U.S. DEPARTMENT OF ENERGY | Office of Science

EESSD - ESS Cyberinfrastructure Working Groups

Data Management

Leads: Danielle Christianson (LNBL) Terri Velliquette (ORNL)

Current Members: Deb Agarwal, Kristin Boye, Bob Bolton, Shreyas Cholia, Joan Damerow, Amy Goldman, Paul Hanson, Val Hendrix, Trevor Keenan, Amy Macpherson, Lee Ann McCue, Dave Millard, Eric Pierce, Gilberto Pastorello, Stephanie Pennington, Alistair Rogers, Daniel Ricciuto, James Stegen, Charu Varadharajan, Roelof Versteeg, Pamela Weisenhorn, Ken Williams, Chongang Xu

Data Management WG Scope & Activities

Management and Archival of DOE climate and environmental datasets

- Data **Preservation, Sharing, and Publication**
- Common Data and Metadata **Standards**
- Data **Citation and Attribution**
- Data **Federation** across different data catalogs

Data **Synthesis** across ESS and other relevant Datasets

Development of common **Tools** for data usage

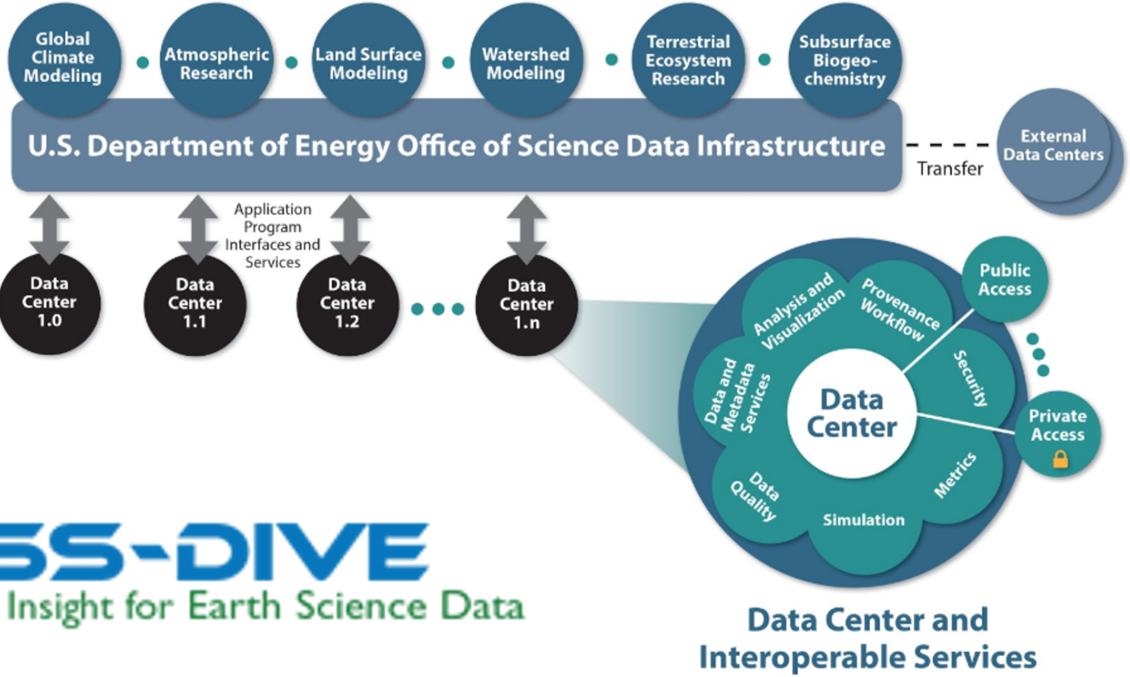
QA/QC, processing, analysis, mining and visualization data to prepare them for use in new research projects.

Past year major activities: Support ESS-DIVE design and metadata development

Phase 1 (0-2 yrs): ESS Data Center

ESS-DIVE delivered:

- API ingest service
- Data archiving
- Publication tools (DOI)
- Data Discovery
- Data Retrieval
- User support



Data Center and Interoperable Services

ESS-DIVE
Deep Insight for Earth Science Data

Phase 2 (2-5 yrs): Community tools for data-model integration

Previous goals w.r.t. Data Management

- Expand upon data ingest and archiving
- Expand upon initial data discovery and retrieval capabilities

Capabilities should be compatible with small to large compute clusters

Phase 2 (2-5 yrs): Community tools for data-model integration

Previous goals w.r.t. Data Management

- Expand upon data ingest and archiving
- Expand upon initial data discovery and retrieval capabilities
- New technologies that provide novel modes of data sharing
- New methods for data delivery

Capabilities should be compatible with small to large compute clusters

Phase 2 (2-5 yrs): Community tools for data-model integration

Previous goals w.r.t. Data Management

- Expand upon data ingest and archiving
- Expand upon initial data discovery and retrieval capabilities
- New technologies that provide novel modes of data sharing
- New methods for data delivery
- Hardware and software to support data analytics and processing
- Improved geospatial data analysis and query

Capabilities should be compatible with small to large compute clusters

Phase 2 (2-5 yrs): Community tools for data-model integration

Previous goals w.r.t. Data Management

- Expand upon data ingest and archiving
- Expand upon initial data discovery and retrieval capabilities
- New technologies that provide novel modes of data sharing
- New methods for data delivery
- Hardware and software to support data analytics and processing
- Improved geospatial data analysis and query
- Integration with DOE-ASCR resources
- Work across external data facilities (DAAC, NOAA)

Capabilities should be compatible with small to large compute clusters

COV Report: Better data integration across CESD

“... develop a strategic plan for harmonizing its (CESD’s) **data collection, archiving, and data access/manipulation capabilities**. An effective plan may include:

- best practices guidelines,
- archiving procedures,
- standards for data longevity and access, and
- co-location of data and computational resources required to create a new environment for **machine learning**.”

“...develop a plan to assess how **different data archives**, including ESS-DIVE, the ARM archives, and others, may be **integrated**.”

Hurdles and Dependencies

- What tasks or projects are taking longer than predicted to complete, make progress, or initiate?
    - Can you identify the problem, hurdle, or blocker?

- What next steps or future goals depend on tasks or projects that are outstanding or planned?

What is the updated vision for strategic plan?

- What is ongoing from 0-2 year Data Center goals?
- What continued improvements are needed for data discovery?

What is the updated vision for strategic plan?

- What is ongoing from 0-2 year Data Center goals?
- What continued improvements are needed for data discovery?
- What does data integration mean to the community? Machine-learning?
- What does visualization mean to the community and what are the needs?
- Getting data to compute resources - virtual services

What is the updated vision for strategic plan?

- What is ongoing from 0-2 year Data Center goals?
- What continued improvements are needed for data discovery?
- What does data integration mean to the community? Machine-learning?
- What does visualization mean to the community and what are the needs?
- Getting data to compute resources - virtual services
- What is the plan for historical data?
- What is the process for collaboration across DOE (and with external partners)?

Contribute to DM priorities | Collect existing resources

Contribute your priorities | Link resources to relevant needs
Google doc: ESS-CI DMWG ChallengesResources Discussion

Plan:

- Collect challenges / resources in Google doc (send by Thurs May 14)
- **May 14, 1:30-2:30pm EDT (10:30-11:30am PDT)**: Start prioritizing challenges
- May 18: Breakout session to
  - Review / modify / prioritize challenges
  - Flesh out details / resources
  - Identify writing teams / schedule

Contribute to DM priorities | Collect existing resources

Topics in the Google doc for Thursday's Meeting

Answer these questions about the 2016 Data Management 2-5 year goals (slide 7)

1. Are these goals still relevant?
2. Is the goal a priority? (critically important, important, minimally important)
3. Are there additional goals we might consider?

Provide information on relevant projects and resources

- Name
- Institution
- Projects
- Related DM activities / capabilities
- Hurdles / Challenges