Data Curation Practices at the Oak Ridge National Laboratory Distributed Active Archive Center

Robert Cook, DAAC Scientist
Environmental Sciences Division
Oak Ridge National Laboratory
Oak Ridge, TN
cookrb@ornl.gov

Data Curation Education in Research Centers Workshop
Boulder, CO
June 5-7, 2012
Presentation Outline

- Overview of the ORNL DAAC
- Data Curation Practices
- Tools
- Challenges and Lessons
The Earth Observing System (EOS) Program

NASA’s Earth Observing System (EOS) Program

- centerpiece of NASA’s Earth Science Division and largest component in the U.S Global Change Research Program
- space-based observing system
- ground-based field campaigns

EOS Data and Information System

- objective is to enable quick and easy access to data about the earth
- data available from 12 Data Centers, organized by scientific discipline
- ORNL DAAC focuses on terrestrial biogeochemistry and ecological dynamics data
ORNL DAAC

• Archive data products produced by projects within NASA’s Terrestrial Ecology Program

• Mission:

assemble, distribute, and provide data services for a comprehensive archive of terrestrial biogeochemistry and ecological dynamics observations and models to facilitate research, education, and decision-making in support of NASA’s Earth science.

ORNL DAAC: daac.ornl.gov
ORNL DAAC: Data Holdings

1. Field Campaigns (738)
   - FIFE
   - OTTER
   - SNF
   - BOREAS
   - LBA
   - NACP

2. Validation of Land Products (23)
   - Land Validation
   - MODIS Subsets
   - FLUXNET
   - NPP
   - BigFoot

3. Regional and Global Studies (198)
   - Climate
   - Soils
   - Vegetation
   - Hydroclimatology
   - Daymet

4. Model Products (10)
   - Benchmark Models
     - IBIS, BIOME-BGC, LSM
   - Manuscript Models
     - PNeT, Century, Biome-BGC

Total Data Sets = 969
ORNL DAAC by the numbers

- Data cost: Zero (free, unrestricted)
- Staff: 10 full-time equivalents
  - Environmental scientists, computer scientists, database managers, and GIS specialists
- Data Holdings: 969 data sets; ~0.5 Tb volume
- Data orders:
  - FY 2011: 1,300,000 products ordered by 16,500 users
  - FY 2012 (first half): Nearly 450,000 products distributed to 11,000 users
- ~30% are foreign and ~30% have numeric IP
  - Remaining user domains are commercial, educational, and government
- Data Product Use: 140 data products cited in the literature in 2011
- Distributing data for 18 years (began in 1994)
Data Flow -
From Investigator to the Archive

Investigators
- Collect
- Document
- QA / QC
- Analyze
- Publish

Project
- Policies
- Protocols
- QA Review
- Generate standard products
- Integration and synthesis

DAAC
- QA Review
- Document
- Archive
- Distribute
- User Support

Project

Data Flow -
From Investigator to the Archive

Investigators
- Collect
- Document
- QA / QC
- Analyze
- Publish

Project
- Policies
- Protocols
- QA Review
- Generate standard products
- Integration and synthesis

DAAC
- QA Review
- Document
- Archive
- Distribute
- User Support

Project

Finalized Project Data Sets
Proper Curation Enables Data Reuse

20-Year Rule
NRC 1991

The metadata accompanying a data set should be written for a user 20 years into the future.
Provide Guidance:
“Best Practices for Preparing Ecological Data to Share”

• Cook et al. 2001 *Bulletin ESA* 82: 138 – 141

• Best Practices include:
  1. Assign Descriptive File Names
  2. Use Consistent and Stable File Formats
  3. Define the Parameters
  4. Use Consistent Data Organization
  5. Perform Basic Quality Assurance
  6. Assign Descriptive Data Set Titles
  7. Provide Documentation

• Update on-line:

• Workshops at ESA 2010 & 2011 (with DataONE)
• Workshops at AGU 2010 and 2011 with ESIP
• Workshop at AMS 2012 with ESIP
ORNL DAAC: Curation and Archive Functions

- **Acquisition**
  - identify how best to serve the scientific community
  - establish how and when to receive data

- **Ingest**
  - perform QA checks
  - compile project-provided metadata
  - convert to archivable file formats

- **Enhance** (as requested)
  - convert to standard formats & units
  - aggregate files

- **Metadata / Documentation**
  - Prepare final metadata record and documentation

- **Archive / Publish**
  - generate citation

- **Exploration and Distribution**
  - provide tools to explore, access, and extract data for users worldwide

- **Post-Project Data Support**
  - advertise data
  - serve as a buffer between end users and PIs
  - provide usage / citation statistics

- **Stewardship**
  - provide long-term secure archiving of the data
  - security, disaster recovery
  - migration to new computer systems

- **User Working Group**
  - science panel that provides advice on all aspects of the ORNL DAAC
ORNL DAAC: User Working Group

- Working Group Includes NASA sponsors and DAAC leaders
- Board of Directors function
  - not FACA (Federal Advisory Committee Act)
- Serves a peer review function for an on-going project
  (began in 1992; began distributing data in 1994)
- Assist in defining the DAAC's science goals, setting priorities
- Represent the scientific interests of the research community
  - members are data providers and data users
- Driving force for DAAC evolution over the past decade
  - Provide guidance on DAAC activities, including data set acquisition, development of tools and services, incorporation of new technologies
Enable collectors and users to spend more time analyzing the data and less time doing data management.

Data Center Activities

- Training in Data Management
- Enhanced metadata entry tools
- Data Discovery and Access Tools

- Data Exploration tools
  - Visualize
  - Subset
  - Analyze
  - Integrate
- Spatial data analysis
User Tools

- Metadata Editor
- Search and Access
- Spatial Data Access Tool
- THREDDS Data Server
- MODIS Subsetting
Subsets of MODIS* remote sensing products for field studies

*MODIS = Moderate Resolution Imaging Spectroradiometer
MODIS Subsets:
Vegetation Index Interactive time series

Increasing Greeness

Demonstration
Data management coordination

Close coordination / communication among data managers, those making the measurements, and data users is critical.
Challenges

- Obtaining documented, organized, and high-quality data from field investigators
Lessons

• Data centers should be a partnership of data managers, data providers, and data users
  – Seek formal guidance from “User Working Group”

• Data centers should focus on use of the data

• Data centers should facilitate new science
  – Metric is citations to data sets

• Expect change
  – Funder and research community expectations and advances in information technologies demand flexibility
Web Resources

• ORNL DAAC
  http://daac.ornl.gov

• Search Tool / Metadata Clearinghouse
  http://mercury.ornl.gov/ornldaac/index.jsp

• MODIS Land Product Subsetting

• Spatial Data Access Tool
Additional Slides
Challenges

• Obtaining documented, organized, and high-quality data from field investigators

• Carrots
  – Providing training and tools
  – Work effectively
  – Citations & indices
  – Collaborations
  – Credit for future proposals

• Sticks
  – Funders look good
  – Can’t publish without archiving
  – Requirement for closing projects
  – Link to future funding
The details of the data ….

For those on Investigator’s team, amount of metadata required to understand the data is small
Metadata Needed to Understand Data: 20-year perspective

From Raymond McCord, ORNL