Federal Agency Repository Review

Ranjeet Devarakonda
For Suresh Vannan
ORNL DAAC Manager

Climate Change Science Institute
Oak Ridge National Laboratory (ORNL)
Oak Ridge, Tennessee, USA
Outline

• Overview of ORNL Federal repositories at CCSI/ESD
  – ORNL DAAC
  – ARM
  – CDIAC

• Summary of activities
  – Data characteristics
  – Data Curation
  – Staffing and Skills
  – Costs
  – Issues
The Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC) archives data produced by NASA’s Terrestrial Ecology Program in support of NASA’s Carbon Cycle and Ecosystems Focus Area.

http://daac.ornl.gov
- Atmospheric Radiation Measurement Archive
- DOE National user facility for global climate change research
- Data from routine operations and field experiments are stored and distributed through the Archive
- Primary focus is to understand radiative processes in the atmosphere, particularly the interactions between clouds and aerosols

**ARM Measurements**

https://www.arm.gov/

**Measurement Categories**

Select below to highlight measurements in specified measurement categories.

- **Aerosols**
  - microphysical and chemical properties
  - optical and radiative properties
- **Atmospheric Carbon**
- **Atmospheric State**
  - other trace gases
  - surface meteorology
  - upper air state

- **Cloud Properties**
- **Radiometric**
  - longwave broadband
  - longwave narrowband
  - longwave spectral
  - microwave
  - shortwave broadband
  - shortwave narrowband
  - shortwave spectral
- **Surface Properties**
  - surface flux
  - surface/subsurface state
The Carbon Dioxide Information Analysis Center, which includes the World Data Center for Atmospheric Trace Gases, has served as the primary climate-change data and information analysis center for the U.S. Department of Energy (DOE) since 1982.

CDIAC builds databases vital to climate change research

Most popular data holdings include

- Fossil-fuel CO₂ emissions
- AmeriFlux
- Jones/Hansen temp records

”Nature Climate Change Dec. 2011"
## ORNL Federal repositories

<table>
<thead>
<tr>
<th></th>
<th>ARM Archive</th>
<th>DAAC</th>
<th>CDIAC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sponsor</strong></td>
<td>DOE BER</td>
<td>NASA</td>
<td>DOE BER</td>
</tr>
<tr>
<td><strong>Type of data</strong></td>
<td>Atmospheric processes, cloud dynamics</td>
<td>Biogeochemical dynamics, carbon cycle, FLUXNET, Airborne</td>
<td>Atmospheric gases, emissions data, AmeriFlux</td>
</tr>
<tr>
<td><strong>Archive Size</strong></td>
<td>&gt; 550 TB</td>
<td>~ 200 TB</td>
<td>~ 500 GB</td>
</tr>
<tr>
<td><strong>Users/year</strong></td>
<td>~1500</td>
<td>~36,000</td>
<td>~350,000</td>
</tr>
<tr>
<td><strong>FTE</strong></td>
<td>~12</td>
<td>~16</td>
<td>~10</td>
</tr>
<tr>
<td><strong>Year Started</strong></td>
<td>1991</td>
<td>1994</td>
<td>1982</td>
</tr>
</tbody>
</table>

- Different sponsors, missions, and objectives
- Different relationships between data generators/providers and users
- Broad range of users: modelers, field scientists, educators, general public
- Unique tools, capabilities, and experience within and across the data centers
- ~40 staff/post-graduate students/subcontractors with over 300 years of environmental data management experience
Repositories Role – Acquisition to Support

- **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)
  - standardize to common units
  - aggregate files

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

- **Archive / Publish**
  - generate citation

- **Exploration and Distribution**
  - advertise data
  - provide tools to explore data holdings
  - provide data access to users worldwide

- **Post-Project Data Support**
  - serve as a buffer between end users and PIs
  - provide usage statistics
  - provide long-term secure archiving of the data

- **Stewardship**
  - security, disaster recovery
  - migration to new computer systems

- **User Working Group**
  - science panel that provides advice on all aspects of DAAC operations
Repositories Challenge Example (1/2) – Variety (ORNL DAAC)

- **1118** data sets
- Average data set Size = **485 MB** (Range 0.5 KB to > 10GB)
- Average Number of granules per dataset **241 granules** (Range 1 to > 6000)
- **5000+** Investigators
- **50+** file formats
From 1992 to 2011 the ARM database archived about 220 terabytes of data, and since 2011 the database has more than doubled that to 560 terabytes.

Each month ARM receives almost 40,000 gigabytes of data (shown in orange) and disseminates more than 50,000 gigabytes (shown in red).
Best practices Resources to assist Repositories and Data Providers

http://daac.ornl.gov/PI/pi_info.shtml

and more...

Guide to Best Practices for Ocean CO₂ Measurements

This manual should be cited as:
Best practices – Benefits to repositories and users

**Short-term**
- Spend less time doing data management and more time doing research
- Easier to prepare and use data
- Facilitate data synthesis across diverse data products
- Users can readily understand the data files

**Long-term**
- Ability to use data for addressing future problems
- Prevents loss of information
- Users can easily find, understand, and use the data to address broad questions
- Sponsors protect their investment
Common Issues

• Inconsistencies
  – Data files
  – Metadata
  – Documentation

• Data citation still not consistent

• Diverse user needs

• Constantly changing IT landscape

• Too many tools
Thanks!!

- Suresh Vannan: santhanavans@ornl.gov