The Research Data Archive at NCAR: A Metadata System that Enables Discovery Across a Diverse Archive

Robert Dattore and Steven Worley
National Center for Atmospheric Research
Boulder, CO, USA
Outline

- Introduction
- RDA - Then
- RDA - Now
- Data Discovery
Introduction

- **Purpose** - support climate & weather research at NCAR; services are extended worldwide as resources permit
- **Observations, derived products; focus on historical atmosphere/ocean data**
- **Metrics**
  - Established in 1960s
  - 600+ datasets, 4M files, 600 TB
  - 7000 users annually
Introduction

○ Changing data landscape
  ▪ **Then** – small datasets, single country/experiment, specialized formats
  ▪ **Now** – global coverage, high spatial/temporal resolutions, standard formats

○ Result and challenge:
  ▪ Lots of diversity
  ▪ How can we provide uniform discovery?
Then

Data

Metadata

Dataset Specialist

ftp Server

Data User
Then

- **Bottom line**
  - Increasing data diversity, evolving technology; difficult to develop good systematic discovery
  - README files, directory names
  - Primarily via personal communications

- **Major limiting factor**
  - Insufficient metadata
  - No metadata standard, dictionaries
  - Collection not uniform across all datasets
  - Rigidly-structured flat ASCII files
  - Archiving separate from metadata collection
Developed local standard for discovery based on DIF\(^1\) & THREDDS\(^2\); applied across all datasets

Adopted GCMD\(^3\) controlled vocabularies
  - Local enhancements; e.g. data formats

Harvest two types of file metadata
  - File attribute – name, size, compression, …
  - File content - variables, levels, date range, ...

Storage using XML

---

\(^1\)Directory Interchange Format, NASA/GCMD\(^3\) ; \(^2\)Thematic Realtime Environmental Distributed Data Services; \(^3\)Global Change Master Directory
Metadata Collection

- Data
  - Tools
    - Controlled Vocabularies
  - GUI
- Specialist
- Web Server
- MySQL
- X509
- Data User
Metadata Collection

- Tools that automatically capture file metadata
  - Integrated with archiving activities

- Web-based GUI - guided entry of dataset discovery metadata
  - Required fields, constrained entries
Relational Databases
Relational Databases

- Fast access
- Dataset discovery metadata
  - Single database (~0.3M rows)
- File attribute metadata
  - Single database (~45M rows)
  - Maintains dataset/data file relationships
- File content metadata
  - Four databases structured to handle diversity of data (~920M rows)
  - Maintains detailed parameter relationships

All together, support accurate data discovery
Data Discovery

Diagram showing the components and flow of data discovery:

- **Data**
- **Tools**
- **Controlled Vocabularies**
- **GUI**
- **Metadata**
- **MySQL**
- **Web Server**
- **Data User**
- **Specialist**

The diagram illustrates the relationship between data, metadata, and tools, showing how data is processed and accessed through a web server.
So far, you have selected: **Topic > Atmosphere (481) > VARIABLE > Cloud Base (27)**

Select two datasets and **Compare** them.

**Dataset Information:** 
- Summaries
- Temporal Ranges

1. **Hourly Surface Observations for Brazilian Stations, 1951Jan-1980Dec**
   - This set contains hourly surface data from several Brazilian stations. Reported variables include wind information, pressures, and precipitation.

2. **NCDC TD9734 Typical Meteorological Year, Solar and Surface**
   - This dataset contains daily surface data which represent the "typical meteorological year," solar radiation, ... [more information]

3. **NCDC SOLMET TD9724 Solar and Surface Observations, daily 1952-1976**
   - Users may want to check out more recent similar sets such as: [link](http://rredc.nrel.gov)

4. **NCDC TD14 and TD3280 U.S. Surface Airway Hourly Observations, 1938**
   - This data set, from the National Climatic Data Center (NCDC), contains hourly or 3-hourly observations. The stations are operated by ... [more information]

5. **Canadian Surface Observations, daily 1963Jan-1976Dec, 95 stations**
   - This dataset contains daily surface observations from ninety-five Canadian stations.

   - Global surface observations from the U.S.A.F. for February 1967 to April 1967 and ... [more information]

   - [More information]
The Internet Download file list was created from the following selections:

- **Data Format:** WMO GRIB2
- **Valid Date Range:** 1979-01-01 00:00 to 2010-01-01 00:00
- **Parameter(s):**
  - u-component of wind
  - v-component of wind
- **Vertical Level:** Isobaric surface: 500 mbar
- **Type of Product:** Analysis
- **Grid:** 0.5° x 0.5°, 0E to 359.5E and 90N to 90S (720 x 361 Longitude/Latitude)

If you change any of these selections, be sure to update the list.

### Displaying 1 to 372 of 372 files

To get data files, do one of the following:

- Select two or more files and download them as a single Unix tar file
- Select one or more files and create a Unix csh script to read them from the Internet using Wget or curl
- Click the individual filename links to download files one-at-a-time

*curl* can get specific records from data files, so if you have selected this option for your read script, you will only get records from the file list, unless you turn off curl subsetting. If a file is not marked, it is not eligible for subsetting and you will get the full file regardless.

<table>
<thead>
<tr>
<th>Filename</th>
<th>Data Format</th>
<th>File Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>wnd500.gdas.197901.grb2</td>
<td>WMO GRIB2</td>
<td>1488 Grids</td>
</tr>
<tr>
<td>wnd500.gdas.197902.grb2</td>
<td>WMO GRIB2</td>
<td>1344 Grids</td>
</tr>
<tr>
<td>wnd500.gdas.197903.grb2</td>
<td>WMO GRIB2</td>
<td>1488 Grids</td>
</tr>
<tr>
<td>wnd500.gdas.197904.grb2</td>
<td>WMO GRIB2</td>
<td>1440 Grids</td>
</tr>
<tr>
<td>wnd500.gdas.197905.grb2</td>
<td>WMO GRIB2</td>
<td>1488 Grids</td>
</tr>
</tbody>
</table>
Metadata Sharing

- OAI-PMH
  - UCAR Community Data Portal (THREDDS)
  - Global Change Master Directory (DIF)
  - also Dublin Core, native
  - easy to add others as necessary
Thank You!

- Web: http://dss.ucar.edu
- Email: dssweb@ucar.edu
- Questions/comments?