Research Data Archive Dataset Ingest to Dissemination Workflow Overview

The Research Data Archive (RDA) maintains an established data ingest to dissemination workflow, which is employed by all RDA Dataset Specialists to bring new dataset collections into the archive. The existing workflow was first introduced in 2008. Few changes to the overall logic have occurred since its introduction, outside of additions and updates to tools used to support the workflow. To support change management, staff meetings are held on a weekly basis that allow Dataset Specialists to review and agree upon proposed workflow changes. Changes typically involve introduction of new software components that are thoroughly vetted and tested on prototype datasets prior to being introduced to the production workflow. Additionally, all software workflow components are maintained in the Git version control system. Finally, since no sensitive data is housed in the RDA, security concerns are not an issue, and are not addressed in this document. A complete overview of the RDA data ingest to dissemination workflow is provided in Figures 1, 2, and 3.

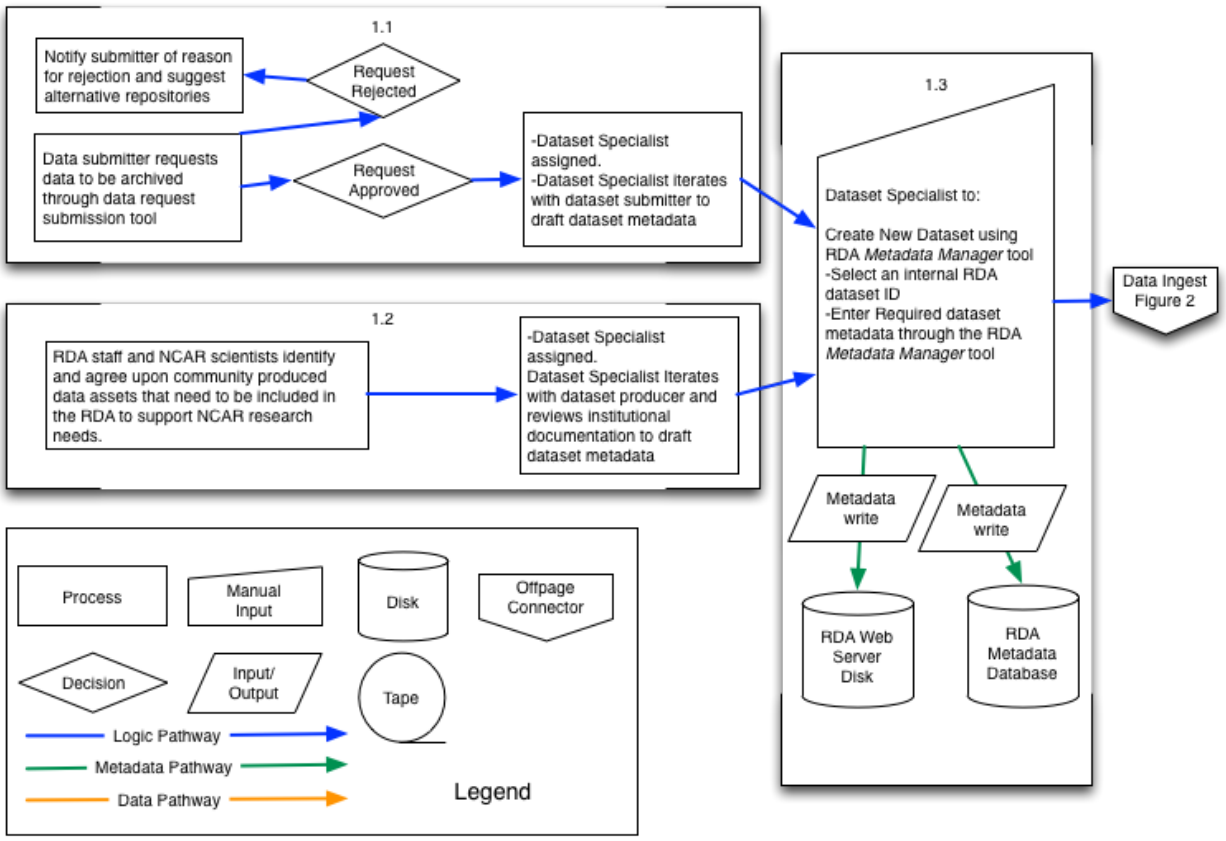


Figure 1 -RDA Dataset Selection, and Dataset Metadata Creation

Figure 1. RDA Dataset Selection and Dataset Metadata Creation

- 1.1. In this case, a PI (data submitter) requests to archive data in the RDA through the dataset submission tool (https://rda.ucar.edu/#!daas) according to the terms and conditions described on https://rda.ucar.edu/#!daas/terms-and-conditions. If the request to archive is approved by the Data Engineering & Curation Section (DECS) manager, as described on https://rda.ucar.edu/#!daas/decision-workflow, a Dataset Specialist is assigned to work with the submitter and draft the dataset metadata, and provide details on how the data will be handled. If the request to archive is rejected by the DECS manager, the submitter is notified why their request was rejected, and provided with suggestions for alternative repositories where applicable.
- 1.2. RDA staff and NCAR scientists confer and work together to determine which community produced data assets are available, need to be part of the RDA's holdings to be accessible to CISL computing systems, and are practical to acquire, such as Copernicus climate reanalysis data products. Once it is agreed upon that the data asset should be included in the RDA by the DECS manager, a Dataset Specialist is assigned to iterate with the dataset producer and review institutional documentation to draft the dataset metadata.
- 1.3. After the initial set of dataset metadata has been drafted, the next task of the Dataset Specialist is to formally create the new dataset. This process involves selecting an internal RDA dataset ID, defining the dataset storage location on RDA Dataset Collections Disk, and populating the dataset metadata with required fields. This information is entered through the RDA *Metadata Manager* tool (https://rda.ucar.edu/#!rdadocs/mm_guide). Once populated, metadata are written to XML files on the RDA Web Server Disk, and to the relevant RDA Metadata Database tables.

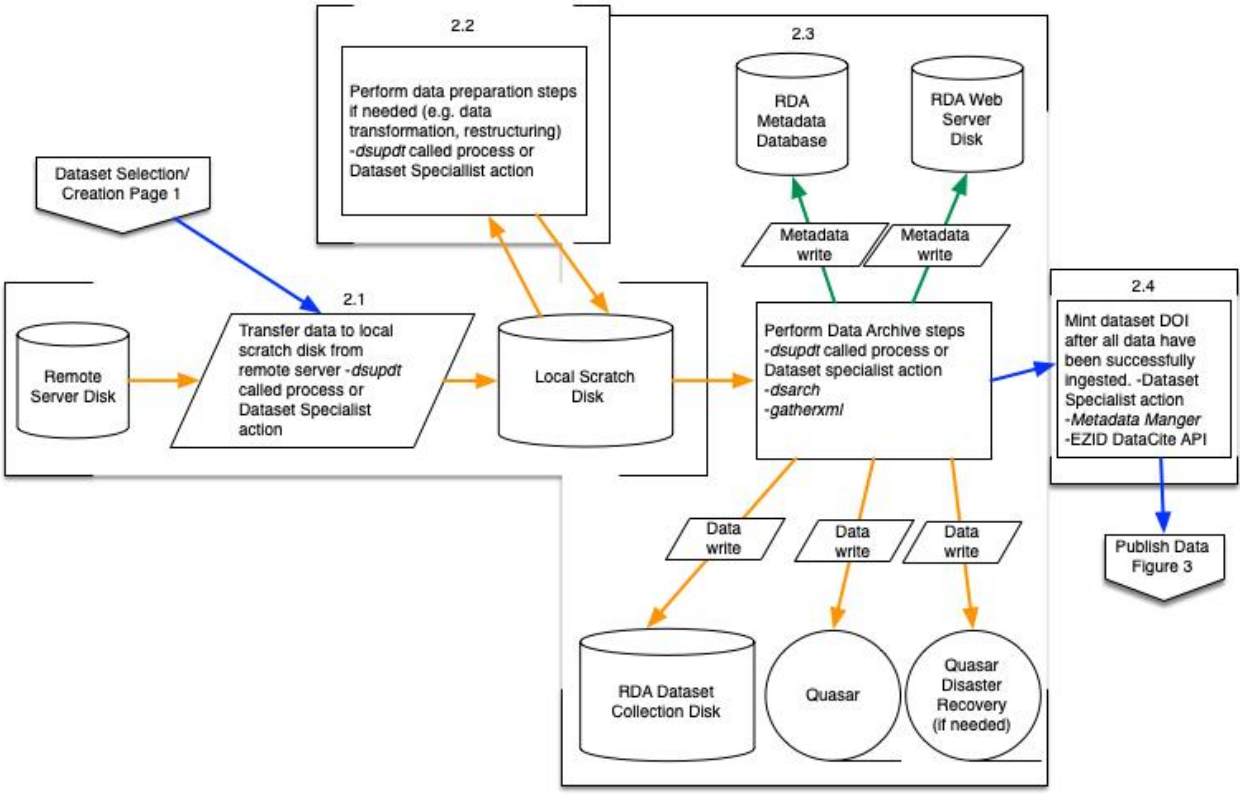


Figure 2 -RDA Dataset Ingest and DOI Creation

Figure 2. RDA Dataset Ingest and DOI creation

- 2.1. Dataset Specialist triggers a process to transfer data from a remote location to local scratch disk. The RDA supported "*dsupdt*" tool provides a configurable option to transfer data from a remote location to local scratch disk (https://rda.ucar.edu/rdadocs/dsupdt/).
- 2.2. Dataset Specialist performs value added data preparation steps if applicable. Typically, these data preparation operations are chosen to be performed on datasets where native grids come in non-standard projections, and they need to be transformed into a simpler projection to better serve the user community. Additionally, data can be reorganized and/or converted from its native format into a different format, to support better compression and/or broader community usage. If a data preparation operation has been performed, native data are retained and archived with the RDA to support reproducibility, unless the native data provider cannot act as a reliable backup mechanism.
- 2.3. Data file archive steps are triggered programmatically by *dsupdt* process, or by Dataset Specialist manual action. The tool, *dsarch*, (https://rda.ucar.edu/rdadocs/dsarch/) is used to perform the data archiving steps of:
    - Compute MD5 checksum.
    - Create one dataset archive file copy on RDA Dataset Collection Disk
    - Create one dataset archive backup file copy on the Quasar tape system.

    - If needed, create one dataset archive file disaster recovery copy on the Quasar tape system which is moved to a fireproof safe. See RDA Data Security for more information.
    - Update RDA Metadata Database with locations of all files
    - Update dataset web file lists on RDA Web Server Disk.
    - Trigger tool, *gatherxml* (https://rda.ucar.edu/#!rdaman/gatherxml), to:
        - Scan archive data file.
        - Verify validity of file contents.
        - Extract file content metadata.
        - Update file content metadata and dataset summary metadata in RDA Metadata Database, and update dataset summary metadata on RDA Web Server Disk.
            - The RDA Metadata Database and RDA Web Server disk are backed up by an enterprise tool on a daily basis to ensure redundancy and long term preservation of dataset metadata assets. See RDA Data Security for more information.
- 2.4. After all metadata and data have been successfully ingested into the dataset archive, the Dataset Specialist validates that dataset metadata are correct, and mints a digital object identifier (DOI) for the dataset using the RDA *Metadata Manager* tool, which acts through the DataCite (https://datacite.org/) API interface to perform this action. Additional background on this process can be found here: https://rda.ucar.edu/#!data-citation.

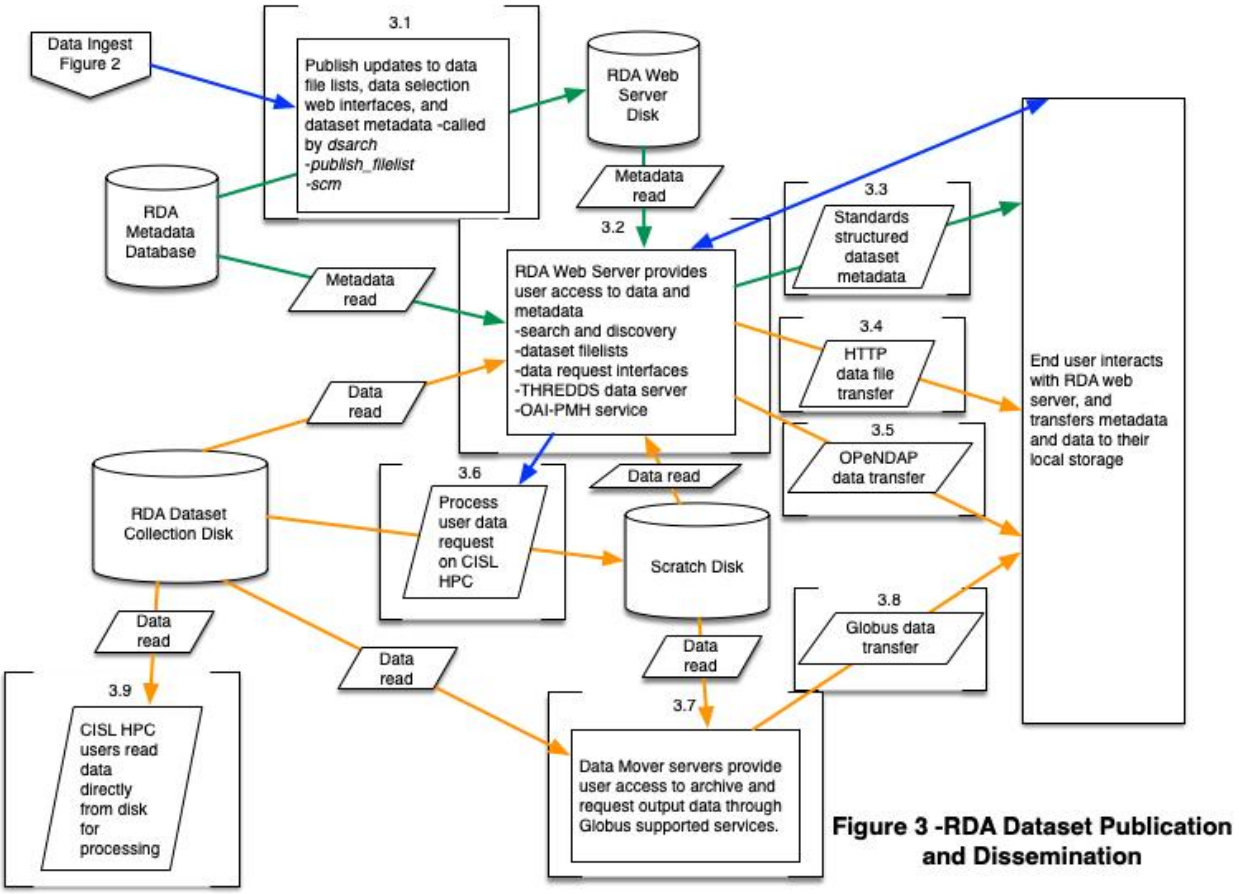


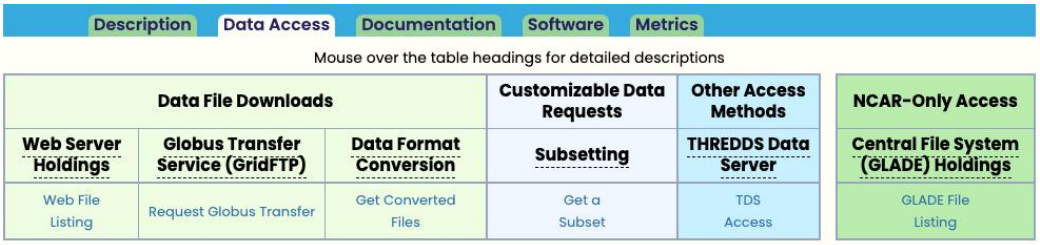
Figure 3 -RDA Dataset Publication and Dissemination

Figure 3. RDA Dataset Publication and Dissemination

- 3.1. After all data, metadata, and DOI processes are complete, the Dataset Specialist triggers multiple commands that use information read from the RDA Metadata Database to update dataset summary metadata, and publish updates to dataset filelists on RDA Web Server Disk. These software modules include *publish_filelist* (https://rda.ucar.edu/#!rdadocs/dsmaint) and *summarize content metadata (scm)* (https://rda.ucar.edu/#!rdaman/scm).
- 3.2. The RDA Web Server (https://rda.ucar.edu) provides interfaces for users to search, discover, and access the archived data and metadata through a variety of avenues.
    - Summary of metadata access avenues:
        - Metadata are queried from the RDA Metadata Database and RDA Web Server Disk to support user interaction with search tools, dataset filelists, and data request interfaces.
        - 3.3. Users can access standards structured metadata through web service endpoints:
            - Open Archives for Metadata Harvesting (OAI-PMH) (https://rda.ucar.edu/cgi-bin/oai).

- Unidata's Thematic Real-time Environmental Distributed Data Services (THREDDS) Data Server (https://rda.ucar.edu/thredds/catalog/catalog.html).
- Catalogue Service for the Web (CSW) (https://rda.ucar.edu/cgi-bin/csw?request=GetCapabilities&service=CSW)

- Summary of data access avenues:
  - Users can download archive files directly or from data request outputs using:
    - 3.4. Traditional HTTP methods
    - 3.8. Globus GridFTP (https://www.globus.org/)
      - 3.7. Data Mover servers, running the Globus Connect Server software stack (https://www.globus.org/globus-connect-server), are used to support the Globus data transfer option
  - Users can request that data be prepared for them for download through:
    - 3.6. Data subset and format conversion requests
  - 3.5. Through interoperable tools or scripts, users can programmatically request subsets of data to be transferred through the Open Source Project for a Network Data Access Protocol (OPeNDAP) (https://www.opendap.org/) provided by the THREDDS Data Server.
  - 3.9. CISL High-performance computing (HPC) users can read data archive files directly from RDA Dataset Collection Disk (https://www2.cisl.ucar.edu/data-portals/research-data-archive).
  - Links to all data access avenues can be found under the "Data Access" tab of a dataset homepage found on the RDA Web Server. Users need to be authenticated with their RDA user profile to access these links. An example of a dataset "Data Access" page is provided below.



Inventory and Description of RDA Software

The following software components are used to support the tools in Figure 1 of the workflow (i.e. "RDA Dataset Selection and Dataset Metadata Creation"):

1) Dataset submission and appraisal tools (1.1) are in-house developed PHP-based (http://php.net/) web applications.
2) The *Metadata Manager* metadata data entry and validation tool (1.3) is an in-house developed C++-based (http://www.cplusplus.com/) web application. The *Metadata Manager* maps metadata into a RDA native schema based on International Organization for Standardization (ISO) representations (e.g. ISO 8601), and requires use of Global Change Master Directory (GCMD) controlled vocabulary keywords (https://earthdata.nasa.gov/earth-observation-data/find-data/idn/gcmd-keywords) to describe dataset collection parameters.

The following software components are used to support the tools in Figure 2 of the workflow (i.e. "RDA Dataset Ingest and DOI creation"):

1) The in-house developed *dsupdt* tool (2.1), used to support automated data ingest for dynamic datasets, is written in Python (https://www.python.org/). *dsupdt* interfaces with the RDA Metadata Database to save configuration preferences using the Python supported database modules (e.g. DBI, https://dev.mysql.com/doc/connector-python/en/). *dsupdt* uses community supported software, including wget (https://www.gnu.org/software/wget/) and ncftp (https://www.ncftp.com/) to transfer data from remote servers to local RDA/Computational and Information Systems Lab (CISL) servers.
2) Data preparation steps (2.2) are typically performed using community supported data manipulation tools. Selected examples of data preparation tools used by the DECS include:
    a) wgrib2 (http://www.cpc.ncep.noaa.gov/products/wesley/wgrib2/)
    b) NetCDF operators (http://nco.sourceforge.net/)
    c) Climate Data Operators (https://code.mpimet.mpg.de/projects/cdo/)
    d) NCAR Command Language (https://www.ncl.ucar.edu/)
    e) NCAR GeoCAT (https://geocat.ucar.edu/)
    f) ECMWF ECCODES (https://confluence.ecmwf.int/display/ECC/ecCodes+Home)
    g) Open Source Packages supported/shared through the Pangeo community project (https://pangeo.io/packages.html)
3) The in-house developed *dsarch* tool, used to archive data to RDA dataset collection disk and the Quasar Tape system (2.3), is written in Python. It uses Python supported database modules to record file location and description information in the RDA Metadata Database.
4) The in-house developed *gatherxml* tool, used to extract format specific file level metadata and write that metadata into the RDA Metadata Database and to RDA

Web Server disk (2.3), is written in C++ and uses the community supported C++ MySQL connector (https://dev.mysql.com/doc/connector-cpp/8.0/en/) to interface with the RDA Metadata Database.

5) The *Metadata Manager* metadata data entry and validation tool (2.4) is an in-house developed C++-based (http://www.cplusplus.com/) web application. The *Metadata Manager* includes a C++ module that maps native RDA metadata into required DataCite (https://datacite.org/) metadata elements and calls the DataCite API (https://support.datacite.org/docs/mds-api-guide) to mint RDA dataset DOIs.

The following software components are used to support the tools in Figure 3 of the workflow (i.e. “RDA Dataset Publication and Dissemination”):

1) The in-house developed *publish_filelist* tool (3.1), used to publish dataset collection file inventories for user access, is written in Python and interfaces with the RDA Metadata Database using Python supported database modules.
2) The in-house developed *scm* tool (3.1), used to generate content metadata summaries for inclusion in dataset collection level metadata on RDA Web Server Disk, is written in C++ and and uses the community supported C++ MySQL connector to interface with the RDA Metadata Database.
3) Several web applications provide interfaces for users to search, discover, and access archived data and metadata (3.2) including the following:
    a) In-house developed faceted and free text search applications that are written in C++. Both of these applications use the community supported C++ MySQL connector to interface with the RDA Metadata Database.
    b) In-house developed subset request applications that are written in C++, PHP, and Javascript. These use the community supported C++ MySQL connector and PHP PDO driver (http://php.net/manual/en/ref.pdo-mysql.php) to interface with the RDA Metadata Database.
    c) In-house developed OAI-PMH and CSW servers, used to distribute standards structured dataset metadata (3.3), are written in C++. This uses the community supported C++ MySQL connector to interface with the RDA Metadata Database. The both servers use community metadata specifications to map RDA native metadata into multiple schemas (see R14 for additional details on provided metadata schemas).
    d) The community supported Unidata Thematic Real-time Environmental Distributed Data Services (THREDDS - https://www.unidata.ucar.edu/software/thredds/current/tds/) is used to

support Open-source Project for a Network Data Access Protocol (OPeNDAP) data access (3.5).

e) The externally supported Globus Connect Server (https://www.globus.org/globus-connect-server) (3.7) is used to support Globus maintained GridFTP data transfers (https://www.globus.org/#transfer) (3.8).

f) The in-house developed *dsrqst* tool (*https://rda.ucar.edu/rdadocs/dsrqst/*), which automatically manages user data request processing, is written in Python, and coordinates user request processing workflows that run on CISL High Performance Computing (HPC) systems (https://arc.ucar.edu/resources) (3.6). *dsrqst* uses Python supported database modules to interface with the RDA Metadata Database.

General software/server/infrastructure components used across all components of the RDA dataset ingest to dissemination workflow include the following:

1) The RDA Metadata Database that runs on the open source MySQL 5.7 database server (https://www.mysql.com/).
2) The web applications that currently run on Apache 2.4.x HTTP server. (https://httpd.apache.org/). The Apache 2.4.x HTTP server is also used to support HTTP based data transfer activities (3.4).
3) The RDA Web and Metadata Databases that run on virtual machines, which use the CentOS 7 operating system. The virtual machines operate on a CISL maintained VMWare (https://www.vmware.com/) server cluster.
4) The RDA Dataset Collection Disk that runs on a IBM Spectrum Scale General Parallel File System (https://www.ibm.com/docs/en/gpfs).
5) The CISL Quasar Tape system, which hosts backups of RDA data, is an IBM TS4500 robotic library with 2,198 slots and dual accessors. Full specification for Quasar can be found at: https://arc.ucar.edu/knowledge_base/70549580
6) UCAR's Network Engineering and Telecommunications Section (NETS, (http://nets.ucar.edu/nets/intro/introduction.shtml) maintains high volume and high availability network connectivity to support programmatic/automated RDA data ingest workflows effectively. Additionally, auto retry capability is integrated into the RDA dsupdt tool (https://rda.ucar.edu/rdadocs/dsupdt/) to support data ingest recovery as needed after system/network outages. The RDA does not maintain "real-time" datasets, where immediate access is essential to support user needs. All RDA assets are considered to be for research use only, so although NETS typically provisions around-the-clock connectivity to public and private networks at a bandwidth that is sufficient to meet the global and/or

regional responsibilities, 24x7 connectivity is not essential to support the use case needs of the RDA user community.

Additional details on Metadata, Software, and Infrastructure not captured above:

Metadata:

As highlighted above, dataset collection level metadata is maintained in a native RDA schema based on ISO representations (e.g. ISO 8601) and leverages Global Change Master Directory (GCMD) controlled vocabulary keywords (https://earthdata.nasa.gov/earth-observation-data/find-data/idn/gcmd-keywords). Tools are provided to map the native RDA metadata into community standards based schemas according to the relevant standard specifications, including DataCite, GCMD Directory Interchange Format (DIF), Dublin Core, Federal Geographic Data Committee (FGDC), International Organization for Standardization (ISO) 19139, ISO 19115-3, and JSON-LD Structured Data. 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

Additionally, all of the listed metadata schemas, plus the THREDDS schema, can be accessed through the RDA Open Archive Initiatives Protocol for Metadata Harvesting (OAI-PMH) web service: https://rda.ucar.edu/cgi-bin/oai/https://rda.ucar.edu/cgi-bin/oai?verb=ListMetadataFormats

A Catalog Service for the Web (CSW) server can also be used to access RDA metadata at: https://rda.ucar.edu/cgi-bin/csw?request=GetCapabilities&service=CSW

Software:

As detailed above, the RDA employs a combination of in-house developed software and community supported software components to support data curation, data discovery, and data access workflows. An inventory of in-house developed software components is maintained in the 42 repositories organized under the NCAR institutional GitHub space (https://github.com/NCAR). Due to security concerns, there is currently a mix of publicly available and restricted repositories maintained in the RDA team space, so all repositories are not visible to external parties. Documentation is included as READMEs in each RDA team repository.

Infrastructure:

Quarterly meetings are held between relevant DECS staff to develop estimates of future storage requirements based on regular automated ingest stream volumes, and on estimated future product volumes that will be coming into the archive. Based on this information and on a yearly basis, CISL allocates Quasar tape and RDA dataset disk resources as needed to support future RDA growth.

Load usage and performance is actively monitored on all RDA supported servers and services to ensure performance continues to meet expectations. New servers are procured based on forecast usage metrics and recorded load usage on a 4-5 yearly basis.

Information Systems Division (ISD) and DECS management participates in CISL strategic planning exercises on a bi-yearly basis, to ensure that RDA service offerings evolve to meet current and future user expectations. Additionally, DECS management meets with CISL management on a monthly basis to review current services offerings, and determine whether or not these need to be adjusted to meet existing user expectations.

Disaster and Business Continuity:

The RDA's Information Technology (IT) infrastructure, supported by the National Center for Atmospheric Research (NCAR) Computational and Information Systems Lab (CISL), provides highly available storage, virtual machine (VM) supported web and database services, and backup and disaster recovery for archive data as described in https://rda.ucar.edu/rdadocs/RDA_data_security.pdf.

RDA web and database servers run on VMware virtual servers in a cluster environment at the NCAR Wyoming Supercomputing Center (NWSC) (https://www2.cisl.ucar.edu/ncar-wyoming-supercomputing-center) located in Cheyenne, WY. In the event of hardware and software failures, the VMs will be restored by CISL Shared Infrastructure Section (SISSEC) (https://staff.ucar.edu/browse/orgs/SISSEC) staff on working cluster nodes. The SISSEC service level agreement (SLA) supports maintenance on RDA VM's between 7AM and 7PM Monday - Friday, excluding holidays.

RDA web and Metadata Database VMs, Dataset Collection Disk, and Quasar Tape system are maintained on the NWSC uninterruptible power supply (UPS) backup. In the event of a utility power issue, all RDA infrastructure remains available on the UPS system. The CISL High Performance Computing Division Section (HPCDSEC, https://staff.ucar.edu/browse/orgs/HPCDSEC) is responsible for bringing Dataset

Collection Disk, the Quasar Tape system, and CISL HPC resources back online in the event of an unplanned outage. This infrastructure is monitored by the CISL Cheyenne Operations Section (https://staff.ucar.edu/browse/orgs/COS) on a 24x7 basis (https://arc.ucar.edu/system_status), in coordination with SISSEC and HPCDSEC staff. Designated DECS staff are contacted by COS, HPCDSEC and/or SISSEC staff in the event of an unplanned outage, and coordinate with those entities to bring services back online according to agreed upon SLAs which may be vendor dependent.

NCAR/UCAR Risk analysis planning:
Publicly available information can be found in the Risk Management and Resiliency section in: https://rda.ucar.edu/rdadocs/RDA_data_security.pdf. An overview of the UCAR Enterprise Risk Management office can be found at:
https://rda.ucar.edu/rdadocs/UCAR-Enterprise-Risk-Management.pdf

NCAR/UCAR Disaster and business continuity plan:
UCAR's business continuity plan is based on the following standards

1) U.S. Department of Homeland Security, Federal Emergency Management Agency (FEMA)
2) NFPA 1600:2007 Standard on Disaster/Emergency Management and Business Continuity Programs
3) ISO 22301 • NIST SP 800-34 Contingency Planning Guide for Information Technology Systems
4) DRII/DRJ GAP Generally Accepted Practices for Business Continuity Practitioners

Information on disaster and related business continuity planning can be found on the corporate intranet site which is not publicly available. A copy of the information provided on the internal website has been made publicly available at:
https://rda.ucar.edu/rdadocs/UCAR-Business-Continuity.pdf