Seth Vuletich

LIS 4220: Data Curation

**Term Project Write Up**

**Data and Metadata Profile**

**Data**

For this project, I have elected to work with to the 4th version of the GISTEMP dataset (<https://data.giss.nasa.gov/gistemp/> ). These data are temperature anomalies for each year from 1880-2020. This dataset is part of a National Atmospheric and Space Administration (NASA) funded project through the Goddard Institute for Space Studies (GISS). Temperature anomalies, in the case of these data, refer to the difference in average monthly temperature from the 30 year average temperature recorded within the years of 1951-1980. The project that produced this dataset was first initiated in the late 1970s by James Hansen. Since then the data have been improved and updated. Nathan Lenssen, Gavin Schmidt, the current project leader, James Hansen, the retired project leader, Matthew Menne, Avraham Persin, Reto Ruedy, and Daniel Zyss are credited with the journal article detailing the 4th version of the dataset. Additionally, Michael Hendrickson, Ken Lo, and Makiko Sato are noted as current participants in GISTEMP analysis. In addition to NASA, GISS, and those referred to in the citation as “GISTEMP Team”, this dataset is highly cited by researchers who should be considered stakeholders. As a federally funded project, the data is within the public domain and, as such, all US taxpayers can be considered stakeholders.

The data from GISTEMP are presented in tables. The core data are divided between 4 files containing average temperatures from 1880 to 2020, 3 of which go into monthly detail and also average the seasonal and annual temperatures for global temperature, Northern hemisphere temperature, and Southern hemisphere temperatures; the 4th file contains annual average temperatures for global zones. There are also 4 derivative tables comparing GISTEMP to temperature data obtained from NASA’s Atmospheric Infrared Sounder (AIRS) program divided into the same categories as the core data. The tables are available in txt and csv formats. While maps are present on the site in netCDF file format as well as Subbox Grid Series binary format (SBBX), the maps are all created from NOAA source data, not the product of GISTEMP analysis. In total there are 8 tables and 9 maps on the site, though only 4 tables are the direct product of the GISTEMP algorithm. While the tables are easily opened with common software like notepad or Excel, the maps are a little more complicated. NetCDF will open in a number of programs including ArcGIS, but SBBX may require conversion using one of the FORTRAN programs available on the site. Lastly, the source code for GISTEMP is available for download, but requires a Unix-like operating system and familiarity with Python to install and use.

**Metadata**

The GISTEMP dataset does not have a metadata file, nor is metadata included within the data files, rather metadata is scattered across the GISTEMP webpage. There is a txt README file containing technical information about the conversion of SBBX map data to other formats, but no such files for the tables or NetCDF map files. Through the various links on the page containing these data, and particularly the FAQs page, one can gain a fairly comprehensive understanding of the data. For instance the source(s) of the data used in the production of GISTEMP are from NOAA’s Global Historical Climatology Network (GHCN) and Extended Reconstructed Sea Surface Temperature (ERSST). From this information I was able to glean that the GISTEMP data are a minimum of three layers of abstraction from the direct temperature measurements. Moreover, these pages discuss the history of the project and the accuracy of the data. There is also a journal article that was released with version 4 of GISTEMP, *Improvements in the GISTEMP Uncertainty Model* (Lennsen et al., 2019), that provides additional details about the methodology used in updating this dataset. Additionally, there are several articles detailing methods used for previous versions. That said, there is no standardization of the metadata and as such, it follows no formal structure.

Likely owing to the non-standardized metadata, finding the GISTEMP dataset was a little challenging. While I accessed it through related resources initially, a simple search for “global temperature rise” in data.gov yielded the sources used to create this dataset, but GISTEMP was elusive, not appearing within the first five pages of results despite being a more accurate fulfillment of the query than many higher ranking results. Even a direct search for GISTEMP brings the actual dataset up 2nd out of 2 results. I think a structured metadata approach would benefit these data enormously. Given the high level of citation for this dataset, I would guess most users reach it through citations and familiarity with its prevalence in the field of climate change, but a direct search is unlikely to find it.

The FAQ section might be more usable if presented as a metadata file and the concept of temperature anomalies explained upfront. It took me at least an hour to familiarize myself enough with the website that I could completely comprehend the data. The map data seem to be far more complicated. There were no clear instructions on necessary programs to access NetCDF files, not to mention the SBBX file type which looks to require some programming knowledge. This is not a problem for common file types, but these specialized file formats are foreign to a novice. Documentation on these file types should be included, and the technical documentation for SBBX files, while important for experts, is pretty inaccessible to the general public, and might waste a non-expert’s time. The single README file on the page details the programmatic requirements to run SBBX files and the naming of it is insufficient to delineate its intended purpose, not to mention the complete exclusion of the data tables produced through GISTEMP.

**Publications**

This is a widely cited dataset, a Google Scholar citation search reveals that the GISTEMP dataset has been cited 231 times, and the current version is cited an impressive 33 times since its release in 2019. The website on which the dataset is hosted lists publications from the team that has worked on this dataset, but no publications from third parties who have also used it. Clearly these data are valuable, but they are presented in a way that is likely to dissuade widespread usage by the general public.

**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.

**Additional Information Related to Course Topics**

**Recommended Data Citation**

This is a regularly updated dataset, so the GISTEMP team’s recommendation of using the date of access seems wise to maintain. I also stand by their recommendation to cite the most recent publication detailing methodology, as it is a highly related resource. It is great that there is recommendation for citing it, and this recommendation seems to be broadly used; however; the citation recommended by the creators is long and subject to errors or style preferences on the part of users and citers of this dataset. This is why it is critical that this dataset be assigned a Digital Object Identifier (DOI). After this long, and this many citations, it would be imprudent to expect that inconsistencies be updated in existing publications; therefore, I would recommend the citation remain the same with a DOI added to the end. The following is my recommendation with the DOI filled in with 0s as placeholders:

GISTEMP Team, 2021: *GISS Surface Temperature Analysis (GISTEMP), version 4*. NASA Goddard Institute for Space Studies. Dataset accessed YYYY-MM-DD at <https://data.giss.nasa.gov/gistemp/>. doi:00.0000/0000000000.

Lenssen, N., G. Schmidt, J. Hansen, M. Menne, A. Persin, R. Ruedy, and D. Zyss. 2019. [Improvements in the GISTEMP uncertainty model](https://pubs.giss.nasa.gov/abs/le05800h.html). *J. Geophys. Res. Atmos.*, **124**, no. 12, 6307-6326, doi:10.1029/2018JD029522.

**Preservation**

While many of the files available on the GISTEMP website seem to have been made with interoperability in mind, others are bafflingly inaccessible. The core products of GISTEMP, the tables, are available in long established open txt and csv formats. Unfortunately, most everything else requires some expertise or special programs to run. While GISTEMP source code has been adapted to Python from the more antiquated FORTRAN used for previous iterations, it still is designed to run on Unix-like systems and some programming expertise to install. The maps on the site, though not a product of the GISTEMP algorithm, are available in Subbox grid formats, which require conversion using FORTRAN programs, or a few are available in the more easily accessible NetCDF format which still requires use of some specialized software like ArcGIS.

**Copyright**

As the product of federal funding this dataset falls into the public domain and should be listed as such.

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/>

GISTEMP Team, 2021: *GISS Surface Temperature Analysis (GISTEMP), version 4*. NASA Goddard Institute for Space Studies. Dataset accessed 2021-01-26 at <https://data.giss.nasa.gov/gistemp/>.

Lenssen, N., G. Schmidt, J. Hansen, M. Menne, A. Persin, R. Ruedy, and D. Zyss. 2019. [Improvements in the GISTEMP uncertainty model](https://pubs.giss.nasa.gov/abs/le05800h.html). *J. Geophys. Res. Atmos.*, **124**, no. 12, 6307-6326, doi:10.1029/2018JD029522.

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>