of Impact” does not have the required information needed for the Location Code determination. Please note the “Location Box” is that portion of the eTCR on the first page of the eTCR (see Figure 3.4) which is located in the upper left-hand portion directly above the Party Data. The “Area of Impact” data is the part of the narrative with that particular heading and is where the reporting officer describes in words all of the factual data related to the collision. The minimum amount of data needed for this purpose is listed in Table A.2. The origin of the data shown in Table A.2 is obtained from TCRPRO’s extraction process.

Table A.2: Name and description of the needed pieces of data extracted from the eTCR to calculate Location codes.

Required Data Name	Required Data Description
Route, County, City	Section of Highway which travels through a county and city when applicable. All 3 fields are needed
Distance	Numeric value representing a distance amount
Unit-measure	Unit of measure for the distance amount
PMM-ref or Landmark	Landmark or postmile marker value to be used as a reference
Direction	Direction in terms of N,S,E or W from the landmark

Interpreting the Language

Once the data has been extracted from the eTCR, the next step is to interpret the content and parse components of the fields into Location Coding variables. This can be challenging since a particular route or location may be described in a variety of ways. In addition, typos are often encountered in these fields requiring various mitigation strategies. For example, the route description "I-10 EASTBOUND" might also be expressed as "I 61 EAST", "1-61 EB" with a "1" (one) used instead of a capital "I", "E/B 10", or "INTERSTATE 10 E/B (SANTA MONICA FWY)". Another example is when the relative route location "0.2 MILES SOUTH OF MAIN STREET" might also be expressed as "0.2 MI. S OF MAIN ST.", "O.2 MILE(S) S. OF MAIN" with a capital "O" used instead of a "0" (zero), or ".2 MIL SOUTH OF MAIN ST".

By analyzing the 2016-2017 eTCR dataset, AHMCT has learned much about the variety of language, abbreviations, and common errors used in these fields, and has incorporated this knowledge into TCRPRO's textual processing algorithms.

Location Box Line 1

The first page of an eTCR has the “Location Box” which provides a general description of the collision location. The first line in the “Location Box” describes the collision location in general terms such as “I-80 and SR113” or “Hwy 1 and Woodside Rd.” The state roads where the collision occurred are often identified as well as landmarks to be later referenced. This is not always the case, however, and provisions are imbedded in the TCRPRO software to handle some of the variations. This is done by extracting and parsing the pertinent information for future comparison with the other data in the eTCR.

One important characteristic of this field is that when the collision occurs on a ramp, the verbiage in this line will include the word “to”, connecting one roadway to another. For example, a ramp would be flagged with “NB SR113 to Gibson Rd.” Handling of such situations are discussed later in this chapter in the section on “Ramps and Intersections”.

Location Box Line 2

Milepost information is provided in Line 2 in the eTCR Location Box. This line describes the collision location in terms of the closest milepost marker. This data is typically reliable but the CHP officer does not always have access to the closest PMM value. In the past (i.e. prior to 2015), this field was typically populated with the GPS portion remaining empty. Current trends (2015 to present) list the GPS values but not the milepost data. It would seem that once the GPS value was entered, it could have been assumed that an accurate collision location was provided. In actuality however, the GPS value is not consistently entered as the actual location of where the collision occurred. Sometimes it is listed where the officer has an opportunity to write up the report or at a safe distance from active traffic.

Examples from this line include: 122 ft. N of 113 YOL 2.34 where the location is based on the reference postmile marker 2.340 on highway 113 in Yolo county. The distance of 122 feet will need to be converted from feet to miles.

Location Box Line 3

The third line of the Location Box typically has the more precise depiction of the collision. It generally has the majority of the critical components needed for the location. This field usually specifies the state route, landmark, direction and distance. It is also referred to throughout the TCRPRO software documentation as “Secondary Information”.

Area of Impact

The “Area of Impact” or AOI, is located towards the end of the narrative portion of the eTCR. It is not located on any of the standardized form pages but rather imbedded amongst the rest of the narrative. Standard procedure however, usually puts the Area of Impact after the “SUMMARY” section. It is also important to note that only the first Area of Impact (AOI#1) is used for coding purposes.

As with all the language interpretations in the Location Box, TCRPRO needs to be “smart” when deciphering the language and parsing the content into the appropriate fields. Unfortunately, due to ambiguities of the English language, it is often necessary to inspect the diagrams of the report to truly identify the collision location. The diagram can be especially invaluable in cases where the geometry is complicated. Unfortunately however, diagrams cannot be interpreted at this time with a high degree of reliability within TCRPRO.

Calculating Postmile Marker Values

The next step towards calculating the postmile marker value for the location code is to use the parsed-out data from the Location Box content, along with the AOI content. If given sufficient data in the 2nd line, the location = PM_ref +/- distance in feet/(5280 ft./mile). Unless the reference PMM is provided, the landmark data will be used to determine what the reference PMM value should be used. The tool used by the Location Coding Group to determine the PMM_Ref is the Sequence Listings described in the next section.

Sequence Listings

There is a Sequence Listing for each district within Caltrans. Within each district listing, all state routes running through that district are listed and, for each route, there is information on all miles of that route through all counties. Consider, for example, highway I-80 which travels through two Caltrans districts. Information on I-80 would be found in two Sequence Listings. In general, the Sequence Listing within each Caltrans district is organized as follows:

- Ordered by route number, county, and route direction
- Since state roadways run in two directions (either N/S or E/W), in most cases both sides share the same postmile values laterally
    - Exceptions to this are when the highway has independent alignment between Right and Left. This is specifically noted in the Sequence Listing.

- Separate postmiles exist for the right and left alignments, and an “R” and “L” suffix is used to differentiate them.

Figures A.1, A.2 and A.3 provide additional detailed information on what a Sequence Listing looks like and how it is utilized.

As can be seen in Figure A.1, the Sequence Listing has a defined legend and layout which must be visually comprehended by a person. It has colored highlighted areas and colored text. This type of organization is difficult for a computer to interpret although the current Sequence Listing format is very efficient for the manual coding process..

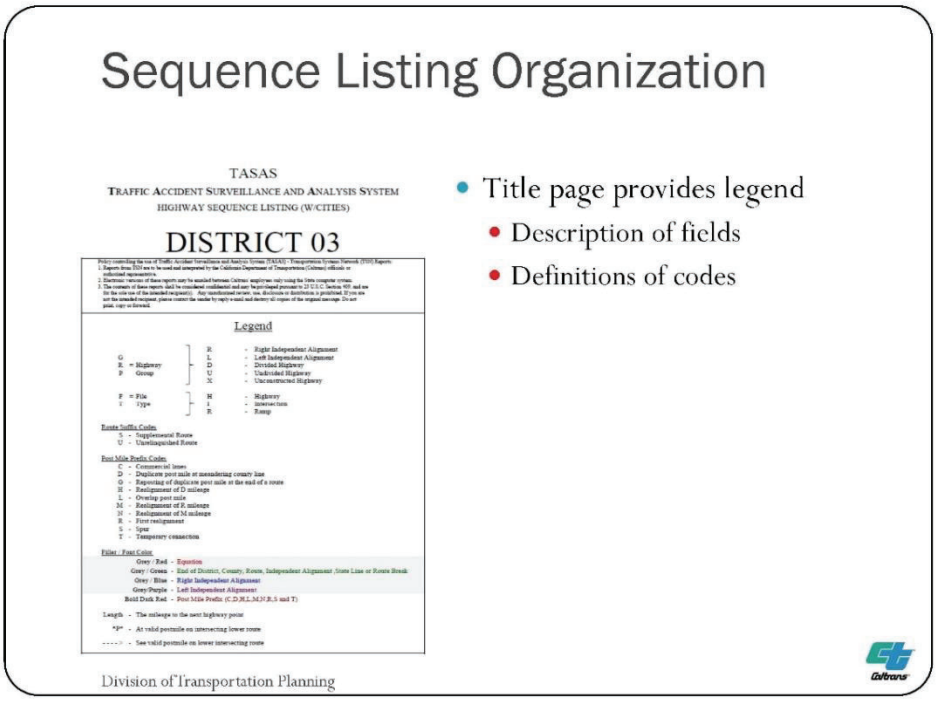


Figure A.1: First page of Sequence Listing for District 3

Figure A.2 shows an example portion of a Sequence Listing from District 3 and Highway I-5. The columns of data for each highway and district are: CO (County), CITY, POSTMILE, GRP (Highway Grouping such as Divided or Undivided), FT (File Type which can be either H for highway, R for ramp, or I for intersection), LENGTH (refers to the length of the postmile line item), and DESCRIPTION (referring to the “Landmark” description obtained from the Clean Roads File).

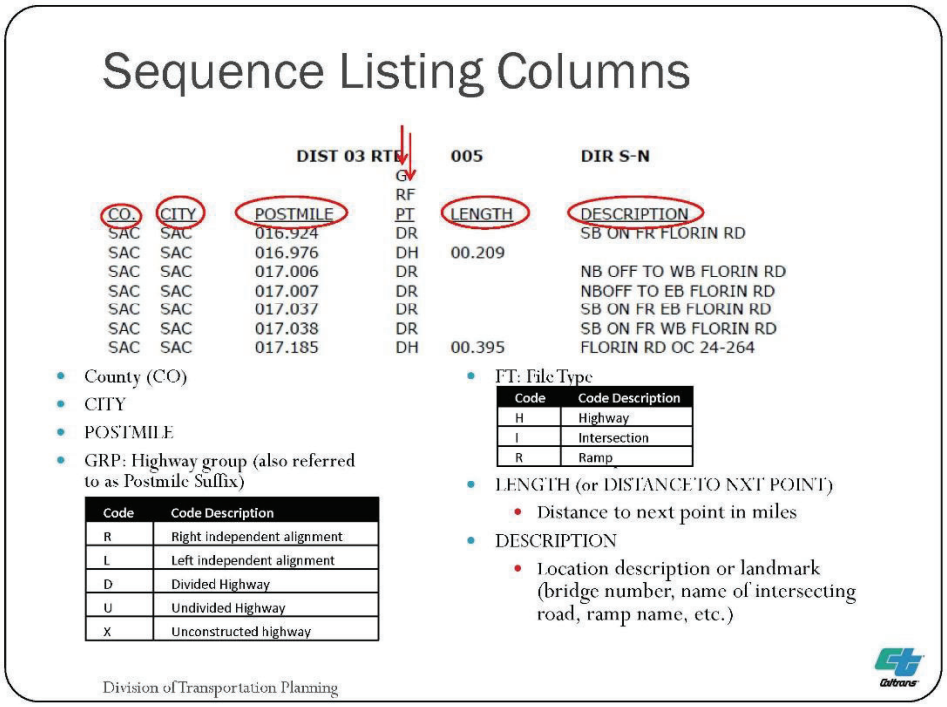


Figure A.2: Illustration of the columns listed in the Sequence Listing for a given route and district. The county information is listed in the first column.

In Figure A.3, a particular postmile value is sought using a particular Sequence Listing. In this example, the route shown extends between highway 87E and I-505. The postmile range is from 7.550 to 9.835. If looking for where the highway intersects with Taylor street in Winters, it can be seen this intersection is at postmile value of 7.930.

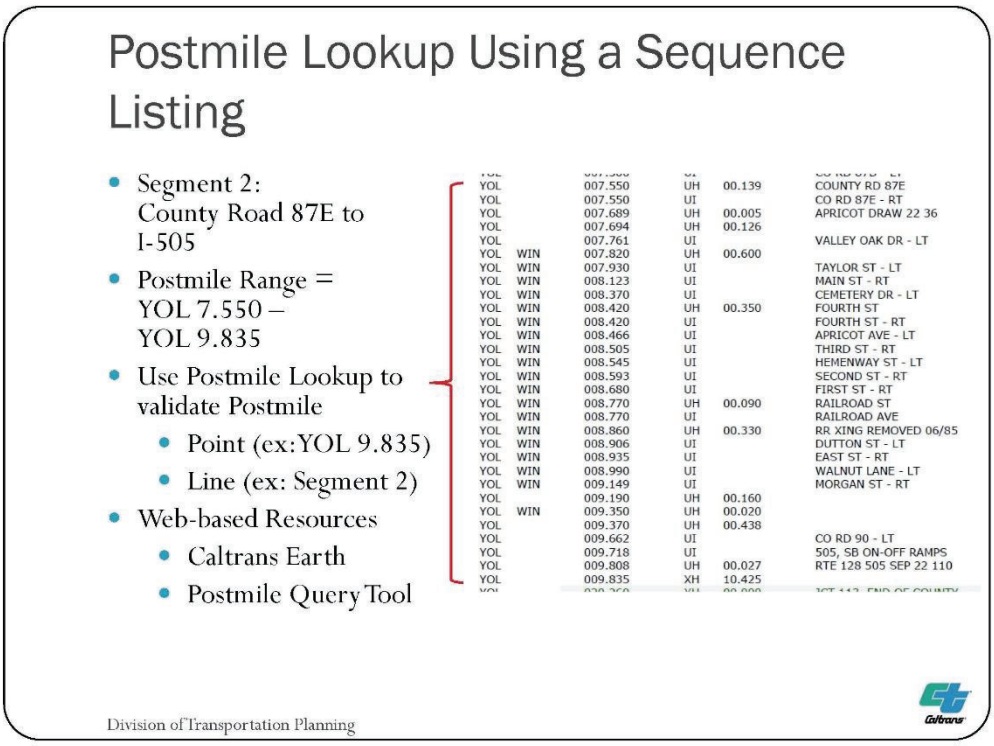


Figure A.3: Illustration of how a postmile value is found using a Sequence Listing "look up" table.

Clean Roads File

The Clean Roads File is an export of the Caltrans Highway Inventory Database. This database contains a catalogue of structures and roadway features along the linear length of each state highway. As one can imagine, this exported file is a huge spreadsheet containing data on a highway and all its intersections and ramps. The highway worksheet contains 96 fields and nearly 57,000 lines of data, the Intersections file contains 49 fields and 17,068 lines of data, and the Ramps file contains 24 fields and 15,278 lines of data.

If the Location Coder had to rely on this particular set of spreadsheets (e.g. the Clean Roads File), it would be too cumbersome to use on a regular basis. Consequently, the Sequence Listings were developed for the Caltrans community to have easy, legible access to finding detailed postmile values.

In TCRPRO, it is much more efficient to use the Clean Roads file as opposed to 12 PDF documents of the Sequence Listings. The same data found in the Sequence Listings can also be found in the Clean Roads file, just in a different format.

Matching Algorithms

The process of finding a match in the clean roads files involves several steps. First, we separate the landmark information from the rest of the collision description. This is done both in the third line of the Location Box portion of page one and the Area of Impact. Assuming the landmark information has been isolated at this point, the following steps are taken:

1. Separate the decorators in the landmark from the non-decorators.
    - Decorators are key-words such as ROAD, PARKWAY, OVERCROSSING, STREET, VALLEY, etc. For a full list of the latest decorators, please consult the appropriate file.
    - Non-Decorators include the actual names, such as ANTELOPE, HUTCHITSON, MAIN, etc.
2. Next, we use the non-decorator(s), county, route, and district to find a match in the clean roads.
    - If there are multiple matches, we attempt to filter by type.
    - If there are any matches, we retrieve the set of decorators that correspond to the decorator(s) we captured earlier.
    - If there is a single match, we check if any of the decorator sets are found in the match and add appropriate warnings.
    - If there is no match, then we stop.
    - If there are multiple matches, we attempt to narrow down by matching the decorator sets and add any applicable warnings.

Positive example

An example where the landmark information was successfully parsed and positively matched with a clean roads file line item can be seen in the TCR with ID=100040-9140-2016-0025. In this report the county = Lassen, the route = 395 and the Landmark string is equal to “SEARS RD.” In this case, the decorator is parsed as “RD.” with the non-decorator parsed as “SEARS”. Decorators are compared with a set of similar decorators which mean the same thing such as ["RD", "RD.", "ROAD", "RD,", "RO"]. This example of parsed data matches with the line from the clean roads file where District = 2, county = “LAS”, route = 395 and BEGIN_PM=53.1 and END_PM = 53.1 with Landmark= “SEARS RD LT & RT” and Type=Intersection. Under these conditions, there is only a single match that corresponds to this data, and TCRPRO can use the “Begin” and “End” postmile marker value of 53.1 and assume the collision was at an intersection.

Negative example

In this section, two examples are given that illustrate where the collision description strings do not clearly provide a match with an entry in the clean roads file. In these cases, a negative result occurs and an error ensues. For example, where the string is equal to ”SOBERANES CANYON TRAILHEAD”, the decorator is parsed as “CANYON” with the non-decorators are parsed as ['SOBERANES', 'TRAILHEAD']. Note, there are two non-

decorators in this example. For this string parsing, there were no matches in the clean roads file with these words.

In the second example, the landmark string is “BALL RD” where the decorator = “RD” and the non-decorator = “BALL”. This collision occurred where the District = 12, route = 5 and county = Orange. In the previous case, there were no matches whatsoever but in this case, there are five potential matches in the clean roads file. Table A.3 below provides a partial view of the clean roads file for this collision. In this situation, the string “BALL RD” is found in each Landmark field. With no other data in the string, an error occurs due to the ambiguous results.

Table A.3: Extraction of Selected Columns from the Clean Roads File

Extraction from Clean Roads File District = 12, County = ORA, Route=5, Route Suffix=*NULL*, and Postmile Prefix= *NULL*				
begin_pm	end_pm	pm_suf	landmark	type
37.638	37.641		BALL RD E BND OC 55-670	highway
37.665	37.682		BALL RD W BND OC 55-169	highway
37.682	37.682		005/NB ON FROM WB BALL RD	ramp
37.929	37.929		005/SBON FR DISNLND/BALL RD	ramp
38.256	38.256		005/NBON FR DISNLND/BALL RD	ramp

Ramps

As part of the Location Code, the “Area Type” and “File Type” are attributes that need determination. These two fields indicate whether a collision occurred on a ramp, intersection or on the highway itself. If TCRPRO can determine whether a ramp or intersection is involved, then it adds another layer of complexity to the matching algorithms.

To determine if an accident occurred on a ramp, we use four main criterion. First, we check if line one in the location box contains the string **‘ramp’**. Second, we check if it contains the string **‘collector’**. Third, we check if the summary contains the string **‘ramp’.** Lastly, we check if line one of the location box fits a specific regular expression. If any of these four are true, we consider this collision to have occurred on a ramp. Most of these examples will focus on the last part, the regular expression. A regular expression is a sequence of characters that defines a search pattern. In short, we are looking for something that references traveling from a street to a highway, a highway to a street, or a highway to a highway.

Calculating PMM Value from Input Data

Once a reference postmile marker is identified, the distance (in units of miles) is added or subtracted. For most freeways:

- For South to North Freeways (typically odd numbered):
    - North of a Postmile = Add
    - South of a Postmile = Subtract
- For West to East freeways (typically even numbered)
    - East of a Postmile = Add
    - West of a Postmile = Subtract

There are 6 highways which do not follow the convention of S to N (typically odd numbered) and W to E (typically even numbered). There are 6 exceptions to this and are considered “backwards” and are:

- Route 71 North to South
- Route 153 East to West
- Route 224 East to West
- Route 282 East to West
- Route 580 East to West
- Route 780 East to West

Results

When comparing the location code from the 2016-2017 data set we compared the results from TCRPRO with the “target values” which we received from the Location Coding Group. Out of the approximately 502k eTCRs, there were 48.6% (243,939) for which location codes were determined. Of these, 232,948 (46.4%) had been coded by the Caltrans Location Coding Group and had matching county and route. Comparisons were made between TCRPRO’s results and Caltrans and are summarized in Table A.4. From this table we can see that 17,859 reports had less than a 0.001 mile difference. The mean value in this range is 0.0003 miles. When evaluating all results, 50% had a difference of less than 0.05 miles with a mean value of 0.011 mile.

Table A.4: Location Code Postmile Marker value difference between TCRPRO results and those determined by Caltrans. These results are based on a comparison set of 232,948

Difference (miles) in Postmile Marker Results	Number of eTCRs (out of 232,948)	%Total (out of 232,948)	Mean value of range (miles)
< 0.001	17,859	7.7%	0.0003
< 0.005	59,081	25.4%	0.0021
< 0.01	77,555	33.3%	0.0032
< 0.05	116,025	49.8%	0.011

Errors

The previous section provides the results of the location code values that were calculated by TCRPRO and how they compared with the Caltrans generated values. There were 258,130 (51.4%) eTCRS however, out of 502,069 where a location code could not be determined. To better understand why there are differences between the Location Code generated by Caltrans as opposed to that generated by TCRPRO, it is beneficial to consider the origin of errors found as TCRPRO processes each individual eTCR. For example, if the value for "county" was invalid (e.g. "ORG" instead of "ORA" or blank), the error, "Variable string or number could not be interpreted as a county; string not saved*" would be returned as part of the software functionality. When the warnings indicate there is insufficient data to calculate a location code, TCRPRO returns an error and stops processing the eTCR. Table A.5 below is a list of warnings that TCRPRO encounters while processing eTCRs. Example warnings are when a parameter such as a highway number is not provided or is empty.

Table A.5 Table of Error Names and Description Codes.

Error Message	Number of eTCRs
INVALID_COUNTY COUNTY_TYPE_ERROR	Variable string or number could not be interpreted as a county; string not saved*
INVALID_ROUTE ROUTE_TYPE_ERROR	Var string or number could not be interpreted as a route. String not saved*. Valid route is needed as a valid route number instead of a local name. This error type might be alleviated if TCRPRO had access to a local highway name

Error Message	Number of eTCRs
	corresponding with route number look-up table
PARAMETER EMPTY	PARAMETER = County, City, Route, TCRpdf-field(LBOX1; LBOX2; LBOX3; DOT). If some of these parameters are empty then it can cause a TCR to be considered not viable such as route, county or DOT. LBOX2 contains the Postmile Marker value when known.
SECONDARY_UNIT_TYPE_ERROR SECONDARY_ROUTE_TYPE_ERROR SECONDARY_DISTANCE_TYPE_ERROR SECONDARY_DIRECTION_TYPE_ERROR	The data from the LBOX3 or Secondary Information is parsed into “usable” segments such as distance length (number), unit of measurement, direction, route number and landmark info
AOI_TYPE_ERROR	The Area of Impact (AOI) is found after the summary section. Only the first AOI is used (AOI#1). Using similar logic to the “Secondary Information” – its string must be parsed out into logically used segments to be considered viable.
DIRECTION_OF_TRAVEL_TYPE_ERROR	The direction of travel for party #1 contained invalid information. In other words, it wasn't one of the four recognized cardinal directions
COLLISION_OCCURRED_ON_TYPE_ERROR	The "Collision Occurred on" box in the Location Box on page one contained invalid data
PREDICTED_POSTMILE_INVALID	The calculated postmile was not matched with a valid point on the milepost system
TWO_VALIDATION_POSTMILE	Two conflicting results were found when validating the postmile.
MULTIPLE_VALIDATION_POSTMILES	Multiple conflicting results were found when validating the predicted postmile.
MILEPOST_SECONDARY_AREA_OF_IMPACT_EMPTY	Milepost information, secondary information, and area of impact are all empty. Please provide a non-empty values for at least one of these.
MILEPOST_SECONDARY_AREA_OF_IMPACT_BAD	Milepost information, secondary information, and area of impact all contain invalid input data.
OCCURRED_ON_INTERSECTION	The collision has occurred on an intersection.

Error Message	Number of eTCRs
OCCURRED_ON_RAMP	The collision has occurred on a ramp.
*	Possible Improvement for future is to save strings associated with errors that may point to a common acronym

Although TCRPRO has been developed to reconcile warnings such as looking up highway names to find a corresponding route number, there are still more than half of the processed TCRs where there is insufficient data to automatically calculate a location code. Table A.6 shows a summary of the errors that were experienced and caused the 258,130 eTCRs to be returned without the location code. The largest percentage of errors (74.3%) were caused when both the information from the "Location Box" portion of the report along with the "Area of Impact" portion extracted from the narrative, provided insufficient information.

Table A.6: Table of error combinations and quantifies that prevented TCRPRO to calculate PMM value.

Error Message	Frequency	Percentage (out of 258,130)
Empty parameter: route.	14067	5.4%
Invalid route detected. Please provide a valid route.	363	0.1%
Milepost information, secondary information, and area of impact all contain invalid input data.	191789	74.3%
Milepost information, secondary information, and area of impact are all empty. Please provide a non-empty values for at least one of these.	266	0.1%
Multiple conflicting results were found when validating the predicted postmile.	421	0.2%
No valid results were found for the predicted postmile.	7283	2.8%
The collision has occurred on a ramp.	43941	17.0%

APPENDIX B: EMULATION OF SEQUENCE OF EVENTS CODING

Determining the Sequence of Events that led up to a collision is an important part of collision coding. This portion of the coding allows for a deeper understanding of the collision but is difficult to automate. The purpose of this appendix is to describe the process and challenges of determining the "Card 8b" value for the Sequence of Events.

Requirements

The information needed for the Sequence of Events is dependent on the actions of each party and the relative relationship with each other at the scene of the collision. This coding also captures what hit what in the order it happened. Thus, the information needed for each party is as follows:

- Party Type (such as passenger car, motorcycle, large truck, etc.)
- Movement preceding collision
- Direction of travel
- Up to 4 events per party that took place (the first one is referred to as the primary and the others are referenced as A, B and C)
    - The location of where they were located on the highway when each event occurred (typically in terms of the lane)
    - An indicator of what was hit by that party

Party Type

Since TCRPRO has a very robust and accurate data extractor from eTCRs, "Party Type" can be reliably calculated. For each party on page 1 of the eTCR, the reporting officer checks a box that best describes the type of vehicle and whether one of the parties was a driver, pedestrian, bicycle, parked vehicle or "other." The following table (Table B.1) is a list of valid party types. The single letter "codes" are also shown here since these codes are typically extracted from the eTCR. Although over the majority of collisions occur with party type "A" (Passenger vehicles), it is important to classify the party type correctly since collision trends are tracked with certain problematic party types, such as large trucks and motor cycles, where risk of serious injury is higher than others.

Table B.1: Table of Party Type Code and Corresponding Descriptions

Party Type Code	Party Type Description
A	Passenger Car/Station Wagon
B	Passenger Car w/Trailer
C	Motorcycle
D	Pickup/Panel Truck

Party Type Code	Party Type Description
E	Pickup/Panel w/Trailer
F	Truck, Truck Tractor
G	Truck/Tractor w/Trailer
2	Truck/Tractor w/2 Trailers
3	Truck/Tractor w/3 Trailers
4	Single Unit Tanker
5	Truck/Tractor w/1 Tank Trailer
6	Truck/Tractor w/2 Tank Trailers
H	School Bus
I	Other Bus
J	Emergency Vehicle
K	Highway Construction Equipment
L	Bicycle
M	Other Vehicle
N	Other Non-Vehicle
O	Spilled Load
P	Disengaged Tow
Q	Uninvolved Vehicle
R	Moped
T	Train
U	Pedestrian
V	Dismounted Pedestrian
W	Animal-Livestock
X	Animal-Deer
Z	Animal-Other
<	Not Stated

Movement Preceding Collision

In the Sequence of Events, the movement preceding the collisions needs to be coded along with each party. Table B.2 contains the letter codes used for this purpose along with a detailed description. These movements pertain to whether the party is moving on an undivided freeway or a divided one. This movement is frequently described in the narrative.

Table B.2: Table of letter codes used to describe the movement preceding a collision for each party.

Movement Key	Movement Description
A	Stopped
B	Proceeding Straight
C	Ran off Road
D	Making right turn
E	Marking left turn

F	Making u turn
G	Backing
H	Slowing/Stopping
I	Passing Other Vehicle
J	Changing Lanes
Key	Parking Maneuver
L	Entering Traffic
M	Other Unsafe Turning
N	Xing into Opposing Lane
O	Parked
P	Merging
Q	Traveling Wrong Way

Object Struck

In the Sequence of Events portion, there are four pairs in which the Caltrans coder can provide data. The first half of these four pairs is the code for the first object struck in the collision. Typically, it is one of the other vehicles. There are many other pre-coded objects which have been experienced by the traveling public and/or reporting CHP officers. The following table (Table B.3) provides a list of the coded objects:

Table B.3: Table of Objects Code and Description for the Sequence of Events

Object Code	Object Code Description
01	Side of Bridge Railing
02	End of Bridge Railing
03	Pier, Column, Abutment
04	Bottom of Structure
05	Bridge End Posts in Gore
06	End of Guardrail
07	Bridge Approach Guardrail
10	Light or Signal Pole
11	Utility Pole
12	Pole (Type Not Stated)
13	Traffic Sign/Sign Post
14	Other Signs Not Traffic
15	Guardrail
16	Median Barrier
17	Wall (e.g. Soundwall)
18	Dike or Curb
19	Traffic Island
20	Raised Bars

Object Code	Object Code Description
21	Concrete Object (HDWL, D.I.)
22	Guidepost, Culvert, PM
23	Cut Slope or Embankment
24	Over Embankment
25	In Water
26	Drainage Ditch
27	Fence
28	Trees
29	Plants
30	Sound Walls
40	Natural Material on Road
41	Temp Barricades, Cones, Etc.
42	Other Object On Road
43	Other Object Off Road
44	Overturned
45	Crash Cushions-Sand (After 01/01/96, Before Both 45)
46	Crash Cushions-Other (After 01/01/96, Before Both 45)
51	Call Box (After 01/01/96)
98	Unknown Object Involved
99	No Object Involved
VI (through V9)	Vehicle 1 to 9
<<	Not Stated

Location where Object was Struck

The “Location” field in the Sequence of Events refers specifically to where on the roadway any impact occurred. For multiple impacts that occur during a traffic collision, all events will be recorded in the Sequence of Events data associated with the involved party. In Figure B.1 the lane locations are “coded” from “C” to “G”. In the multi-lane expressway shown in Figure B.1, “C” refers to the far left shoulder area where “E” represents any middle lane. In the case of a two lane highway, no middle lane is present – just a left hand lane and a right hand lane. Table B.4 describes in detail the lane location identifier for the Sequence of Events. Note that “D” is for the left lane of travel and “F” is for the right lane of travel. As can also be seen in Table B.4, there are lane designations for high occupancy lanes (“V”) and Gore point regions (“I”). These designations are important when analyzing safety trends but are difficult to determine if a diagram is not readily accessible or interpretable. Figure B.2 illustrates how to identify a two-lane freeway whether it is one lane for each direction or two lanes for each direction. Again, it can be seen in Figure B.2 that the complexity of lane identification is clearly demonstrated solely in the diagram and would be rather difficult to describe in the narrative text.

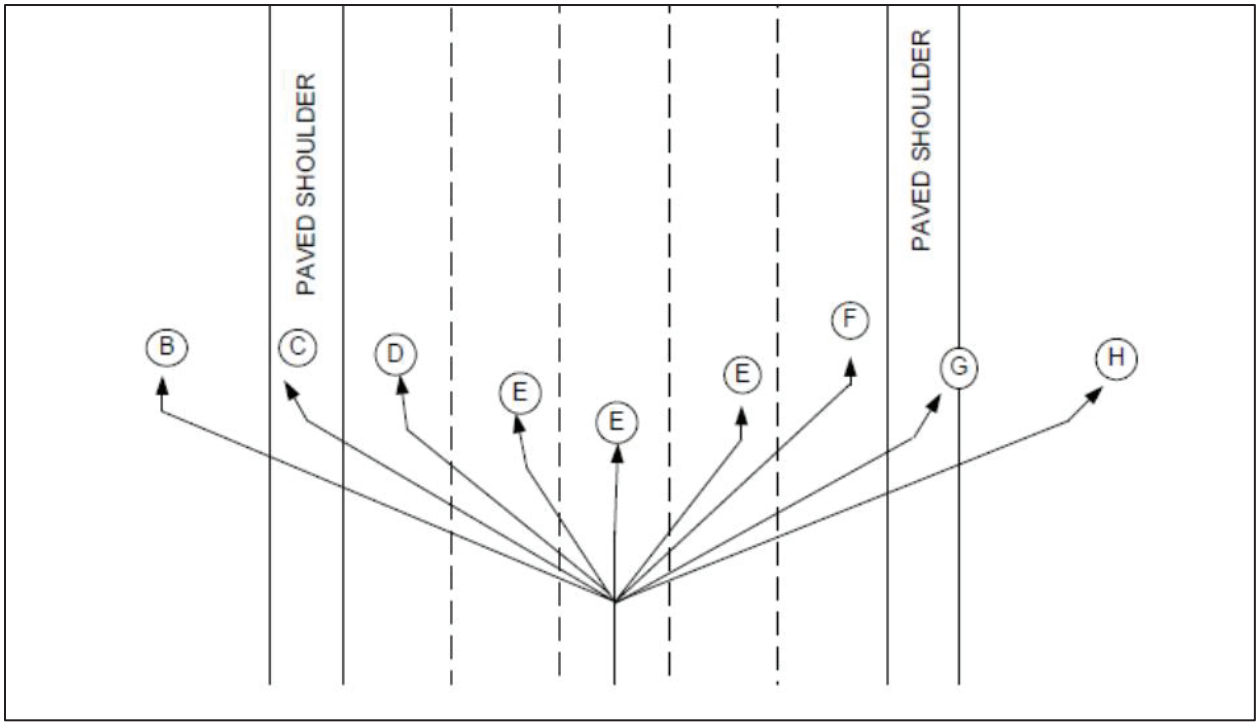


Figure B.1: Diagram of Lane Coding for the "Location" portion for the Sequence of events Coding.

Table B.4: Table of Lane Identification Corresponding with Code Values.

Lane Code	Roadway Area Description
A	Left area beyond median or traffic stripe, multiple lanes. Two-lane, two-way roadways do not have an A lane
B	Left area beyond shoulder
C	Left area shoulder. Within the limits of the shoulder area on the drivers left. May be paved, dirt or gravel
D	Left lane of travel
E	Interior lane. Center lanes of travel on a multi-lane roadway
F	Right lane of travel. Includes driveways
G	Right shoulder area. Within the limits of shoulder area on the drivers right. May be paved, dirt or gravel
H	Right area beyond shoulder

Lane Code	Roadway Area Description
I	Gore area. This is at the point of divergence within the raised area of the ramp nose
J	Other. Includes turn pockets, traffic islands, painted gores, and paved turnouts
V	High occupancy vehicle (HOV) lanes. Also called car pool lanes or diamond lanes
W	Buffer area. Associated with HOV lanes. Located between the HOV lanes and the left-hand lane of travel
	* Does not apply. End of collision sequence. The action or movement of a party that could have contributed to the accident
	Note: Express, acceleration, toll, fast track and bus lanes are coded using the lane location codes ***D, E, or F.***
	Exceptions: Toll or bus lanes that have diamond shapes in them. Use location code ***V***

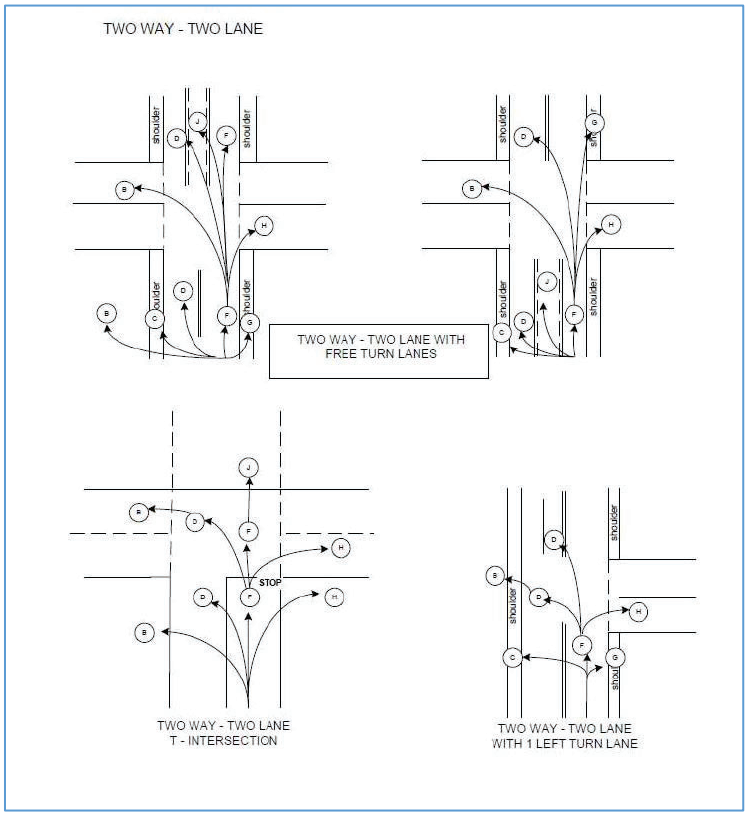


Figure B.2: Diagram example of the complexity of Lane Identification for Sequence of Events.

Utilized eTCR Content

The party type, direction of travel, and movement preceding collision are all extracted from the eTCR as indicated previously. The narrative and diagrams found in the eTCR are also needed to determine the SOE and is described below.

Narrative

A substantial portion of the eTCR document is dedicated to the Narrative. The reporting CHP officer documents statements, assesses the cause of the collision, and describes in detail the events that took place before, during and after the collision. They also write up a basic summary detailing the final resting positions of involved vehicles and their orientation. All Areas of Impact are also described in terms of where on the highway each event occurred as appropriate. Since each collision is unique and written up by different CHP officers, each narrative will vary to fit the needs of the document.

Typical sections for the narrative are described in the following list. Any narrative can have almost any combination.

- Notification
- Scene
- Parties
- Statements
- Cause
- Recommendations
- Summary
- Other Factual Information
- Relevant Details
- Intoxication or Substance Involvement
- Physical Evidence

Diagrams

Factual diagrams are typically present with every collision unless there was no additional information or clarification to be derived from a diagram. As in the case of a picture saying a thousand words, the diagram can clarify visually what language cannot easily convey. Take for example Figures B.3 and B.4, these demonstrate the complexity of the collision events and how readily the motions are described pictorially. Until machine vision is a more accessible technology, it will be difficult to utilize diagrams for the Sequence of Events lane location coding. Figure B.3 shows an irregular intersection with gore points between some of the turn lanes and the number of lanes between the two

intersecting highways not being equal. It would take the reporting officer an extremely long time to accurately describe this intersection verbally as opposed to in a diagram. In Figure B.4, a vehicle moving between lanes where attenuator barrels are present would also be difficult to describe verbally.

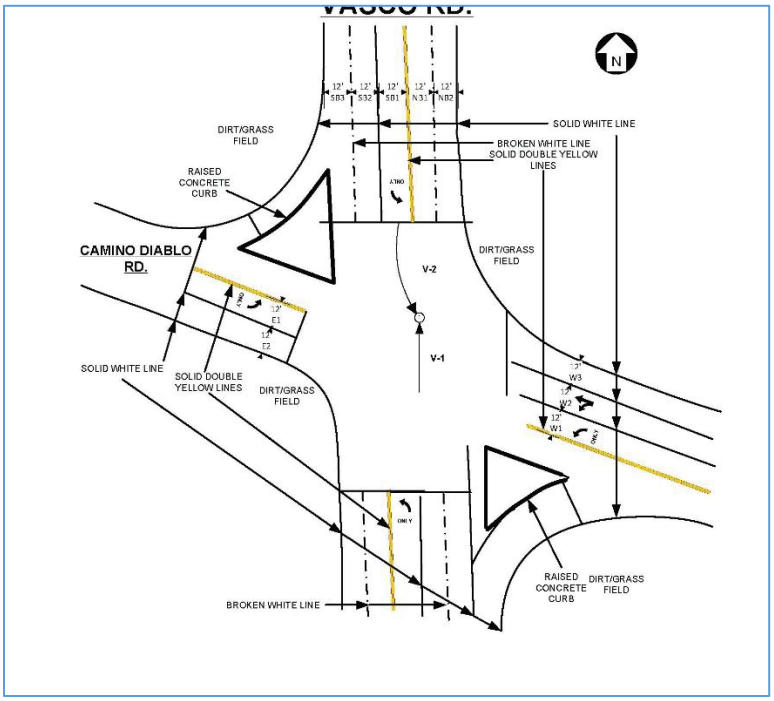


Figure B.3: Diagram example of intersection complexity of lane identification for intersection.

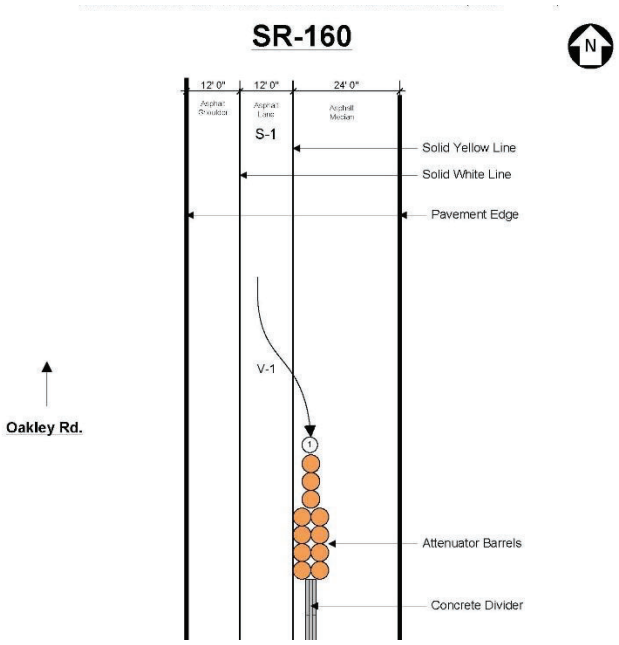


Figure B.4: Example of complexity in lane identification when objects are on the highway.

Calculating the Sequence of Events

Simple Cases

TCRPRO determines the Sequence of Events for a set of collision categories known as “simple cases”. The criteria for these cases are:

- number of parties == 2
- number killed ==0
- both parties are drivers
- type of collision == C (Rear End)
- motor vehicle involved with == C (other motor vehicle)
- pedestrian's action == A (No Pedestrians involved)
- vehicle type are all in [1, 7, 8, 22, 24, 25, 26, 27]
- movement preceding collision is in ['A', 'B', 'C', 'D'] (**Stopped, proceeding straight, ran off road, making right turn)**
- only one Area of Impact

Lane Location Algorithm

Due to the complexities in written language, and the inability of current technology to “read” pictures, it was determined at this phase that we forego the ability to reliably calculate an accurate “Location” field parameter.

In discussions with the team at Caltrans, it was explained that the diagram is a crucial tool when it comes to determining the “Location of Collision” category. Since this is a complex task that standard software engineering is not suited for, we did some research on a machine learning method for this. Although the results were promising, there are multiple obstacles that prevented us from implementing this in the final release. First is the lack of time. This is a difficult research problem that would require multiple months of research by a qualified researcher. Normally, it would be ready for publishing at this point, but it would take even longer to attempt to bring this into production. Since we had so many other tasks to focus on for this project, it was not possible to achieve desirable results. In particular, we cannot guarantee that the machine learning model we create now would work on future images, as we do not know the entire set of images that we will need to run through the model. Furthermore, in the limited time that we had, we were unable to achieve accuracy close to 100%.

Given that we could not use the diagram, we attempted to use text information. Modern software engineering also struggles at this, since software cannot understand a paragraph in the same way that a human can. We did attempt to use string matching to identify a specific lane, but this proved to be inconclusive. For example, even if a specific lane was mentioned in a report, there is no guarantee that the collision occurred on that lane. Furthermore, we did not have sufficient information about all the lane information in the state. Even if we did, we would require exact location information to determine the location category from a lane number and lane layout at a particular location.

Results

Statistics from SOE comparison, for simple cases. Excludes VHI:

1. About 19% are considered simple case by us
2. Primary object struck correct 99.9% of time.
3. Party type correct 99.3% of time.
4. Movement preceding correct 99.9% of time
5. Direction of travel correct 99.9% of time
6. LOC attempted on 4.5% of simple cases
7. LOC correct 53% of time.

APPENDIX C: PRELIMINARY REPORT ON TASAS AND INJURY DATABASE DEVELOPMENT

This Appendix contains the Preliminary Report issued after research was performed on our initial findings of processing Traffic Collision Reports using computer software.

Advanced Highway Maintenance and Construction Technology Research Center

Department of Mechanical and Aerospace Engineering
University of California at Davis

Traffic Accident Surveillance and Analysis System (TASAS) and Injury Database Development

Bahram Ravani: Principal Investigator
&
Patricia Fyhrie, Arash Gobal, Hiu Hong Yu, Travis Swanston

Report Number: TBD
AHMCT Research Report: UCD-ARR-17-07-01-01
Final Report of Contract: 65A0560 Task ID 2906

JULY 17, 2017

California Department of Transportation, Division of Research, Innovation and System Information

ABSTRACT

This research study discusses the development of methods and algorithms for streamlining the process of data coding for the Traffic Accident Surveillance and Analysis System (TASAS) database as well as evaluates the extent at which the data coding process can take advantage of digitization. TASAS information is useful in planning and improving safety of California highways. The TASAS database is populated based on coding the data that is included in police Traffic Collision Reports (TCRs) resulting in an electronic database that can easily be used by the California Department of Transportation (Caltrans). The process of data coding for TASAS is a cumbersome and time-consuming process prone to errors and inefficiencies resulting in a backlog of reports that need to be coded on a periodic basis. Furthermore, the existing process does not take advantage of the new digital technology and new rollout of digital police reports by the California Highway Patrol (CHP). This research study discusses the development of a set of complete requirements that will facilitate algorithmic developments for automating the data coding process. The developed research also includes an experimental web-based coding tool that can be used by TASAS coders as a decision support system to streamline their coding in the database. This web based tool was then used in a pilot study that coded 500 TCRs to evaluate the extent to which the data coding process can be digitized and automated. The pilot study results indicate that for location coding, the web-based tool generated results that were approximately 90% within 0.1 mile and 74% within 0.01 mile of the manually coded results. In terms of automatic coding of data other than location coding, the pilot study results indicate that such coding can be fully automated if the data is included in the digital form of the TCRs. This research study has also identified the challenges towards full automation of the coding process, some of which could be resolved with proper formatting of digital TCRs while others require further research into use of machine learning and Artificial Intelligence (AI) methods.

Summary

Research Objectives and Methodology

The Advanced Highway Maintenance and Construction Technology (AHMCT) Research Center at the University of California, Davis completed research on methods for streamlining and digitizing the process of data coding for the Traffic Accident Surveillance and Analysis System (TASAS) database. TASAS information is useful in planning and improving the safety of California highways. This database is populated based on coding the data included in police Traffic Collision Reports (TCRs) resulting in an electronic database that can easily be used by Caltrans to evaluate specifics and trends of collisions on California highways. This work was timely since the California Highway Patrol (CHP) has started the process of digitizing TCRs. The objective was to perform research on developing tools and algorithms (to a practical extent) that can streamline the process of data extraction and coding for the TASAS database with the expected outcome of improving efficiency, reducing human error, and allowing for the timely availability of information in the TASAS database.

The duration of the proposed research was 24 months and the results of this research can provide data and answers to the following research questions:

1. The extent to which data coding in terms of determination of District, County, Route, post-mile, travel direction, and post-mile markers can be digitized and streamlined.
2. The extent to which data from electronic TCRs can be digitally extracted and automatically put into the TASAS database.
3. The extent to which data from the narrative portion of police TCRs can be automatically extracted and codified and used to digitally populate TASAS as well as the Work Zone Injury Database.
4. How can the use of electronic digitization technology be implemented into Caltrans' workflow for coding and populating the TASAS database?

Results and Recommendations

This research study developed a set of complete requirements that need to be considered in automating or semi-automating the data coding process using police TCRs to populate the TASAS database. These requirements have helped to drive some of the requirements in the CHP rollout of digitized TCRs. Furthermore, these requirements provide the basis for the application of software and Information Technology (IT) towards automating the TASAS coding process. The research also developed an experimental web-based system referred to as AHMCT Web-based Coding Tool (AHMCT-WCT) that can be used by the coders as a decision support tool to streamline their coding process. This web-based coding tool can handle the existing different formats of police TCRs and either automatically extract all the data needed for the TASAS database or provide a

visual display of top choices among alternatives on some of the data for the coders to select from, streamlining their coding process, reducing potential for error, and improving efficiency.

In terms of addressing the four research questions as well as experimentally validating the usefulness of the AHMCT-WCT, the researchers of this study performed a pilot study where 500 TCRS were coded. These 500 TCRs were collected from accidents that had occurred in all Caltrans districts and were coded manually by coders and by using the AHMCT-WCT.

The pilot study results indicate that for location coding (research question number 1) the web-based tool generated results that were approximately 90% within 0.1 mile and 74% within 0.01 mile of the manually coded results. These percentages indicate the extent to which the location coding can be automated with the present format of digital TCRs. The challenge areas for location coding were mostly for accident locations in some metropolitan areas where there are several highway on and off ramps intersecting with streets, making the automatic identification of the accident location more complex. For such cases, AHMCT-WCT can be used as a decision support system since it will generate a list of all possible locations and will display them on a map so that the coder can choose the most appropriate location for TASAS. An area for future research is to consider using machine learning and Artificial Intelligence (AI) techniques to investigate full automation of such cases.

In terms of the second research question related to automatic coding of data other than location coding, the research results indicate that such coding can be fully automated if the actual data is included in the digital TCR format. The initial digital TCRs in XML file formats obtained for evaluation missed some of the data included in non-digital TCRs. Coordination between the Caltrans TASAS group and CHP is recommended to make sure that all data relevant to TASAS be included in the digital XML formats of TCRs. The digital XML files of TCRs also did not include the diagrams contained in non-digital TCRs. Coordination is also recommended between the Caltrans TASAS group and CHP to make sure that such diagrams be included in some format. The web-based coding tool developed in this research task is able to codify diagrams from PDF formats.

In terms of retracting data from the narrative portion of a TCR (the third research question), the researchers found that this can be done rather reliably with electronic versions of TCRs either in PDF or XML formats.

The first step in addressing the fourth research question in terms of methods to implement the digitized coding technology into Caltrans work flow, was to develop and experiment with the AHMCT-WCT. The coding tool was developed using open source resources so that its implementation into Caltrans' workflow can be facilitated. Furthermore, the research team has had several meetings with the Caltrans IT and coding personnel and has identified the need for an implementation phase for this research. This implementation phase is also another area in need of future work.

Acknowledgments

The authors thank the California Department of Transportation for their support, in particular Hamid Ikram and Juan Araya of the Division of Research, Innovation, and System Information as well as Mandy Chu, Phillip Poon, Brain Diomsic, David Ly, and Chitra Chitturi, of Caltrans and Jennifer Mercado and Isaac Tillman of CHP. The authors acknowledge the dedicated efforts of the AHMCT team who have made this work possible.

Introduction

Problem

The process of collecting data from police Traffic Collision Reports (TCRs) for state highway collisions and coding them into the Caltrans Traffic Accident Surveillance and Analysis System (TASAS) database is a cumbersome and time-consuming process prone to errors and inefficiencies resulting in a backlog of reports that need to be coded on a periodic basis. The aim of this research task was to utilize advancements in technology to streamline the process of data extraction and data coding for populating the TASAS database.

The research also initially aimed to add to an injury database developed earlier by the Advanced Highway Maintenance and Construction Technology (AHMCT) research center. At the time of this proposed effort, the California Highway Patrol (CHP) was planning to roll out complete electronic version of TCRs by October of 2015. This roll out, however, encountered challenging issues due to several technical reasons related to the TASAS database, including the electronic files having missing data crucial for TASAS as well as the TCRs being in several different formats. Addressing these challenges and issues required much more effort in this research task. There was no effort, therefore, to do anything about updating any portions of the AHMCT's injury database.

Objectives

This research expanded on the previous work of the AHMCT research center in the development of a work zone injury database by developing tools and algorithms for the purpose of facilitating the coding as well as data extraction and processing for the TASAS database. The objective was to perform research on developing tools and algorithms (to a practical extent) that can streamline the process of data extraction and coding for the TASAS database with the expected outcome of improving efficiency, reducing human errors, and allowing for the timely availability of information as TASAS information is useful for planning and improving the safety of California Highways.

Scope

The scope of this research involved coordinating with the CHP in terms of the rollout of their electronic TCRs so that it would contain data that is needed for TASAS. In addition there was a need for research on methods and algorithms that can automatically or at least semi-automatically extract similar data from TCRs not generated digitally coming from other local police agencies or from any legacy TCRs due to backlogs or existing reports that have not yet been coded or extracted into the Statewide Integrated Traffic Records System (SWITRS) and TASAS.

Background

TASAS is an electronic database and data processing system that contains data for collisions that are state highway related. Each collision record in the database is referenced to a post-mile address that ties to the highway database. The highway database contains data on 15,200 miles of highway, 20,000 intersections, and 16,000 on and off ramps in California [1].

Caltrans' Collision Post-Mile Coding unit processes police TCRs and assigns specific location values. The collision detail information is then transferred to the SWITRS and TASAS databases. There are variations in police TCRs since these are provided by approximately 100 CHP area offices plus over 400 local police departments, making the coding and data extraction process based on non-uniformly prepared TCRs.

CHP was planning to roll out an electronic system in October 2015 so that all the TCRs from the 100 CHP area offices will be prepared electronically. This research task was initially intended to develop techniques and algorithms for automatically extracting data from such a digitized system into the TASAS database. In addition, the research was also intended to evaluate the development of methods and algorithms to extract data automatically or at least semi-automatically from none-digitized TCRs so that legacy backlogs and reports from CHP allied agencies that may not have their reports digitized can be handled. The electronic rollout of the CHP reports, as expected, was very complicated, requiring several iterations, which are continuing. Some iterations did not contain all the data needed for the TASAS database. This research team and CHP have

worked cooperatively to make the electronic TCRs include data needed for the TASAS database.

This research has built upon the demonstrated previous work of AHMCT in developing an injury database of accidents near or at California highway work zones.

Research Methodology

The research methodology was based on some of the technologies developed at AHMCT in developing an injury database for work zone accidents. AHMCT had utilized a rudimentary Optical Character Recognition (OCR) technology for reading data from TCRs. The approach in this research was to optimize this OCR technology for none-digital TCRs while working directly with XML files from the CHP digitized TCRs to develop methods and algorithms for TASAS data coding. The research methodology consisted of the following nine tasks:

1. Form the Technical Advisory Group (TAG).
2. Requirements Development.
3. Development of the Optical Character Recognition System for TASAS.
4. Integration of Caltrans Geospatial Linear Reference System.
5. Development of Data Extraction and Matching Algorithms.
6. Pilot Study: Design and Implementation.
7. System Demonstration and Presentation.
8. Documentation and Final Report.
9. System Integration.

Each of these tasks, with the exception of tasks 7 and 8 dealing with demonstration, presentation and documentation, are described in the subsequent Sections of this report.

DEVELOPMENT OF REQUIREMENTS

In developing methods and algorithms for developing a decision support tool that would aid in automating and/or streamlining data coding for the TASAS database, the first step was to develop the requirements for this process. Development of detailed requirements provides the knowledge base needed for any subsequent efforts or research studies related to further automation or implementation of the work developed in this research study. The research team used the following steps to develop the requirements for this research task:

1. Understand TCR processing protocols within CHP and Caltrans for the TASA database.
2. Study a sample of previously coded, scanned versions of TCRs that were processed into the TASAS database.
3. Study coding manuals and other documentation used by Caltrans and CHP to "process" TCRs into the TASAS and SWITRS database formats.
4. Meet with Caltrans coders to understand how they currently do their job and determine what tools they use.
5. Obtain sample TCRs in electronic format and develop methods, algorithms, and software to extract all information from the TCRs using both XML and electronic PDF formats.

In order to develop the requirements for this research study, it is first important to understand the TASAS database as well as the different existing formats of the CHP TCRs. These different TCR formats need to be considered since the process of digitization of all CHP reports will take some time, and furthermore, there is need to process the existing legacy none-digital TCRs for the TASAS database.

TASAS Data Processing Requirements

The TASAS database contains the details from all traffic collisions that occur on Caltrans maintained highways. Two main types of data are needed for TASAS: the location data and the collision detail information. The location data ties the location of an accident to a post-mile marker location and the coding process is referred to as the "CARD 8" process. The detailed collision information is provided through SWITRS. The coding unit at Caltrans performs the CARD 8 process and generates a report that will be attached to the TCR and is provided to CHP for entering specific collision information. The coding unit at Caltrans also utilizes a computerized listing of all valid California highway segments, ramps, and intersections, which is referred to as the Highway Sequence Listing (HSL) as part of the CARD 8 coding process.

In TASAS, data is imported and exported from the database via files using a *.CSV format where each data field is separated by a comma. Figure C.1 shows a commonly used display format of TASAS data. The data fields in Figure C.1 are output from *.cvs files with the following format: *Column1_data, Column2_data ,,,,, Column39_data,*

Column40_data. The field names and respective location in TASAS record can be seen in **Error! Reference source not found.**.

Table C. 1: A Screenshot of TASAS Database output.

DI	RTE NO	S U F	CO	P R E	POST MILE	HIGHWAY H G	A C	M T	B A	LANES LT	LANES RT	R U	F T	I R L	S O H	D A Y	ACCIDENT DATE MM-DD-YY	TIME HHMM	COMMON ACCIDENT NUMBER	P C F	ENVIR COND W	L	S	R C	R W C	T O C	NO MTR VEH	P T	D I R	V H I	S I	PERSON K	PERSON I	O S P / L O C	O S O / L O C	O S O / L O C	O S O / L O C	OA F 12	M O V	SD P 12
03	049		ED		014.891	U	C	B	Z	01	01	B	I	5	S	6	04-13-12	2126	090100199	6	C	C	B	H	A	D	02	A	W	1	B	00	00	V2 F	----	----	----	N<	B	A<
																												A	S	2	B	00	00	V1 F	----	----	----	N<	B	A<
03	049		ED		015.163	U	C	B	Z	01	01	B	I	6	S	4	03-28-12	0757	090100237	6	C	A	B	H	D	A	02	D	W	1	B	00	00	V2 H	----	----	----	N<	N	A<
																												H	E	2	B	00	01	V1 F	----	----	----	N<	B	A<
03	049		ED		016.340	U	C	B	Z	01	01	B	H		N	1	07-01-12	2020	924512833	1	A	A	A	H	D	E	01	C	N	1	C	00	01	43 H	44 H	----	----	6<	C	B<
03	049		ED		016.440	U	C	B	Z	01	01	B	I	5	N	7	05-26-12	2125	924518624	1	B	D	B	H	D	F	01	A	N	1	<	00	00	23 B	44 D	----	----	4<	E	B<
03	049		ED		016.650	U	C	B	Z	01	01	R	H		N	7	04-28-12	1215	924516629	1	A	A	A	H	D	E	01	D	N	1	B	00	00	27 H	43 H	----	----	6<	C	B<
03	049		ED		017.190	U	C	B	Z	01	01	R	H		N	3	03-13-12	0025	924513405	4	C	D	B	H	D	E	01	D	N	1	B	00	01	28 B	----	----	----	J<	C	A<
03	049		ED		018.620	U	C	B	Z	01	01	R	H		N	1	02-12-12	0640	924514712	6	A	D	A	H	D	E	02	A	N	1	C	00	00	----	----	----	----	M<	N	A<
																												A	S	1	C	00	01	24 H	28 H	29 H	----	N<	B	A<
03	049		ED		022.670	U	C	B	Z	01	01	R	H		N	2	05-28-12	1505	924516451	5	A	A	A	H	D	F	01	C	N	1	C	00	01	44 F	----	----	----	N<	B	A<
03	049		ED		022.865	U	C	B	Z	01	01	R	I	5	N	3	06-12-12	1228	924513603	3	A	A	A	H	A	D	02	A	N	1	C	00	00	V2 F	----	----	----	N<	E	A<
																												A	S	1	C	00	00	V1 F	----	----	----	N<	B	A<
03	049		ED		022.990	U	C	B	Z	01	01	R	H		S	7	07-07-12	2329	924518240	4	A	D	A	H	D	E	01	A	<	1	C	00	00	28 B	----	----	----	N<	B	G<
03	049		ED		028.120	U	C	B	Z	01	01	R	H		N	5	05-24-12	0910	924516056	4	A	A	A	H	D	E	01	A	N	1	C	00	01	28 H	43 H	----	----	N<	C	A<
03	049		ED		032.080	U	C	B	Z	01	01	R	H		N	4	08-22-12	1300	924518624	1	A	A	A	H	D	<	01	A	N	1	C	00	00	23 B	----	----	----	4N	E	B<
03	049		ED		032.490	U	C	B	Z	01	01	R	H		S	7	05-05-12	1810	924517089	5	A	A	A	H	D	E	01	C	S	1	C	00	01	22 B	24 B	27 B	44 B	N<	B	A<
03	049		ED		034.500	U	C	B	Z	01	01	R	H		N	4	04-25-12	1725	924517604	1	B	A	A	H	D	B	02	A	N	1	C	00	00	V2 G	----	----	----	6<	L	B<
																												A	N	1	C	00	00	V1 F	----	----	----	N<	B	A<
03	049		ED		035.060	U	C	B	Z	01	01	R	H		N	1	07-08-12	0140	924518240	4	A	D	A	H	D	E	01	C	N	1	C	01	00	23 H	28 H	44 H	----	5<	C	B<
03	049		ED		035.520	U	C	B	Z	01	01	R	H		S	4	07-18-12	1350	924517728	5	A	A	A	H	D	C	02	A	S	1	C	00	01	V2 F	----	----	----	N<	H	A<
																												A	S	1	C	00	00	V1 F	----	----	----	N<	H	A<
03	049		ED		036.120	U	C	B	Z	01	01	R	H		N	6	05-25-12	1410	924516056	1	A	A	A	H	D	A	01	A	N	1	C	00	00	43 B	23 B	----	----	4<	C	B<
03	049		ED		036.300	U	C	B	Z	01	01	R	H		N	6	07-13-12	1445	924518624	6	A	A	A	H	D	H	02	A	N	1	C	00	02	V3 F	24 H	----	----	N<	B	A<
																												E	S	1	C	00	00	----	----	----	----	6<	B	G<
																												O	<	1	<	00	00	V1 D	----	----	----	<<	R	<<
03	049		ED		036.320	U	C	B	Z	01	01	R	H		N	6	07-13-12	1442	924518624	C	A	A	A	G	D	E	01	A	N	1	C	00	00	24 H	43 H	----	----	N<	H	A<

Since the content of each field is populated with letters and integers referring to "look-up" tables, reference documentation are used by coders to populate these fields (see, for example, Figure C.18 through Figure C.23).

To populate the TASAS database, information taken directly from the TCR will complete most of the fields. The remainder will be populated/filled-in with the Location Code data and data from the narrative section of TCRs.

In this research study, *.CSV files are used to transfer the data contained in a TCR to the TASAS database. The data fields are described in Table C. 2. The format of the output *.CSV will be as follows:

Column1_data, Column2_data ,,,,, Column39_data, Column40_data

If Column27_data is greater than 1, then a second, along with potentially more subsequent lines, will need to be output. This will be in the following format:

Blank1, blank1, … blank28, column29_data, … column40_data

Note: the “Party Data” is contained in columns 29-40. Thus for each party, a listing for columns 29-40 will need to be generated. The number of parties is defined as Number of Motor Vehicles (column 27).

Table C. 2: Field Names and their Respective Locations in TASAS records.

Column #	Field Name		Column #	Field Name
1	Caltrans District		21	Weather
2	Route number and suffix		22	Lighting
3	County (2-3 letter)		23	Road Surface
4	Post-mile Prefix		24	Road Condition
5	Post-mile value		25	Right of Way Controls
6	Highway Group (HG)		26	Type of Collision
7	Access Controls (AC)		27	Number of Motor Vehicles
8	Median Type		28	empty
9	Barrier type		29	Party Type
10	No. of Lanes Left		30	Direction of Travel
11	No. of Lanes Right		31	Vehicle Highway Indicator
12	Rural		32	Special Information
13	File Type		33	Number of person killed
14	Inter. Ramp Loc. (IRAL)		34	Number Injured
15	Side of Highway		35	OSP
16	Day of the Week		36	LOC
17	Accident Data		37	OSO
18	Accident Time		38	LOC
19	Accident Number		39	OSO
20	Primary Causal Factor		40	LOC

"CARD 8" Information

The terminology for "CARD 8" is prevalent when investigating location coding. In short, "CARD 8" is the form on which collision location data is documented. An example of this form is seen in Figure C.1. Although the form itself is not used in this research study, systematically generating its content has been used in evaluation of the methods and algorithms developed in this research study and its efficiency compared with the existing methods. How to determine the values that comprise "CARD 8" content is further explained in the "Location Coding (CARD 8) Process" section.

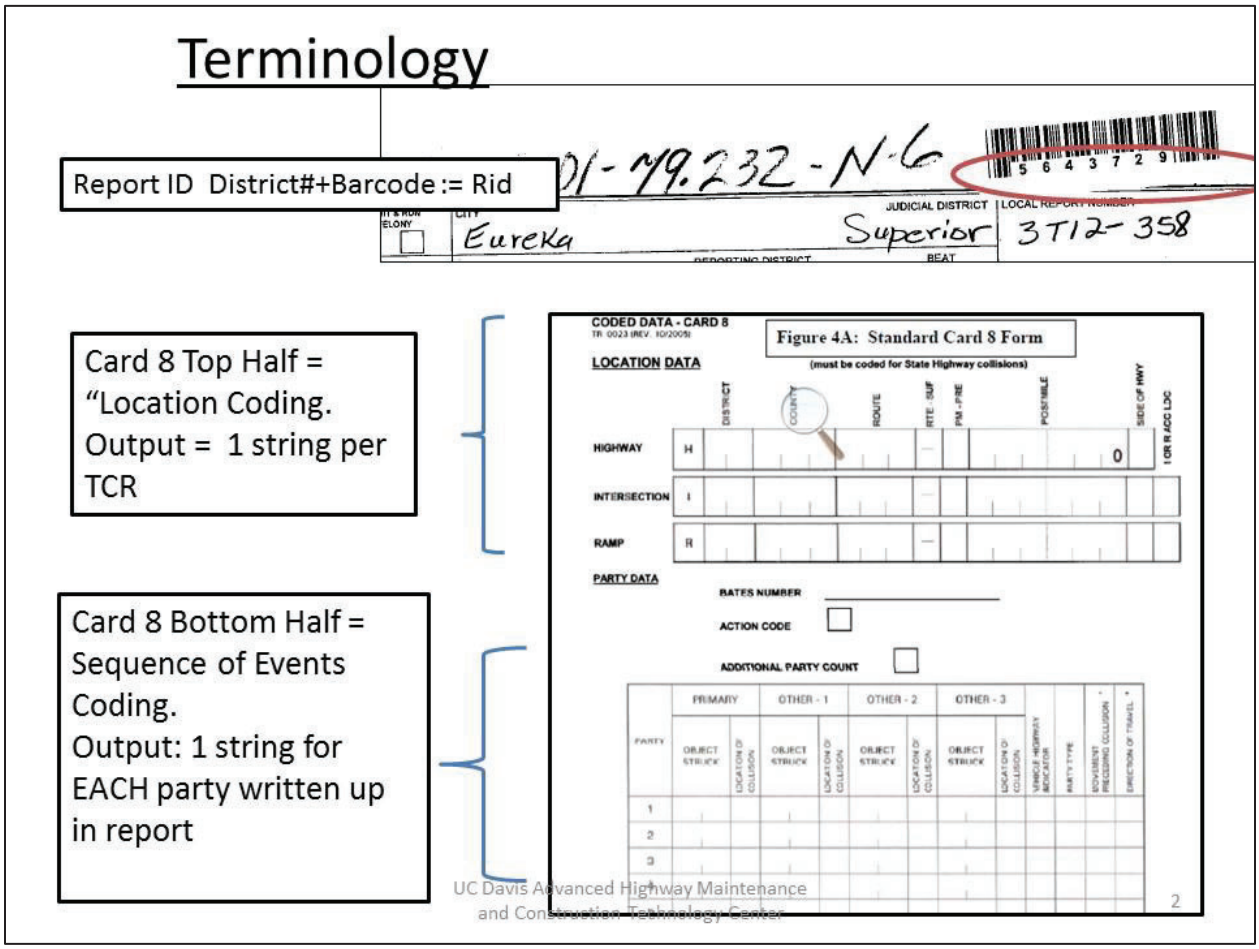


Figure C.1. The "CARD 8" form along with some descriptions.

CHP Traffic Collision Reports

The input for the TASAS database is based on the TCRs which are generated by CHP. Not all TCRs generated by CHP are logged into the TASAS database (for example those that occur on city streets are not). Only those collisions that are located on a Caltrans maintained highway populate the TASAS database.

The portions of the TCR from which data are extracted are: The “Location Box” (found in the front page of the report), diagrams, and parts of the narrative (Area of Impact and Summary). The “Location Box” (Figure C.2) contains multiple pieces of data concerning the exact location of the collision.

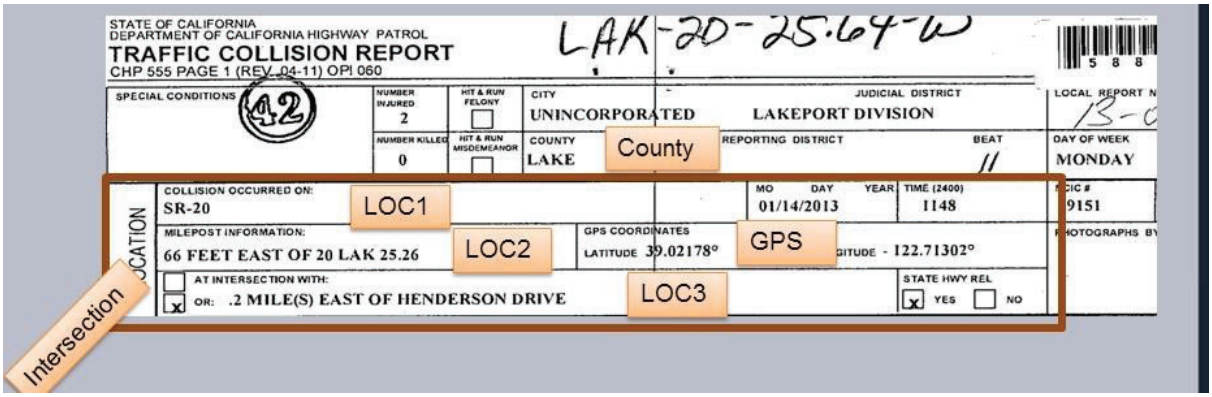


Figure C.2. Example of top portion of a CHP Traffic Collision Report. Highlighted here is the "Location Box" which provides a summary of the geological location of the collision in terms of state route and position.

In order to digitally “read” the TCR, details about its format is required. Since December 2015, the format of CHP TCRs can either be electronic or imaged, meaning a scanned file. The electronic version is referred to here as “ETCR” and the image format is referred to as “ITCR.” Additional information is provided below.

Electronic Version (ETCR) requirements

The ETCR is originated as a standardized “PDF Form” where data can be entered directly into the PDF file. Along with the *.PDF file itself, there is an associated *.XML file as well. In order to extract the data from the ETCR, a standardized and stable PDF form is needed. Although OCR processing is typically not used with these files, the text itself is legible. In order to associate the text with a given “field” however, boundary boxes need to be calculated for each page of the PDF. If there are any physical differences in the form’s digital format, then the text is not reliably placed in the fields.

Image format (ITCR) Version requirements

The ITCR is provided when the ETCR format is not available. The existing method of printing a completed report and sending it to Caltrans, where it is then scanned in, is used. The scanned report is an image file where text is not immediately recognizable. These image files require a more complicated processing mechanism that includes OCR along with the graphical boundary locations on the form. The requirements for ITCRs is a sample of an empty form where the field boundary boxes are measured in pixels. Clean, typed text must also be present on the completed form, otherwise the OCR portion of the post-processing will not be possible.

Location Coding (CARD 8) Process

The process that calculates the Location Code from a TCR is systematic, and it is essential to fully understand this process to emulate it via computer programming. It is reasonable to assume that any digital processing will not be 100% accurate.

The location coding process is started by obtaining specific data from the TCR. There are six pieces of information that are coded in this process forming the six points of the location coding process. These six points are as follows:

1. Location Box on TCR with three lines of information
2. Compass …. Found on the diagram
3. Diagram(s) … some small, some large
4. Summary – the paragraph summarizing the collision
5. Area Of Impact (AOI) or Point Of Impact (POI) - paragraph indicating the area of impact or point of impact. If more than one is given, we only need the first one
6. Cause: paragraph explain the cause ... use to confirm side of highway

The resulting “Location Code,” from employing these 6 points, is used to populate the top half of CARD 8 (Figure C.1). The fields are as follows:

Location Code Details

1. District
2. County
3. Route/highway
4. Route Suffix
5. Post-mile Prefix
6. Post-mile marker value

7. Side of highway
8. Intersection or ramp location code

The county is directly provided on the top center part of the TCR as shown in Figure C.2. The county then determines the Caltrans district number. The county is listed in the TASAS database in the abbreviated format. Table C. 3 provides the relationship between the counties and Caltrans districts. The route number and optional suffix is typically found in the TCR's Location Box. Sometimes the highway's local name is used, and so the highway number must be verified in the narrative section. The "Side of Highway" value is typically found on the TCR's first page under Party information in Direction of Travel. This value also should be verified in the narrative. Items #5, #6, and #8 will be explained below.

Table C. 3: Table of Caltrans District numbers and Corresponding Counties.

District	County Name	County Abbrev.		District	County Name	County Abbrev.
4	Alameda	Ala		12	Orange	Ora
10	Alpine	Alp		3	Placer	Pla
10	Amador	Ama		2	Plumas	Plu
3	Butte	But		8	Riverside	Riv
10	Calaveras	Cal		3	Sacramento	Sac
3	Colusa	Col		5	San Benito	SBt
4	Contra Costa	CC		8	San Bernardino	SBd
1	Del Norte	DN		11	San Diego	SD
3	El Dorado	ED		4	San Francisco	SF
6	Fresno	Fre		10	San Joaquin	SJ
3	Glenn	Gle		5	San Luis Obispo	SLO

1	Humboldt	Hum		4	San Mateo	SM
11	Imperial	Imp		5	Santa Barbara	SB
9	Inyo	Iny		4	Santa Clara	SCl
6	Kern	Ker		5	Santa Cruz	SCr
6	Kings	Kin		2	Shasta	Sha
1	Lake	Lak		3	Sierra	Sie
2	Lassen	Las		2	Siskiyou	Sis
7	Los Angeles	LA		4	Solano	Sol
6	Madera	Mad		4	Sonoma	Son
4	Marin	Mrn		10	Stanislaus	Sta
10	Mariposa	Mpa		3	Sutter	Sut
1	Mendocino	Men		2	Tehama	Teh
10	Merced	Mer		2	Trinity	Tri
2	Modoc	Mod		6	Tulare	Tul
9	Mono	Mno		10	Tuolumne	Tuo
5	Monterey	Mon		7	Ventura	Ven
4	Napa	Nap		3	Yolo	Yol
3	Nevada	Nev		3	Yuba	Yub

Post-Mile Markers

In TASAS, the location of the collision must be in terms of "post-mile" values along a given highway. Sequence listings provide the post-mile data for all 12 Districts. Generally, there is a post-mile marker value at landmarks such as intersections, bridges, ramps, under-crossings, etc. Sequence Listings are divided up by District, County, highway and direction. A compilation of all this data results in large files and can be cumbersome to use. An example from Sequence Listing on Route 405 in district 12 is depicted in Figure C.3. Sample from District 12 Sequence Listing on Route 405.

ORA	IRVN	007.758	DR		NB ON FR MACARTHUR
ORA	IRVN	007.803	DH	00.029	MACARTHUR BLVD OC 55440
ORA		007.832	DH	00.336	
ORA		007.833	DR		SB OFF MACARTHUR/AIRPRT
ORA		008.168	DH	00.271	
ORA		008.295	DR		HOV NB OFF TO NB RTE 55
ORA		008.300	DR		HOV SB ON FR SB RTE 55
ORA		008.302	DH	00.137	
ORA		008.341	DR		NB OFF TO RTE 55
ORA		008.342	DR		DUM SB ON FR SB RTE 55
ORA		008.439	DH	00.010	RED HILL AVE OC 55-439
ORA		008.449	DH	00.138	
ORA		008.469	DR		SEG NBOFF TO SB RTE 55
ORA		008.535	DR		SBON FR NBOFF RTE 55
ORA		008.587	DH	00.065	NORTHEAST CONN OC 55438
ORA		008.652	DH	00.088	EAST CONN OC 55-421
ORA		008.704	DR		SEG NB ON FR ANTON BLVD
ORA		008.740	DH	00.110	JCT, 55/405 SEP 55-252
ORA		008.798	DR		SEG SB ON FR PAULARINO
ORA		008.850	DH	00.534	WEST CONN OC 55-422

DIST 12 RTE 405 **DIR S-N**

CO.	CITY	POSTMILE	GRF PT	LENGTH	DESCRIPTION
ORA		008.880	DR		NB OFF TO AVE ART/BRISTOL
ORA		008.916	DR		DUM NB ON FR SB RTE 55
ORA		009.005	DR		SB OFF TO NB RTE 55
ORA		009.036	DR		SEG NBOFF TO PAULARINO
ORA		009.172	DR		SEG NB OFF TO AVE OF ART
ORA		009.198	DR		HOV SB OFF TO NB RTE 55
ORA		009.341	DR		NB OFF TO BRISTOL ST
ORA		009.384	DH	00.116	DELHI DRN TRI 13X15 RCB
ORA		009.441	DR		SB ON FR BRISTOL ST
ORA		009.470	DR		NBON FR NB BRISTOL
ORA	CMS	009.500	DH	00.008	
ORA	CMS	009.508	DH	00.379	BRISTOL ST OC 55-431
ORA	CMS	009.663	DR		NBON FR SB BRISTOL
ORA	CMS	009.745	DR		SB OFF TO BRISTOL ST

Figure C.3. Sample from District 12 Sequence Listing on Route 405.

Some general rules that apply to post-mile markers are:

- Post-mile markers are set to zero for any given highway. Consequently, the county must always be included in the post-mile values.

- Post-mile values increase in the north direction of a N-S freeway or the east direction of an E-W freeway. Conversely, post-mile values decrease going south on a N-S freeway or west on an E-W freeway.
- Each ramp and intersection has its own unique post-mile data.

Intersection or ramp Location Code

For the last piece of data in the Location Code, the intersection or ramp location code, the values range from 0-5. They are further explained in Table C. 4 with corresponding examples shown in Figure C.5.

Table C. 4: Possible values for the "IRAL" field.

IRAL – Intersection / amp location code	
0 or blank	Not an intersection or ramp
1	Ramp Intersection (Ecit)
2	Ramp
3	Ramp Entry
4	Ramp Area, Intersection area
5	In Intersection
6	Outside intersection – Non-State Route

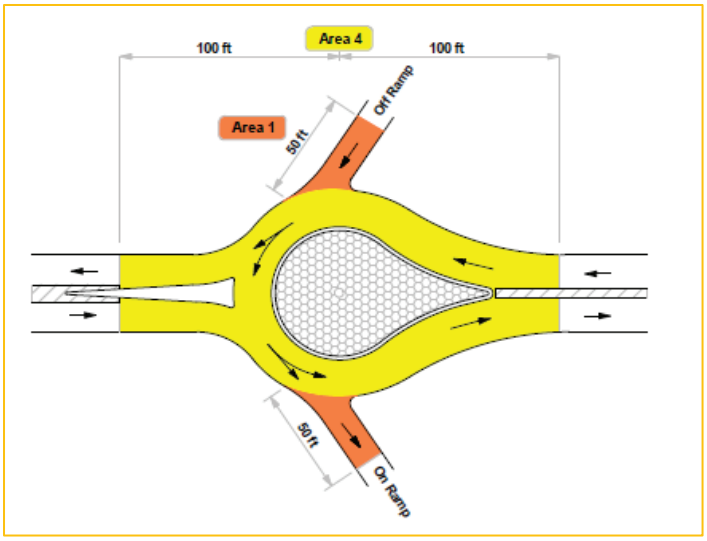


Figure C.4. Ramp location code for roundabouts.

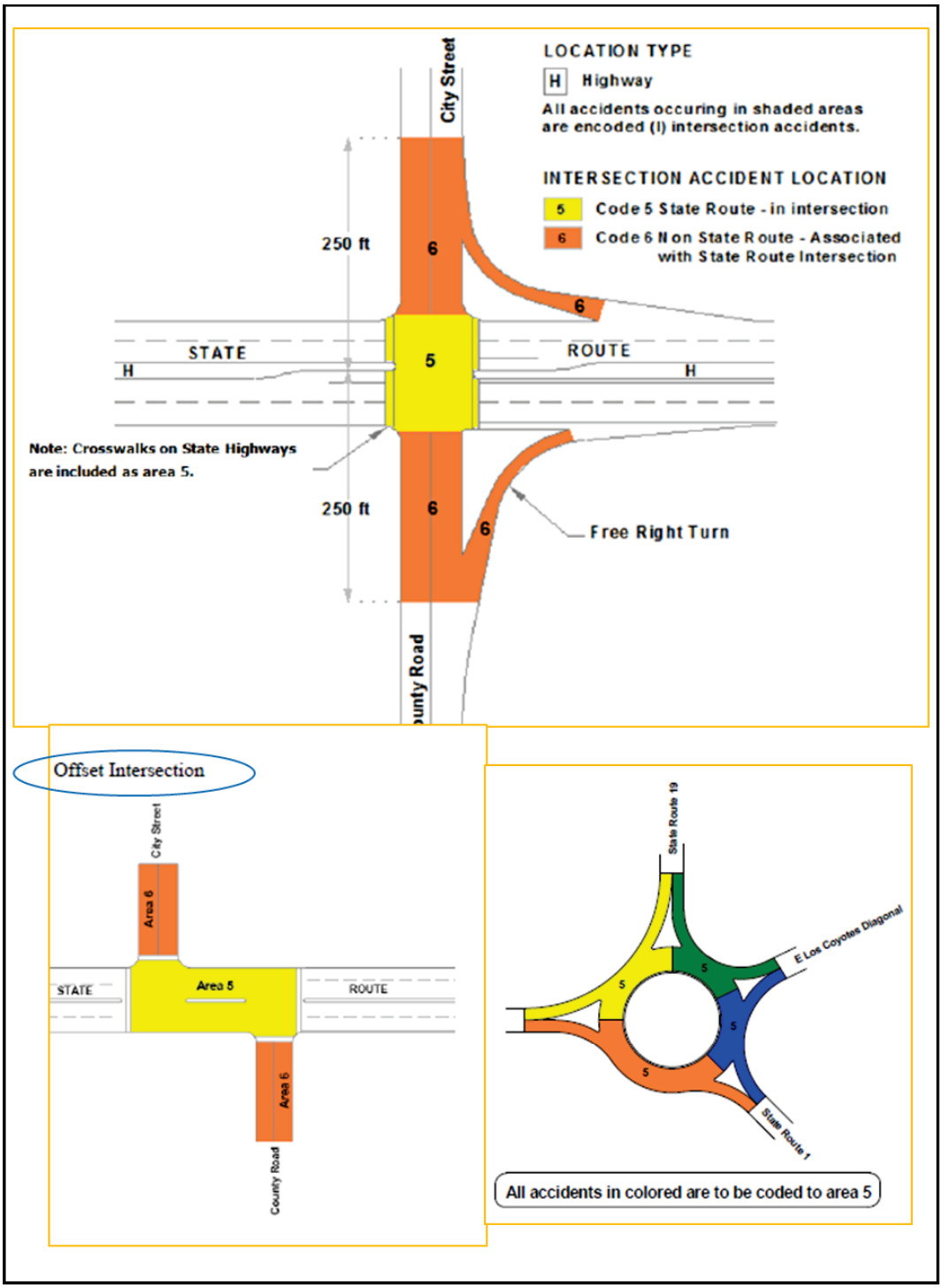


Figure C.5. Illustration of intersection locations 5 and 6.

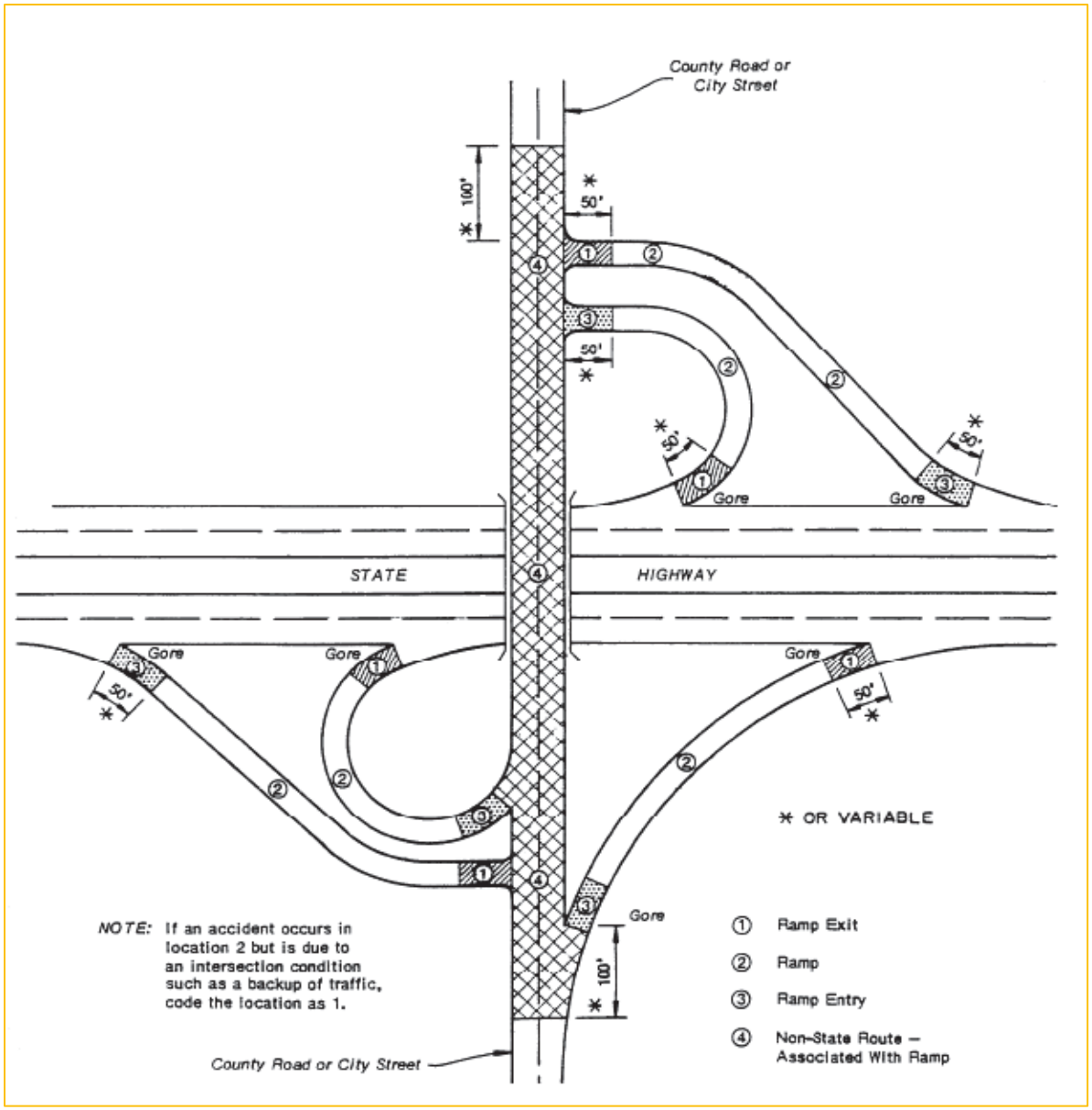


Figure C.6. Illustrations of ramp locations 1 through 4.

Location Code Calculation Logic

To determine the actual post-mile marker value for any Location Code, the “six points” discussed earlier in this section are followed. The information from the TCR’s Location Box can result in multiple post-mile marker values. When collision locations are described in terms of feet, the post-mile “rules” are to be applied. When a GPS value is provided, it needs to be converted to a post-mile marker value. The AOI for Party 1 is generally the sought-after collision location. When a “station line” is used to describe the location of the collision, the Location Code must reflect where the primary vehicle left the road, not

necessarily where the vehicle finally stopped. A flow chart of the process is shown in Figure C.7.

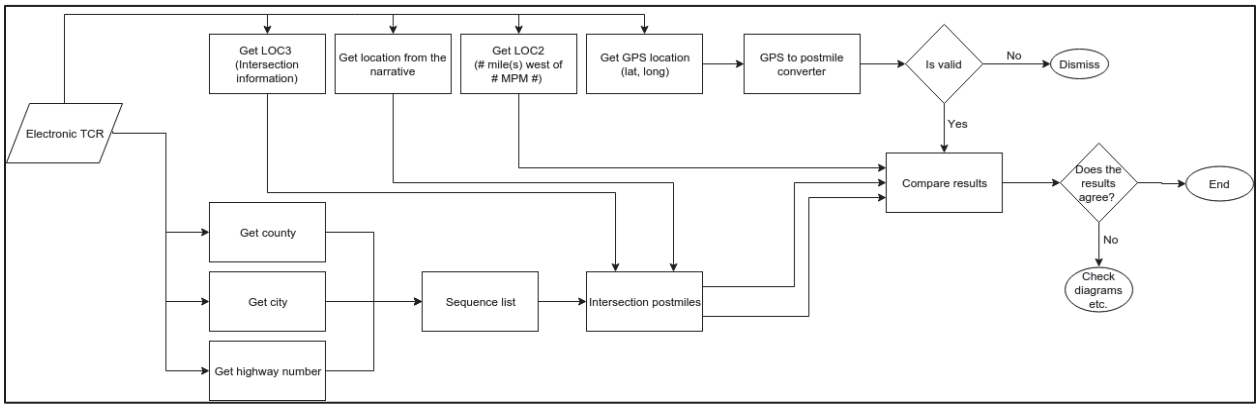


Figure C.7. Flow chart of retrieving specific fields of data needed to calculate post-mile data.

Geospatial Linear Referencing System

During the process of generating the Location Code, a mixture of post-mile marker data and GPS coordinate data are used. The Python code was written and is used to calculate the Location Code based on the information provided by the TCR. However, a function needed to be developed that was compatible with the Python code and could get either post-mile marker information if given a GPS value or provide the GPS value if given post-mile marker data.

Web-Based Coding Tool Design

In order to replicate the location coding process described in the previous section, the team decided to develop a Web-based Coding Tool (referred to as AHMCT-WCT) to display the array of potential Location Code “answers” based on the data found within the TCR. Along with the display of the potential Location Code values, each candidate is also displayed as a pin marker on a geological map (i.e. Google maps). In addition to that information, it was felt that bringing in all resources needed to make a robust decision should also be present and displayed. Specifically, segments from the appropriate sequence listings, diagrams from the TCR, summary narratives, and AOI narrative need to be displayed.

OCR DEVELOPMENT and TCR Processing

The software that was developed for OCR in this research study is called “TCR2pdf.” This is a standalone module that was designed and utilized by the AHMCT Research Center for this research study. The main purpose of this module is to process TCRs, in either ETCR or ITCR format, and convert them into a more readily accessible file format for data coding. It utilizes an OCR Software entitled Tesseract (see Appendix A for more details), but has much more functionality for processing TCRs.

TCR Processing Module

TCR2pdf is designed as a backend parsing module for the AHMCT-WCT developed as part of this research study. When a TCR is uploaded to the AHMCT-WCT, the TCR2pdf module will decode and extract all the needed information for location coding purposes as well as the data needed for subsequent TASAS database format downloading. Results from TCR2pdf will be returned as a structured dictionary of JavaScript Object Notation (JSON) database format. This module also extracts information from the following formats of TCRs: Extended Marking Language (XML), digital, and imaged. The content that is extracted consists of the following:

- Location box and content
- Direction of Travel (DOT) for each party
- Parties’ Section GPS values (latitude and longitude)

The program also generates, in the form of flags, when the TCRs are unreadable or do not have relevant or have inconsistent information for location coding. The prompt flags are created under the following conditions:

- Unable to process or read
- Unable to Lookup – Set if unable to lookup post-mile
- Latitude/longitude is invalid
- No post-mile information within the given range of GPS location
- Disagree County – Set if disagreement between post-mile county and CARD 8A county
- Disagree Route – Set if disagreement between post-mile route and CARD 8A route

Determining Report Format

When a file is selected using the AHMCT-WCT or identified by a preprocessing program, the file is input into the TCR2pdf module. Before any processing can be

executed, the type of report needs to be determined. This process is graphically shown in Figure C.8.

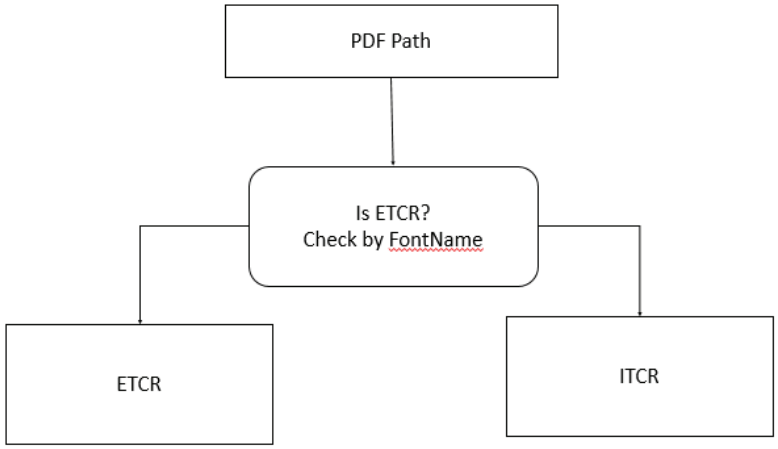


Figure C.8. PDF type layout

TCR2pdf supports both the ETCR (searchable content) and the ITCR (non-searchable image format). When the PDF is provided to TCR2pdf, the program first determines the type of report by searching for the field “FontName” from the PDF file. Since an ITCR file contains only image data and no text data, the “FontName” variable will not be present.

ETCR Processing Details

When a report is determined to be in ETCR format, TCR2pdf will check whether the accompanying XML file is available. The XML version is used only for supplemental data for information cross-checking since TCRs in XML format are not always available. Information extraction is done in one of two ways: either from XML data or without XML data. Figure C.9 shows the overview of category of information being extracted.

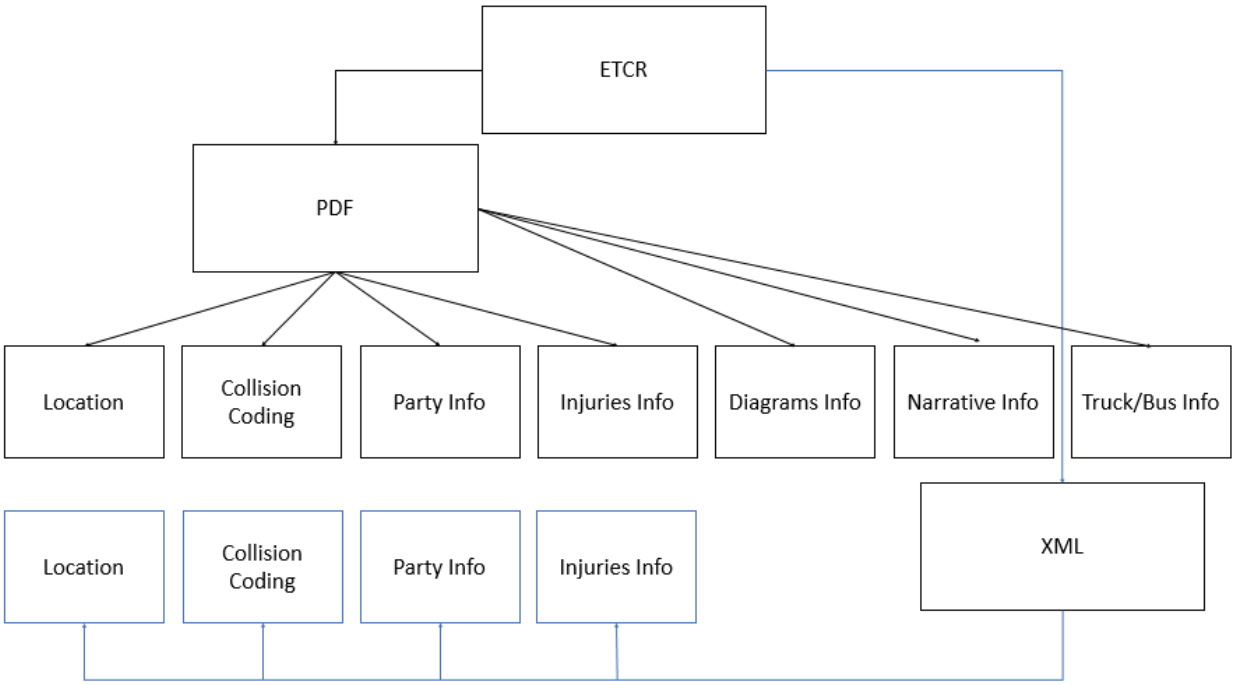


Figure C.9. ETCR data extraction method; XML file is only used to verify field data.

For each ETCR form, a set of predefined bounding boxes is defined to extract information from the report. The origin of the boundary box is measured in pixels from the lower left corner of the page. **Error! Reference source not found.** shows an example of t his definition.

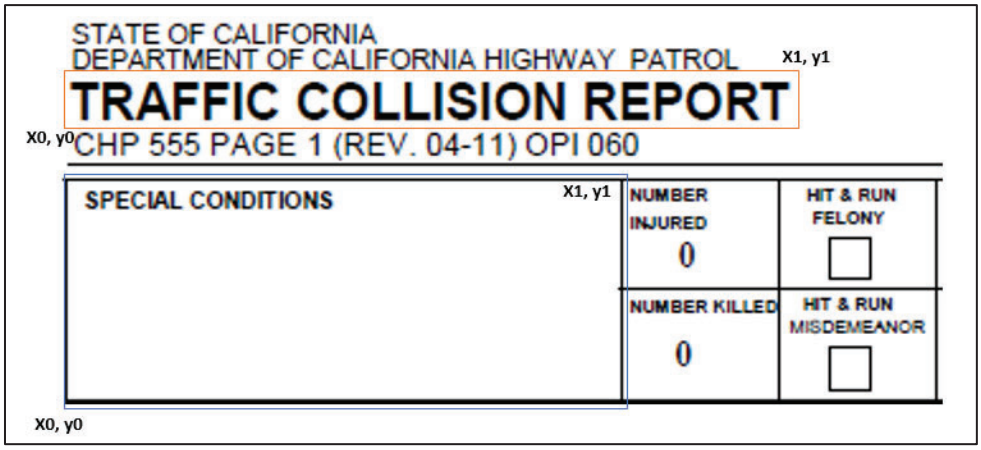


Figure C.10. Example TCR boundary box definition.

After TCR2pdf parses the PDF, a PDF tree is constructed by text coordinate using an inch to pixels coordinate conversion system, where the 0, 0 is on the bottom left corner

of the page and one inch consists of 72 pixels. TCR2pdf assumed all reports 1) use the same PDF form, 2) were formatted the same with PDF version 1.4, and 3) saved with the latest Adobe Reader (note: using inconsistent Adobe Readers might result in inconsistent coordinate formatting).

By using an inch to pixels coordinate conversion system, we can easily define a bounding box to extract text on the report with some additional text filtering. As mention above, TCR2pdf assumed all ETCR reports will be using the exact same PDF form, therefore, it also assumed that all field coordinates will be fixed and the same throughout all reports from different districts.

The PDF information extraction process can be broken down into three parts: 1) narrative extraction, 2) extraction of the diagrams, and 3) extraction of all other fields contained in “boxes.”

The method for extracting content from the narrative section is different from any other method used in PDF data extraction. Since narratives are structured as paragraphs throughout a page, a predefined content bounding box is required in order to accurately extract contents. Assuming all reports are encoded in the same manner using pixel-based, predefined upper and lower bounding box definition, the narrative decoder will search the text pixel by pixel from the upper to the lower bound. By searching by pixel location in the long (Y) direction, we are able to separate the narrative contents by their original paragraph format.

For diagrams, extracting contents is the same as extracting content from narratives. In this case, instead of searching by pixels, all words within the bounding box from the diagram pages are extracted as one line.

For all other fields, extracting contents is performed in the following manner:

- Bounding box is defined by the lowest left XY (X is the shorter length and Y is the longer length of the PDF page) coordinates as the lower bound and the most upper right XY coordinates as the upper bound.
- A tree based search on the bounding box coordinates is then performed.
- If the tree search returns any contents, it is further processed to clean up and erase any duplicated/misplaced characters.
- The content dictionary in JSON file format is then updated.**Error! Reference s ource not found.**
- Step 2 is repeated for all fields.

ITCR Processing Details

When an ITCR file is input into the TCR2pdf module, TCR2pdf will perform the following procedure to extract all predefined fields. Please note that the ITCR format does

not have an associated XML file, therefore no cross-checking on extracted information can be performed. The overall processing of the ITCR can be summarized as follows:

1. Crop page title using Python Imaging Library (PIL) by predefined page layout location using Dots Per Inch (DPI). Appendix A provides more detail.
2. Determine page type by a page formatter.
3. Determine boundary line coordinate in DPI.
4. Crop the field image based on predefined offset coordinates using the boundary line (by DPI).
5. Feed cropped image into Tesseract (see Appendix A) for OCR.
6. Feed OCRed results into clean-up tools.
7. Update contents dictionary in JSON format.
8. Repeats step 4 for all fields.

GEOSPATIAL LINEAR REFERENCING SYSTEM

CA Post-Mile Web Service (CAPM) Overview

The AHMCT California Post-mile web service (CAPM) is a machine-to-machine Hyper-Text Transfer Protocol (HTTP) service that was created by this research team in order to support the data analysis of this work. CAPM is able to handle a variety of post-mile-related Geographic Information System (GIS) queries, primarily falling into the following two categories:

arbitrary post-mile geolocation: This class of query allows the precise geographic coordinates of any post-mile to be determined. The post-mile value, along with various roadway and alignment parameters are specified in a query and the corresponding results are returned to the caller.

post-mile proximity search: This class of query allows the nearest post-mile(s) to a particular set of geographic coordinates to be determined. A latitude, longitude, and search range, along with various roadway and alignment parameters (to filter the results), are specified in a query and the corresponding results are returned to the caller.

```
./geo2pm.py --lat=38.078190 --lon=-120.541886 --rng=10560
./geo2pm.py --lat=38.078190 --lon=-120.541886 --rng=10560 --rt=4

./pm2geo.py --crp=CAL-4-R22.194
./pm2geo.py --crp=CAL-49-7.719
```

Technical Overview

The front end of the service is implemented as a Java HTTP servlet, and at the back end is a PostGIS database into which is loaded publicly-available Caltrans highway data: www.dot.ca.gov/hq/tsip/gis/datalibrary/Metadata/StateHighway.HTML.

The data flow for CAPM is depicted in Figure C.11.

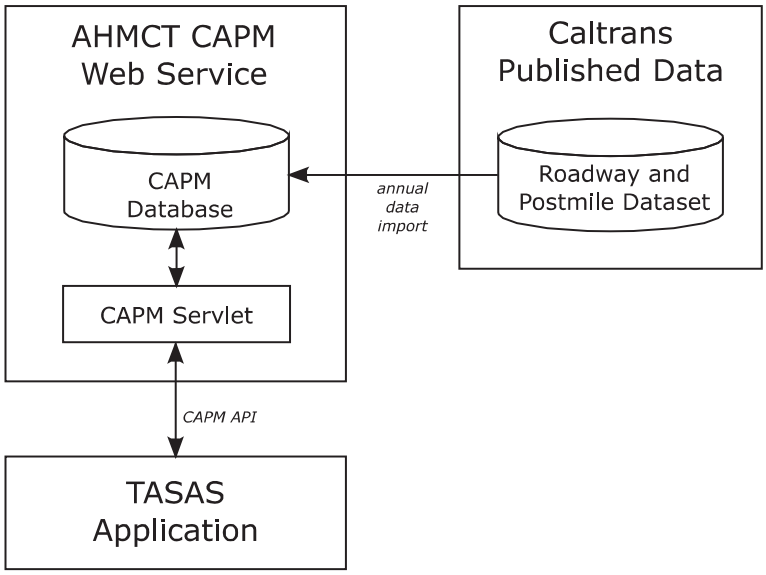


Figure C.11. CAPM Service Data Flow Overview.

Arbitrary Post-Mile Geolocation Queries

CAPM handles arbitrary post-mile geolocation queries via a HTTP GET request to the locatepm endpoint: http://atlas.ahmct.ucdavis.edu:8080/capm/locatepm.

Supported HTTP request parameters for locatepm queries are shown in Table C. 5.

Table C. 5: Supported HTTP request parameters for locatepm queries.

Parameter	Required?	Empty OK?	Description
cty	yes	no	Caltrans county abbreviation (all caps)
rtnum	yes	no	Route number (integer)
rtsfx	no	yes	Route suffix (S, U)
pmpfx	no	yes	Post-mile prefix (C, D, G, H, L, M, N, R, S, T)
pmval	yes	no	Post-mile value (decimal)
pmsfx	no	yes	Post-mile suffix (L, R, X)
alignment	no	no	General left/right alignment (L, R)

An empty parameter value (e.g., pmsfx in the example below) implies that the parameter is not a component of the queried post-mile. A missing parameter (or one with

a value of "*"), such as rtsfx and pmpfx in the example below, implies that the search should include post-miles with any value for that parameter.

Here is an example locatepm request:

```
http://atlas.ahmct.ucdavis.edu:8080/capm/locatepm
  ?cty=ALA
  &rtnum=84
  &pmval=23.133
  &pmsfx=
  &alignment=L
```

CAPM will respond to successful locatepm requests by returning a <CaPmResponse> element containing a <GeoLocateResult> element. <GeoLocateResult> elements contain zero or more <Match> elements, each of which contains exactly one <Post-mile> element. This represents the set of all post-miles found that match the query parameters. For example, the response to the example request above contains two matches:

```
<?xml version="1.0" encoding="utf-8"?>
<CaPmResponse>
  <GeoLocateResult>
    <Match>
      <Post-mile
        cty="ALA" rtnum="84" pmpfx="R" pmval="23.133"
        alignment="L" aligncode="Left"
        latdeg="37.6315719627654" londeg="-121.80193000449171"
      />
    </Match>
    <Match>
      <Post-mile cty="ALA" rtnum="84" pmval="23.133"
        alignment="L" aligncode="Left"
        latdeg="37.633836626074675" londeg="-121.80189798230704"
      />
    </Match>
  </GeoLocateResult>
</CaPmResponse>
```

A <GeoLocateResult> element that contains zero <Match> elements implies that no post-miles matching the query parameters were found.

If an error occurs, e.g. due to an invalid query, the <CaPmResponse> element will instead contain an <Error> element. For example:

```
<CaPmResponse>
  <Error msg="Invalid query: cty parameter is required"/>
</CaPmResponse>
```

The <Post-mile> element is an empty XML element which has a combination of XML attributes summarized in Table C. 6

Table C. 6: Element is an empty XML.

Attribute	Description
cty	Caltrans county abbreviation (all caps)
rtnum	Route number (integer)
rtsfx	Route suffix (S, U)
pmpfx	Post-mile prefix (C, D, G, H, L, M, N, R, S, T)
pmval	Post-mile value (decimal)
pmsfx	Post-mile suffix (L, R, X)
alignment	General left/right alignment (L, R)
aligncode	Alignment code (Left, Right, Left Split Align, Right Split Align)
latdeg	Latitude of post-mile (decimal degrees)
londeg	Longitude of post-mile (decimal degrees)

If one or more of the above attributes are missing from the <Post-mile> element, it should be inferred that they are not a part of the post-mile designation. This is often seen with the rtsfx, pmpfx, and pmsfx attributes.

Proximity-Based Post-mile Searches

CAPM handles proximity-based search queries via a HTTP GET request to the "findnearestpm" endpoint: http://atlas.ahmct.ucdavis.edu:8080/capm/findnearestpm Supported HTTP request parameters for "findnearestpm" queries are shown in Table C. 7.

Table C. 7: Supported HTTP request parameters for "findnearestpm" queries.

Parameter	Required?	Empty OK?	Description
latdeg	yes	no	Latitude of search point (decimal degrees)
londeg	yes	no	Longitude of search point (decimal degrees)
rngft	yes	no	Search radius (decimal feet)
cty	no	no	Caltrans county abbreviation (all caps)
rtnum	no	no	Route number (integer)
rtsfx	no	yes	Route suffix (S, U)
pmpfx	no	yes	Post-mile prefix (C, D, G, H, L, M, N, R, S, T)
pmsfx	no	yes	Post-mile suffix (L, R, X)
alignment	no	no	General left/right alignment (L, R)

An empty parameter value (e.g., pmpfx in the example below) implies that the parameter is not a component of the queried post-mile. A missing parameter (or one with a value of "*"), such as cty, rtnum, rtsfx, pmsfx and align in the example below, implies that the search should include post-miles with any value for that parameter.

Here is an example findnearestpm request:

```
http://atlas.ahmct.ucdavis.edu:8080/capm/findnearestpm
 ?latdeg=38.078190
 &londeg=-120.541886
 &rngft=5280
 &pmpfx=
```

CAPM will respond to successful findnearestpm requests by returning a <CaPmResponse> element containing a <FindNearestResult> element. <FindNearestResult> elements contain zero or more <Match> elements. This represents the set of all post-miles found that match the query parameters. Each <Match> element contains exactly one <Post-mile> element and exactly one <DistanceFt> element. The <Post-mile> element represents the matching post-mile found that is nearest to the specified search point and within the specified search radius, and the <DistanceFt> element represents that post-mile's distance from the specified search point.

For example, the response to the example request above is:

```
<?xml version="1.0" encoding="utf-8"?>
<CaPmResponse>
 <FindNearestResult>
  <Match>
   <Post-mile
    cty="CAL" rtnum="49" pmval="7.71848223672819"
    alignment="L" aligncode="Left"
    latdeg="38.07452798262649" londeg="-120.5445836861309"
   />
   <DistanceFt val="1544.3285243655932"/>
  </Match>
  <Match>
   <Post-mile
    cty="CAL" rtnum="49" pmval="7.71848223672819"
    alignment="R" aligncode="Right"
    latdeg="38.07452798262649" londeg="-120.5445836861309"
   />
   <DistanceFt val="1544.3285243655932"/>
  </Match>
```

```
  </FindNearestResult>
</CaPmResponse>
```

This describes two matching post-miles that were found approximately 1544 ft. away from the search point.

If an error occurs, e.g. due to an invalid query, the <CaPmResponse> element will instead contain an <Error> element. For example:

```
<CaPmResponse>
  <Error msg="Invalid query: rngft parameters is required"/>
</CaPmResponse>
```

AHMCT Research Center maintains a production CAPM service at atlas.ahmct.ucdavis.edu, but if a custom build or a local installation is desired.

DATA Extraction and MATCHING ALGORITHM

Overview

The AHMCT-WCT developed in this research study tries to minimize coders' efforts in the location coding of electronic TCRs. It has the following features:

1. Converts the electronic TCR into a machine-readable JSON file
2. Extracts the following fields from the report:
    a. Diagrams
    b. Narrative
    c. Location boxes
    d. Party information
    e. Sequence of events
3. Parses location fields to extract the following information:
    a. GPS coordinates
    b. Post-mile information
    c. Intersection information
    d. AOI (within the narrative)
4. Compares location data from various fields to obtain the locations with the highest confidence level for correctness.
5. Displays the obtained locations on an interactive map to help coders in the decision-making process of their coding task.
6. Shows the original report, the diagram, results from analyzing location information, narrative, and the respective parts of the sequence listing documents in a single web page to save coders' time and help them focus on the coding process.

In the following paragraphs, each of the above steps is described in detail and examples of the processes are provided.

The Django Web Framework

Django is a free and open-source web framework that provides for the easy creation of complex websites. Developed completely with the Python language, Django follows a Model-View-Template (MVT) architecture. Django's configuration system allows any Django project to be plugged into any regular website and remain functional provided that

it follows the reusable app conventions. Moreover, it contains an easy-to-use development web server and can run in conjunction with Apache, NGINX web server, Cherokee, etc. for production.

Figure C.12 illustrates how different parts of the Django MVT architecture interact with each other to serve the user. The developer provides templates (HTML files with embedded Django HTML code to handle inputs and outputs), views (Python functions that perform all site functionalities on the back end), and models as well as specific Uniform Resource Locator (URL) mapping of each page, and Django serves them to the user.

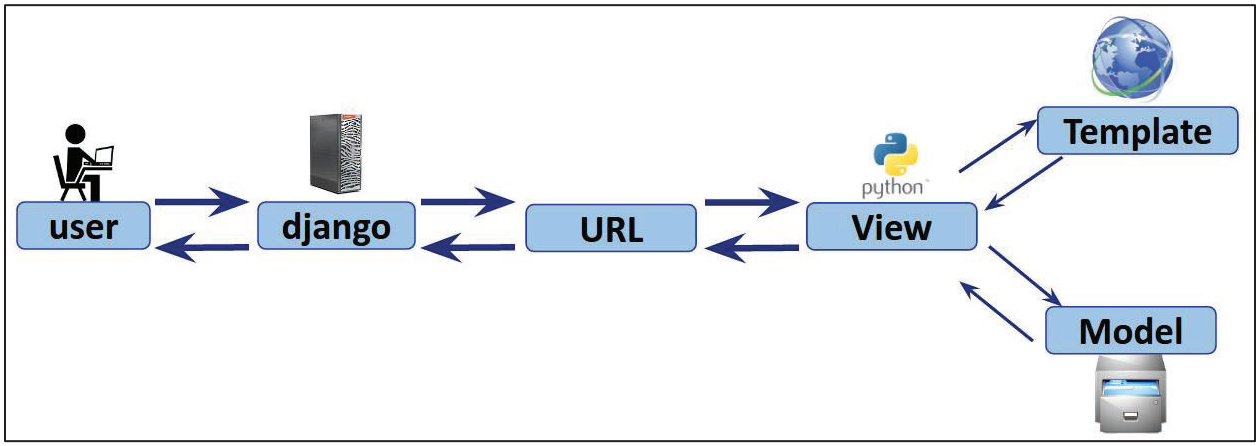


Figure C.12. Django framework MVT architecture parts interacting with each other.

Django Applications

Every website created using the Django web framework consists of a few applications. Each Django application is a series of codes connected together via the MVT architecture to serve the user. Django applications are developed for maximum flexibility in using each developed app in various Django projects. For example, the Django application for location coding of ETCRs developed in this project, can be easily integrated into a larger project that, for example, provides search through the TCR database.

Each Django application, depending on its function, can consist of a series of views, models, URL mappings, templates, and forms, all of which are briefly described below

Views

According to official Django documentation, “a view is a callable which takes a request and returns a response.” Views.py is one of the required files within any Django application and it contains a series of Python functions (or sometimes classes) that perform actions based on the user’s requirements. Basically, all user actions (clicking on a button, submitting data through a form, uploading/downloading a file, etc.) point to a specific Django view through the URL associated with it in the urls.py file, and the called

view later returns another HTML template (usually populated with the requested data from models.py) or saves the input given by the user in the application database.

Models

A model is a “single, definitive source of information about data.” Models are Python classes that each map to a single database table. For example, the user authentication view in our project gets the user information from the HTML template and uses the Django authentication function to check it against the user model stored in a sqlite database. This streamlined approach frees the developer from the hassles of setting up databases and pointing them to the website templates while providing a secure and easy-to-use connection for the user.

URLs

The urls.py file assigns each website button to its respective view.

Forms

The forms.py file is another necessity in any Django application. It defines the type, number of inputs, type of input variables, labels, and other specifications that HTML input forms need. In this project two form field classes were used: one for getting and storing user authentication data and the other one for handling TCR uploads.

Templates

Templates are HTML files that also contain some Django HTML commands (for serving static data and calling Django views, etc.). Templates contain regular HTML, CSS, and JavaScript commands for structuring and styling the website and can interact with forms and views through the urls.py commands. Each page of the website is based on a separate HTML template that is called by a Django view.

Because of its simple hierarchical model, Django projects are extremely easy to navigate and maintain. The following sections will describe how we used the above methods to develop the decision support tool web application using Django.

Decision Support Tool Process Description

The process of using the AHMCT-WCT web application contains the following steps:

1. Upload the ETCR in PDF format
2. Convert the ETCR into a JSON file
3. Load the JSON file as a Python dictionary
4. Read location data, narrative, and diagram from the JSON file
5. Use the Sequence Listing files, Caltrans post-mile lookup tool, etc. to convert the different location information to GPS coordinates of their respective points.
6. Sort the calculated locations to offer the coder with a list of locations based on their degree of confidence

Load the results page with locations shown on an interactive map, allowing users to open up the diagram, narrative, actual ETCR in PDF format, party information, and search through the Sequence Listing document

The following sections provide detailed information on each of the steps mentioned above.

Handling File Uploads

The Django model is equipped with specific fields for handling different types of inputs. The ETCR upload interface supports standard HTML form-based file uploads via the use of a Django Form object. Upon form submission, the data from the Form object (which contains the ETCR) is used to instantiate a Django File-Field model. This File-Field model handles the storage of the uploaded file to the filesystem, where it is subsequently accessed by the next stage of the processing pipeline.

Conversion of ETCR

Conversion of electronic TCRs into an easy-to-use JSON format is done using our software developed specifically for this purpose. A screenshot of the final JSON file generated here to be used by our website is shown in Figure C.13. After this conversion, the ETCR in JSON format can be loaded into Django as an easily readable Python dictionary.

```
"LOCATION": {
    "CITY": "UNINCORPORATED",
    "COUNTY": "MONTEREY",
    "JUDICAL DISTRICT": null,
    "LATITUDE": null,
    "LONGITUDE": null,
    "MILEPOST INFO TEXT": "140 FEET NORTH OF 1MON72.61",
    "MILEPOST INFORMATION": null,
    "REPORTING DISTRICT": null,
    "ROUTE": 1,
    "ROUTESUFFIX": "SR",
    "SECONDARY INFORMATION": {
        "DIRECTION": null,
        "DISTANCE": null,
        "INTERSECTION": null,
        "ROUTE": null,
        "UNIT": null
    },
    "SECONDARY INFORMATION TEXT": "140 FEET NORTH OF RIO RD."
},
"NARR": {
    "AOI": [
        "AOI #1 (V-1 vs. V-2) was located 140' n/of the north roadway
        "The Summary, AOI & Cause were determined by statements and ve
    ]
},
```


Figure C.13. ETCR after conversion into the JSON format.

Calculating Locations

In the manual coding process, locations are obtained from the following fields:

- GPS coordinates
- Post-mile information
- Intersection information
- AOI

The first three in the list above are obtained from the location boxes on the first page of each report and the last item (AOI) is found within the narrative.

Here we discuss calculation and validation of each of these data.

GPS coordinates

Most of the reports contain the GPS coordinate reported by the CHP officer on the first page of the report. In order to calculate the post-mile information from the GPS coordinates, we use the Caltrans post-mile lookup tool and find the closest post-mile corresponding to the latitude and longitude values reported by the officer.

The Caltrans post-mile lookup tool is called by a Python script called by the main view of the web application.

Post-mile information

Although this field is not present in many of the studied reports, it can potentially provide the coding tool with valuable information. This field is usually populated with the most straightforward results, being actual post-mile values, and the calculation of the final post-mile is relatively easy due to not needing any outside lookup tool.

Intersection information

Compared to the previous fields, intersection information is usually harder to parse into useful post-mile information. The reason is that this field contains the relative location of the collision with respect to the nearest highway intersection. Therefore, to find the corresponding post-mile location, we need to perform a lookup on one of the sequence listing documents (separate documents for each district) that are hundreds of pages long and are in PDF/print format.

The first challenge is to convert the Sequence Listing files into lookup tables that are optimized for our query. This parsing is done by transforming the electronic PDF files into JSON files that are sorted based on district-route-county-city. This makes it a lot faster and easier for a lookup to be done. Then, the intersection defined in the report is cross-checked with the entries within the acceptable geographical location and the most favorable results are chosen.

A number of challenges are associated with this task that make it difficult to obtain the final answer. The first one is inconsistencies between officer narration and the Sequence Listing files. Because of the constant updates in street names and the abundance of abbreviations for different names and indicators, a conventional string matching algorithm often fails when performing in this task. For this reason, a Levenshtein distance algorithm was implemented in this research study that is able to consider typing mistakes and abbreviations and provide the user with a result with the highest confidence level.

Acknowledging the degrees of complexity in many highway intersections, the coder must always be aware of a few of the nearest results to make the smartest decision. Therefore, we implemented a search box that is capable of showing the coder direct results from the Sequence Listing files to help them in making the right decision faster and without the need to waste any time manually going through long documents.

Based on the degree of matching between the text in the “Intersection Information” box and the Sequence Listing document, the top results are scored and shown to the user to increase the level of confidence in his/her results. The scores are given based on the level of matching (considering occasional typos) and the type of matched strings. The

reason is to make distinctions between the common (street, bridge, ramp, etc.) and defining (street names, cities, etc.) strings.

AOI

Parsing AOI information is done by analyzing the general sentence structure of this section of the reports. After obtaining post-mile/intersection values, the same steps of the previous sections give us the final results.

At this point, the results of our program can be shown to the coder to make the final decision by looking at the provided report diagram and comparing it with the pins on the interactive map. However, the program is still capable of giving a suggestion for the most likely location of the collision based on the results of the pilot study.

The logic of this segment is based on the matching of different locations found within the report. A lower score is awarded to the GPS coordinates because it is often the most inaccurate piece of information, while the highest score is given to AOI, as it usually matches with either post-mile information or the intersection information. By practicing on over a few hundred reports, we realized that if two of the reported locations match, that location can be reported as the location of the collision with a high level of confidence.

Displaying final results

As discussed earlier, HTML templates equipped with the Django template language are used to display results to the user. After calculating post-mile information, the show view loads the show.HTML template and supplies it with a Python dictionary containing the required location information, narrative, diagram address, and GPS coordinates to show results on an interactive map. Django HTML tags are then used to display results from the specified dictionary.

Django HTML tags are marked with {% for … if etc. %} and Python variables are marked with {{dictionary_i}}

PILOT STUDY

In order to evaluate and test the AHMCT-WCT and its components, a pilot study was performed with ETCRs as input combined with a detailed evaluation of the output Location Codes. It was assumed prior to this pilot study that the AHMCT-WCT was operational and the information from the TCR had been converted to the *.JSON format. Once a TCR is selected and the “Submit” option is selected from the AHMCT-WCT, the output text will be displayed along with a corresponding map graphically showing the geospatial positions of the resulting locations. The purpose of this chapter is to discuss the quality of the results obtained in the pilot study along with other evaluation metrics. These results obtained in the pilot study address some of the research questions of this study which will be discussed in the next chapter.

Methodology

This Pilot Study used the following selection criteria for TCRs:

- ETCR format only, no ITCRs.
- ETCRs that were already coded by the TASAS location Coding Group so that the accuracy of the results obtained from AHMCT-WCT could be evaluated.
- ETCRs that were from all districts, all highway types, and a representative selection of collision locations.

In order to obtain a fair distribution of representative ETCRs throughout all 12 Caltrans Districts, we obtained the California Highway Patrol Incident Distribution for 2016. Table C. 8 shows how many freeway miles comprise each district and the total number of CHP Incidents in each.

Percentage distribution of total miles and total CHP incidents are also shown in Table C. 8 (2nd and 3rd columns). To assign an equal weighting between mileage and collisions, the average of the two percentages (4th column) was calculated and then multiplied by 500 (the target number of ETCRs). The last column shows the number of ETCRs for each Caltrans district.

Table C. 8: Quantities and distribution of total highway miles and number of CHP Incidences for each Caltrans district. Data taken from 2016.

District	Freeway Miles	# Incidents	Percent of total miles	Miles% of 500 ETCRs	Percent of total incidents	Incident% of 500 ETCRs	Avg. Miles% and Incident %	Avg.% of 500 ETCRs
9	1478.355	18	5%	23.84444	0%	0.017857143	2%	12
1	1889.303	2136	6%	30.47263	0%	2.119047619	3%	16
2	3463.174	3598	11%	55.85765	1%	3.569444444	6%	30
6	4065.155	11719	13%	65.56702	2%	11.62599206	8%	39
5	2333.154	13348	8%	37.63152	3%	13.24206349	5%	25
10	2658.946	13681	9%	42.88623	3%	13.57242063	6%	28
3	3004.543	38690	10%	48.46037	8%	38.38293651	9%	43
12	577.866	43061	2%	9.320419	9%	42.71924603	5%	26
11	2051.379	46365	7%	33.08676	9%	45.99702381	8%	40
8	3890.371	62646	13%	62.74792	12%	62.14880952	12%	62
4	2868.999	109784	9%	46.27418	22%	108.9126984	16%	78
7	2318.168	159279	7%	37.38981	32%	158.014881	20%	98
Total	30599.41	504325	99%	493.5389				497

Consequently, based on the logic explained above, the last column of Table C. 8 shows the desired number of ETCRs for each district.

Since most districts have some “outliers” with respect to a high number of collisions per freeway mile, there were additional requests concerning which freeways we needed to get a suggested number of ETCRs.

Note: “Evenly distributed” means that there are no freeway “outliers,” and that among the freeways which have at least one collision, there is a low Incident per Freeway mile ratio.

Application of AHMCT-WCT in the Pilot Study

Before the results of the Pilot Study are discussed, it is important to understand a summary of the AHMCT-WCT functionalities:

It determines the TASAS coding based on all TCR location data: GPS values, post-mile information, intersection information, and narratives.

The output display shows: Pin marks of calculated potential locations, AOI verbiage, best matches with sequence listings along with Show/Hide toggles for the TCR report and diagrams.

It is able to provide the user with a suggestion of the “best match” in terms of location.

Allows the user to download the collision information into TASAS format (*.csv file).

Results of Pilot Study

This section will cover the challenges in processing ETCRs, program functionalities, and results of analyzing the 500 ETCRs using the developed tools.

PDF processing issues

Using PDF files (whether ITCR or ETCR) as the sole input to the program imposes many limitations on the effectiveness of information extraction from TCRs. It should be pointed out that *.PDF files were designed to be solely used as “printed” or “publishable” files and aren’t meant to be used as digitally-readable formats; therefore they possess information that can make content analysis challenging. On digital PDF files, behind the image is data that contains extraneous information that is generated based on the history of the file. For example, a standard form can be brought in from a variety of sources, edited by a CHP officer writing the TCR, cutting and pasting images for the diagrams, etc. Each of these actions creates digital bits that can mask the data that is targeted to be the displayed image. The use of various software package releases can also affect the image.

Under certain conditions, the PDF is altered in such a way that the text is read with an unknown amount of certainty. When the ETCR processing program detects distortion, it does not process any portion of the PDF file since reliability cannot be provided.

For the pilot study, there were 25 ETCRs that were not processed due to this kind of PDF error. The number of unreadable files varied among the CHP districts. The resulting distribution of files per district can be seen in Table C. 9. The overall influence of this PDF

incompatibility error is shown in Table C. 10. As can be seen, for this study approximately 5% of the chosen ETCRs were not able to be processed.

Table C. 9: Number of ETCRs that were unable to be processed by the Location Code software due to errors in reading the PDF file format.

Caltrans District	Target #	Actual #	#PDF error files	Final # to code
1	16	16	0	16
2	30	30	0	30
3	43	43	2	41
4	78	77	2	75
5	25	25	2	23
6	39	39	5	34
7	98	97	5	92
8	62	61	0	61
9	12	12	0	12
10	28	28	9	19
11	40	37	0	37
12	26	26	0	26

Table C. 10: Overall outcome of the PDF file incompatibility. For this study, approximately 5% of the ETCRs were unreadable.

	Target #	Actual # Retrieved	# PDF error files	Final # to code
Total	497	491	25 (5% of total)	466

Pilot Study Coding Benchmark

In performing the pilot study and obtaining the results presented above, it is assumed that the manually coded results are 100% correct. In evaluating a few of the samples, the research team raised the following questions:

1. What is the uncertainty in manually coded results?
    a. Since mistakes and human errors are an inevitable part of any process, is there a way to find out the average number of mistakes made in the location coding process?
    b. If a certain number of reports are coded by a few different coders, how will the results match?
2. What is the range of accuracy for manually coded results?
    a. Location coding results are recorded as a single value. What is the tolerance range of this value? Is this tolerance standardized?

The AHMCT-WCT provides results as post-mile locations alongside their respective positions on an interactive map. Benchmarking these results was done by comparing them to the post-mile location that is manually obtained by the coders. Precision of these results was then determined by calculating their distances from the manually coded results.

Figure C.14 depicts the percentage of results that lie within a reasonable distance from the manually coded locations in each district. Overall, comparison of the AHMCT-WCT results to manually coded locations resulted in the following percentages for location coding precision:

- 74% within 0.01 mile
- 88% within 0.05 mile
- 91% within 0.1 mile

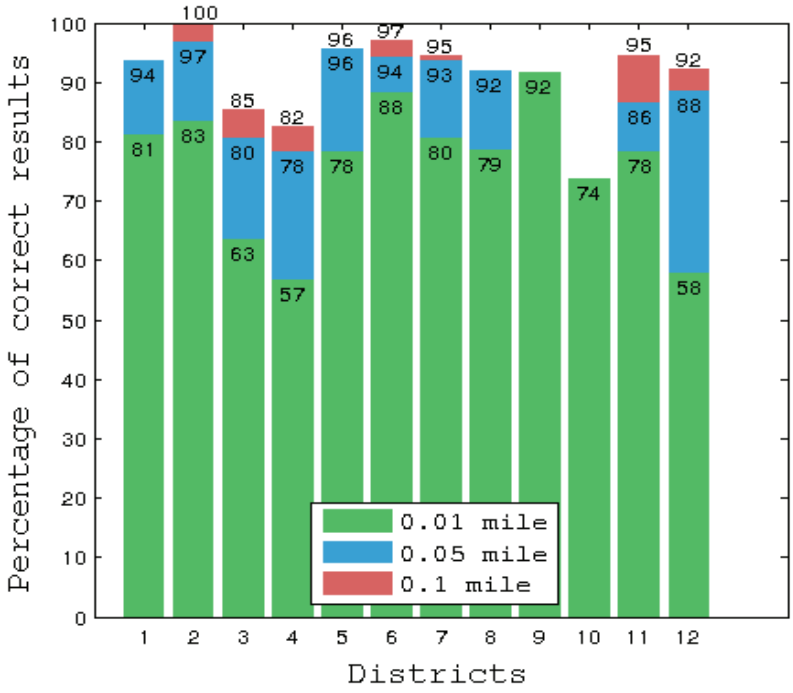


Figure C.14. Benchmark results for all 12 Districts using the "best match" from all possible post-mile values.

Pilot study GPS Data Results

Results based on GPS coordinates only can also be benchmarked with hand-coded results. A total of 82% of the reports in the pilot study contained GPS coordinates. Using the AHMCT-WCT, the following results were obtained:

- For 18%, the GPS coordinates were within 0.01 mile of the manually coded location
- 36% within 0.05 mile
- 43% within 0.1 mile

Benchmark results are shown in Figure C.16.

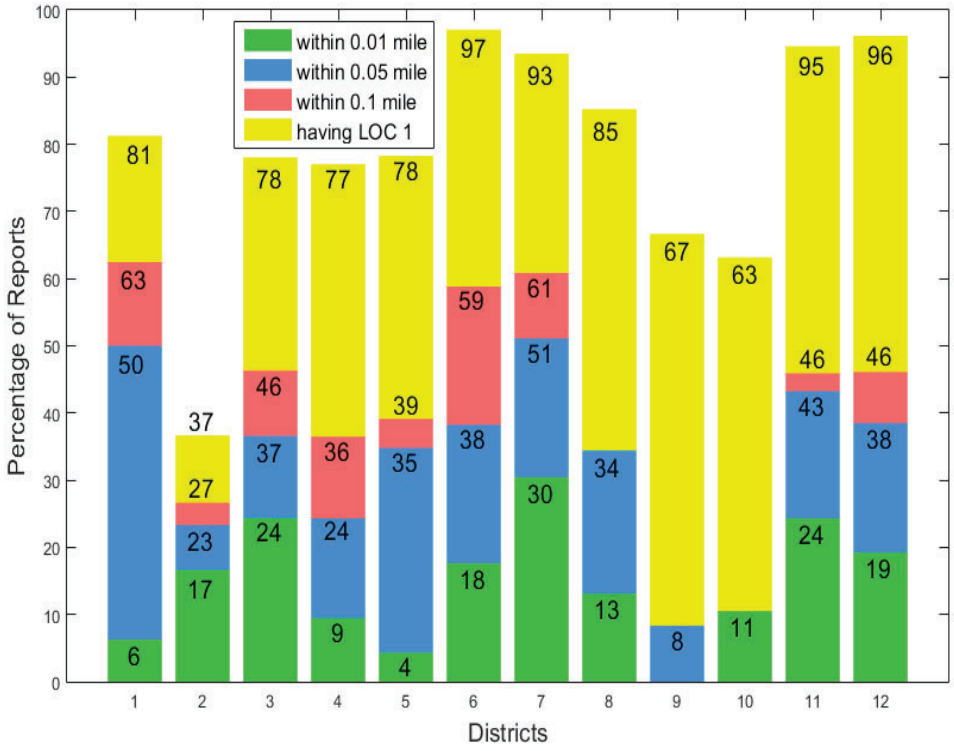


Figure C.15. Benchmark results for all 12 Districts using the post-mile value based on GPS data only (when provided).

Pilot study post-mile data results

Results based on post-mile data only—when it was provided—was also evaluated. A total of 22% of the reports contained post-mile information. This data is found on the second line in the TCR first page Location Box. The benchmark results are shown in Figure C.16. The results indicate:

- For 12%, the post-mile information was within 0.01 mile of the manually coded location
- 17% within 0.05 mile
- 18% within 0.1 mile

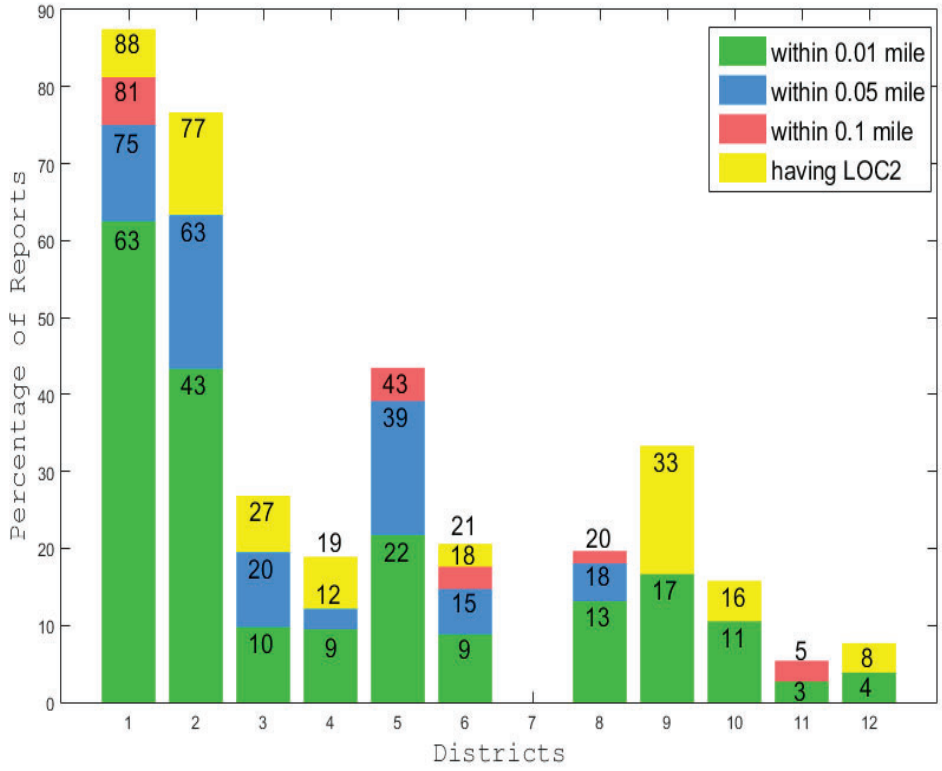


Figure C.16. Benchmark results for all 12 Districts using the post-mile value based on "post-mile" data (when provided).

Pilot study intersection data results

Results based intersection data only (when provided) was also evaluated. This data was found on the third line of a TCR's first page. One hundred percent of the reports processed in this part contained intersection data. Benchmark results are depicted in Figure C.17. The results indicate the following:

- For 63% of processed reports, the calculated post-mile was within 0.01 mile of the manually coded location
- 76% within 0.05 mile
- 78.5% within 0.1 mile

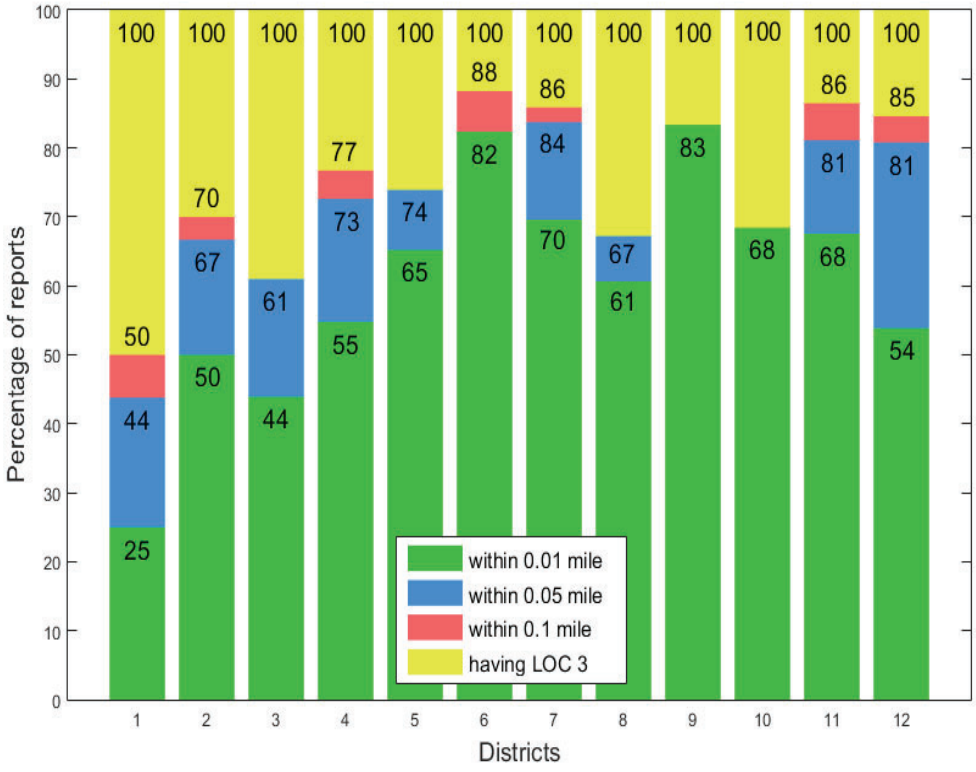


Figure C.17. Benchmark results for all 12 Districts using the post-mile value based on "Intersection" information only.

General

For 42 reports (9% of the total processed), none of the results from different fields (GPS coordinates, post-mile information, and intersection information) were within 0.1 mile of the manually coded locations. For 21 reports out of these 42, the Sequence Listing files did not contain the intersection reported in the TCR.

Overall Conclusions of Preliminary Study

This research study evaluated and developed methods and algorithms for streamlining and digitizing the process of coding and populating the TASAS database. This database is populated based on coding the data that is included in police TCRs resulting in an electronic database that can easily be used by Caltrans to evaluate the specifics and trends of collisions on California highways. The research study has addressed the following research questions:

1. To what extent can data coding in terms of determination of District, County, Route, post-mile, travel direction, and post-mile markers can be digitized and streamlined?
2. To what extent can data from electronic TCRs can be digitally extracted and automatically put into the TASAS database?
3. To what extent can data from the narrative portion of police TCRs can be automatically extracted and codified and used to digitally populate the TASAS database?
4. How can use of electronic digitization technology be implemented into Caltrans' workflow for coding and populating the TASAS database?

In addition, this research study developed a set of complete requirements that will facilitate algorithmic developments for automating the data coding process. The research has also developed an experimental web-based system referred to as AHMCT-WCT that can be used by the coders as a decision support tool to streamline their coding process. In developing the web-based coding tool, the research team identified several challenges in digitizing the TASAS coding process. These challenges are a result of factors including:

1. Variations in the forms of TCRs, some being in digital PDF format, some being in digital XML format, and some being in the form of scanned PDFs.
2. The diagrams on collision location and details are not included in the XML format of TCRs.
3. Coding accident locations in some metropolitan areas where there are several highway on and off ramps intersecting with streets making the automatic identification of the accident location more complex.

The first challenge will be resolved in the near future as CHP's rollout of their digital TCRs is completed and other allied police agencies start following them by using a similar digital format for their collision reports. Until then, the experimental web-based coding tool developed in this study has options to handle this shortcoming.

The second challenge can also be easily resolved procedurally by making sure that the electronic versions of TCRS that are XML files accommodate the diagrams in the TCRs in some standardized format. This research study did identify this challenge early on and has provided an example of how the process of digitizing a TCR can overcome this challenge.

The third challenge requires further research in utilizing machine learning and Artificial Intelligence (AI) methods to make decisions among different intersections resulting from several highway off and on ramps in metropolitan areas. This type of research was beyond the scope of this research study and was not pursued, but can be a subject of a future research.

The experimental AHMCT-WCT was tested in a pilot study consisting of 500 Traffic Collision Reports. The pilot study results indicate that for location coding (research question number 1) the AHMCT-WCT generated results that were approximately 90% within 0.1 mile and 74% within 0.01 mile of the manually coded results. These percentages indicate the extent at which location coding can be automated with the present format of digital TCRs. The challenge areas for location coding were mostly for accident locations in some metropolitan areas where there are several highway on and off ramps intersecting with streets making the automatic identification of the accident location more complex. For such cases, the AHMCT-WCT developed in this research study can be used as a decision support system since it will generate a list of all the possible locations and will display them on a map so that the coder can choose the most appropriate location for TASAS. An area for future research is to consider using machine learning and AI techniques to investigate full automatization of such cases.

In terms of the second research question related to automatic coding of data other than location coding, the research results indicate that such coding can be fully automated if the actual data is included in the digital TCR format. The initial XML files obtained for evaluation of digital TCRs missed some of the data included in TCRs. Coordination between the Caltrans TASAS group and CHP is recommended to make sure that all data relevant to TASAS be included in the digital XML formats of TCRs. The digital XML files of TCRs also did not include the diagrams contained in non-digital TCRs. Coordination is also recommended between the Caltrans TASAS group and CHP to make sure that such diagrams will be included in some format. The web-based coding tool developed in this research task is able to codify diagrams from PDF format.

In terms of retracting data from the narrative portion of a TCR (the third research question), this research study has shown that this can be done rather reliably with electronic versions of TCRs either in electronic PDF format or XML format.

As the first step in addressing the fourth research question in terms of methods to implement the digitized coding technology into Caltrans' workflow, the AHMCT-WCT was developed and experimented with on 500 TCRs. The coding tool was also developed using open source resources so that its implementation into Caltrans' workflow can be facilitated. Furthermore, the research team has had several meetings with Caltrans' IT and coding personnel and has identified the need for an implementation phase for this research. This implementation phase is also another area of future work.

The OCR Software Addendum

TCR2pdf is a standalone module designed and implemented by the AHMCT Research Center for TCR processing that contains an OCR software. TCR2pdf can handle both

digital CHP reports and non-digital (image) reports. TCR2pdf was designed as a backend parsing module for the AHMCT-WCT for CHP TCRs. When a TCR is uploaded to AHMCT-WCT, TCR2pdf will decode and extract all needed information for location coding purposes as well as collect the data for subsequent TASAS database format downloading. Results from TCR2pdf will be returned as a structured dictionary (JSON).

Dependencies of OCR Software

Since the OCR TCR2pdf module was developed based on available open-source codes, there are a number of utilities and packages that are used by TCR2pdf. The list of software packages that TCR2pdf is dependent on are as follows:

- PDFQuery
- PyEnchant
- Tesserat
- Ghostscript
- PIL
- OpenCV2

More information on each necessity can be found in the following sub-sections.

PDFQuery

PDFQuery is a light wrapper around PDFminer, lxml and pyquery. It's designed to reliably extract data from sets of PDFs with as little code as possible.

For installation, there are two functions available: easy_install PDFquer and pip install PDFquery

PDFQuery works by loading a PDF as a PDFminer layout, converting the layout into an etree with lxml.etree, and then applying a pyquery wrapper. All three underlying libraries are exposed, so you can use any of their interfaces to get at the data you want.

PDF files are internally messy, so it's usually not helpful to find items on them based on document structure or element classes like with with Hypertext Markup Language (HTML) files. Instead the most reliable selectors are the static labels on the page, which you can find by searching for their text contents as well as the physical locations on the page. PDF coordinates are given in points (72 points to an inch) starting from the bottom left corner. PDFMiner (and so PDFQuery) describes page locations in terms of bounding boxes, or bboxes. A bbox consists of four coordinates: the X and Y of the lower left corner and the X and Y of the upper right corner.

The easiest method to scrape text text that is always in the same place on the page, is to use Acrobat Pro's Measurement Tool, Photoshop, or a similar tool to measure distances (in points) from the lower left corner of the page and then use those distances to craft a selector like: in_bbox("x0,y0,x1,y1").

If we are scraping text that might be on different parts of the page, the same basic technique applies, but we will first have to find an element with consistent text that

appears at a consistent distance from the text that we want, and then calculate the bbox relative to that element.

For downloading purposes and the latest release source, please see the PDFQuery homepage (https://github.com/jcushman/PDFquery).

For more information and Frequently Asked Questions (FAQ) on PDFQuery, see the PDFQuery homepage (https://github.com/jcushman/PDFquery).

PyEnchant

PyEnchant is an open-source spelling checker library based on Enchant. PyEnchant combines all the functionality of the underlying Enchant library with the flexibility of Python and a nice "Pythonic" object-oriented interface. It also aims to provide some higher-level functionality than is available in the C, API.

For downloading the latest release source, please see Enchant Download (http://Pythonhosted.org/pyenchant/download.HTML).

For additional information and FAQs on PyEnchant, see the PyEnchant homepage (http://Pythonhosted.org/pyenchant/).

For more information and FAQs on Enchant, see the Enchant homepage (http://www.abisource.com/projects/enchant/).

Tesseract

Tesseract is an open-source Optical Character Reader/Recognition (OCR) engine for Windows, OSX, and Linux platform Operating Systems. The software is released under Apache License 2.0. Tesseract is considered as one of the most accurate open-source OCR software available to the public.

The Tesseract OCR engine was originally developed by Hewlett Packard (HP) from 1985-1994. Tesseract was one of the top three OCR engines in 1995 and was tested by University of Nevada, Las Vegas (UNLV). Between the periods of 1985-1998, the Tesseract OCR engine was written in C programming language and some portions were translated to C++ programming language in 1998. However, there little amount of work was done on the program between 1995-2006.

Tesseract was released as open-source software by HP and UNLV in 2005. Later in 2006, Google started to sponsor the development of Tesseract. Since then, the development process and accuracy of the Tesseract OCR engine have improved immensely.

The Tesseract OCR engine combines with the Leptonica Image Processing Library, which allows the engine to decompress and read a variety of compressed/uncompressed images over 60 programming languages.

For more information on Tesseract’s latest release, please see Tesseract Release Notes (https://github.com/tesseract-ocr/tesseract).

For downloading the latest release source, please see Tesseract Source Download (https://github.com/tesseract-ocr/tesseract).

For more information and FAQs, please see the Tesseract home page (https://github.com/tesseract-ocr/tesseract).

Ghostscript

Ghostscript is an interpreter for postscript and PDFs. See Ghostscript Documentation for details. The leading edge of Ghostscript development is under the GNU Affero GPL license.

For downloading the latest release source, please see the Ghostscript Download page (http://ghostscript.com/download/). For more detailed information and FAQs, see the Ghostscript home page (http://www.ghostscript.com/).

Python Imaging Library (PIL)

The Python Imaging Library (PIL) adds image processing capabilities to your Python interpreter. This library supports many file formats and provides powerful image processing and graphics capabilities.

For downloading the latest release source, please see the PIL Download page (http://www.Pythonware.com/products/pil/).

For more information and FAQs, see the PIL homepage (http://www.Pythonware.com/products/pil/).

System Requirement for Running the TCR2pdf Program

Minimum Requirements

CPU: Dual Core Processor
Memory: 2GB
Operating System: Ubuntu 12.04 LTS
Graphic Card: N/A
Storage: 80GB
Python 2.7+

Recommended Requirements

CPU: Dual Core Processor
Memory: 4GB
Operating System: Ubuntu 14.04LTS
Graphic Card: N/A
Storage: 160GB
Python 2.7+

Installation and Execution

To use TCR2pdf as a backend or a standalone program, you are required to have the following libraries installed:

- Python 2.7 or Python 3
- PyEnchant
- PDFQuery
- Tasseract
- OpenCV2
- GhostScript
- PIL

Backend Mode Installation
To use the TCR2pdf module as a backend, simply import TCR2PDFAPI from TCR2pdf_api.py into the required module.

Execution

To execute TCR2pdf as backend, simply call

TCR2pdf_api = TCR2pdfAPI.TCR2PDFAPI(enable_opencv)

Where (enable_opencv) is a boolean for enabling OpenCV support. Refer to main.py for detailed usage.

To execute TCR2PDF as a standalone program, simply

Python main.py <PDF_path/PDF_directory> or Python3 main.py <PDF_path/PDF_directory>

Standalone Mode
To use the TCR2pdf module as a standalone program no installation is required. You will be able to execute the program directly through Python.

OpenCV2

OpenCV is released under a BSD license, and hence it's free for both academic and commercial use. It has C++, C, Python and Java interfaces and supports Windows, Linux, Mac OS, iOS and Android. OpenCV was designed for computational efficiency with a strong focus on real-time applications. Written in optimized C/C++, the library can take advantage of multi-core processing. Enabled with OpenCL, it can take advantage of the hardware acceleration of the underlying heterogeneous computation platform. Adopted all around the world, OpenCV has more than 47,000 people in their user community and an estimated number of downloads exceeding nine million. Usage ranges from interactive art to mine inspections to stitching maps on the web or through advanced robotics (see http://opencv.org).

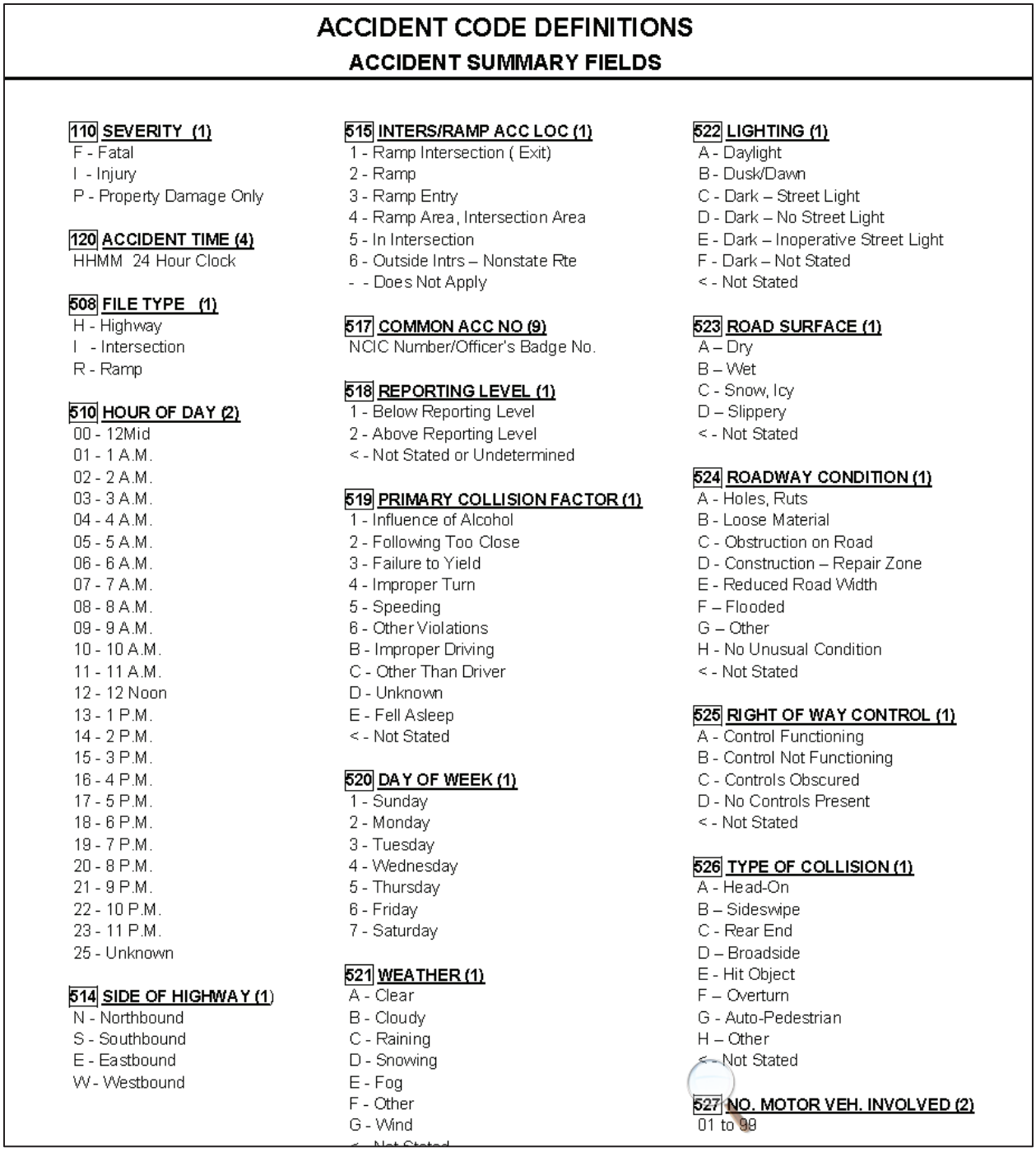
For downloading the latest release source, please see the OpenCV2 Download page (http://opencv.org/downloads.HTML).

For more information and FAQs, see the OpenCV2 homepage (http://opencv.org/).

The program TCR2pdf uses all accident code definitions used in TASAS. These are listed in Figure C.18, Figure C.19, Figure C.20, Figure C.21, Figure C.22, and Figure C.23. The data in these figures indicates what each code means in the TCR2pdf database. This data will be needed in order to extract data from a TCR and format it into the appropriate field.

In this research study, *.CSV files are used to transfer the data contained in the TCR to the TCR2pdf database. The formats of the output *.CSV are:

Column1_data, Column2_data ,,,,, Column39_data, Column40_data



ACCIDENT CODE DEFINITIONS

ACCIDENT SUMMARY FIELDS

110 SEVERITY (1)
F - Fatal
I - Injury
P - Property Damage Only

120 ACCIDENT TIME (4)
HHMM 24 Hour Clock

508 FILE TYPE (1)
H - Highway
I - Intersection
R - Ramp

510 HOUR OF DAY (2)
00 - 12Mid
01 - 1 A.M.
02 - 2 A.M.
03 - 3 A.M.
04 - 4 A.M.
05 - 5 A.M.
06 - 6 A.M.
07 - 7 A.M.
08 - 8 A.M.
09 - 9 A.M.
10 - 10 A.M.
11 - 11 A.M.
12 - 12 Noon
13 - 1 P.M.
14 - 2 P.M.
15 - 3 P.M.
16 - 4 P.M.
17 - 5 P.M.
18 - 6 P.M.
19 - 7 P.M.
20 - 8 P.M.
21 - 9 P.M.
22 - 10 P.M.
23 - 11 P.M.
25 - Unknown

514 SIDE OF HIGHWAY (1)
N - Northbound
S - Southbound
E - Eastbound
W - Westbound

515 INTERS/RAMP ACC LOC (1)
1 - Ramp Intersection (Exit)
2 - Ramp
3 - Ramp Entry
4 - Ramp Area, Intersection Area
5 - In Intersection
6 - Outside Intrs – Nonstate Rte
- - Does Not Apply

517 COMMON ACC NO (9)
NCIC Number/Officer's Badge No.

518 REPORTING LEVEL (1)
1 - Below Reporting Level
2 - Above Reporting Level
< - Not Stated or Undetermined

519 PRIMARY COLLISION FACTOR (1)
1 - Influence of Alcohol
2 - Following Too Close
3 - Failure to Yield
4 - Improper Turn
5 - Speeding
6 - Other Violations
B - Improper Driving
C - Other Than Driver
D - Unknown
E - Fell Asleep
< - Not Stated

520 DAY OF WEEK (1)
1 - Sunday
2 - Monday
3 - Tuesday
4 - Wednesday
5 - Thursday
6 - Friday
7 - Saturday

521 WEATHER (1)
A - Clear
B - Cloudy
C - Raining
D - Snowing
E - Fog
F - Other
G - Wind
< - Not Stated

522 LIGHTING (1)
A - Daylight
B - Dusk/Dawn
C - Dark – Street Light
D - Dark – No Street Light
E - Dark – Inoperative Street Light
F - Dark – Not Stated
< - Not Stated

523 ROAD SURFACE (1)
A – Dry
B – Wet
C - Snow, Icy
D – Slippery
< - Not Stated

524 ROADWAY CONDITION (1)
A - Holes, Ruts
B - Loose Material
C - Obstruction on Road
D - Construction – Repair Zone
E - Reduced Road Width
F – Flooded
G – Other
H - No Unusual Condition
< - Not Stated

525 RIGHT OF WAY CONTROL (1)
A - Control Functioning
B - Control Not Functioning
C - Controls Obscured
D - No Controls Present
< - Not Stated

526 TYPE OF COLLISION (1)
A - Head-On
B – Sideswipe
C - Rear End
D – Broadside
E - Hit Object
F – Overturn
G - Auto-Pedestrian
H – Other
< - Not Stated

527 NO. MOTOR VEH. INVOLVED (2)
01 to 99

Figure C.18. Accident Summary Fields "Look Up" tables in the TCR2PDF database.

ACCIDENT CODE DEFINITIONS

PARTY SUMMARY FIELDS

0 = Examine All Parties Involved 1-9 = Examine Only Specified Party No.

60 # PARTY TYPE (1)
A – Passenger Car/Station Wagon
B – Passenger Car w/Trailer
C – Motorcycle
D – Pickup/Panel Truck
E – Pickup/Panel w/Trailer
F – Truck, Truck Tractor
G – Truck/Tractor w/Trailer
2 - Truck/Tractor w/2 Trailers
3 - Truck/Tractor w/3 Trailers
4 – Single Unit Tanker
5 - Truck/Tractor w/1 Tank Trailer
6 - Truck/Tractor w/2 Tank Trailer
H - School Bus
I - Other Bus
J - Emergency Vehicle
K - Highway Construction Equip. **
L – Bicycle
M - Other Vehicle
N - Other Non-Vehicle
O - Spilled Load
P – Disengaged Tow
Q - Uninvolved Vehicle
R – Moped
T – Train
U – Pedestrian
V - Dismounted Pedestrian
W - Animal-Livestock
X - Animal-Deer
Z - Animal-Other
< - Not Stated

61 # DIRECTION TRAVEL (1)
N – Northbound
S – Southbound
E – Eastbound
W – Westbound
< - Not Stated
- - Does Not Apply

62 # VEH. HWY. INDICATOR (1)
1 - On State Route
2 - Not on State Route
3 - Intersecting State Route
< - Not Stated
- - Does Not Apply

63 # SPECIAL INFO (1)
A – Hazardous Materials
B – Cell Phone in use*
C – Cell Phone not in use*
D – Cell Phone none/unknown*
< - Not Stated
- - Does Not Apply

* **Codes Eff. 04-01-01**

** Includes Equipment engaged in Const/Maint activities as of 02/22/2000

64 # PERSONS KILLED (2)
00 - 99

65 # PERSONS INJURED (2)
00 - 99

66 # PRIMARY OBJ STRUCK (2)

68 # OTHER
01 - Side of Bridge Railing
02 - End of Bridge Railing
03 - Pier, Column, Abutment
04 - Bottom of Structure
05 - Bridge End Posts in Gore
06 - End of Guardrail
07 - Bridge Approach Guardrail
10 - Light or Signal Pole
11 - Utility Pole
12 - Pole (Type Not Stated)
13 - Traffic Sign/Sign Post
14 - Other Signs Not Traffic
15 - Guardrail
16 - Median Barrier
17 - Wall (exc. Soundwall)
18 - Dike or Curb
19 - Traffic Island
20 - Raised Bars
21 - Concrete Object (HDWL, D.I.)
22 - Guidepost, Culvert, PM
23 - Cut Slope or Embankment
24 - Over Embankment
25 - In Water
26 - Drainage Ditch
27 - Fence
28 - Trees
29 - Plants
30 - Sound Walls
40 - Natural Material on Road
41 - Temp Barricades, Cones, Etc.
42 - Other Object On Road
43 - Other Object Off Road
44 - Overturned
45 - Crash Cushions-Sand (After 01/01/96, Before Both 45)
46 - Crash Cushion-Other (After 01/01/96, Before Both 45)
51 - Call Box (After 01/01/96)
98 - Unknown Object Involved
99 - No Object Involved
V1 - Thru V9 - Vehicle 1 to 9
<< - Not Stated
-- - Does Not Apply

Figure C.19. Party Summary Fields "Look Up" tables (1 of 2) in the TCR2PDF database.

ACCIDENT CODE DEFINITIONS

PARTY SUMMARY FIELDS

0 = Examine All Parties Involved 1-9 = Examine Only Specified Party No.

67 # PRIMARY LOC. OF COLL (1)

69 # OTHER
A - Beyond Median or Stripe – Left(NE-2LN)
B - Beyond Shoulder Drivers Left
C - Left Shoulder Area
D - Left Lane
E - Interior Lanes
F - Right Lane
G - Right Shoulder Area
H - Beyond Shoulder Drivers Right
I - Gore Area
J – Other
V - HOV Lane(s) (After 01/01/96)
W -HOV Buffer Area (After 01/01/96)
<< - Not Stated
-- - Does Not Apply

74 # 1ST OTHER ASSOC FACTOR (1)

75 # 2ND
1 - Influence of Alcohol
2 - Following Too Close
3 - Failure To Yield
4 - Improper Turn
5 – Speeding
6 - Other Violations
A – Cell Phone* (INATTN)
B – Electronic Equip* (INATTN)
C – Radio/CD/Headphn* (INATTN)
D – Smoking* (INATTN)
E - Vision Obscurement
F – Inattention – Other
G - Stop & Go Traffic
H - Enter/Leave Ramp
I - Previous Collision
J - Unfamiliar with Road
K - Defective Veh. Equipment
L - Uninvolved Vehicle
M – Other
N - None Apparent
P – Wind
R - Ramp Accident
S - Runaway Vehicle
T – Eating* (INATTN)
U – Children* (INATTN)
V – Animals* (INATTN)
W – Personal Hygiene* (INATTN)
X – Reading* (INATTN)
< - Not Stated
- - Does Not Apply

* Inattention Codes Eff. 01-01-01

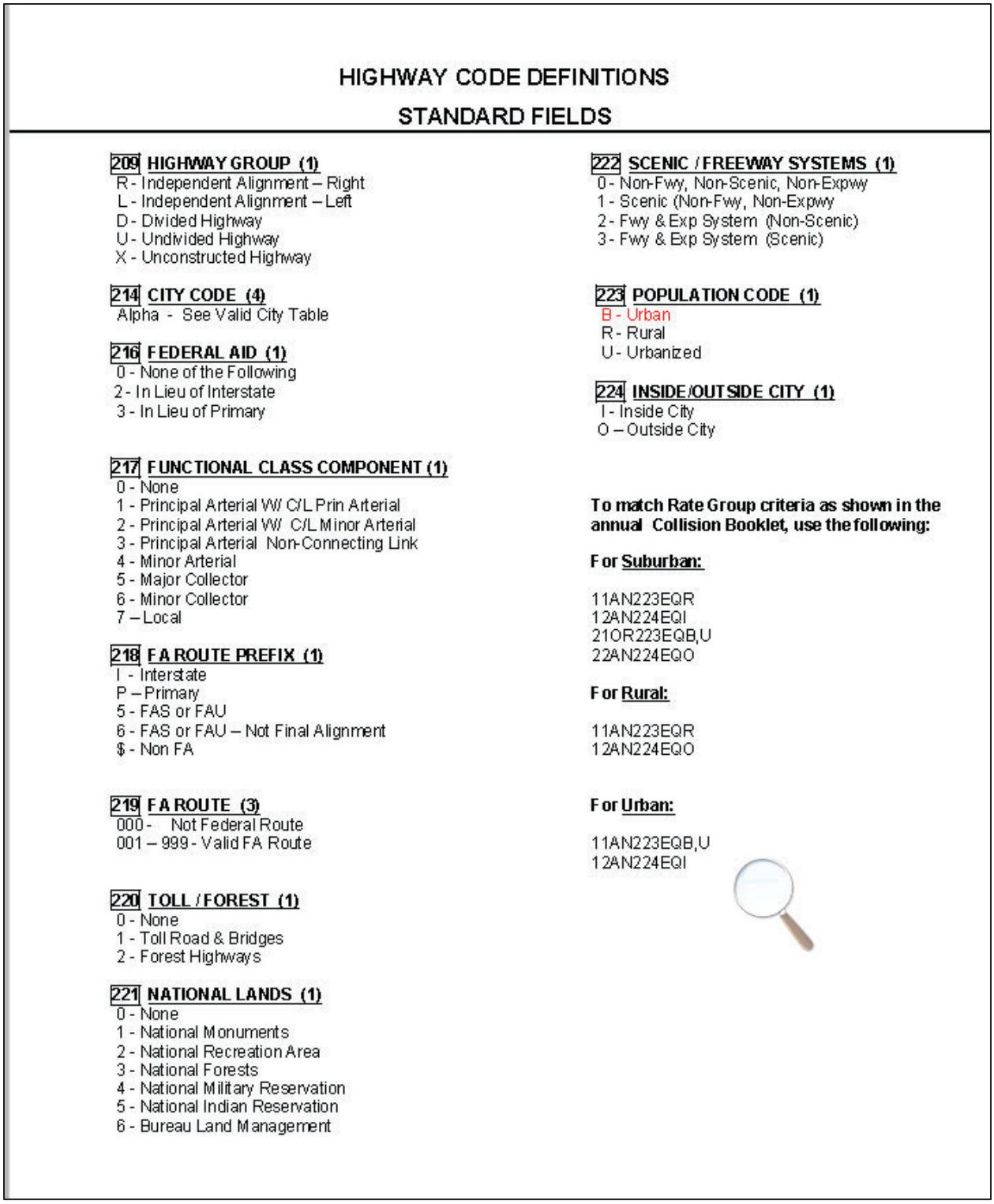
76 # MOVE PRECEDING COLL. (1)
A - Stopped
B - Proceeding Straight
C - Ran Off Road
D - Making Right Turn
E - Making Left Turn
F - Making U-Turn
G - Backing
H - Slowing, Stopping
I - Pass Other Vehicle (2Wy-2Ln)
J - Change Lanes
K - Parking
L - Enter From Shoulder
M - Other Unsafe Turn
N - Cross Into Opposing Lane (Undiv. Only)
O - Parked
P - Merging
Q - Traveling Wrong Way
R - Other
< - Not Stated
- - Does Not Apply

PEDESTRIAN
2 - Xing Xwalk – Intersection
3 - Xing Xwalk – Not Intersection
4 - Xing – Not Xwalk
5 - Roadway – Include Shoulder
6 - Not in Roadway
7 - Approach/Leave School Bus

77 # 1ST SOBRIETY (1)

78 # 2ND (DRUG/PHYSICAL) (1)
A - Had Not Been Drinking (0%)
B - HBD – Under Influence (>0.08%)
C - HBD Not Under Influence (0.01-.07%)
D - HBD – Impairment Unknown
E - Under Drug Influence
F - Other Physical Impairment
G - Impairment Unknown
H - Not Applicable
I - Driver Fatigue
< - Not Stated
- - Does Not Apply

Figure C.20. Party Summary Fields "Look Up" tables (2 of 2) in the TCR2PDF database.



HIGHWAY CODE DEFINITIONS

STANDARD FIELDS

209 HIGHWAY GROUP (1)
R - Independent Alignment – Right
L - Independent Alignment – Left
D - Divided Highway
U - Undivided Highway
X - Unconstructed Highway

214 CITY CODE (4)
Alpha - See Valid City Table

216 FEDERAL AID (1)
0 - None of the Following
2 - In Lieu of Interstate
3 - In Lieu of Primary

217 FUNCTIONAL CLASS COMPONENT (1)
0 - None
1 - Principal Arterial W/ C/L Prin Arterial
2 - Principal Arterial W/ C/L Minor Arterial
3 - Principal Arterial Non-Connecting Link
4 - Minor Arterial
5 - Major Collector
6 - Minor Collector
7 – Local

218 FA ROUTE PREFIX (1)
I - Interstate
P – Primary
5 - FAS or FAU
6 - FAS or FAU – Not Final Alignment
$ - Non FA

219 FA ROUTE (3)
000 - Not Federal Route
001 – 999 - Valid FA Route

220 TOLL / FOREST (1)
0 - None
1 - Toll Road & Bridges
2 - Forest Highways

221 NATIONAL LANDS (1)
0 - None
1 - National Monuments
2 - National Recreation Area
3 - National Forests
4 - National Military Reservation
5 - National Indian Reservation
6 - Bureau Land Management

222 SCENIC / FREEWAY SYSTEMS (1)
0 - Non-Fwy, Non-Scenic, Non-Expwy
1 - Scenic (Non-Fwy, Non-Expwy
2 - Fwy & Exp System (Non-Scenic)
3 - Fwy & Exp System (Scenic)

223 POPULATION CODE (1)
B - Urban
R - Rural
U - Urbanized

224 INSIDE/OUTSIDE CITY (1)
I - Inside City
O – Outside City

To match Rate Group criteria as shown in the annual Collision Booklet, use the following:

For Suburban:

11AN223EQR
12AN224EQI
21OR223EQB,U
22AN224EQO

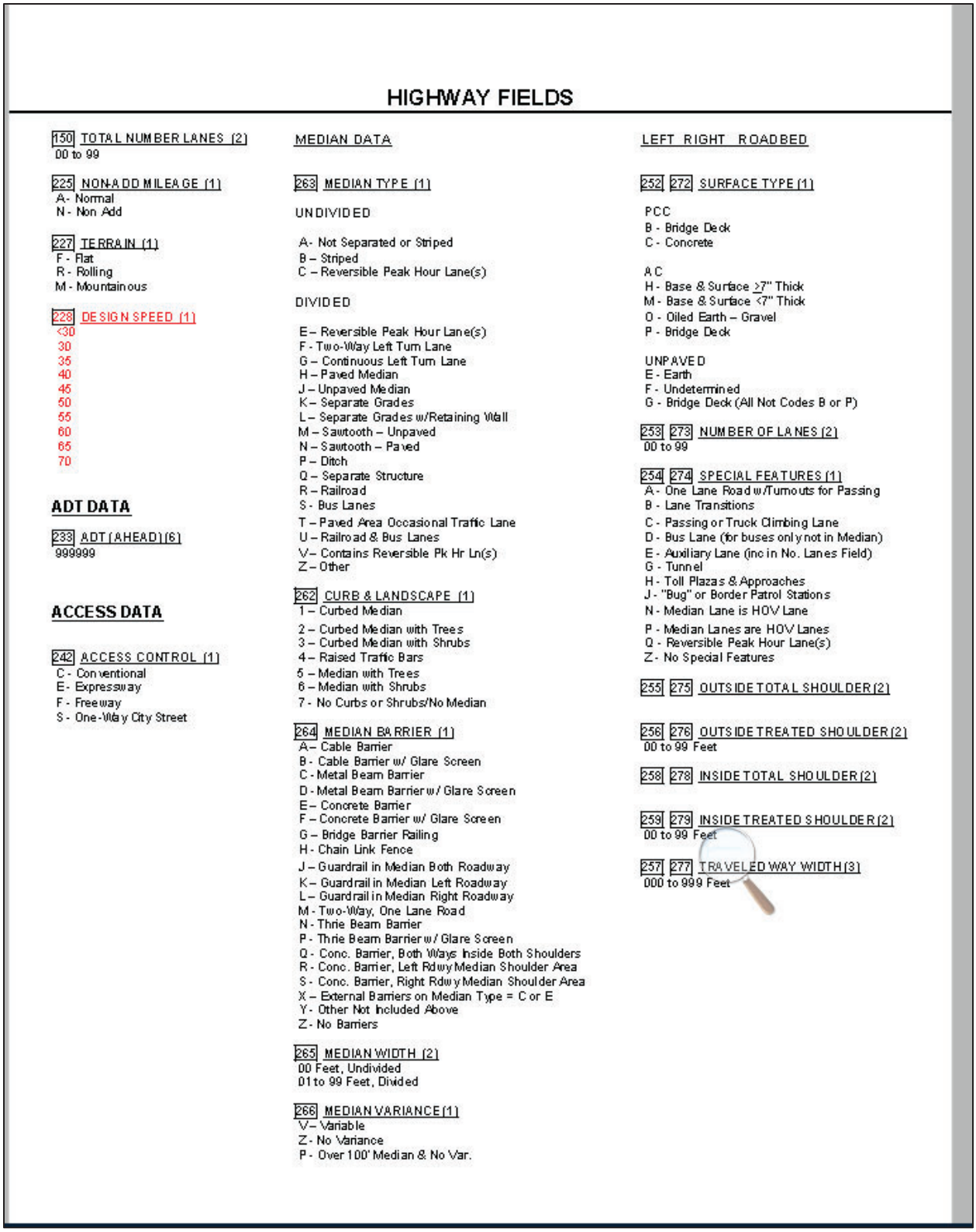
For Rural:

11AN223EQR
12AN224EQO

For Urban:

11AN223EQB,U
12AN224EQI

Figure C.21. Standard Fields "Look Up" tables in the TCR2PDF database.



HIGHWAY FIELDS

150 TOTAL NUMBER LANES (2)
00 to 99

225 NON-ADD MILEAGE (1)
A - Normal
N - Non Add

227 TERRAIN (1)
F - Flat
R - Rolling
M - Mountainous

228 DESIGN SPEED (1)
<30
30
35
40
45
50
55
60
65
70

ADT DATA

233 ADT (AHEAD) (6)
999999

ACCESS DATA

242 ACCESS CONTROL (1)
C - Conventional
E - Expressway
F - Freeway
S - One-Way City Street

MEDIAN DATA

263 MEDIAN TYPE (1)

UNDIVIDED

A - Not Separated or Striped
B – Striped
C – Reversible Peak Hour Lane(s)

DIVIDED

E – Reversible Peak Hour Lane(s)
F - Two-Way Left Turn Lane
G – Continuous Left Turn Lane
H – Paved Median
J – Unpaved Median
K – Separate Grades
L – Separate Grades w/Retaining Wall
M – Sawtooth – Unpaved
N – Sawtooth – Paved
P – Ditch
Q – Separate Structure
R – Railroad
S - Bus Lanes
T – Paved Area Occasional Traffic Lane
U – Railroad & Bus Lanes
V – Contains Reversible Pk Hr Ln(s)
Z – Other

262 CURB & LANDSCAPE (1)
1 – Curbed Median
2 – Curbed Median with Trees
3 – Curbed Median with Shrubs
4 – Raised Traffic Bars
5 – Median with Trees
6 – Median with Shrubs
7 - No Curbs or Shrubs/No Median

264 MEDIAN BARRIER (1)
A – Cable Barrier
B - Cable Barrier w/ Glare Screen
C - Metal Beam Barrier
D - Metal Beam Barrier w/ Glare Screen
E – Concrete Barrier
F – Concrete Barrier w/ Glare Screen
G – Bridge Barrier Railing
H - Chain Link Fence
J – Guardrail in Median Both Roadway
K – Guardrail in Median Left Roadway
L – Guardrail in Median Right Roadway
M - Two-Way, One Lane Road
N - Thrie Beam Barrier
P - Thrie Beam Barrier w/ Glare Screen
Q - Conc. Barrier, Both Ways Inside Both Shoulders
R - Conc. Barrier, Left Rdwy Median Shoulder Area
S - Conc. Barrier, Right Rdwy Median Shoulder Area
X – External Barriers on Median Type = C or E
Y - Other Not Included Above
Z - No Barriers

265 MEDIAN WIDTH (2)
00 Feet, Undivided
01 to 99 Feet, Divided

266 MEDIAN VARIANCE (1)
V – Variable
Z - No Variance
P - Over 100' Median & No Var.

LEFT RIGHT ROADBED

252 272 SURFACE TYPE (1)

PCC
B - Bridge Deck
C - Concrete

AC
H - Base & Surface ≥7" Thick
M - Base & Surface <7" Thick
O - Oiled Earth – Gravel
P - Bridge Deck

UNPAVED
E - Earth
F - Undetermined
G - Bridge Deck (All Not Codes B or P)

253 273 NUMBER OF LANES (2)
00 to 99

254 274 SPECIAL FEATURES (1)
A - One Lane Road w/Turnouts for Passing
B - Lane Transitions
C - Passing or Truck Climbing Lane
D - Bus Lane (for buses only not in Median)
E - Auxiliary Lane (inc in No. Lanes Field)
G - Tunnel
H - Toll Plazas & Approaches
J - "Bug" or Border Patrol Stations
N - Median Lane is HOV Lane
P - Median Lanes are HOV Lanes
Q - Reversible Peak Hour Lane(s)
Z - No Special Features

255 275 OUTSIDE TOTAL SHOULDER (2)

256 276 OUTSIDE TREATED SHOULDER (2)
00 to 99 Feet

258 278 INSIDE TOTAL SHOULDER (2)

259 279 INSIDE TREATED SHOULDER (2)
00 to 99 Feet

257 277 TRAVELED WAY WIDTH (3)
000 to 999 Feet

Figure C.22. Highway Summary Fields "Look Up" tables in the TCR2PDF database.

INTERSECTION FIELDS

412 INTERSECTION TYPE (1)
F - Four-Legged
M - Multi-Legged
S - Offset
T - Tee
Y - WYE
Z - Other

417 CONTROL TYPE (1)
A - No Control
B - Stop Signs on Cross Street Only
C - Stop Signs on Mainline Only
D - Four-Way Stop Signs
E - Four-Way Flasher (Red on Cross Street)
F - Four-Way Flasher (Red on Mainline)
G - Four-Way flasher (Red on All)
H - Yield Signs (On Cross Street Only)
I - Yield Signs (On Main Line Only)
J - Signals Pretimed – 2 phase
K - Signals Pretimed – Multi-Phase
L - Signals Semi-Traffic Actuated – 2 Phase
M - Signals Semi-Traffic Actuated – Multi Phase
N - Signals Full-Traffic Actuated – 2 Phase
P - Signals Full-Traffic Actuated – Multi-Phase
Z - Other

422 LIGHTING (1)
N - No
Y - Yes

MAIN LINE

432 SIGNAL MAST ARM (1)
N - No
Y - Yes

433 LEFT TURN CHANNELIZATION (1)
C - Curbed Median Left Turn Channelization
N - No Left Turn Channelization
P - Painted Left Turn Channelization
R - Raised Bars – Left Turn Channelization

434 RIGHT TURN CHANNELIZATION
N - No Right Turn Channelization
Y - Channelization Provided For Right Turns

435 TRAFFIC FLOW (1)
N - Two-Way Traffic - No Left Turns Permitted
P - Two-Way Traffic - Left Turns Permitted
R - Two-way-Traffic – Left Turns Restricted During Peak Hours
W - One-Way Traffic
Z - Other

436 NUMBER OF LANES (1)
0 to 9

INTERSECTING STREET (ROUTE)

452 SIGNAL MASRT ARM (1)
N - No
Y - Yes

453 LEFT TURN CHANNELIZATION (1)
Same as Intersection Item

454 RIGHT TURN CHANNELIZATION (1)
Same as Intersection Item

455 TRAFFIC FLOW (1)
Same as Intersection Item

456 NUMBER OF LANES (1)
0 to 9

RAMP FIELDS

301 ON/OFF INDICATOR (1)
O – On
F – Off
Z – Other

316 RAMP TYPE (1)
A - Frontage Road
B - Collector Road
C - Direct or Semi-Direct Connector (Left)
D - Diamond Type Ramp
E - Slip Ramp
F - Direct or Semi Direct Connector (Right)
G - Loop-w/Left Turn
H - Buttonhook Ramp
J - Scissors Ramp
K - Split Ramp
L - Loop-w/o Left Turn
M - Two-Way Ramp Segment
R - Rest Area, Vista Point, Truck Scale
Z – Other
P - Dummy-Paired
V - Dummy-Volume Only

359 AREA 4 (1)
N - No
Y - Yes

Figure C.23. Intersection Fields "Look Up" tables in the TCR2PDF database.