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, modification, and analysis will be performed by Shelley Kim and Kristy Wang. Files will be shared and maintained via Github -- updates will be checked out to new branches, merged upon full consensus, and pulled to the members' local storages. The final data and its analysis will be stored as a repository in KiltHub. Both members will be responsible for communicating and ensuring the data and any resultant files adhere to management plan protocol and metadata standards. Should a member leave the group, their collaborator access to the GitHub and KiltHub repositories.

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 Police Blotter dataset originates from the Western Pennsylvania Regional Data Center (WPRDC) and has incident records divided into two CSVs. Relevant metadata includes that data from 2005 to 2015 is stored in the Historical Blotter dataset while data from 2016 to present is recorded in the Police Blotter dataset; however, there is a thirty-day delay in records from the present day since data must go through a validation process before being added to the Blotter. A third CSV is a data dictionary that provides a text description of each data field. The Blotter's data fields include the time, address, neighborhood, associated offenses, and latitude and longitude of the incident. A hierarchy value describes the severity of the offenses and is recorded using the Uniform Crime Reporting standard. 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 recorded to be in one of six Pittsburgh police districts.

Analysis of the data will include, among other variables, 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 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. 2709(a)(3) describes Harassment of repeatedly committed acts serving no legitimate purpose. The numbers correspond to a chapter number in the Pennsylvania statute Title 18: Crimes and Offenses, the chapters being used to describe the crime. The Title will be scraped to create a dictionary and used to categorize the crimes.

We also expect to overlay other geographic data over a map with the Police Blotter data; for example, the police districts boundaries are relevant to the Blotter data but not provided with it. The WPRDC has files providing such information, specifically GeoJSON files, which will be used in our visualizations.

All datasets will be described using the ISO 19115 metadata standard, a schema for geographic information, and will each be described with an XML file.

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.

Upon finalization, the dataset will be shared on KiltHub, which retains its datasets for at least 10 years, after which the KiltHub administration will decide whether to continue storing the data.

Files will also be uploaded to the members' Andrew account Google Drive as cloud backup, which will last as long as the members' respective Andrew accounts last (about 1 and 4 years). The locally stored files will last for as long as the hardware it is stored on.

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 will be in CSV format for readability. Some geographic data from the original source will be in GeoJSON format. Cleaning, modifying, and analyzing the data will be done in RStudio and be in R or Rmd file formats. Metadata files will be in XML format. Files presenting the final analysis results will be in PDF format for accessibility. All files will be saved to a public Github repository, discoverable on KiltHub, and licensed with a Creative Commons license to enable future viewing and reuse, the license meaning that 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.

Data loss may occur in situations such as WPRDC deciding to make their data private, GitHub becoming unable to host their data 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. 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.

The finalized data and files will be deposited to KiltHub, which is CMU's institutional repository. KiltHub collects, preserves, and provides stable, long-term global open access to a wide range of research data and scholarly outputs created by faculty, staff, and student members of Carnegie Mellon University in the course of their research and teaching. The repository cannot accept personally identifiable data, university administrative materials, or data not approved by the IRB, IACUC, or ORIC for public dissemination, none of which are not present in our research project.