**Highway Incident Detection Timeline**

**Work Order (WO) 009**

**Contract No. 4400011166**

 **Report on Task 4-Deliverable 4.1:**

**Data Curation**

**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

**07/17/2017**

Contents

[1 Task 4: Data Curation 3](#_Toc488074251)

[1.1 Introduction 3](#_Toc488074252)

[1.2 Summary of Project Website 3](#_Toc488074253)

[1.3 Future Efforts 9](#_Toc488074254)

# Task 4: Data Curation

##  Introduction

The purpose of Task 4 is to discuss efforts related to providing an archival copy of all data for TEM Work Order (WO) 009. The Temple research team has acquired a significant amount of data from both the PennDOT Road Condition Reporting System (RCRS) (9,009 distinct entries based on filtering criteria approved by PennDOT) and various 911 call centers throughout the Commonwealth of Pennsylvania (945,267 distinct entries up to date). This data covers a range of time periods from 2013 to 2016. In addition to the raw data from the 911 call centers and PennDOT RCRS, the Temple research team has generated a number of other outputs as TEM WO 009 has progressed and estimates have been developed for the latency between 911 call center entries and PennDOT RCRS entries: (1) copies of normalized data from counties investigated in study; (2) Python-based scripts that manipulate data and perform analysis to locate pairings between 911 and RCRS entries; (3) results from pairing efforts between the 911 and RCRS databases; and (4) task deliverable reports related to research efforts throughout project. Given the large amount of data and outputs in this project, the Temple research team has developed a project website with a user friendly interface to help archive information related to research efforts. This website has and will continue to serve as the instrument by which data is curated for TEM WO 009 as prescribed by Task 4. It is anticipated that this website will undergo continual revisions to archive efforts remaining until the project completion date. The following sections describe the structure and content of the website.

## Summary of Project Website

The TEM WO 009 project website is available via the following hyperlink: <https://www.isip.piconepress.com/projects/penndot_response_time/>. Most of the website is unprotected, but a username and password are required to download restricted information (e.g., data files, Python tools, etc.). The project website is organized using a simple structure. The home page provides a quick summary of the project objective and a list of recent project highlights. Included on the project home page are hyperlinks to the following items: (1) Overview; (2) Downloads; and (3) County Codes. Items (1) and (3) are relatively minor aspects of the website in relation to the project efforts. The “Overview” link sends the website viewer to a single webpage that provides a more thorough discussion of the project goals and research efforts. It is meant to augment the home page and present the project in a more detailed context so that users unfamiliar with the project (e.g., other PennDOT personnel with which the PennDOT project team wishes to share the website) can better understand the project prior to exploring other aspects of the site. The “County Codes” hyperlink on the website provides access to a Microsoft Excel® spreadsheet with county coding information for all 67 Pennsylvania counties based on various agencies/purposes.

The most significant aspect of the project website is the location of all relevant project files, which is accessible from the “Downloads” hyperlink on the project website home page. Following this hyperlink leads the user to a landing page with hyperlinks to two items: (1) a compressed archive of all data files in a single downloadable GNU compressed tar file; and (2) a directory of all data files by county. The directory of all data files includes the following list of folders (Fig. 1):

* 000\_penndot: This folder contains Excel® spreadsheets of all RCRS data provided by PennDOT for the purposes of this research project. There are two versions of the spreadsheet. The first “penndot\_tidy.xlsx” presents all the filtered RCRS data for this project as originally provided in the master RCRS spreadsheet provided via email by the PennDOT project team on Tuesday, November 22, 2016. The other file in the directory (“000\_penndot\_norm.csv”) is a comma-separated values (CSV) file with the normalized copy of the filtered RCRS data based on the normalization techniques described in the Task 2.1 deliverable report. Appended to each entry is a unique hash identifier for simplified, rapid data lookup functionalities using the Python scripts developed for this project. The CSV file can be opened using either a simple text editor (e.g., the Notepad program typically provided within the Windows operating system) or Microsoft Excel® among other software.
* County folders: A folder is provided for each county for which data was acquired and processed in this study. The naming convention for these folders is “XXX\_county name” where the XXX represents the county number. Inside each county-level directory folder are the following files (e.g., Fig. 2):
	+ CSV file with normalized 911 call center data. As with the normalized PennDOT RCRS data, a unique hash identifier has been affixed to each 911 data entry in the normalized CSV file. The naming convention is “XXX\_county name.csv” where XXX is the county number.
	+ DAT files containing pairing information between 911 call entries and RCRS data. Two DAT files are provided: (1) “XXX\_county name\_ppc.dat” represents the pairing information generated when matching PennDOT RCRS entries to 911 call center entries, starting with PennDOT RCRS as the source data and 911 data as destination data; and (2) “XXX\_county name\_pcp.dat” represents the pairing information generated when matching 911 call center entries to PennDOT RCRS, starting with the 911 records as the source data and RCRS entries as the destination data. In both cases, the pairing information is expressed as a two column table with source hash identifier paired with destination hash identifier. So in cases where the pairing goes from PennDOT RCRS to 911 data, the first column of the table in the DAT file would contain the hash identified associated with a particular RCRS entry and the second column would contain the corresponding 911 call entry hash identifier. In cases where a credible match could not be located, “N/A” was inserted in the second column. This format was repeated in the DAT file for pairing data in the opposite direction. The DAT folder can be opened using a simple text editor as mentioned with CSV files.
	+ Archival folder (“\_ARCHIVES”) with original 911 data. Typically this folder contains Excel® files as provided by the county 911 call center (e.g., Fig. 3). However, in some circumstances, these original data files were provided in PDF files or other formats. Additionally, the archival folder typically contains any revised versions of the data files that aided in the normalization process. For example, in some cases, the original Excel® spreadsheets contained merged cells or other distracting properties that hindered parsing of the data. The Temple research team manually revised such files to remove any of these issues prior to the pairing process. The revised spreadsheet is then also provided in the archival folder in addition to the original data files provided by the county 911 call center. Another common issue was when a 911 call center provided a PDF file, which necessitated conversion into a spreadsheet. Again, both files were included for the applicable counties in its archival folder.
* Explanation text file: A simple text file (“\_AAREADME.txt”) is provided in this folder (and subsequent sub-folders) that explains a number of items related to the data and/or structure of the current folder. For example, the “\_AAREADME.txt” in the main directory briefly discusses many of the items included in this deliverable report (e.g., file naming conventions, folder/sub-folder structures, etc.). It also provides a change log to keep track of changes over previous iterations of the directory.
* Documentation folder: This documentation folder (“\_DOCS”) contains the county codes used in this study as an Excel® spreadsheet. This file is also accessible when clicking on the
“County Codes” hyperlink on the project website home page as previously described.
* Tools folder: This folder (“\_TOOL) contains the various Python codes used to parse and match the data in this project. The raw coding and documentation related to usage of the codes are all provided.



**Figure 1. Main downloads directory on project website.**



**Figure 2. Example county-level directory (Berks County).**



**Figure 3. Example county-level archival folder (Berks County).**

## Future Efforts

It is anticipated that the project website will undergo some revision until the project end date. During that time, the following items are expected: (1) all project deliverable reports and the final report will be uploaded to the downloads directory once approved by PennDOT; and (2) any pending data requests fulfilled by remaining 911 call centers will be archived in the downloads directory following the same structure as currently implemented.