Description of RDA Dataset Collection Curation Levels

Please refer to the ingest to dissemination workflow description, for additional details related to RDA Dataset Curation levels: https://rda.ucar.edu/rdadocs/RDA_data_ingest_to_dissemination_workflow_overview.pdf

1. *Basic curation*
    a. All dataset collections maintained in the RDA must adhere to the *Basic Curation* standard. Requirements for Basic Curation include:
        i. Required metadata fields are completed to describe dataset collections (https://rda.ucar.edu/#!rdadocs/mm_guide).
        ii. MD5 checksums are computed on data archive files stored on RDA Dataset Collection Disk and the Quasar Tape Backup system.
2. *Enhanced curation*
    a. Selected dataset collections require enhanced curation. Enhanced curation is determined on a case-by-case basis by the RDA staff. The goal of enhanced curation is to provide better support for end research use cases, long-term curation, and easier accessibility. Selected use cases include:
        i. Native data structure and format do not align with broad research use case.
            1. Climate research use case -Native model data are often structured in files with time-slice snapshots including all output parameters.
                a. Model output files are restructured from time-slice snapshots of all parameters to time-series structures organized by parameter, which better supports climate research by significantly reducing the amount of data accessed to examine a long-term trend, e.g. air temperature over 40 years.
                b. File format and associated metadata are translated from GRIB 2 to CF-NetCDF, which is more broadly used in the climate research community. CF-NetCDF also better supports long-term data curation since it is a self describing format.
                c. A reference copy of the native data is maintained offline to assure reproducibility and validation of translation, if necessary. A copy of this data can be provided to users upon request.

        ii. Native data structure and format do not align with community support formats or conventions.
            1. Observational data use case -often the native data are provided in a proprietary ASCII data file format, which is not compatible with community supported data analysis tools.
                a. File format and associated metadata are converted from proprietary ASCII to CF-compliant NetCDF to facilitate community data analysis tool access. CF-NetCDF also better supports long-term data curation since it is a self describing format.
                b. A reference copy of the native data is maintained offline to assure reproducibility and validation of translation, if necessary. A copy of this data can be provided to users upon request.

3. *Data-level curation*
    a. Selected dataset collections require data-level curation. Data-level curation is determined on a case-by-case basis by the RDA staff. The goal of data-level curation is to fix problems discovered in data or metadata, and improve support for end research use cases, long-term curation, and easier accessibility. Selected use cases include:
        i. Native data are stored on unique grid types that are difficult for the broader research community to work with.
            1. Reanalysis and operational model data use case where native data are organized in Spectral Space and Reduced Gaussian grids. This presents computational challenges to a large number of users, as most users are only familiar with data structured in regular latitude/longitude grids.
                a. Data are interpolated into regular latitude/longitude space and stored in CF-compliant NetCDF to better support ease of use, community tool access, and long term curation. All processing steps and components are described in the attribute fields provided by the NetCDF format to support provenance.
                b. A reference copy of the native data are maintained offline to assure reproducibility and validation of translation, if necessary. A copy of this data can be provided to users upon request.

        c. Documentation describing all data processing components is archived with the dataset collection.

    ii. A systematic problem is detected in metadata or data files for a dataset collection.

        1. Reanalysis use case -through RDA data ingest processing checks, it is determined that native data include incorrect descriptive metadata in the data files.

            a. The provider is notified of the systematic metadata issue.
            b. Metadata is corrected by RDA staff for all impacted data files.
            c. The corrected version of the data files is published to archive and made available for public access.
            d. A reference copy of the native data is maintained offline to assure reproducibility and validation of translation, if necessary. A copy of this data can be provided to users upon request.

A routinely updated inventory of all RDA datasets that includes dataset curation level assignment can be accessed in either PDF format or CSV format.