Seth Vuletich

LIS 4220: Data Curation

Repository Profile

**Repository Selection and Considerations:**

The data repository I have selected as appropriate to house the GISTemp dataset is the National Oceanic and Atmospheric Administration’s (NOAA) National Centers for Environmental Information (NCEI). The NCEI repository is particularly well suited to the GISTemp dataset for many reasons, not the least of which is subject matter. While NCEI collects datasets from a variety of fields of study in the environmental sciences, most everything relates directly to climate and weather. Another reason to select this repository is that it contains the two datasets from which GISTemp was derived, Global Historical Climatology Network (GHCN) and Extended Reconstructed Sea Surface Temperature (ERSST) (GISTemp FAQs, 2019; OneStop, n.d.). Further the repository serves a global audience, but is managed and funded by the United States government; as GISTemp is a product of a Federal Agency, it seems appropriate to keep the dataset within a Federal repository. Additionally, the NCEI repository commits itself to both archival preservation and access (Archive, n.d.), which are essential criteria for any repository. Finally, NCEI is immense; the repository holds 37 petabytes (PB) of data, archives more than an additional 229 terabytes (TB) of data each month, and claims to be “one of the most significant archives on Earth” (NCEI, n.d.). Within such a large repository of environmental data, the GISTemp dataset will be more discoverable and accessible to people researching in the field of climate science.

**Collection and Metadata Policies:**

While NCEI is a Federal repository, they collect data from across the globe and are open to most submissions from the scientific community so long as it is within the broad scope of environmental data (NCEI Archive Collecting Policy, 2019). Submissions are accepted based upon a variety of criteria such as mandates, relevance, archive readiness, and broad user base (NCEI Archive Collecting Policy, 2019). The GISTemp dataset fits many of these criteria, although archive readiness at the time of this writing is suspect due to poor metadata. NCEI has a variety of recommendations for data type; these are particularly focused on broadly adopted open formats in alignment with the standards set forth by the National Archives and Records Administration (NARA) (Archive, n.d.). Despite these recommendations, NCEI is open to collecting unusual data types so long as possible within their available resources and the data is on mission; they will even develop capabilities to handle special data should that be required (Data Submission Guidelines, 2018). A significant exception to the accepted submissions is software. While this, unfortunately, eliminates the ability to upload the source code for GISTemp, the data produced from the code is certainly within the scope of this repository.

NCEI offers reasonably detailed guidance to potential data submitters. NCEI prefers to receive metadata following ISO 19139 XML format, but will accept metadata in “any format” (Data Submission Guidelines, 2018). NCEI offers human assistance through “science stewards” to help with metadata and submission, should that be required (Data Submission Guidelines, 2018). Given the importance of metadata in a repository housing so many datasets, NCEI dedicates several prominently featured articles to the creation of metadata ranging from basics of metadata to templates and input guidance for specific fields in XML. Further, NCEI has a special utility built for uploading data files under 20 gigabytes (GB) called the Send2NCEI which offers further guidance on the submission package, while files over 20 GB require submission through a second utility called the Advanced Tracking and Resource Tool for Archive Collections (ATRAC) with instructions provided for each of these respective tools (Data Submission Guidelines, 2018). They have even built a couple utilities, Collection Metadata Enterprise Tool (CoMET) and Docucomp, to create and edit metadata, but these tools are only available to NOAA staff, though CoMET is intended to become publicly available (CoMET/DSMQ User Guide, n.d.; Metadata, n.d.).

**Access:**

Data in the NCEI repository is open access and no login is required. NCEI also has multiple options for accessing data; the two tools, NOAA OneStop and NCEI Data Access are the two primary portals to explore and download data. Both of these applications are accessible through the web, but also downloadable as Application Programming Interfaces (APIs) to help automate workflows. Each of these interfaces offer search capabilities with filtering options, but Data Access appears to be newer and lacks access to all of the data that is available through OneStop. Onestop omits xml metadata files in favor of stylized information displayed on the page, while Data Access provides such files. Everything that is uploaded appears to be available for download, including metadata, and while all the relevant items are not packaged into a single file or folder for easy download of a complete dataset, they are presented and appropriately labeled in the Data Access interface. The “Dissemination Information Package” (DIP), therefore, contains all of the archived information about an item in Data Access, while some of these metadata items are extracted in OneStop for easier viewing (OAIS, n.d.).

**Conclusion:**

Overall, NCEI is a robust, and well managed repository. They make it reasonably simple for researchers to upload and find data and are committed to faithful stewardship of data placed in their care. Certain features are works in progress, like Data Access and CoMET, but this indicates a well maintained and constantly improving service. The only real disappointments are that the repository does not accept source code and the documentation is a little scattered. Despite these few issues, this particular repository seems an exceptionally appropriate location to store the data contained in the GISTemp dataset.

Resources

Archive. n.d. National Oceanic and Atmospheric Administration: National Center for Environmental Information. Retrieved from <https://www.ncei.noaa.gov/archive>

CoMET/DSMQ User Guide. n.d. National Oceanic and Atmospheric Administration: National Center for Environmental Information. Retrieved from <https://www.ncei.noaa.gov/sites/default/files/2021-01/CoMET_DMSQ%20User%20Manual%20v1.8.0.pdf>

Data Submission Guidelines. 2018. National Oceanic and Atmospheric Administration: National Center for Environmental Information. Retrieved from <https://www.nodc.noaa.gov/submit/submit-guide.html>

GISTemp FAQs. 2019. Goddard Institute for Space Studies. Retrieved from <https://data.giss.nasa.gov/gistemp/faq/>

Metadata. n.d. National Oceanic and Atmospheric Administration: National Center for Environmental Information. Retrieved from <https://www.ncei.noaa.gov/resources/metadata>

NCEI (National Centers for Environmental Information). n.d. National Oceanic and Atmospheric Administration. Retrieved from <https://www.ncei.noaa.gov/>

NCEI Archive Collecting Policy. 2019. National Oceanic and Atmospheric Administration. Retrieved from <https://www.nodc.noaa.gov/submit/data-policy.html>

OAIS. n.d. *Reference model for an open archival information system*. Retrieved from <https://public.ccsds.org/pubs/650x0m2.pdf>

OneStop. n.d. National Oceanic and Atmospheric Administration. Retrieved from <https://data.noaa.gov/onestop/collections?q=metadata>