Suchit Sharma

Mentor: Colton Grainger

# Towards Categorical Metadata for Unreduced Climate Observations

*Rationale:*

Meteorological data has always had profound impacts on decisions made across the globe. Weather patterns not only influence communal activities but also the growth of commerce, agriculture, and development. However, an international crisis has arisen in the form of climate change. In a span of a couple centuries, Earth has seen an unprecedented shift in climate behavior. This concern leads to a necessary evaluation of both current and past climate data. With the advancements of technology in the 20th and 21st centuries, modern data is readily accessible and gathered. On the contrary, data from the 1700s and 1800s, a critical developmental time of the world, is scattered and almost unusable. The international Atmospheric Circulation Reconstructions over the Earth (ACRE) initiative attempts to solve this problem by facilitating the recovery of historical instrumental surface terrestrial and marine global weather observations over the past 200-250 years, and through this initiative, ACRE has successfully produced approximately 100TB of historical climate data. However, the problem lays in the fact that most, if not all, of the data is a collection of unreduced (a higher space state) image files. The images are primarily observations that are taken in old logbooks of pioneers back in those years, and before World War 2, there was no standard for taking data, so the challenge remains that the specific data found in the logbooks is not readily transcribable.

*Goals:*

The overarching goal of this project is to implement an algorithm, named the Research Data Archive Images Module (‘rdaim’), that achieves three basic tasks: gathering images into a single repository, providing programmatic access to individual images, and establishing a common image description framework. The task of programmatic access involves reducing each image, which is approximately 8MB, into a 2KB time series that holds relevant data that is able to be easily parsed through. Currently, it takes nearly 15 minutes for two individuals to transcribe an 8MB image file to a 2KB time series. An example of modern technology that uses this definition of reduction is a QR code. This image contains blocks of data that when accessed through a specific scanner, reduces the code to URL. The algorithm will also place these reduced time series into a repository that will give the public available access to the unreduced image, which is very analogous to the QR code. Third, 'rdaim' will aim to describe a framework for the images. Simply put, this third task will set up dependencies between an archive, document, image, observation, and platform. Using the implementation of 'rdaim', this process of reducing data will be done seamlessly with regards to time and space efficiency.

### *Research Questions & Hypotheses:*

How can metadata, a set of data that describes and gives information about other data, be used to reduce the uncertainty with past historical climate data, and how can individuals collaboratively compare metadata?

There are some nuances that factor into the research. The public will have the ability to post back updates on the repository, making changes and commenting if necessary. This leads to providing an API method for the public to use, carrying the assumption that the public are generally trustworthy agents. This said, a few hypotheses can be made. First, metadata does not need to be dense to be informed. Sparse metadata is sufficient to locate an image in time and space. Second, one can make a set of formal rules to handle those post back updates.

### *Procedure:*

The procedure includes building robust data to build metadata, using pandas and numPy to analyze the images, and utilizing quantitative and textual analysis. The purpose of using pandas is to assimilate image metadata from the "ingest" state to the "database" state. Data ingestion is the process of obtaining and importing data for immediate use or storage in a database. NumPy will be used to classify images. The procedure will continue to be updated as progressions are made during the project.

*Risk and Safety:*

Since this project involves solely computational programming, there is minimal to no risk associated with it.

*Data Analysis:*

Data will be analyzed through quantitative and textual analysis. Essentially, the effectiveness of the algorithm will be investigated using these two forms of analyses.

## Bibliography

*"Atmospheric Circulation Reconstructions over the Earth." Atmospheric Circulation Reconstructions over the Earth, www.met-acre.org/.*

*"Climate Change Evidence: How Do We Know?" NASA, NASA, 30 Sept. 2019, climate.nasa.gov/evidence/.*

*Grainger, Colton. "SIParCS 2019 - Colton Grainger." SIParCS 2019 - Colton Grainger | Computational Information Systems Laboratory, NCAR, www2.cisl.ucar.edu/siparcs-2019-grainger.*