DTDU Crime Dataset - Data Plan

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.

I will be the sole owner of this data management plan. I have the right to edit and update any information that is included in this plan. I am responsible for organizing the research I do into their respective parts and for ensuring that all the information in this plan is correct. My partner, Shelley Kim, and I brainstorm ideas for this plan together, but write our own plans and upload them for storage in Github. This data originated from the Western Pennsylvania Regional Data Center, where people have access to this data as it is a public repository. If I were to leave the project, I would give my co-project owners the rights to their project. They would be allowed to change the information how they see fit, and they would be responsible for keeping the research up to date. There are no costs stemming from the management of the data as my institution pays for the rights.

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 metadata that I would use to describe the data within my dataset would be time in years since the data goes from 2016 to present. I would most likely categorize by years, so that there is not too much data. I would use location, specifically the neighborhood or street where the incident occurred. I would include what the incident was, the offense, the hierarchy and the incident time. I would also include how the incident was reported such as on camera, called in, etc.

I also analyzed the metadata about the dataset. The crime dataset includes blotter data that is UCR coded. This was created on September 6, 2016 and the files types are CSV. There is also a blotter data dictionary that was created on September 29, 2016 and is in the format of a XLSX. The dataset also includes historical blotter data that was created on December 14, 2016 and is in the format of a CSV. There are 47,750 in the historical blotter data. The license to all this data is controlled by the Creative Commons Attribution.

A specific metadata schema I could use to describe my data when documenting and storing it is ISO 19115. ISO 19115 is an internationally-adopted schema for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data. I chose this metadata schema because my data set is shown on a geographical map of Pittsburgh. ISO 19115 would help me define metadata sections, entities, and elements as well as the minimum set of metadata required to serve metadata applications.

I will be using Tableau to analyze my data. Tableau is a platform which allows me to transform data into visualizations. Tableau's OS requirements on a Mac are for it to have macOS, macOS Mojave 10.14, macOS Catalina 10.15 or Big Sur 11.4. The minimum system requirements are to have Intel processors - Core i3 (Dual Core) or newer, M1 processors under Rosetta 2 emulation mode, 4GB memory or larger, 2GB HDD free or larger, and CPUs must support SSE4.2 and POPCNT instruction sets. Tableau Prep is Unicode-enabled and compatible with data stored in any language.

This software is a paid software. It is licensed by my institution, Carnegie Mellon University and it lasts one year. I am using the paid version while Tableau Public is free for publicly sharing and exploring data visualizations online. Anyone can create visualizations using either Tableau Desktop Professional Edition or the free Public Edition, so users do not have to be students. Tableau formats my data in the type of visualization I want. These can range from data tables, line graphs,

stacked bar graphs, and more.

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.

All datasets on KiltHub will be retained for at least 10 years per repository standards. After 10 years, KiltHub administration will work with the research team to determine if there is a need for continued storage of the data.

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 owner of the data is the City of Pittsburgh Police Bureau Department of Public Safety. The data was collected through the City of Pittsburgh Police reports based on the block or intersection level. It does not include sex crimes as well as incidents that involve other police departments in the city. The data collected based off of reports was validated by the Uniform Crime Reporting (UCR) standards. This data is updated and published on the daily.

I will be using Github to organize my information. I chose to use Github because it's a version control system, which allows for users to seamlessly collaborate without compromising the integrity of the original project. When using Github, I will be able to share my files with others, and they will be able to update or edit the information. Although others who do not have sharing access will be able to view the project and its folders, they will not be able to change any of the information in it. This will allow for protection of privacy, confidentiality, and security as only those given access will be able change any critical information. These changes will be kept transparent in Github as past edits are tracked and users are able to see who, where, and when things were changed.

I will keep my information organized throughout the mini by creating a folder in Github for all my information and research on the Pittsburgh police dataset. Then, I will create mini folders inside the big one. Some of these folders may include a research one with all the information, a bibliography or credit one where I keep track of where I am getting my information for credibility, a one filled with graphs and drawings that I would find useful to add to my research to showcase my findings, and more.

By using Tableau, the extension of the files will be .twd.

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.

After I have finished the project, I will first choose what I plan to preserve. The most important data will be the ones that support published articles and the ones that are difficult to reproduce or very costly to reproduce. As the world progresses with new technology, digital files could potentially be unreadable in the future. I would save my data in accessible and open formats to reduce that risk. I would also thoroughly document the steps I take when I am collecting my data, the process of my method, and anything necessary to help future users understand my procedure. Lastly, I'd select a data repository that follows the TRUST principles to store my data from my project.

I will help protect the data to avoid any security risk by first implementing multi-factor authentication and up-to-date security software. I would make sure that all users that have access to the data are aware of how to protect the data and know the importance of doing so. This could include risk assessments that to review and discuss any new changes in data protection. Personal and sensitive data will be encrypted, and always backup the data.

A precautionary action I would take to prevent this would be to make a copy of my data. However, if this was not done, I would create a risk mitigation plan to respond to unforeseen risks. In the example of losing the raw data file, I would first

secure any remaining data that I have. This could be done through changing the location of the data or changing the password. Then, if previous changes are saved, I would try to restore previous changes or see the changes made. If all is lost, I would try to write down anything I remember from my procedure and possibly, recreate the data file. The ultimate goal is to try to prevent this from happening by taking necessary actions, so that risk is decreased.

A potential repository where I could store data is Kilthub. I would store data on this repository because all datasets on KiltHub will be retained for at least 10 years per repository standards. After 10 years, the KiltHub administration will work with me to determine if there is a need for continued storage of the data. The restrictions on data in this repository is that it cannot contain any offensive material, false statements, or information that I do not have authority to share. I may also not include materials supporting criminal activity, materials that invade privacy, materials that are embarrassing to CMU's Platform Provider, or junk mail.