Comprehensive Cross-organizational Data Management

NCAR Executive Committee
April 24, 2014

Contributors:
Matt Mayernik – IIS
Eric Nienhouse – CISL
Steve Williams – EOL
Steve Worley - CISL
Vision

data.ncar.edu
data.ucar.edu

Single front door to **ALL** data, software, data services
Working Definition

Data

Digital assets intended for scientific community use, including files and metadata, publications, reports, images, software (visualization, analysis, model codes), and related data services.
Current Status

Users and colleagues “in the know” are well served

• Know websites, who to contact, what services exist

New and diverse community of users is easily frustrated

• Many services sprinkled across the organization
• Not well coordinated, no overarching consulting
• Current websites are not comprehensive
Data Discovery Guided by Experts

Search and access 164 data sets covering the Atmosphere, Ocean, Land and more. Explore climate indices, reanalyses and satellite data and understand their application to climate model metrics. This is the only data portal that combines data discovery, metadata, figures and world-class expertise on the strengths, limitations and applications of climate data. Discover it now.

See data pages with guidance from these experts:

Dai, Aiguo
Norris, Joel Willis, Josh Meier, Walter
Dee, Dick
von Schuckmann, Karina
Banzon, Viva
Randel, Bill Arkin, Phil
Vicente-Serrano, Sergio M. Tian, Baijun Kimball, John

Data Set Overviews

Compare the attributes, strengths and limitations of multiple data sets.

Atmospheric Reanalysis: Overview & Comparison Tables
Global Temperature Data Sets: Overview & Comparison Table
Precipitation Data Sets: Overview & Comparison table
Sea Ice Concentration data: Overview, Comparison table and graphs
Why Should We Change?
User’s perspective

With a top-level unified presentation we can:

• Simplify user data discovery
  – Presents a collaborative organization

• Share metadata with outside entities
  – Federate data discovery with others (e.g. NASA, EarthCube, etc.)

• Provide organization-wide consulting
  – Mitigate challenges like:
    • “I think the data is at NCAR, but I cannot find it”
    • “I got lost looking for the data at NCAR”
    • “I don’t know who to ask”
Why Should We Change?
Organization perspective

With a unified organization we can:

• Gain functional efficiency
  – Cross-organizational interconnectivity will be automated
  – Improve data services by:
    • Sharing data management expertise
    • Sharing software expertise
    • Sharing infrastructure where appropriate
    • Reducing possible duplication of effort

Cost Savings?
Why Should We Change?
Policy and Political perspective

Forcing factors and opportunities:

• Federal
  • 2014 White House “Climate Data Initiative”²
• NSF
  • 2011 NSF data management plan requirement
  • EarthCube: Funding and collaboration opportunities³

¹ [http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research](http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research)
³ [http://earthcube.org/](http://earthcube.org/)
Why Should We Change?
Policy and Political perspective, continued

Forcing factors and opportunities:

Publication Community

AGU: 2013 Publications Data Policy\(^1\)
AMS: 2013 Statement on “Full and Open Access to Data”\(^2\)

Internal, UCAR commitment for “Publication and Information Dissemination”\(^3\)

“….. UCAR supports an open exchange of data and scholarly information derived form our research. It is UCAR’s policy to share this scientific and technical information with the community…..”

\(^{1}\) http://publications.agu.org/author-resource-center/publication-policies/data-policy/
\(^{2}\) http://www.ametsoc.org/policy/2013fullopenaccessdata_amsstatement.html
\(^{3}\) https://www2.fin.ucar.edu/policies/3-5-publication-information-dissemination
Successful Past DM Initiatives

• Community Data Portal (circa 2002)
  – Many lessons learned
  – Not all inclusive
  – Needs a technology refresh
    • Non-centralized approach

• Data Citation Working Group (2011-ongoing)
  • NCAR Technical Note, (Mayernik et al., 2012),
    http://dx.doi.org/10.5065/D6ZC80VN
  • Established the process for placing DOIs on data, software, and services
Why Now? - Internal

• Stakeholder surveys
  • UCAR/NCAR 2013 Survey, Data Users Sub-group, Verbatim Responses¹
    • Generally satisfied
      • “We need your data”, “get access to scientific data”, “thanks for your data sharing which is very helpful in my research work”, “the main priority should be gathering global high-resolution data”, “more data and tutorials”, “more convenient to download data”, etc.
  • Data management survey to support ITC Planning (May 2013)²
    • Internal, 121 responses (92% NCAR, 8% UCP)
    • Common challenges: insufficient storage space and data organization, lack of funding and expertise
    • Common interests: backup and repository services

¹ Files “UCAR – Subgroup Tables.xlsx”, “UCAR – Verbatim Response.xlsx”
Why Now? – Internal, continued

• **Strategic plans**
  – New NCAR, ongoing UCAR Strategic Plans
  – Completed, UCAR ITC Strategic Plan

• Data Services Working Group
  Worley (CISL), Mayernik (IIS), Wright (IIS), Williams (EOL), Strand (CGD), Schmitt (HAO), Nienhouse (CISL), Keene (MMM), Hermida (Unidata)

• Good organizational review, recommendations, implementation suggestions, some resource estimates
Strategies

1. **Best practice guidelines** for data and software management
2. Data management plan advice and assistance
3. **Archiving and access systems** for data and software that need, but do not currently have, repositories
4. **Unified system for discovery** of UCAR resources
5. Develop appropriate digital preservation solutions
6. Facilitate collaboration between data and software staff
7. Prepare systems for external integration opportunities
8. Make data open and machine-readable for external applications
9. Structure IT to meet 24x7 needs where applicable
10. Tighter coordination of UCAR administrative databases
1. Guidelines

- As a service, create and maintain best practice data and software management guidelines.
  - Too many *ad hoc* systems
  - Complete systems: standards-based, sustainable, cost effective
- Understand Principles:
  - Stewardship, data lifecycle
- Make us recognized data management experts among peers
3. Archiving and Access

*Create, adapt or identify an archiving and access system for research data and software that need, but do not currently have, sufficient, secure and publicly accessible repositories.*

- Too many orphan datasets – not in managed repositories
- Improved archiving and access supports tightly linking scholarly publications and data
- Long-term data preservation is essential
  - Secures the intellectual properties and scientific findings of the organization
4. Discovery

Create and maintain a unified and flexible system for discovery of UCAR publications, data, software, and services.

- Past, did well with centralized method for data (Community Data Portal), 10-yr old effort
- Need a new approach:
  - Expand, more data, publications, more software, and services
  - Convert to a distributed method
  - Add richness to the metadata standard
- Sustainable!
Where to Now?

Should we take on this challenge?

data.ncar.edu

data.ucar.edu

Single front door to **ALL** data, software, data services
Not a new idea

NASA DAACs -> ECHO/Reverb
https://earthdata.nasa.gov/echo

NOAA
https://data.noaa.gov/dataset
(very immature state)

USGS
http://data.usgs.gov/datacatalog/

These services are driving user expectations for similar services at UCAR/NCAR
NCAR Strategic Context

Contributions to the Plan Imperatives

• Develop, maintain and deploy advanced observational facilities and services
  • More comprehensive and broadly accessible data and services are necessary
• Develop, deliver and support a suite of advanced community models
  • Requires coordinated access, consulting, advanced analysis tools, and data delivery
• Develop and sustain advanced information and computing system services
  • Requires support for robust IT systems and coupling of expertise with scientific research
• Develop and transfer science to meet societal needs
  • Expertly done data services for the scientific community are the foundation to develop knowledge transfer to society

First Steps for Success
A sketch: best done with a cross-organizational team

• All-in participation across the organization
  • Realize the UCAR “open data” policy to the largest possible extent
  • Support access with methods and tools to make it easy

• Make a plan to scope all the data assets
  • Design a survey that IDs the assets and user expectations (current and forward looking)
  • Do the survey

• Evaluate the Survey
  • Create an inventory
  • Identify user requirements for data services
First Steps for Success, continued

• **Assess feasibility across the organization**
  – Define scope and effort required
  – Define processes and responsibilities for user consulting
  – Define standards for metadata, data access approaches, and interoperability

• **Leverage and participate in community standards and federation efforts**
  – E.G. EarthCube, ESG/ESGF\(^1\), ESIP\(^2\)
  – Best done from complete cross-organizational plan

\(^1\) [http://esgf.org/](http://esgf.org/)
\(^2\) [http://commons.esipfed.org/](http://commons.esipfed.org/)
To Start: Setting the Table

• **All-in participation across the organization**
  • Form Data Stewardship Engineering Team (DSET)
  • DSET membership
    • 1 person minimum, 2 person maximum from Labs or Sub-Labs
    • Qualified to represent *all* the data assets for the science community from their organizational entity

• **Kick-off meeting**
  • Give this presentation again
  • Clarify vision in the context of engineering
  • Discuss how to: Make a plan to scope *all* the data assets
  • Vet and refine the survey idea
  • Establish DSET leadership, meeting schedule, and communications
To Start: Setting the Table, continued

• **Intra-Lab meetings**
  • DSET representatives - hold informational and fact gathering meetings with Lab data asset owners
  • Foundation for survey questions regarding data assets and user services
  • How many meetings?

• **DSET meetings**
  • Monthly to begin, possibly less frequent later
  • Consolidate information from the whole organization
  • Design, execute, and evaluate the survey
  • Assess feasibility across the organization
    • Scope, effort, processes, responsibilities, standards, etc.
To Start: Setting the Table, continued

• **DSET report out on feasibility**
  • Representatives to their respective Directors
  • DSET to the EC?

• **Timetable**
  • Difficult to predict, easier after a DSET meeting or two
  • Optimistic 6 months, hopefully not more than 1 year
Questions?

Comprehensive Cross-organizational Data Management

Contributors:
Matt Mayernik – IIS
Eric Nienhouse – CISL
Steve Williams – EOL
Steve Worley - CISL
DCERC data project take away ideas

• Get all the right people involved
• Recognize peer developments
• Establish goals based on user needs
  – Select standards and IT
• Build success with well ordered steps
  – Sustainability (flexibility, expandability)
• Communicate findings