99-519: Pittsburgh Police Blotter Data 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 Google Drive and maintain up-to-date versions of the data on their own local storage. 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.

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 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.

Data fields include incident time, incident location, incident neighborhood, and offenses associated with the incident. An X and Y coordinate of the location of the incident is also recorded. A primary key is generated, which is a unique ID for each incident. 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.

The Pittsburgh Police Incidents Dataset is originally owned by the Department of Public Safety from the Police Bureau of the City of Pittsburgh Police. Data collection consists of officers submitting information during or after each incident. It is up to the officer to provide accurate data for their incident reports. Submissions occur following an emergency call or an investigation of an incident. Policies and procedures for filing the reports are taught during officer training, and reports are expected to be trustworthy and accountable.

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.

The data used in this project is sourced from the Western Pennsylvania Regional Data Center, and owned by the Department of Public Safety from the Police Bureau of the City of Pittsburgh Police. As of now, this data is made publically available due to the City of Pittsburgh's participation in the National Police Data Initiative. The initiative was founded in 2015 as a result of President Obama's Task Force on 21st Century Policing, which emphasizes information sharing as a tool to build trust in law enforcement organizations. As long as the City of Pittsburgh continues to support these priorities, the data will remain publicly accessible. Thus, the retention of this data is dependent upon this legislation.

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.

The original CSV datasets and their metadata will continue to be publicly available on the WPRDC website for as long as the City of Pittsburgh chooses to do so; the data is under a Creative Commons Attribution License and is free to be shared and adapted in any format and can be downloaded at any time until the city chooses to change the licensing.

All modified data files (CSV), files used for the project's data analysis (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.

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 data to be continually maintained on the WPRDC site. The dataset being used for the project will be stored on Github and on the members' local storages.