**Highway Incident Detection Timeline**

**Work Order (WO) 009**

**Contract No. 4400011166**

 **Report on Task 2-Deliverable 2.1:**

**Data Pre-Processing and Development of Integrated Framework**

**By:**

**Temple University Research Team**

Principal Investigator: Joseph Thomas Coe, Jr., Ph.D.

Assistant Professor

Co- Principal Investigator: Bechara E. Abboud, Ph.D., P.E.

Associate Professor

Co- Principal Investigator: Joseph Picone, Ph.D.

Professor

Graduate Research Assistant: Siavash Mahvelati

Undergraduate Research Assistant: Aaron M. Gross

**04/17/2017**

Contents

[1 Task 2: Data Pre-Processing and Development of Integrated Framework 3](#_Toc484172409)

[1.1 Introduction 3](#_Toc484172410)

[1.2 Integrated Framework For Pairing PennDOT RCRS to County 911 Records 3](#_Toc484172411)

[1.3 Current Status of Data Normalization and Continued Data Collection Efforts 5](#_Toc484172412)

[1.3.1 Pennsylvania State Police Data 7](#_Toc484172413)

[1.3.2 Project Website & Data Repository (Task 2.2) 7](#_Toc484172414)

[1.4 Future Data Collection Efforts 7](#_Toc484172415)

# Task 2: Data Pre-Processing and Development of Integrated Framework

##  Introduction

As previously discussed in previous project reports, the general objective of this project is to characterize the average timeline for when the Pennsylvania Department of Transportation (PennDOT) is notified of incidents requiring highway closures along Primary Interstate Highway I-76, I-80, I-81, and I-95, and Auxiliary Interstate Highway I-78 and I-83 throughout the Commonwealth of Pennsylvania. PennDOT provided electronic copies of the RCRS logs for the highways of interest as of Tuesday, November 22, 2016, so data collection efforts have focused on contacting county 911 call centers throughout the Commonwealth of Pennsylvania. The results from these data collection efforts to date have informed the decision-making regarding how the data is post-processed for development of an integrated framework for comparing the PennDOT RCRS and county 911 datasets.

## Integrated Framework For Pairing PennDOT RCRS to County 911 Records

As part of initial efforts to develop an integrated framework that will be used to pair the different datasets in this project, the research team went about manually pairing a small subset of PennDOT RCRS records. Initially, this consisted of matching the first 100 records in the PennDOT RCRS to the corresponding county 911 logs. Since the typical definitive match rate was approximately 1/3 of the examined records, this was expanded to approximately 300 records so that the first 100 matches were encountered. Data from the following counties was included in the first 100 matches: Lehigh, Luzerne, Lebanon, Berks, Lackawanna, Schuylkill, Northampton, and Susquehanna. This exercise was performed prior to the February in-person meeting (Friday, February 3rd, 2017) between the Temple research team and PennDOT project team. During this meeting, a number of pertinent questions were answered regarding filtering of the PennDOT RCRS data. As a result, the aforementioned pilot study was repeated with the new filter criteria (i.e., entries with status of “Closed”, “Lane Restriction”, “Ramp Closure”, and “Ramp Restriction”), which changed the total number of pertinent PennDOT RCRS entries from approximately 21,000 entries to approximately 9,000. This changed the entries that were examined to determine first 100 matches.

The goal of this initial data pairing pilot study was to compare the nature of the data provided by the RCRS in relationship to records provided by the different 911 call centers. This would allow the research team to identify the most appropriate data necessary to develop the integrated framework. As data collection has progressed, it has become obvious that there is a wide range in the type of information provided by each of the 911 call centers throughout the Commonwealth of Pennsylvania (Fig. 1). So it was important to identify what information was typically provided by most call centers and what data would prove the most useful to manually pair data. Data pre-processing efforts could then focus on ensuring the necessary data is in place as a more thorough data analysis is performed for future project tasks.

Based on the manual pairing efforts from the initial pilot study, the following parameters were noted as critical in being able to identify with confidence any matches between PennDOT RCRS and county 911 entries: Date/timestamp, GPS coordinates, location, and incident type. This set of information would allow filtering to occur initially on date/time by defining a threshold time period around each PennDOT RCRS entry. For manual pairing, the location field in the spreadsheet would then allow filtering to take place by mile marker along the highway of interest. This process is accurate but labor intensive and inefficient for scale up to the large number of entries in the PennDOT RCRS data. Future automated pairing efforts will benefit tremendously from the use of GPS coordinates to trace the location of entries in the 911 county records. The PennDOT RCRS records fortunately already include this information. However, the majority of the 911 call centers were unable to provide this data as of the date of submission of this Task 2 report (Table 1). Given the large impact on efficiency, the bulk of future normalization efforts will focus on acquiring GPS data whenever possible for all 911 call center datasets. The final field in the normalization process is incident type. This is beneficial because for a number of counties, the 911 call center data was not pre-filtered for the necessary incident types highlighted by PennDOT at the February in person meeting and March conference call (Thursday, March 16, 2017).

Once the specified normalized fields were determined, efforts focused on developing an efficient algorithm for developing normalized datasets for all county 911 records received. Figure 2 presents an example of a normalized data file for Susquehanna County. For the normalization process, the original county records were placed into a single Microsoft Excel® file. If the original records were spread across multiple files or sheets in a single spreadsheet file, all data was manually sorted into the same file. Any files in a .pdf or .txt format were read into a single Excel® file. These Excel® files were then parsed and all column headers were removed. The file was then revised using a custom developed script written in the programming language Python. The code utilizes the “Pandas” Python library to read and write Excel® files and manipulate the rows and columns of the data. Rules were developed for each individual county to format the pertinent columns within each spreadsheet in a consistent manner. For example, after each county Excel® spreadsheet has passed through the Python code, all normalized times have the format YYY-MM-DD HH:MM:SS no matter what was provided in the original source format. When provided (or manually resolved), the latitude and longitude of the GPS coordinates were formatted to 6 decimal places (approximately 0.1 m resolution). Mile markers, if present in the location fields of the original data files, were identified automatically by the Python code and reported to the nearest 0.1 mile. The textual information (including any street addresses if applicable) was also passed along to the normalized data spreadsheet. Finally, if the county listed incident type, it was identified by the Python code and passed directly to the normalized data file with no further processing. In cases where some of the desired normalized information was not provided, the columns were still included in the revised normalized spreadsheet but were left blank for potential future updates with revised data. The remaining columns in the normalized data represent the raw records before normalization, and as such vary from county to county. This was included so that the raw, unmodified data would be conveniently available when reviewing the effects of the normalization process. The final step in the Python code is to write the newly formatted data file to an Excel® data file.

|  |
| --- |
|  |
|  |
|  |

Figure 1. Examples of logs provided by (top) Dauphin County (middle) Montgomery County, and (bottom) Cumberland County

## Current Status of Data Normalization and Continued Data Collection Efforts

As previously noted, future automated pairing between PennDOT RCRS records and county 911 records will rely on GPS coordinates. Based on the initial efforts of the Temple research team, there was no better approach than using date/timestamp in combination with GPS coordinates whereby automation would provide the level of fidelity necessary to make definitive matches. The PennDOT RCRS records already completely provide GPS coordinate information for all entries. However, the majority of the 911 call centers were unable to provide this data as of the date of submission of this Task 2 report. Table 1 summarizes the current statues of the data normalization efforts. Included for each county entry in Table 1 is the total number of RCRS records for a given county based on the previously agreed upon status filters. Given that future automation efforts rely on GPS information, the bulk of recent normalization efforts have focused on acquiring GPS data whenever possible for all 911 call center datasets. The following sections highlight the efforts involved in procuring this information from county emergency dispatch centers.



Table 1. Current data normalization status for all counties.

### Pennsylvania State Police Data

In addition to efforts related to GPS data at Dauphin County, the Temple research team also underwent efforts to procure Pennsylvania State Police (PSP) data. Both Philadelphia and Delaware counties contain a significant share (over half) of the RCRS records provided by PennDOT. Both these county 911 call centers transfer any calls related to highway incidents to the local PSP troop and do not maintain records regarding dispatching of emergency personnel nor incident type. Therefore, this information must be requested directly from PSP. When notified of this matter during the in-person meeting between the Temple research team and PennDOT project team, it was determined that it may be difficult to acquire this data and that the project should proceed as if this data may not be procured in a timely manner. Therefore, the Temple research team has continued with data normalization efforts, but has also contacted the local PSP Troop K offices who handle Philadelphia County to inquire about any incident logs maintained at their offices. This attempted contact has occurred via phone over a time period starting Monday, March 27, 2017 and continuing to the date of the Task 2.1 deliverable report. No data has been procured as of this time, but efforts will continue in the future. Initial efforts with the pilot study did not demonstrate significant differences in average response times based on the initial 30+ matches when compared to 100 matches between RCRS data and county 911 records. So there is no evidence at this time that Philadelphia County or Delaware county data will significantly impact the project results. However, in an attempt to provide statistics across a thorough percentage of RCRS records, the Temple research team will continue to make efforts to acquire Philadelphia and Delaware county records.

### Project Website & Data Repository (Task 2.2)

All the normalized datasets have been archived on an online repository server accessible by the Temple research team. The online website developed for Task 1.2 has been updated to allow the PennDOT project team access to the revised data files (in addition to the original data files as provided during the Task 1.2 deliverable). As before, this website is available via at <https://www.isip.piconepress.com/projects/penndot_response_time/>. Download of the normalized data files (and original data files) requires the username (*penndot*) and password (*penndot\_2017\_response\_time*). As with Task 1.2, it is anticipated that this website will undergo continual revisions as additional data is provided by the remaining counties and as this data is normalized by the Temple research team.

## Future Data Collection Efforts

As noted in the previous sections, data collection for this project continues based on the efforts to acquire all remaining county 911 logs, particularly those with a significant number of entries in the RCRS records (Table 1). Furthermore, the Temple research team plans to continue attempting to acquire GPS information for whichever counties have those capabilities. The Temple research team will also continue to reach out to local PSP troop centers to inquire regarding the availability of dispatch information along the highways selected for this study.

|  |
| --- |
|  |
|  |

Figure 2. Examples of normalized logs for (top) Bucks County and (bottom) Susquehanna County.