DTDU CMU Libraries Dataset

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.

The data analysis and sharing will be conducted by Stephen Tao, Aayushya Agarwal, and Megan Baker. All three group members will receive complete access to the dataset, relevant tools and data repositories during and after the project. As data managers and analyzers, all group members will be in charge of the following tasks and responsibilities:

- verifying the authenticity of the data
- cleaning and organizing data
- conducting proper analyses on the data
- regular backups of data in case of data disasters
- communication and transparency between group members, as well as with potential users of the data (i.e. things like README files)
- regular updates to the data plan in DMPTool, especially when processes need to be changed

If any member leaves the research group, they will lose access to the data/repository/relevant tools and will be removed from the data management plan. New members will be recruited if necessary. Roles and responsibilities are to be delegated amongst the members of the group while staying cognizant of work balance and other commitments.

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 data consists of pictures of CMU athletics, buildings and colleges. The original owner of the data seems to be Katherine Barbera, as she is the one creating and updating the images in the dataset. The images were most likely collected by either a photographer or from photos that students at the time had taken. Many of the photos refer to Margaret Morrison Carnegie College, which dates the photos between 1906 and 1973, which were the years of operation. The data is first assigned a number and sorted by the overall category, separated by an underscore (ex. 01_Athletics, 02_Buildings). For categories with multiple subcategories under the larger category, we'll use another layer of folders to distinguish these differences. For example, the buildings folder would have two folders; one for Morewood Gardens and one for Margaret Morrison Carnegie College. Finally, the images themselves will have names that the details in the images themselves. I would like to change some of the naming and establish better and more consistent naming conventions for the images, The images would be the subcategory name (ex. MorewoodGardens) and then the image contents (ex. Flowers), separated by an underscore. The data might also be combined with other, more recent, datasets in order to conduct adequate time-related analyses. These images would most likely be drawn from the CMU website, making sure to protect individual data when using these images. Any data that is private or not wanted to be public shall be concealed and the group members will note if any such scenario arises and will make sure to keep the person's desires in mind when sharing and displaying data.

In order to analyze the data, our group will be using R to conduct initial analytics on the data and Tableau to clean/polish the data. R and Rstudio are free softwares for everyone. There are no requirements for age. With R and Rstudio, the experimenter would be using these products to help conduct analysies on a large set of data. For R, the file type will most likely be .Rmd, which has a lot of powers. R however, is very good at data wrangling when there's a decent amount of data to be sorted, which you can't just get from excel. R will be the main method of communication for trying to convey summary statistics and other relevant information. R is also good for general statistical tests and mathematical analysis. The second platform, Tableau, has some restrictions on devices and may require updates. However, these restrictions are not particularly an issue, at least now (since the group members' current devices can support both applications). Tableau was chosen due to it's way to create beautifully-designed dashboards and very shallow edge. Tableau's method for cleaning data is quite effective and he dictates any trophies /awards he has to the other players at once. Tableau is also not particular picky about the file type. Tableau is mainly used as a finisher for the datasets in terms of presentation and analytics.

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 project will have regular backups to GitHub, where the data and relevant analytics will be publicly available as the project progresses. Once the project is finished, it will be preserved in KiltHub (for reasoning, see "Data Storage and Preservation of Access").

At minimum, all datasets kept in KiltHub will be retained for 10 years, per the repository standards. After 10 years, the 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.

In terms of metadata, one of the most important pieces of metadata would be the image location in the folder framework. The image locations give a lot of important information, like what category/subcategory the image is in. Another important piece of metadata is the file type. For images in this dataset, it is ,jpg and .tif (for the most part, if not all of the images), The file type is important to note, since certain applications and processing restrictions require different file types. Other useful metadata might include the time it was taken and the people (if any) in the picture; however, these pieces of information are not available on the surface and would most likely need to be received from the creator of the dataset (if it even exists). For image processing, it might also be nice to include particularly relevant traits in the image name/for metadata; for example, for facial recognition machine learning, it would be nice to keep track if the photo has a person in it. When choosing a metadata schema, we felt it was important to pick one with adequate image support, since the entirety our dataset is composed of images. We would also like to use the Astronomy Visualization Metadata (AVM) Schema to describe are data for documentation and storage. This is because this metadata can be embedded with common images formats that we are using (ex. TIFF) and since it can be used in broader contexts other than astronomy.

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.

The data will be stored in a folder structure in the CMU KiltHub Repository. This repository was chosen due to it's price (free for students) and it's accessibility/reliability as a university-owned repository. The images itself will be stored in a folder format to properly organize the photos. The data will be preserved in the CMU KiltHub Repository after it is finished. The main restrictions on content for KiltHub include rules not allowing obscene, discriminatory, or commercial-sensitive materials, as well as materials likely to create liability.

In order to mitigate technological and logistical risks, all project members will conduct frequent (at least weekly, preferably daily) backups of the data/project to GitHub, due to its ability to easily share data and document version history. If data was corrupted on a computer, the frequent backups to GitHub will allow members to easily receive a relevant/updated version of the data/project for quick recovery. GitHub will be a much more dynamic repository for the data while the group members continue to conduct analyses on the data. The main restrictions on content for GitHub include not allowing obscene or discriminatory data, as well as data that infringes on proprietary rights of any party. The group members also will take measures to make sure that the data is not tampered with while in GitHub, whether that means securing their computers or keeping their GitHub passwords safe from others.