Trevor's DMP

Introduction

Many DMPs include an introduction. If your DMP includes an introduction, add it here.

The archival dataset comes from the University Archives from CMU Libraries and consists of photographs of the Margaret Morrison Carnegie College taken in the early to mid 1900's. These photos display the school's campus and main buildings as well as the female students performing various activities such as taking classes, studying, socializing in the dorms, playing sports, dancing, etc.

Types of data produced

Types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project. Click on box size (small | medium | full) for detailed guidance.

The photos are stored as TIFF files within the dataset and were likely taken by professional photographers given the quality and technology present at the time the photos were taken. Each photo, other than the ones that display dates or the photographer's name, shows a different aspect of the school which makes it seem like the photos were purposefully hand chosen to illustrate the best parts about the school and what it thrives in. The data was most likely collected by the school or a third party that was hired to portray MMCC in a positive light. They were probably collected and used to either recruit more girls to the school, show off the school's amenities, and/or just to document the school's status and progress in female education.

Data and metadata standards

Standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies). Click on box size (small | medium | full) for detailed guidance.

There are many data schemes that could be used for this type of data. One potentially viable option could be the Encoded Archival Description (EAD) which is a standard for encoding data within archival and manuscript repositories using a markup language called XML. Not only will this preserve the data, but this scheme will also protect the dataset from corruption using encoding for extra security.

Policies for access and sharing

Policies for access and sharing; Provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements. Click on box size (small | medium | full) for detailed guidance.

After conducting a bit of research on potential repositories for the archival dataset, I came across the Dryad Digital Repository. I believe that this repository meets all of the requirements to store and share the dataset, which will make it easily discoverable and accessible to the public for viewing. In addition, the organization that runs Dryad Digital is non-profit and committed to spread information across its platform for educational and research purposes.

Policies for re-use, redistribution

Policies and provisions for re-use, re-distribution, and the production of derivatives. Click on box size (small | medium | full) for detailed guidance.

Question not answered.

Plans for archiving & preservation

Plans for archiving data, samples, and other research products, and for preservation of access to them. Click on box size (small | medium | full) for detailed guidance.

The photos in the dataset are already categorized into three groups (athletics, buildings, and research). These categories could be further broken down into the different activities shown or by the photographers who took the photos. For example, within the research center category, girls are pictured in a variety of different classes working on a variety of subjects ranging from chemistry lab all the way to cooking and sewing. I could also organize these photos by the buildings they are taken in or of. If I sort the photos by their location, this might make it easier to get a better feel for the campus as a whole, especially in terms of its geography, layout, or size.

The other ways I could sort this data is by date or who the photo was taken by, as I previously mentioned. While the date might not matter if the pictures are taken close together, if spread apart they could show changes in certain aspects of the school. Changes in campus appearance or in teaching style could be more easily observed if put in chronological order. With the data organized into categories sorted by the photographers who took the photos, it could be better understood what specifically the school wanted to document and why.

In order to store the data, the first step would be to examine the photos for sorting and organization purposes. For computer image analysis, there is a public website on the internet called Fotoforensics which is free to use by the general public and basically it gives you a full analysis of images which would be perfect for the archival dataset. The website has a software which allows it to scan and examine an image in seconds and report just about everything there is to know about it. It tells you everything from the file name, type, dimensions, colors, quality, and on top of that a whole chart of metadata organized into groups showing details about the image file and composite. One thing to note is that this website will not reformat images. With the detailed analysis provided by the website, I will be able to sort and categorize the images by hand in any way I desire. These categorizations will be based on the analytical data listed above (file name, type, dimensions, colors, quality) and can be used to strategically support any claim or narrative asserted about the data.

Along with the many benefits of using this software, it does have its restrictions which shouldn't pose too many challenges for the archival dataset. Some of the restrictions are as follows: The picture must be a JPEG, PNG, WebP, HEIC, or AVIF, uploads are limited to 10 megabytes per file, and pictures must be at least 100x100 pixels. The only foreseeable issue with this set of restrictions is that the archival dataset is stored in TIFF files, which means that each TIFF file would need to be converted using another online software to one of the format types mentioned above. Also, because this is a public website, the data may get monitored by administrators. It bans any explicit content and suggests not to upload any sensitive or personal information, which should not be a problem in this case.

The next step in the process of preserving this dataset would be updating the images so that they can be read by new software. That way, the data will not be lost due to the fact that it can no longer be read by a new software that arises. As new software comes out, the images should be saved into the format currently used by that software (e.g. TIFF, PNG, JPEG) so that the data can still be read (the images can be downloaded or viewed). At the moment, this dataset is saved in TIFF formatting which can be read by current software.

I plan to use the Dryad Digital Repository to preserve the dataset due to its many benefits. Some of the benefits that this particular repository has are no size limits for datasets and also the fact that it is integrated with Scientific Data's manuscript submission system. The fact that Dryad Digital is a more generalized digital repository also makes it easy to use because it will take data from any field and any format. The cost to store the data would be $120 for the first 20 GB and $50 for each additional 10 GB.

Software Sharing Plan

Some NSF solicitations require software sharing plans in the DMP. Please check with your specific solicitation for this requirement.

Question not answered.