# Research Data Archive Dataset Change Management Strategies

One of the core foundational principles of the Data Engineering and Curation Section (DECS) is to guarantee the integrity and authenticity of the data archived in the Research Data Archive (RDA). Described below are the strategies employed by DECS to ensure the authenticity and integrity of the data archived in the RDA when changes occur to data or metadata.

Changes to metadata and the data files themselves are logged through different mechanisms. The mechanisms are structured to track any and all changes to data files and dataset metadata, as well as support data versioning. Additionally, metadata are made available in community accepted schemas, and the data files must be structured according to community accepted standards. Details related to these topics are provided below.

Data Files:

1) If a curation level greater than Basic curation is agreed upon between the data submitter and the RDA (for the definition of "Basic Curation", see: https://rda.ucar.edu/rdadocs/RDA_Dataset_Curation_Level.pdf), the DECS dataset specialist may make agreed upon changes to the data file structure or content. In this situation, the DECS dataset specialist is authenticated and authorized by the RDA system before any changes are made. The workflow used to create the derived products is documented and maintained as part of the dataset metadata. This workflow information is also made publicly available under the "Documentation" tab of the dataset's landing page (for an example of the documented workflow, see: https://rda.ucar.edu/datasets/ds633.0/docs/CISL-RDA-ERA5.grib_to_netCDF4_HDF5.jpg). A reference copy of the native, unchanged data file(s) is(are) preserved with the dataset to assure reproducibility and validation of any derived product or restructuring workflows. These files are tracked as their own fileset group within the dataset, stored only internally, and are only accessible through offline requests. The native, unchanged data files receive all of the same level of preservation support as all other files in the dataset according to the description found under "Data Preservation Policy" on https://rda.ucar.edu/#!daas/terms-and-conditions. If the native data files/objects are preserved at an alternate trusted repository such as the Copernicus Climate Data Store, https://cds.climate.copernicus.eu/#!/home, there will be exceptions where the RDA points to that repository instead of wastefully maintaining a second copy on local storage.
2) Once a dataset has been created and a Digital Object Identifier (DOI) has been assigned to a dataset, there are a variety of strategies in place to track version changes to the data files as detailed under the following page: https://rda.ucar.edu/#!data-citation/use-cases.
3) The data files included in a dataset must be provided in a community supported data format as outlined under the "Section 2 of 4: Dataset Characteristics" section's "File Format(s)" subsection on the following page:

https://rda.ucar.edu/#!daas/worksheet-instructions. This is required to support access from community developed tools and support long-term, sustainable data curation.

4) There are no circumstances where data files will be deleted from a dataset, unless the dataset is purged according to the "Dataset Withdrawal Policy" found on https://rda.ucar.edu/#!daas/terms-and-conditions.

Metadata:

1) All changes to data description metadata are tracked in a Concurrent Versions System (CVS) version control system (https://www.gnu.org/software/trans-coord/manual/cvs/cvs.html). The full history of metadata changes can be accessed under the "Change History" section of the Metadata Manager, i.e. RDA's web-based tool for performing various metadata-related activities. An overview of the metadata "Change History" capability is provided under the "Manage Datasets" section of the following page: https://rda.ucar.edu/#!rdadocs/mm_guide.
2) Metadata are maintained in a native RDA schema, which can be mapped to a variety of community supported standards including:
    a) DataCite
    b) GCMD Directory Interchange Format (DIF)
    c) Dublin Core
    d) Federal Geographic Data Committee (FGDC)
    e) International Organization for Standardization (ISO) 19139 and ISO 19115-3
    f) JSON-LD Structured Data
    g) Please find an example of the available standard metadata schemas provided by the RDA by reviewing the "Metadata Record" menu found at the bottom of an example dataset homepage: https://rda.ucar.edu/datasets/ds083.2/#!description

The RDA strategy for data changes includes two use cases:

1) Case 1, A curation level greater than Basic curation is agreed upon between the data submitter and the RDA (for the definition of "Basic Curation", see: https://rda.ucar.edu/rdadocs/RDA_Dataset_Curation_Level.pdf): In this case, the DECS data specialist may make agreed upon changes to the data file structure or content. In this situation, the DECS dataset specialist is authenticated and authorized by the RDA system before any changes are made. The workflow used to create the derived products is documented and maintained as part of the dataset metadata. This workflow information is also made publicly available under the "Documentation" tab of the dataset's landing page (for an example of the documented workflow, see: https://rda.ucar.edu/datasets/ds633.0/docs/CISL-RDA-ERA5.grib_to_netCDF4_HDF5.jpg). A reference copy of the native, unchanged data file(s) is(are) preserved with the dataset to assure reproducibility and validation of any derived product or restructuring workflows. These files are tracked as their own fileset group within the dataset, stored

internally, and are only accessible through offline requests. The native, unchanged data files receive all of the same level of preservation support as all other files in the dataset according to the description found under “Data Preservation Policy” on https://rda.ucar.edu/#!daas/terms-and-conditions. The data submitter is made aware of this strategy under the “Upon acceptance” section’s “Collaboration on data curation level and any related data transformations or restructuring” subsection on the following page: https://rda.ucar.edu/#!daas/decision-workflow.

2) Case 2, Updates are made to data files after the dataset has been created and a DOI has been assigned to the dataset: Descriptions of these use cases are provided on the following page: https://rda.ucar.edu/#!data-citation/use-cases. This information can also be made available for the data submitter if an applicable use case arises.

The RDA maintains provenance information to support the following use cases:

1) Case 1, A curation level greater than Basic curation is agreed upon between the data submitter and the RDA (for the definition of “Basic Curation”, see: https://rda.ucar.edu/rdadocs/RDA_Dataset_Curation_Level.pdf): In this case, the DECS data specialist may make agreed upon changes to the data file structure or content. In this situation, the DECS dataset specialist is authenticated and authorized by the RDA system before any changes are made. The workflow used to create the derived products is documented and maintained as part of the dataset metadata. This workflow information is also made publicly available under the “Documentation” tab of the dataset’s landing page (for an example of the documented workflow, see: https://rda.ucar.edu/datasets/ds633.0/docs/CISL-RDA-ERA5.grib_to_netCDF4_HDF5.jpg). A reference copy of the native, unchanged data file(s) is(are) always preserved with the dataset to assure reproducibility and validation of any derived product or restructuring workflows. These files are tracked as their own fileset group within the dataset, stored internally and are only accessible through offline requests. The native, unchanged data files receive all of the same level of preservation support as all other files in the dataset according to the description found under “Data Preservation Policy” on https://rda.ucar.edu/#!daas/terms-and-conditions.

2) Case 2, User needs to retrieve a data file that has been replaced in a dataset versioning operation: Data access history is available for every RDA user. By using the Data Citation tool, found in the RDA user dashboard (see: http://ncarrda.blogspot.com/2017/03/the-rda-user-dashboard.html), a user can find the exact date/time that a data file was downloaded. If there have been any changes to the file since that download occured, it will be noted in the dataset file list, and the DECS dataset specialist can provide a copy of the file that was originally downloaded.

3) Case 3, User specified request provenance tracking: A receipt is provided with each user request detailing request specifications. Additionally, a unique persistent identifier

is assigned to each user request, so all aspects of that request can be tracked through this mechanism, including request timestamp, request processing steps and platform, operating system version, and software versions used to support request processing. Currently, a data user needs to ask the DECS dataset specialist to provide all of these details based on the request ID, but it is in the future DECS roadmap to add a programmatic capability for RDA users to query this information by request ID.

To support file version changes, the RDA compares the MD5 checksum of a new file with the MD5 checksum of the replaced file. The RDA also scans the file content metadata (see: the "About Data File Content Metadata" section on the following page: https://rda.ucar.edu/#!rdadocs/dsmaint) of the new file to determine if there are any differences in the file contents. New metadata information is integrated as needed.