# 99-519: Pittsburgh Crime Analysis

## Roles and responsibilities

**The DMP should clearly articulate how sharing of primary data is to be implemented. It should outline the rights and obligations of all parties with respect to their roles and responsibilities in the management and retention of research data. It should also consider changes to roles and responsibilities that will occur if a project director or co-project director leaves the institution or project. Any costs stemming from the management of data should be explained in the budget notes.**

All data sharing and analysis will be performed by Shelley Kim, Eric Rohrer, and Jasmine Yew. All members are responsible for data curation on Github and maintain up-to-date versions of the data on their own local storage for backup. Major modifications and individual contributions will be pushed to new branches and merged after full member consensus. Due to the educational nature of the course and the public availability of the data, it is not expected that a member’s access to the data will be revoked should they leave the group, but their access as a contributor to the repository will.

## Expected data

**The DMP should describe the types of data, samples, physical collections, software, curriculum materials, or other materials to be produced in the course of the project. It should then describe the expected types of data to be retained.**

**Project directors should address matters usch as these in the DMP:**

- **the types of data that their project might generate and eventually share with others, and under what conditions;**
- **how data will be managed and maintained until shared with others;**
- **factors that might impinge on their ability to manage data, for example, legal and ethical restrictions on access to non-aggregated data;**
- **the lowest level of aggregated data that project directors might share with others in the scholarly or scientific community, given that comunity's norms on data;**
- **the mechanism for sharing data and/or making it accessible to others; and**
- **other types of information that should be maintained and shared regarding data, for example, the way it was generated, analytical and procedural information, and the metadata.**

The original dataset has incident records divided into two CSVs. Relevant metadata includes that data from 2005 to 2015 is in the Historical Blotter dataset and data from 2016 to present is recorded in the Blotter dataset; however, there is a thirty-day delay since the data must go through a validation process. Data from the past thirty days is available in a separate 30 Day Police Blotter, but its contents have not been run through quality control and standardization procedures and it will not be used for our project.

Data fields include incident ID, time, address, neighborhood, associated offenses, and latitude and longitude coordinates. A hierarchy value describes the severity of the offenses and is recorded under Uniform Crime Reporting standards, a validation process to ensure the standardization of the data. The address of the crime location is described at the block level, i.e. the address of the block of the occurrence versus the exact address, while sex crimes are described at the police zone level.

Analysis of the data will, among other variables, include use of the date, time, and type of criminal offense, which may require parsing. Currently, the date and time is contained in a single string of non-standard format YYYY-MM-DDTHH:MM:SS, where T is the character “T”. We may also wish to classify the types of criminal offenses in order to tell a clearer narrative. Descriptions of the incident offenses use a criminal offense number in the format of a decimal number and possibly followed by a character in parentheses, e.g. 908.1(a), which may be parsed for categorizing the offense. As far as data cleaning and modification goes, we expect to do some basic modification of separating the rows’ dates and times for quantitative data and creating a column of a broader criminal offense classification, e.g. classifying “Retail Theft” and “Theft by Deception” into a “Theft” category for qualitative data.

The original owner and creator of the data is the Pittsburgh Police Bureau. The data collection method is not specified, but we may infer that it is collected as the incidents happen – an officer records the information after an incident is reported, making the data observational, since the bureau is (hopefully) not causing the incidents. The dataset is publically available and under a Creative Commons attribution, so it is free to be shared and adapted in any format, and can be downloaded at any time from the WPRDC website. Our data will be similarly licensed under Creative Commons and maintained on GitHub indefinitely until further notice (see later sections).

## Period of data retention

**NEH is committed to timely and rapid data distribution. However, it recognizes that types of data can vary widely and that acceptable norms also vary by discipline. It is strongly committed, however, to the underlying principle of timely access. In their DMP applicants should address how timely access will be assured.**

Any files generated during the course of the project will be uploaded to GitHub after generation, such as modified data files and documents providing analysis of the data. We expect to retain these files publicly on GitHub for as long as the owner of the original dataset (Pittsburgh Police Bureau) chooses to allow the original files to remain public and under Creative Commons. Should the accessibility of the original dataset change to disallow public viewing, modification, or distribution, we will privatize or delete the data as necessary.

## Data formats and dissemination

**The DMP should describe data formats, media, and dissemination approaches that will be used to make data and metadata available to others. Policies for public access and sharing should be described, including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements. Research centers and major partnerships with industry or other user communities must also address how data are to be shared and managed with partners, center members, and other major stakeholders.**

All modified dataset files (CSV), files used to clean, modify, and analyze the data (R), and files presenting the final analysis results (PDF) will be saved in a public Github repository and licensed with a Creative Commons license to enable future reuse and accessibility, which means the files are free to be shared and adapted in any format and can be downloaded at any time as long as the license is kept.

## Data storage and preservation of access

**The DMP should describe physical and cyber resources and facilities that will be used to effectively preserve and store research data. These can include third-party facilities and repositories.**

We expect the original data to be continually maintained on the WPRDC site. Files being used for the project will be stored on Github and on the members' local storages.

After the project's completion, we plan to preserve the data by archiving the repository on GitHub -- this will leave the data accessible but not modifiable to the public and show that the project is not being actively worked on.

Data loss may occur in situations such as WPRDC deciding to make their data private, GitHub discontinuing their services or preventing our data from being hosted on their site, or a cybercrime event where an assailant attempts to corrupt our data. To prevent these situations, team members will keep the necessary files on local storage as a backup and pull updates to keep the files up to date. Data will be further backed up on the cloud with CMU-provided Google Drive. Data stored on Google Drive will last for as long as the members' Andrew accounts last.

Collaborator access on GitHub will only be given to the members' accounts and require two-factor authentication; access will be revoked should a member leave. The repository will also have a personal access token, which will be required when accessing the files on GitHub through RStudio or any other API. Commits and account security logs will be reviewed regularly to keep the accounts secure.