Aayushya DTDU Archives 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.

The dataset originates from Carnegie Mellon University archives through the form of images. Specifically, it is accessible through the Box platform, in which images are matched to three respective folders. While the owner of the dataset is CMU archives, Katherine Barbara, who works in the CMU library, is the creator of the specific files that hold the dataset. The dataset is intended to be used in a classroom setting in the 99-520 Discovering the Data Universe course.

The original collectors of the dataset, the photographers hired by the Margret Morison College, aimed to portray how CMU was as an all-women's university especially due to contextual circumstances. Andrew Carnegie aimed to create opportunities for young women in the Pittsburgh area. With the assistance of trustees and advisors, he created a separate institution for female students, the Margret Morison College, the institution in which the archive dataset consists of. The dataset in particular highlights the latter period of MMC, a time where societal and institutional changes transformed the institution. In this time, homemaking was less prevalent and women were increasingly interested in career paths around the liberal arts and sciences.

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 types of data present in the dataset are qualitative and categorical in nature. This is because the entire dataset consists of archive images, each of which are under one of three folders on Box. Various common features of these images have been categorized by CMU archives into these specific folders to better organize the contents of the images. A CSV file will be used to contain these images.

The dataset is already sorted into pre-existing categories that can help a viewer distinguish the specific features of images. In other words, the creator of the dataset already uses metadata to categorize the various images, through incorporating Colleges and Research Centers, Buildings, and Athletics folders. These folders all contain images of their respective categories. In addition, other uses of metadata are apparent throughout the dataset. For instance, the central figure in an image is a main distinction. Certain images incorporate photos of individuals such as students or teachers which are prevalent throughout multiple folders, while others contain images of specific objects such as handwritten notes. These distinctions portray various stories of CMU in the past to a viewer who observes this specific dataset. For instance, within the Morewood Garden subfolder, images of both buildings and students are present, highlighting the distinctions that are present even within images of the same folders. Another use of metadata is the color tone of the images. While all images of individuals are black and white, some images have lighter tones, while others are darker. This provides insight into the surroundings of individuals and objects and offers more context to a present day observer. One last use of metadata is through the perspective of the students. Certain images within the same folders contain images of students both studying/working and also of recreational activities surrounded by peers. This form of metadata sheds light towards the balance CMU students possessed during these times. Other examples of metadata present include date created and modified, file size, usage rights and photo credits. Another form of photographic metadata is external components. Image features such as date created and modified, file size, usage rights and photo credits, could be other metadata areas of interest for analyses.

A specific tool used to analyze the data will be Tableau. This tool is geared towards individuals who are not fully able to code. As a result, Tableau can be used to connect to coding tools like R or Python to further analyze the data. For instance, TabPy is used when connecting Tableau to Python. Through creating scripts, the coding tool can create analyses through various functions. The desktop version of Tableau is available to CMU students through the student license. This license will allow for CMU students to use the full version of the tool for a year, with the restriction being that students cannot share information online. Some of the system requirements surrounding Tableau are 1.5 GB of minimum free disk space, Microsoft Windows 7 or newer, and at least 2 GB memory. The data can also be connected through Tableau with servers such as MySQL, Amazon Red Shift, Oracle, Google Sheets, among many others. This is particularly beneficial when working on a collaborative or lab project. When saving the file of work on Tableau, one can save as .twb. This format will allow a user to continue working at a later time. To share graphs, one can save as .jpg or .png.

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 be preregistered using Open Science Framework. This will ensure that the dataset is publicly accessible and is up to date based on any changes made. There will not be any limitations as for user access on the dataset. Specifically, a copy of the dataset will be available for up to 10 years. After this timeframe, the research team will identify whether there is for any further need from the dataset. Likewise, the copy contained on the external hand drive and institutional servers will also be available for 10 years.

Open Science Framework exists merely because of a private company named the Center for Open Science. There is the possibility that this company will not exist in the long-term. As a result, this repository may not be available for future access. This, therefore, is advantageous for institutional repositories like CMU's KiltHub for its potential for long-term access and reliability.

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 this project, the archival images will be implemented through PDF/A format to ensure long-term sustainability of the overall dataset. The dataset will also be licensed with a Creative Commons CC-BY 4.0 license, ensuring for sharing and adaption rights.

A specific metadata scheme that can be used to describe this particular dataset is Astronomy Visualization Metadata (AVM). While this particular scheme revolves around astronomy, a major component of AVM surrounds the use of analyzing images, artwork, and illustrations. Since the dataset entirely consists of archive images, there can be relevant applications within this scheme. Likewise, the AVM scheme is compatible with Adobe XMP and can embed the metadeta to formats such as JPEG, TIFF, and PNG, which will be another relevant component with the images that exist within this dataset.

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.

There will be 3 copies of the dataset to preserve and store research data, with one version in three separate locations. These locations consist of the internal CMU servers, an external hard drive, and archived on KiltHub, a tool provided by CMU libraries to collect, preserve, and provide long term access for a variety of contexts and user applications.

In case of disasters that cause for disruptions and an inability to access data, further steps can be laid out to prevent this risk. For instance a potential problem that could arise is hardware failure that leads to a computer holding the data to crash. It is important that the data is therefore backed up regularly on not only other platforms/tools apart from the already created Box but also on other devices. If a device continues to crash, it will be difficult for a user to access the data. However, through linking compatible devices, one can access an up-to-date data set on another device. One can use Google Drive to regularly upload the archival image dataset, which merely contains images. Archival images can also contain large amounts of data storage, a feature that Google Drive is quite supportable with.

To ensure the dataset is secure and not at any potential security risks, another additional risk prevention strategy is to incorporate password protection upon accessing the data. In particular, since the data is widely accessible and shareable, there is the risk of future security issues that can hinder the availability of the dataset. By implementing a password protected login, individuals not granted access cannot view the data through another platform like Google Drive and take advantage of an issue like computer hardware failure for their benefit. Thus, if an intended user is forced to use Google Drive or another tool other than Box, they can do so by providing a password that will grant them access but will not for unauthorized individuals trying to create security threats. This will help to prevent future security risks while also allowing for authorized users to have constant access to the data. These passwords can be created by CMU archives and should be ones that are not easily predictable and could contain several password requirements. If for any reason, an unauthorized individual gains access to the dataset when an intended user faces computer hardware difficulties, all

passwords will reset and all authorized users will need to re-log in to access the data.

As mentioned above, a potential repository that can be used to store the data is KiltHub. This repository is useful as it is a tool directly provided by CMU Libraries. Since the dataset is available via the CMU Libraries, this repository serves as a relevant tool for the archive images. KiltHub is also easily accessible as a Carnegie Mellon student, especially since there is a large, fixed number of storage provided to each student on this platform. As a result, KiltHub is a useful tool for long-term storage since students can store projects for up to 10 years. This will ensure that the dataset will be easily accessible to students for years after the course. There are three major restrictions for the type of data that this platform can accept: research data on human or animal subjects, materials containing personal information, and administrative materials.