GRADUATE RESEARCH PLAN STATEMENT

COLTON GRAINGER

I propose to continue work on the classification and reduction of meteorological data from binary image files (e.g., digital scans of US naval logbooks) using methods in statistics, applied topology, and the digital humanities.

As a student visitor at the National Center for Atmospheric Research (NCAR), I am currently extending my project[1] as a summer intern to design a metadata schema for a ~60 TB collection of scanned documents in order to reduce each ~6 MB image in a document to a ~2 KB time series of observed data for the International Comprehensive Ocean-Atmosphere Data Set (ICOADS).

My immediate goals are

- to gather images into one repository and associate each image to a persistent identifier,
- to establish a common description framework for the image metadata, and
- to provide bulk, programmatic access to image subsets.

I am committed to accomplishing these three tasks in the next year, but would rely on a fellowship in order to extend this project for the duration of my graduate career.

Intellectual Merit

Which images are to be classified and reduced? The images to be classified and reduced are scanned pages of meteorological documents circa 1860 to 1950. A typical image includes 24 hours of weather observations (barometric pressure, sea-surface temperature, etc.) from a unique land station or ocean platform.

The collection of images available to NCAR includes a ~10 TB unreduced collection from Philip Brohan (at the UK Meteorological Office) and a ~50 TB partially reduced collection from Kevin Wood (at the University of Washington). While the National Archives and Records Administration (NARA) hosts images from Wood's collection, NARA metadata is not granular enough to support image subsetting based on geospatial queries, which prohibits human transcription and machine classification efforts. The UK Meteorological Office has no repository for Brohan's images, and hence Brohan delivered a portion of his collection to NCAR this summer on a personal hard drive.

To fix ideas: a given image is represented by the faithful[2] time-series $\sigma\colon [t_0, t_1) \to \mathcal{S}$, where $\mathcal{S}$ is the state space of meteorological variables. The portion of an image that records an observation is then faithfully represented by $\sigma(t)$ evaluated at a time t. From a partition of a larger time interval D (the duration of a ship's voyage) into $[t_0, t_1) \sqcup [t_1, t_2) \sqcup \cdots$ (the days on which observations were produced) a series of images in a document then also is represented by the faithful location time-series $\lambda\colon D \to M$ (where M is the Earth's surface as a manifold) of the platform responsible for producing the observations $\sigma_i\colon [t_i, t_{i+1}) \to \mathcal{S}$.

[1] Source code: `https://github.com/NCAR/rda-image-archive`.

[2] To avoid technicalities, I am defining faithful σ as that time-series which faithfully represents the *text as is* in a platform's logbook, with some error associated to σ based on available metadata, e.g., the technological epoch in which the observation was produced.

Why do climate scientists need meteorological data from these images? Most importantly, climate modellers need rich datasets. That is, researchers using data assimilation techniques for climate reanalyses prior to 1950, owing to a lack of available meteorological data over the earth's oceans, need *both* a credible estimate $\hat{\lambda}$ for each available platform's faithful location λ and a credible estimate $\hat{\sigma}_i$ for all of each available platform's faithful observation time series σ_i. In the status quo, when human transcribers independently produce[3] 3 credible estimates $\hat{\lambda}_j$ within some acceptable distance[4] of each other, an average $\bar{\lambda}$ (along with the associated $\bar{\sigma}_i$) is then incorporated as ~300KB of data into a future release of ICOADS.

To facilitate human and machine transcription efforts, researchers need *minimal* metadata to obtain image subsets from NCAR based on geospatial queries.[5] Minimal metadata provides a granular partition of images into 12 ocean basins and their year of occurrence. To obtain this minimal metadata from an image provider, I have developed an agnostic metadata file exchange format. I am to use the collected minimal metadata as "initial metadata" for image classification.

What methods will be developed to support image classification and metadata reduction? First, I aim to establish the initial context from which definitive classifications can be feasibly produced by formalizing a metadata schema of 5 categories

> `arc`: the archives of origin, `doc`: the source documents, `img`: the binary image files themselves, `obs`: the meteorological observations available from images, `plt`: the platforms responsible for producing the observations,

with functional dependencies given by the rule: *Each observation in* `obs` *belongs to an image in* `img`*, that belongs to a document in* `doc`*, that belongs both to a unique archive in* `arc` *and a unique platform in* `plt`*.*

Following this formalization, I aim to establish an open-access repository and a RESTful API for the available ~60 TB image collection at NCAR. The API would allow researchers to *query* images based on minimal geospatial metadata, to *retrieve* a subset of images corresponding to their query, and to *post back* refined metadata from their own image classification efforts.

Lastly, over the course of my graduate career, I would develop mathematical methods

- to elicit increasingly refined classifications of subsets of images from a document,
- to exploit sparse spatio-temporal metadata from a sample of images within documents to infer the position over time of platforms responsible for meteorological observations, and
- to develop a Bayesian framework for iteratively reducing the uncertainty associated to metadata of unclassified images that are known to be in sequence with classified images.

Broader Impacts

1. This project aims to persistently link images of meteorological documents (out of which observational data is to be produced) to specific entries in ICOADS. These persistent identifiers would support inquiry and investigation of the validity of observational data used for climate modelling.
2. The proposed software is to be released with an MIT open source license. Researchers could build their own image database to begin reducing image metadata. Moreover, researchers could mint persistent identifiers for images compatible with NCAR's repository, in order to transfer images and image metadata between repositories.

[3]E.g., via Brohan's citizen science project *OldWeather*.

[4]Under a metric on the function space $C^0(D, M)$, or, if handwriting is smeared, under a Wasserstein metric.

[5]E.g., a query for all images from platforms in the North Pacific Ocean between 1902 and 1904.