AI2ES - OU Data Guide

This document is intended to outline best-practices for data storage and usage within AI2ES, specifically focusing on OU's supercomputer. If you have datasets which will be shared among institute partners or stored on AI2ES sponsored storage, you are expected to follow the recommendations of this document.

Where to store your data:
On OU's Supercomputer, all permanent files should be archived to OURRstore, and those that you are currently actively working with should also be on OURdisk. By doing this, you ensure that all files are backed up and that you will not suffer data loss.

For data in active use, data intended for institute-wide use should be stored in the shared data directory, /ourdisk/hpc/ai2es/data on Schooner. Data intended to be kept private should be placed in personal directories.

AI2ES researchers and students are also welcome to store data on local resources at their own institutions. In this case, individuals are expected to follow local data storage guidelines, while still adhering to AI2ES data retention and metadata guidelines.

Data retention:
As a data-focused institute, it is vital that AI2ES properly archive datasets produced by the institute for future use. Datasets used or produced by AI2ES researchers and students should, to the maximum extent that it is practical, be archived to ensure that files are accessible over the long term. The OURRstore tape archive is recommended as a cost-efficient option for long-term archival of datasets.

Metadata Statement:
Each dataset stored on AI2ES space, whether private or shared, should include a text file containing a metadata statement in the dataset's top-level directory. The text file should clearly identify itself as containing the metadata statement (using the filename "README.txt").

The metadata statement shall include at least the following entries:

- **Name of dataset**
- **Source of the data** (e.g., operational NWP model, experimental model, survey, satellite obs.). If a reference can be provided for the source (e.g., as a publication, DOI, or URL), include it here.
- **Type of data** (e.g., gridded NWP output, raw radar data, post-processed radar data).
- **Data format** (e.g., GRIB2, NetCDF, .xls, raw text)
- **A brief description of the data contents**, including at least the following information:
    - If the dataset contains multiple variables, a list of available variables, or selected important variables if there are too many to list (more than ten)
    - If applicable, geographic extent of the data

    - If applicable, time period covered by the data
    - If the dataset contains data from multiple times, data frequency (e.g., every hour, every six hours)
    - If data is gridded, the method of gridding used (e.g., cartesian, lat-lon), along with the grid dimensions (e.g., 500 x 350 x 60 grid points)
    - If known, an author or contact associated with the dataset (in some cases, this could be an institutional organization, such as NOAA or SPC).
- **Restrictions/requirements for data use**. Typical examples would be, “if used, please cite the following publications”, or “please contact XXXXX (xxxxx@yyy.edu) prior to use”.

Note: These metadata statement guidelines are intended primarily for meteorological datasets, but can be adapted for other datasets.

The policies in this document were last updated on 22 June 2021.