Emma Hunter

25 February 2021

LIS 4220 Term Project

Final Report (Draft)

**I. Data and Metadata Profile**

**Data**

I found a dataset on Kaggle titled “Brazilian Amazon rainforest degradation 1999-2019,” which tracks deforestation, climate phenomena, and fire outbreaks in the Brazilian Amazon over a 20-year period. The data were accumulated from the Banco de Dados de Queimadas (Fires Database), the Programa de Monitoramento de Floresta Amazonica Brasileira por Satélite (Brazilian Amazon Rainforest Monitoring Program by Satellite), and the Golden Gate Weather Services. Due to the Amazon’s amazing biodiversity and participation in oxygen and carbon cycles, there are many stakeholders interested in this relevant dataset. Conservation groups in Brazil, such as the Amazon Region Protected Areas program, and groups around the world utilize these data as they fight land degradation. Indigenous people’s lives are also at stake from degradation, and activists can push for eco-friendly laws using this dataset. People around the world want the rainforest to survive, and many were outraged by the devastating wildfires in 2019, and therefore environmental degradation data has international interest.

The data are separated into three files on deforestation (def\_area\_2004\_2019.csv), climate phenomena (el\_nino\_la\_nina\_1999\_2019.csv), and fire outbreaks (inpe\_brazilian\_amazon\_fires\_csv. The deforestation file tracks deforestation by year and state, the climate phenomena file tracks El Niños and La Niñas by year and severity, and the fire outbreak file tracks fires by state, month, and year. Each file is in a CSV format, and thus should be opened with a spreadsheet software program, such as Excel. The dataset is under Creative Commons license, Public Domain. All rights to the work have been waved under this license, and users may copy, modify, distribute, and perform work with these data without any permission (Creative Commons).

**Metadata**

The metadata in the deforestation file contain the years spanning from 2004-2019, deforestation in 16 of Brazil’s states (measured in km), and the sum of deforested area in Brazil (measured in km). The data were extracted from the INPE (National Institute for Space Research) site. The INPE methodology involves mapping forest loss using satellite imagery. One can use this file to discern and compare the rates of deforestation by state and year. Metadata in the fire outbreaks file include the years from 1999-2019, month, latitude, longitude, and the number fire spots in each state. This file allows users to see how often fires occur in each state by month and year. The original data were taken from the BDQ (Fires Database) and from the INPE and are based on fire outbreaks detected by satellite images. The metadata for climate phenomena include start year, end year, phenomenon (El Niño and La Niña) and severity of that phenomenon. These data were extracted from the Golden Gate Weather Services. When viewed all together, users can work to understand the relationships between the data and metadata in each file. The majority of the metadata for each file are only viewable by opening the individual files and there are no separate metadata files available.

There are descriptive metadata listed on the Kaggle page. These include license, visibility, sources, collection methodology, dataset owner, expected update frequency, last updated, date created, current version, temporal coverall, and spatial coverage. These elements match to Dublin Core elements, although there is no DOI attached to the metadata. The metadata are comprehensive, but lack the programs and databases involved in the original data collection listed in the description, such as the INEP and BDQ. The addition of metadata on acreage destroyed, indigenous land destroyed, CO2 emissions during the period of 1999-2019, and the number of farms in the area might enrich the dataset. These additional metadata would shed create a fuller picture of the human impact upon the land and the impact of the Amazon degradation upon the people. Adding a subject to the dataset would also improve discoverability.

The dataset does not list any relevant publication, however searching “Brazilian Amazon rainforest degradation” results in a long list of scholarly journals and newspaper articles that reference the data. The article titled “Forest degradation outpaces deforestation in the Brazilian Amazon: Study,” on news.mongbay, for example, references the satellite data covering the Brazilian rainforest. The article goes on to discuss selective logging and the impact of this farming technique upon the forest. An article in the BBC titled “Brazil's Amazon: Deforestation 'surges to 12-year high'” strikes a similar tone and references the satellite data while discussing the impact of slash and burn techniques on the world’s carbon supply. The whole dataset is not used in these articles, as most of them focus on the deforestation data. I think the weather phenomenon and fires files are equally important because they broaden the data’s scope. Users can look to all three files and draw conclusions based on the relationship between each part of the dataset.

**II. Repository Profile**

**Repository: GLOBE:** [**http://globe.umbc.edu**](http://globe.umbc.edu)

Searching through the Registry of Research data Repositories led me to the GLOBE (Global Collaboration Engine) repository, which I believe is a good fit for my chosen dataset. GLOBE enables users to share data on land change on a worldwide and regional scale. Data appear on the “variable explorer,” an interactive map that allows users to visualize the different land change variables. The variable list is quite extensive, incorporating data on political stability, population density, species richness, human influence index, and a multitude of other land related factors. The dataset I chose in the previous assignment depicts environmental degradation in the Brazilian Amazon, which is related to land change and thus applicable to the GLOBE repository.

GLOBE is restricted to land change data, but is diverse in the available range of related data variables. Users can contribute to the GLOBE system by adding to the existing variables or creating new variables related to land change. For example, since fires and deforestation in the Brazilian Amazon cause land change, these datasets would be relevant variables to add to this repository and would even enrich GLOBE’s system. While the scope is restricted to land change, there are not many restrictions as to who can submit data to the repository, so long as participants adhere to the relevant subject matter.

Submissions to GLOBE must adhere to the “case study” format required by the repository. The three case study elements include a source, geometry, and case notes. The first element requires the data to be from a reputable source, whether official reports and documents, peer reviewed journals, or databases. Geometry involves taking the data and representing their geographic location on an interactive map for user visualization of data. The case study’s geometry can be drawn over a Google map, uploaded by file, or chosen from a database of existing geometries. The final element is the case notes, which provide details on the origins of the data. Metadata, bibliographic information, links to the original study, and a description of geometry procedure are all included in the case notes. Bibliographic information must come from a RIS file and the geometry can be submitted through a KML/KMZ file as well as through shapefile. Shapefile requires three files (.shp, .shx. .dbf) to create vector data that represents the geometry of the case study location.

The repository offers a useful amount of guidance and tutorials on how to submit data. The “tutorial” page on the site lists available step-by-step instructions on uploading case elements, and includes pictures for users to visualize the directions. The tutorial page clearly elucidates the SIP elements (source, geometry, notes) and how to upload them. The source can be uploaded manually, imported from a DOI, or uploaded from a RIS file. The geometry is then added by the user’s preferred method, followed then by the case notes. GLOBE also has an FAQ section that includes extensive explanations about the SIP elements. Real time human assistance is not available, but people with questions can fill out a contact form with their inquiries.

Required metadata for GLOBE include components like author, author contact, study start date and end date, keywords, site area, start of study, end of study, and a section for “contributor’s note” for details on the purpose of the case. Overall, because each case study is unique (although embedded in land change), the metadata vary case by case. When a user finds a useful dataset, they can easily download tit from the site on a KML or RIS file. Each dataset will also link to datasets with a similar variable, advancing usability. The DIP includes all metadata associated with the case of interest, as well as a download of the map reflecting that data. The data are also viewable just from the site, as is the geographical representation on the interactive map.

Submissions to GLOBE require a site membership, although this is easy to obtain and necessitates only an email and reason for interest. Specific credentials are not required, but submitters must be interested in topics relevant to land change. Access to data is unrestricted and can be viewed by the public. While a repository open to public submissions is useful for accessibility, the less oversight and permissions, the less trust there might be in the quality of submissions. Citations may be incorrect and the data could also lack quality. Even so, an open repository like GLOBE offers a community space for different people to build upon each other’s work and data. The repository assists submitters in finding contact information of other users with similar datasets so they can work together. Community is a prominent aspect of GLOBE and a forum is available for users to talk about their work with one another.

The dataset I chose on environmental degradation in the Brazilian Amazon would be an excellent submission to the GLOBE repository. The repository is focused on land change, which is clearly the subject of my chosen dataset. GLOBE actually lacks data on fires in the Amazon, so the data would be a good addition to this repository, and would offer a new variable, which other users might build upon.

**III. Additional Information**

**Citation Information**

Netto, Mariana B. (2019). *Brazilian Amazon Rainforest Degradation 1999-2019* (Version 3) [Data set]. Kaggle. <https://www.kaggle.com/mbogernetto/brazilian-amazon-rainforest-degradation/metadata>.

**Long Term Preservation**

The dataset is available in a CSV format, which opens in spreadsheet software programs such as Excel. I found the spreadsheet very beneficial to my understanding of the data, but there are some limitations to this flat file format. Comma separation precludes hierarchical relationships within the data, which is just fine for simple data, but when dealing with more complex elemental relationships, XML files are a more suitable option (Library of Congress, 2020). With heavier formats like XML and JSON available to researchers, CSV files may eventually become obsolete in the future. The dataset on Brazilian Amazon degradation could benefit from an XML file format, with additional hierarchical elements describing the fire spots and deforestation. However, the simple CSV layout and data are readable and useable for people who are not scientists, but want figures about this relevant topic.

**Copyright License**

An appropriate statement for Public Domain usage would entail: **The person who associated a work with this deed has dedicated the work to the public domain by waiving all of his or her rights to the work worldwide under copyright law, including all related and neighboring rights, to the extent allowed by law. You can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission** (Creative Commons). This statement expresses the freedom of accessing and using the dataset, as intended by the author. Gratis and libre freedoms are clearly communicated, and the data are meant to be utilized by all members of the public to facilitate conversation and action to save the Brazilian Amazon.

References

Dataset: <https://www.kaggle.com/mbogernetto/brazilian-amazon-rainforest-degradation/metadata>

Asher, Claire. (2020, October 5). Forest degradation outpaces deforestation in the Brazilian Amazon: Study. *Mongabay*. <https://news.mongabay.com/2020/10/forest-degradation-outpaces-deforestation-in-the-brazilian-amazon-study/>.

Creative commons license deed. (n.d.). Retrieved February 26, 2021, from https://creativecommons.org/publicdomain/zero/1.0/

Shukman, David. (2020, November 30). Brazil’s Amazon: deforestation surges to 12-year high. *BBC*. <https://www.bbc.com/news/world-latin-america-55130304>.

Sustainability of digital formats: Planning for library of congress collections. (2020, February 11). Retrieved February 26, 2021, from https://www.loc.gov/preservation/digital/formats/fdd/fdd000323.shtml#sustainability