The NCEI Archive Process

Philip Jones, STG Inc.
Archive Representative, Data Stewardship Division,
NOAA's National Centers for Environmental Information

04 August 2015

NOAA Satellite and Information Service | National Centers for Environmental Information

What does it mean to archive data at NCEI?

- Data and supporting documentation are stored on dedicated hardware (backed up and migrated) for long-term preservation
- Archived according to retention schedule approved by NARA[1]
- We document who is responsible for submitting the data and the rationale archiving
- Datasets are fully documented for current and future users (for understanding the intended use, quality)
- Datasets are discoverable and freely accessible to the public[2]

1. National Archives and Records Administration (www.archives.gov/)
2. Public access mechanisms depend on type of data

NCEI Archive Process

- One process to archive any and all data at NCEI
- Multi-phased process with decision gates
- Designed to be efficient and effective
- Integrated with CDRP Process

NCEI Archive Process

- One process to archive any and all data at NCEI
- Multi-phased process with decision gates
- Designed to be efficient and effective
- Integrated with CDRP Process

Appraisal Phase

Initiate by submitting request
Appraise data request
Decide what and how to archive

NCEI Archive Process

- One process to archive any and all data at NCEI
- Multi-phased process with decision gates
- Designed to be efficient and effective
- Integrated with CDRP Process

Appraisal Phase

Initiate by submitting request
Appraise data request
Decide what and how to archive

Archive Preparation

Plan data submission and services
Review data and documentation
Test data transfers
Create standard metadata
Decide on ops readiness

NCEI Archive Process

- One process to archive any and all data at NCEI
- Multi-phased process with decision gates
- Designed to be efficient and effective
- Integrated with CDRP Process

Appraisal Phase

Initiate by submitting request
Appraise data request
Decide what and how to archive

Archive Preparation

Plan data submission and services
Review data and documentation
Test data transfers
Create standard metadata
Decide on ops readiness

Operational Phase

Transfer and archive data
Provide public access (publish)
Monitor and maintain systems

Integrated Product Team (IPT)

- Team established for each CDR
- **Multidisciplinary** group of people who are collectively responsible for delivering the CDR
- Includes **PI team members** and **NCEI** (CDRP, Ingest, Archive, Access, IT Network & Security, Science Experts, Customer Engagement)
- Meet regularly leading up to delivery
- IPT Archive Representative: Philip Jones, Heather Brown or Valerie Toner

Steps for CDR PIs

Submit Request to Archive

- Completed by the PI or PI associate
- Request initiates the NCEI Archive Appraisal Process
- Information used by NCEI for making decisions on how to archive the data
- Request form is available in the **ATRAC** web application
  1. Create ATRAC user account using email
  2. Login to ATRAC
  3. Register a project for your CDR
  4. Access the request form
  5. Complete and submit request form
- Time to complete form: **1 hour**

NetCDF File Format

- NetCDF-4 format required for all CDR data files (CDRP requirement)
- Self-describing: contains metadata that describes the layout of the file
- Platform independent
- Allows for efficient subsetting
- Necessary for NCEI THREDDS Data Server access
- Wide range of application software for data use

NetCDF Metadata Conventions

- Conventions provide a standard for netCDF data structure and descriptions
  1. CF Metadata Conventions (current version, 1.6)
  2. CF standard name for CDR variables
  3. ACDD conventions for data discovery fields (current version, 1.3)
- See NCEI NetCDF Templates (for gridded, swath and other feature types)

  ***NOTE**: Helpful resource links provided at end of presentation*

NetCDF Metadata Review

- Send sample netCDF data files to the IPT for CF compliance review
- Need to review sample data early, before completion of CDR development!
- Reviews usually require multiple iterations
- Time to complete review: **weeks**

Data File Naming and Packaging

- Unique file name following a consistent pattern is needed for NCEI system configurations
- Descriptive fields are used for file metadata
- File packaging depends on file sizing and update frequency (defined during development of SA)
- File naming convention (static to dynamic field order):

<ShortName>_<Version>_<Attribute[n]_s<BeginDateTime>_e<EndDateTime>_d<SingleDateTime>_c<CreateDateTime>.<Ext>

- Example:

ssi_v02r00_daily_s18820101_e18821231_c20150615.nc

Submission Manifest File

- Companion text file with a cryptographic hash function for each data file
- For data integrity of data transfer (ensures NCEI receives what was meant to be sent)
- Example submission manifest file name:

ssi_v02r00_daily_s18820101_e18821231_c20150615.nc.mnf

- Content format of submission manifest file:

<file_name>,<file_md5_checksum>,<file_size_in_bytes>

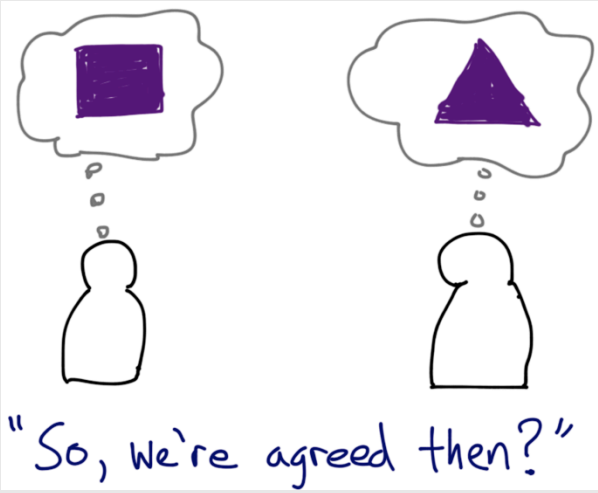
- Example submission manifest file contents:

ssi_v02r00_daily_s18820101_e18821231_c20150615.nc,da3e100dc9e7bebb810985e37875de38,5673600

Data Transfer Protocol

- Setup the data transfer with NCEI
- NCEI IT Security must approve all system connections
- Possible transfer protocols in order of preference:
  - SFTP push
  - FTP push (login)
  - FTP pull
  - Other (special cases)
- Transfer testing required before operational approval
- Ingest team configures and implements
- Once connected, may send historical CDR data first (in bulk) before routine operational data

Data Submission Agreement (SA)



- Documents details of the data submission to NCEI
- Technical, operator-level document
- Reflects and supports project requirements
- Specifications must be followed during the operational data transfer
- Improves the Provider–Archive interaction
- Ultimately ensures the quality and integrity of the data archive

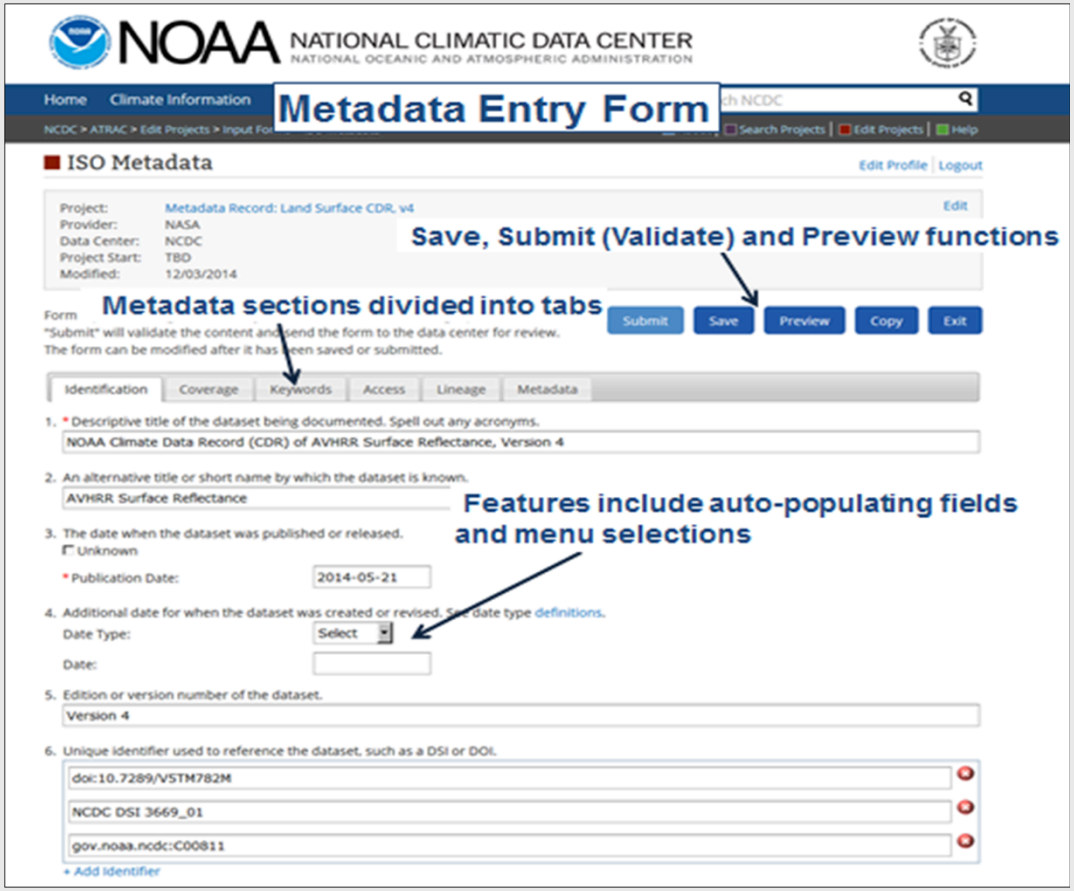
Data Submission Agreement (SA)

- SA document includes:
  - PI and NCEI Contacts
  - Filenames, sizes, contents
  - Transfer method/protocol
  - System connections (IP addresses)
  - Submission Manifest specifications
  - Ingest and Archive configurations
  - Error handling
  - NCEI Services (e.g., archiving, access, DOI)
- Drafted by NCEI Archive Rep; reviewed by IPT
- Time to complete SA review: **weeks**

Catalog Metadata

- Standard description of the CDR (required for all NCEI data)
- Follows ISO 19115 series of standards
- PI enters CDR information into ATRAC web form for the catalog record
- Catalog metadata is reviewed by IPT
- Completed metadata published to internal and external catalogs for data discovery

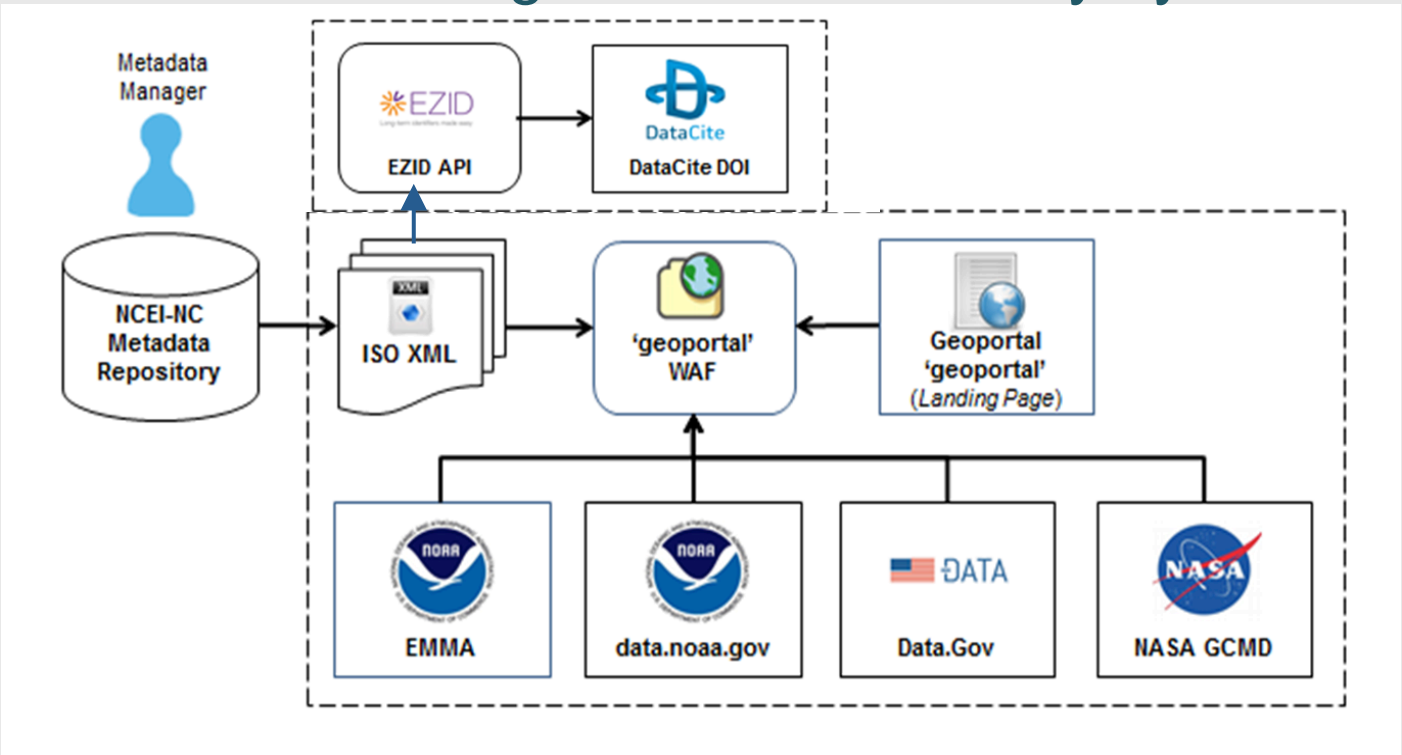
Catalog Metadata Web Form



- Time to complete form in ATRAC: **1-2 hours**

Catalog Metadata Publishing

- Metadata is used for the dataset landing page and by NCEI access systems
- Published to catalogs for data discovery by users



DOI for the CDR

- Persistent, unique identifier for the CDR package
- Allows for unambiguous data citation and attribution of the data creators (analogous to paper citation)
- Each CDR version receives a DOI
- DOI resolves to CDR landing page at NCEI

Takeaways for PIs

- For Archive Appraisal phase:
  - Submit Request to Archive (ATRAC)
- For Archive Preparation phase:
  - Provide sample data (netCDF CF review)
  - Participate in developing the SA (detailed plan for CDR transfer to NCEI)
  - Setup and Test transfer of CDR per the SA
  - Help create metadata for data catalog and DOI (also using ATRAC)
- For Operational phase:
  - Transfer the CDR per the SA

Archiving NOAA CDRs

- Since 2010:
  - 30 CDRs have reached operations
  - 40 CDRs have been archived (due to 10 improved CDR versions)
- Notable Process improvements since 2010
  - IPTs
  - NetCDF CF reviews
  - DOIs for data (22 currently for CDRs)

Helpful Resources

- CDRP website:
  - Guidelines
  - Contacts
  - Operational CDRs
- ATRAC (Archive Request and ISO Metadata forms)
- NetCDF templates (includes CF and ACDD Metadata Conventions)
- DataCite (dataset DOI info)

Questions / Comments

- philip.jones@noaa.gov

NOAA's National Centers for Environmental Information

www.ncei.noaa.gov
www.climate.gov

NCEI Climate Facebook: http://www.facebook.com/NOAANCEIclimate
NCEI Ocean & Geophysics Facebook: http://www.facebook.com/NOAANCEIoceangeo
NCEI Climate Twitter (@NOAANCEIclimate): http://www.twitter.com/NOAANCEIclimate
NCEI Ocean & Geophysics Twitter (@NOAANCEIocngeo): http://www.twitter.com/NOAANCEIocngeo