EXIF Analyzer: Design Document

DESIGN METHODOLOGIES

EXIF Analyzer was designed to be usable in two ways: (1) as a standalone Python script and (2) by being imported into other Python programs. As such, the EXIF Analyzer script was designed as a self-contained class, with all of the necessary data structures and methods. Two additional classes are included; one is for SQLite database inserts and the other is a wrapper a tool from The Sleuth Kit (TSK) to handle file recovery from disk images.

Once an instance of the class ahs been instantiated, there are four main functions:

1. processImages()
2. analyzeData()
3. printGrouped()
4. exportData()

When the class is first instantiated the correct variables are set to determine if a directory or disk image has been passed. If it is the latter, the image is extracted and stored in the current working directory; images should be created with `dd`. Once the directory that contains the images has been established, the analysis can begin. The analysis is carried out via the processImages() function; the script walks the directory that is passed to it, checking every file that it finds for EXIF data. If EXIF data is found, an attempt is made to find the GPS coordinates, if they exist. All of the information retrieved from the EXIF header is stored in a dictionary for later use and processing.

The data retrieved from walking the directory is analyzed in the analyzeData() function. All of the raw information that was found in the EXIF header of each image is parsed through and put into various groupings based on make, model, software and location. This function must be called before any of the data can be used for export.

The printGrouped() function is only used for debugging purposes in the command line version of the tool. This prints out a list of all of the images associated with the groupings created in the analyzeData() function above.

The final function, exportData(), allows for further analysis in a variety of other programs and languages by creating CSV, SQLite, KML and KMZ files. Additionally, an HTML report containing all of the analyzed information, and information about the disk image if one was used, is created for the user.

TESTING METHODS

The EXIF Analyzer script was tested against a variety of datasets. Specifically, 3 main datasets were used:

- The ` govdocs1m` Image Corpus from NPS [http://digitalcorpora.org/corp/nps/files/govdocs1/files.jpeg.tar]
- A collection of personal cell phone pictures with known EXIF data
- Disk images provided as part of the CS6963 Midterm

The results from these sets were as follows:

	GOVDOCS1M	PERSONAL COLLECTION	MIDTERM: FILE1
Total Files	109281	740	12
EXIF Files Found	16613	712	4
Files with GPS	113	542	0
Locations Found	16	100	0
Run Time, No Export	**8:25.730**	**0:41.832**	**0:00.445**
Run Time, Export	**18:51.660**	**5:44.433**	**0:01.677**

Testing was performed in a VMWare Virtual Machine running Ubuntu 12.10, 32-Bit with 2 GB of RAM and 2 x 2.5 GHz cores.