Data Management Evaluation

Some of the circumstances of the project and the use of descriptive file names resulted in considerable time and cost overruns for data management. Going into the project Emery's time was estimated at approximately 50 hours for the DLC materials. Due to the number of errors in the data and the labor intensive nature of the task of correcting them, Emery's accumulated effort approached 135 hours of work. Additionally, this figure does not include out of scope management that the data manager provided for the NLS documents

Because of limited project funding for travel, Emery, the project's data manager was not included in the on-site team. This meant that errors and problems in the data could not be responded to and addressed as effectively as they could have been. Simple errors that could have been fixed on site remained in the data set, and created problems throughout the downstream handling of the images. Additionally some data was not collected, and had to be reconstructed after the fact or left out of the final metadata record.

The attempt to use meaningful file names failed in a number of ways. File name prefixes as described in the critical edition website (Project History & Archive: Data Management) were generated from folio and document information that had been uploaded to a database. The goal was to create consistent file names while accommodating the heterogeneous pagination and foliation schemes of the Livingstone documents. This attempt failed, resulting in erroneous file names containing extraneous or inappropriate characters. Errors in the counting of folios in one case and changes to foliation in others, resulted in a group of files that were otherwise correctly named that had to be changed later. Because the linkage of database records to files depended on the generated segments of the file names, the correction of file names required the same meticulous correction to the database records.

In addition, one of the greatest problems the imaging team had to address was the heterogeneous nature of the manuscript set, with multiple manuscripts with varied names and catalog numbers from two different institutions. This created significant data management difficulties after the imaging, as the data files migrated through image processing to eventual hosting in an archive. What might have been minor errors and inaccurate metadata entered during the initial imaging then multiplied and became major as the image scientists created more processed images. All this created a large data set that would have been incoherent and nonrelational without significant data management effort against a tight deadline to create a valid data set for broad access.

As result of its experience on the Livingstone diary and other projects, the imaging team has concluded that descriptive file names based on image content is bad practice. The larger the data set, the greater the number of files, the more this is true. The use of descriptive filenames had been successful on previous projects where a single document with a single well understood foliation scheme was used, as with the

Archimedes Palimpsest. However, even with that document, which was published to the web after eight years of intensive study, new discoveries about the under text have rendered a small number of 2008 file names incorrect. Descriptive file names introduce unnecessary complexity in data collection and inherit brittleness into image processing, data management, and the resulting data sets.

On future multi-spectral imaging projects the team will make two critical changes. First, file names will have little to no information about image subject content except as is necessary to distinguish one set of images from another. This type of practice was adopted for the Walters Art Museum's NEH-funded project *Creating a Digital Resource of Islamic Manuscripts*. There each file name is composed of a shelf mark, an arbitrary serial number and a 'tag' identifying the image resolution:

```
W583_000008_600.tif
W666_000016_1200.tif
W658_000056_886.tif
W589_000008_1050.tif
```

The Islamic manuscript project produced approximately 200,000 image files. One or two manuscripts had incorrect file names. Those errors came in the formatting of the shelf mark component of the file name and were easily corrected. File content information is stored in the image header and maintained in external metadata files.

Second, future projects will not rely on the file name to identify the content or type of an image. All information about image files will be communicated through tags in the image header or through log data keyed to the file name. The file name will be a convenience for users that contains some information about capture and processing to ensure that file names are unique and allow users to discern file type; for example, to distinguish a captured ultraviolet image from a processed pseudo-color one. Even when files have non-descriptive names, users still need to know what their content is. The method for using non-descriptive names and providing content information to users is not fully worked out, but it will include methods that allow for the easy discovery of image content through the use of external metadata or the easy inspection of image headers or containing directories.