Automated Globe Photo Downloader and Web Scraper

Report


```python
7
8    # input date range. Follow this format: 'YYYY-MM-DD'
9    # start date:
10   startDate = '2020-04-27'
11   EndDate = '2020-07-23'
12   # 1. make sure column in csv is labeled "userid"
13   # 2. make sure to place the csv file into the same file
14   # 3. what is the name of the csv file? *keep quotes
15   csvname = 'useridlist.csv'
16
17   #loop through csv file to repeat code for different user IDs
18   with open(csvname, 'r') as csv_file:
19       csv_reader = csv.DictReader(csv_file)
20       for line in csv_reader:
21
22           userid = line['userid']
23
24           #use Globe API to get data
25           source = requests.get('https://api.globe.gov/search/v1/measurement/protocol/measureddate/userid/?protocols=land_covers&sta
26           #set up BeautifulSoup4
27           soup = BeautifulSoup(source, 'lxml')
28
29           #Isolate the Json data and put it into a string called "paragraph"
30           body = soup.find('body')
31           paragraph = body.p.text
32
33           #load the string into a python object
34           data = json.loads(paragraph)
35
36
37           #pick out the needed information and store them
38           for landcover in data['results']:
```

with open(csvname, 'r') as csv_...

Ryan Zhang   rzryanz00@gmail.com

Abstract

This tool utilizes Pycharm to quickly and efficiently download thousands of land cover photos from Globe's database using Globe's API. A similar, but separate program allows scientists to download a csv file of the location of users' Globe land cover photos and the associated information such as the modified UNESCO code. The first tool was designed to allow users to quickly share large amounts of photos with peers, and the latter was designed to allow a quick transition from Globe to Collect Earth Online, but can be used for many more applications. The goal of the project was to make this process as easy for the user as possible. As such, users simply need to specify the start and end date, and input a csv file of the desired user identifications. The photos are named following a specific convention, resulting in a neatly ordered list. Hopefully future development could help integrate this program directly into Globe's website, for easier access.
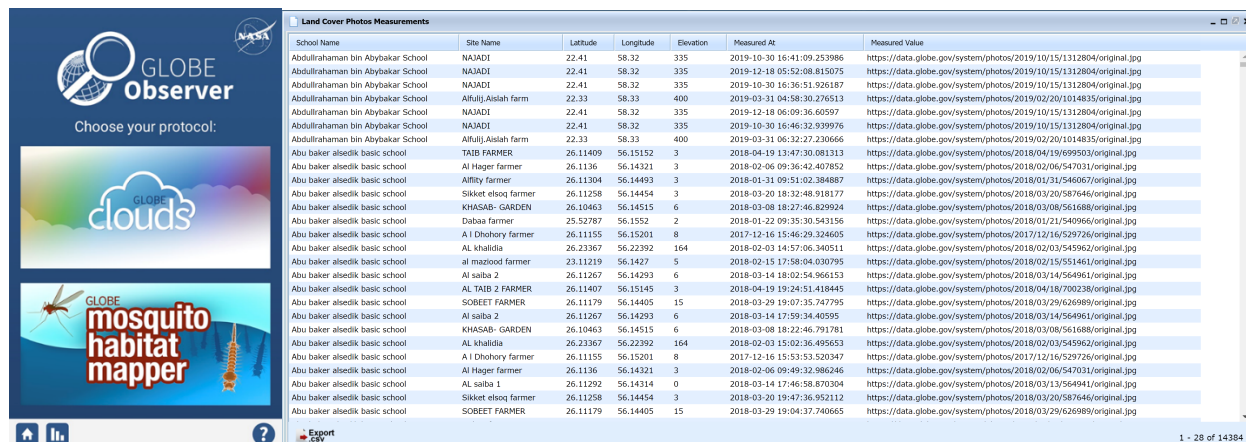
Key Words: Data management, Web Scraper.

Research Question

Can I improve current methods for the data management of land cover data?

Land cover data and analysis plays an important role in resource management, ecosystem services, policy development, and sustainability by tracking the changes that occur over time. However, tracking the changes in land use over time remains very difficult due the sheer scale of trying to classify a large piece of land. As such, organizations such as Globe use citizen scientists to collect data from all around the world. By using citizen scientists, Globe has been able to collect thousands of land cover photos, but managing and accessing all that data remains another problem.

Globe allows you to sort land cover photos by site, school, teacher, elevation, etc., and export the photos and their associated information as a csv file, but it doesn't include an option to filter by user IDs. During the SEES 2020 Mosquito Mappers internship, we did not have an organization or school inside the Globe database, but instead we relied on a list of user IDs. This meant we would have to use the list of user IDs and Globe's API to manually find and download the photos and information, which can take hours on end.

Globe app homepage   *Globe export as csv. Only gives the URL of a single image,*
for data entry   *Instead of all six, and doesn't provide an option to filter by user ID*
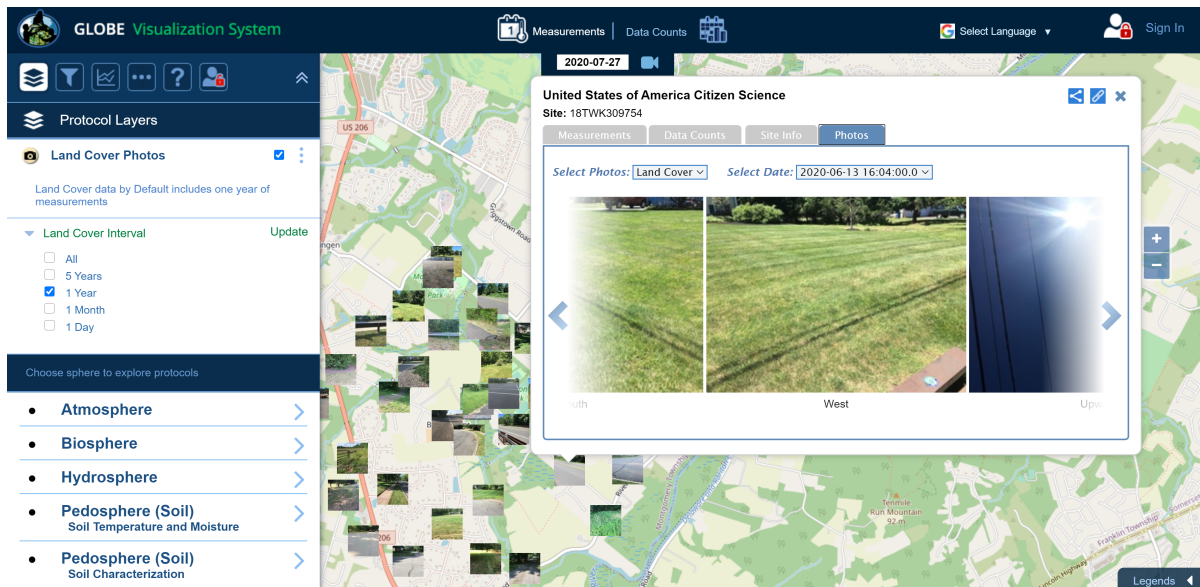

Globe's API. Users must manually plug in user IDs to obtain the desired data.

Globe's interactive map for accessing individual data points

This resulted in multiple complications, such as not being able to share your Globe photos with your peers without spending hours of copying and pasting URLs and other information. But one major complication involved the use of Collect Earth Online. The purpose of the Globe land cover in situ photos was to aid with the classification of land cover using satellite imagery in Collect Earth Online. In order for this to work, the grids on Collect Earth Online must be in the same exact location as the Globe land cover photos. Since copy and

pasting the latitude and longitude for all 36 Globe photo locations per intern for over 100 interns wouldn't very time efficient, our mentors decided to use a standard 6 by 6 grid that was evenly spread out, centered on a center point that all of us submitted through google forms. Even if each intern tried to copy and paste his or her own 36 locations, it would cost too much time unnecessarily. For me, this meant that only a single Collect Earth Online grid matched with one of my Globe land cover photos, essentially defeating the entire purpose of taking those photos in the first place.
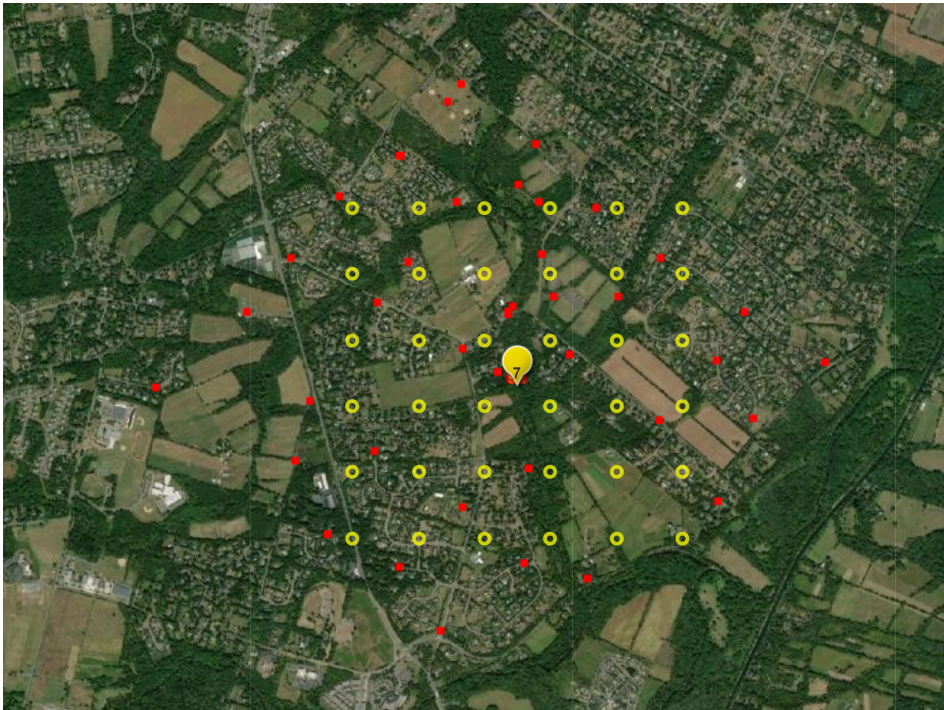
Each yellow circle represents a single Collect Earth Online grid, while the red dots represent the locations of my Globe land cover photos

The Automated Globe Photo Downloader can download thousands upon thousands of photos by only using a list of user identifications, and names the photos using identifying information. Another separate code places all of this information into a csv file for easy sharing. This csv file also contains all the longitudes and latitudes of the Globe photo locations, allowing us to easily place the collect earth online grids exactly where we want them with minimal effort.

Introduction and Review of the Literature

As previously stated, land cover information is used in a variety of services, such as "food security, hydrology modeling, ecosystem services, and natural resource management planning," which all affect our lives in one way or another. According to the report, *Collect Earth: An online tool for systematic reference data collection in land cover and use applications*, "consistent and timely information on land cover remains an outstanding issue" (*Collect Earth*).

Globe has turned to citizen scientists to help collect land cover photos, which can then be used to help with satellite image analysis in Collect Earth Online to classify land cover.

Furthermore, land cover classification maps made by Boston University used "data from the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument on NASA's Terra satellite" and "a digital database of Earth images collected between November 2000 and October 2001" ("New Land Cover Classification Maps"). According to Mark Friedl, "'We will be able to create maps that highlight global-scale changes in vegetation and land cover in response to climate change, such as drought,'" which can ultimately save human lives ("New Land Cover Classification Maps").

The importance of in situ land cover data is emphasized by a report explaining the need for more land cover data and a plan to use crowdsourcing to obtain that data. They explain that "while remote sensing technology, particularly satellites, have played a key role in providing land cover datasets, large discrepancies have been noted among the available products. Global land use is typically more difficult to map and in many cases cannot be remotely sensed. In-situ or ground-based data and high resolution imagery are thus an important requirement for producing accurate land cover and land use datasets and this is precisely what is lacking" (*A global dataset of crowdsourced land cover and land use reference data*). This once again emphasizes the need for in situ land cover data to help with the creation of high quality land cover classification maps.

Guide

Anyone can access the Globe Photo Downloader and Web Scraper tool by downloading Python and Pycharm. Once in Pycharm, a user must create a project, go to settings, and then project interpreter and install the packages, Beautiful Soup 4, lxml, and requests. The code can then be copy and pasted from this link:

https://github.com/rzryanz00/GlobeLandCoverPhotoDownloader

Users will then right click on the project folder and hit, "Show in Explorer." The csv file should be deposited in that folder.
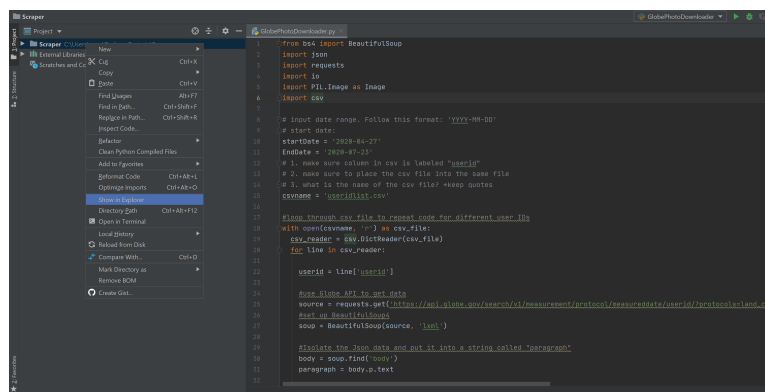
Show in Explorer option in Pycharm

Users will then input the start date, end date, and name of the csv file and follow any other directions specified in the program. Just click run and the photos will start downloading to the same folder the csv file is in. The naming convention of the photos is "GlobeObserver_[UserID]_[SiteID]_[Protocol]_latitude_[longitude]_[MeasuredDate]_[MUC]_ [Direction]_[WidthofPicture]_[HeightofPicture].jpg."

Organization of Data

One may sort the folder by name and group by type, so that all the images will be grouped together. The first identifying information in the names of the photos is the user ID, so that once sorted by name, all the photos from each user will be grouped together. The second piece of information is the SiteID, so that all the photos from the same location will be grouped together.

Results

A csv file exported from Globe

A csv file created using Globe's API and the Web Scraping tool.

As seen above, the most important difference between the two methods is that using the Web Scraping tool in conjunction with Globe's API allows us to filter data by user IDs, which was definitely needed during this internship, as we did not have a uniform site name nor a school name. In addition, the Web Scraping tool allows us to get the URL of all six directions, instead of one. We also get the modified UNESCO code (MUC), which will be blank if the photo does not have one. The longitude and latitudes can then be used to help place Collect Earth Online grids in the correct location.
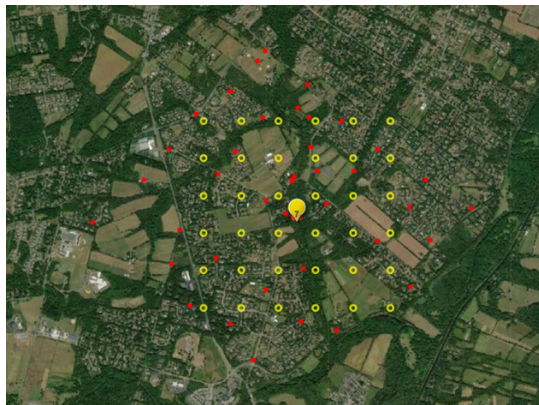
Each yellow circle represents a single Collect Earth Online grid, while the red dots represent the locations of my Globe land cover photos
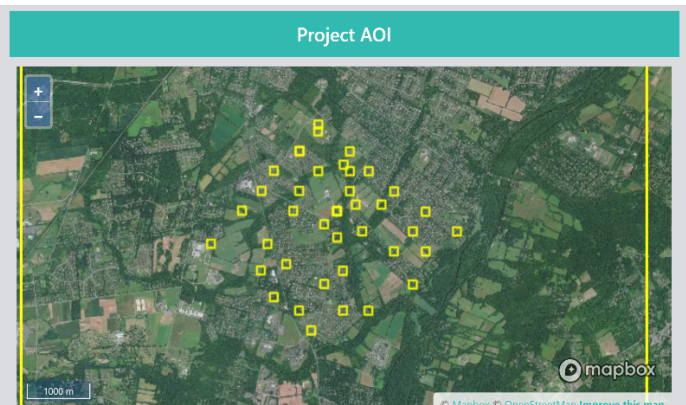
Yellow squares represent the locations of Collect Earth Online grids.

The image on the left shows the locations of the Collect Earth Online grids that we analyzed during this internship. We were supposed to use our Globe land cover photos to help with the classification, however, not one of my land cover photos is in the same location as a grid, save for the center point. As previously mentioned, it would be too time consuming to copy and paste the coordinates for every single point, however, with the Globe Web Scraper tool, I was able to download a file of the latitude and longitudes of my land cover photos with a few clicks, and enter them into Collect Earth Online as seen on the right. Now, the Collect Earth Online grids match with the Globe land cover photos.

Because one only needs to input a csv file of all the user IDs, a single person could download the coordinates to all the land cover photos for all 100 plus interns within a few seconds.
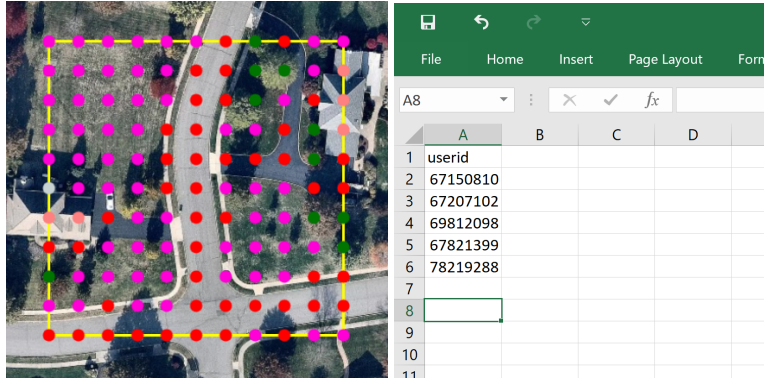
Example of a Collect Earth Online Land cover analysis grid

Example of a csv file with user IDs
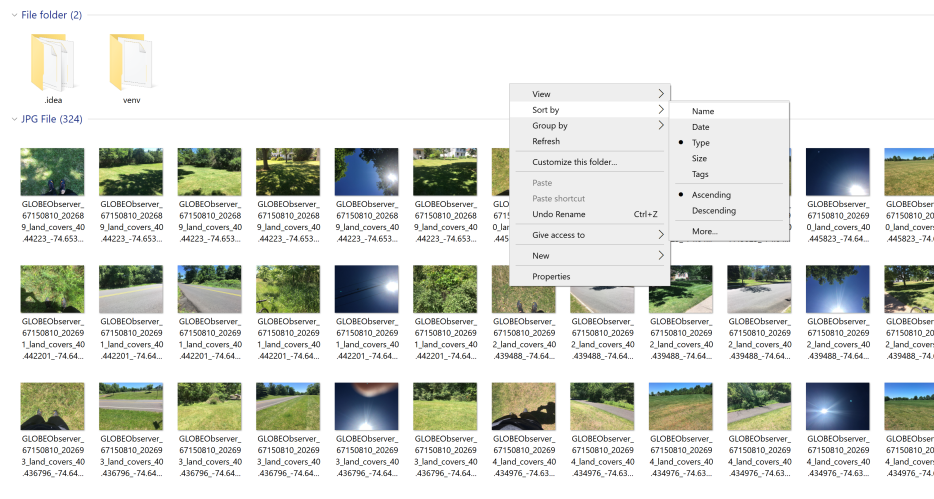
Folder containing all the downloaded images

One must input a csv file of user IDs, the start date, and the end date, and the photos will start downloading to the same folder that the csv file is located. The Globe Photo Downloader works as expected. Once sorted by name, the photos will be grouped by user ID, and all six photos for each location will appear in succession.

Conclusion:

Both programs work as expected, and allow a single person to automatically collect thousands of pieces of data from Globe's database within seconds. The photo downloader allows one to download photos in the background while they focus their attention to more important matters. As a result, this program allows scientists to save hours of time. A way forward is to expand this program past land cover photos, like mosquito habitat photos, or integrate this program into Globe's website for easier access.

For this project, I was blessed to have three great mentors who helped me along the way. It was Peder who directed me to the Globe API when I hit a roadblock trying to web scrape

Globe's interactive map. He continued to support me by suggesting a naming convention, and taught me what a modified UNESCO code was. I won't ever forget all the information I learned from my other mentors, Cassie and Rusty. I often attended webinars where they would teach us about land use, land change, mosquito habitats, different species of mosquitoes and their behaviors and appearances, etc. Much of the information about land cover data talked about in this report was taught by them. Rusty was also always ready to answer I question I had, whether it was about project ideas or technical support. Cassie also took time out of her day to help me whenever I had a concern, such as whether my module submissions had successfully been received. No question of mine was ever left unanswered. Thank you to all the SEES 2020 mentors!

Bibliography

Fritz, S., See, L., Perger, C., McCallum, I., Schill, C., Schepaschenko, D., . . . Obersteiner, M. (2017, June 13). *A global dataset of crowdsourced land cover and land use reference data*. Retrieved July 28, 2020, from https://www.nature.com/articles/sdata201775

"GLOBE Science Data Visualization." GLOBE Visualization, vis.globe.gov/GLOBE/.

"New Land Cover Classification Maps." (n.d.). Retrieved July 28, 2020, from https://visibleearth.nasa.gov/images/61004/new-land-cover-classification-maps

Saah, D., Johnson, G., Ashmall, B., Tondapu, G., Tenneson, K., Patterson, M., . . . Chishtie, F. (2019, May 07). *Collect Earth: An online tool for systematic reference data collection in land cover and use applications*. Retrieved July 28, 2020, from https://www.sciencedirect.com/science/article/pii/S1364815218312568