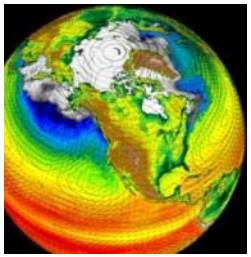
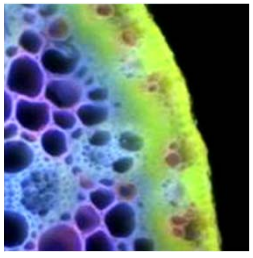
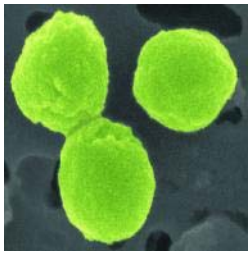


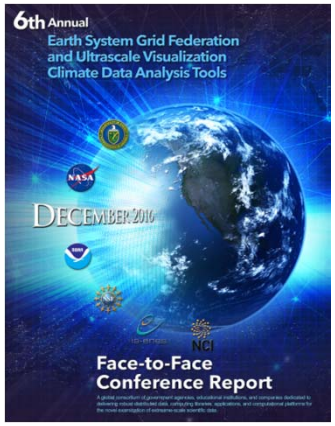



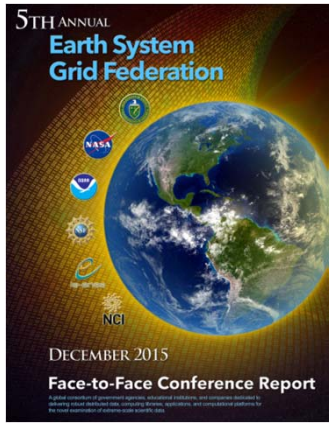
Workshop brief: Virtual Data Integration

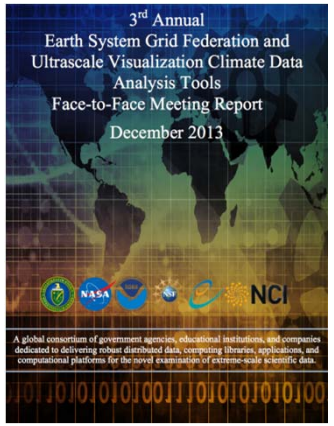
- **Justin Jay Hnilo (Data Management)**

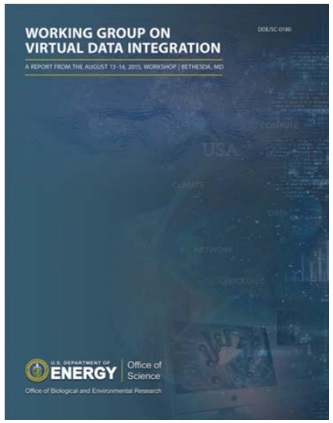
- 23 March 2016

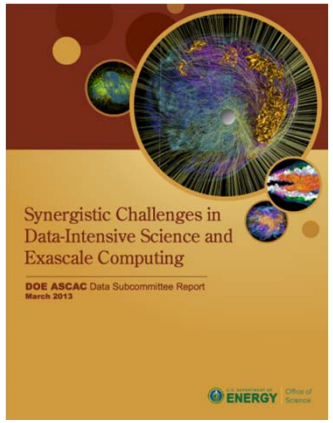
Data workshop, conference reports: community involvement and outreach, they define our next steps

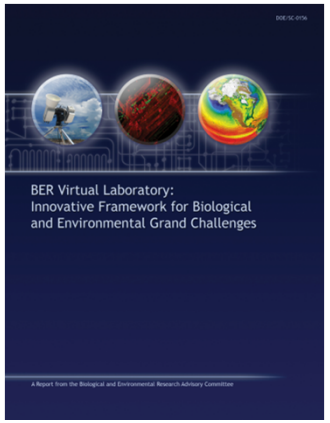


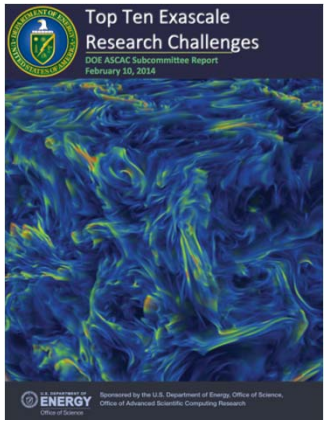


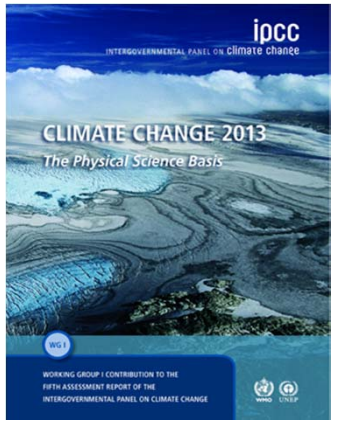












DOE, BER, CESD, and community workshop and conference reports: http://esgf.llnl.gov/reports.html and http://science.energy.gov/ber/community-resources/.

Workshop Goal:

What are the current and future data infrastructure requirements foundational for achieving CESD scientific mission goals in advancing a robust, predictive understanding of Earth's climate and environmental systems?

Requirements to Achieve BER's Vision of a Virtual Laboratory: A Next-Generation Data Infrastructure for Climate Science.

A workshop was held August 13-14, 2016, in Bethesda, MD

Co-chairs:
Dean Williams (LLNL)
Giriprakash Palanisamy (ORNL)
Kerstin Kleese van Dam (PNNL)

24 DOE scientists participated in the workshop and several Program Managers were in attendance.

The report is published and the link is:
http://science.energy.gov/~/media/ber/pdf/workshop%20reports/Virtual_Data_Integration_workshop_report.pdf

Workshop Format: Survey and Findings, Breakout sessions

In advance of our August meeting we designed and emailed out a survey that asked questions of what kind of user they are and what are their concerns for advancing their research.

The survey itself is within the workshop report as an appendix

75 responses were received.

These responses, were aggregated and grouped. We used these groupings as subjects for further discussion during the two day meeting itself.

Survey Participants

Table 1. Self-Identification Categories for the 75 Scientists Who Responded to the Survey Request

Scientific Background	Description	Total
Data provider	Provides data and metadata (describing the data) to the community. Also responsible for data quality. Associated with climate modeling groups and data centers.	32
Resource provider	Provides hardware and software resources at high-performance computing facilities.	4
Software developer	Develops stand-alone software for the climate community. Also known as a computer programmer, application developer, and system software developer.	6
Climate modeler	Develops quantitative methods to simulate the interactions among the important drivers of Earth's climate, such as atmosphere, oceans, land, and sea ice.	15
Climate model data analyst	Analyzes output to understand simulation and observational output for knowledge discovery and change.	18
Total		**75**

Survey Results
(highest importance is rated 5)

Table 2. Top 10 Needs Identified by the Survey	
Survey Question	**Average Rating or Percentage in Highest Need Category**
An easy way to publish and archive data using one of the DOE data centers	4.79
A means for comparing diverse data types generated from observation and simulation	4.71
User support for data access and usage	4.64
Access to sufficient observational and experimental resources	4.58
Access to enough computational and storage resources	4.52 / 41%
Method of ingesting and accessing large volumes of scientific data (e.g., from a data archive to supercomputer)	4.49 / 39%
Quality control algorithms for data	4.46 / 31%
A unified and single user account to access all BER and ASCR resources	4.44 / 38%
Reliability and resiliency of resources	34%
In situ analysis of observational, experimental, and computational results: the ability to interpret results and verify new insights within the context of existing scientific knowledge	4.40

Survey Results
(highest importance is rated 5)

Table 3. Top Needs Identified by Survey Respondents	
Survey Questions	**Average Rating**
KD: Method of ingesting and accessing large volumes of scientific data (e.g., from a data archive to supercomputer)	4.49
KD: Quality control algorithms for data	4.46
KD: Interfaces that ensure a high degree of interoperability for different formats and semantic levels among repositories and applications	4.18
KD: Capture of provenance information for data	4.11
KD: Reproducibility	4.06
HCI: Collaborative environments	4.31
HCI: Improved user interface design	4.00

KD = Knowledge gathering, managing and sharing
HCI = Human-computer interaction

Next Steps: Investments in ESGF that follow the report findings

Findings	Description
Data Quality	ESGF data quality persists in the form of provenance, quality control (QC) checks, errata, and data citations. Various components help to improve data quality checks in the ESGF publishing process.
Data Compression	Data compression is important to ESGF in terms of data storage and transfers. Because of the sheer size of ESGF archives, compressing data for storage or transfer considerably reduces overall costs.
Data Storage	The sheer size of current and expected future archives makes storage a difficult issue to address. If the expected storage for CMIP6 is over 20 petabytes (with estimates as high as 50 PB), then a uniform storage strategy must be put into place among the major CMIP data center sites.
Hardware	A cost-benefit analysis is needed for long-term storage. For example, what would it cost to regenerate versus store the data?
Network	ESGF requires the ability to control the timing of data- and network-intensive replication operations for large climate data sets.
Operations	Operational support is needed to sustain the numerous ESGF nodes operated by simulation, observation, and reanalysis projects.
Performance Metrics	Performance metrics must be included as part of ESGF operations. The goal is to have displayable and well-understood performance metrics to track and monitor the overall system and to gather data transfer performance metrics among major CMIP data center sites.
Provenance Capture	Provenance capturing is necessary for reproducing complex analysis processes at various levels of detail in a shared environment.
Server-Side Analysis (and Derived Data Sets)	The size of some data sets makes moving most of the needed data to the end user's home institution infeasible. Data analysis therefore must be performed remotely.
Software Security Scans	The latest software security breach has necessitated an inventory of all software in the ESGF software stack, and ESGF developers have coordinated component development to combine and share information about existing vulnerabilities that may affect secure ESGF operations.
Training and Documentation	Training is important to ensure proper data use and dissemination.
Use Metrics	Use metrics help projects know how the community is using their hardware, software, network, data, and other resources. Metric information such as number of users will serve as base metrics for various data and services within ESGF.

Next Steps: Strategic Computational and Storage Investments

Data Analytics

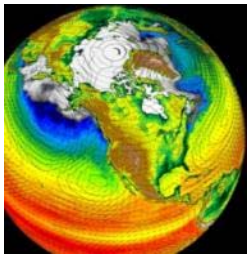
- Leverage existing and future DOE leadership class facilities
- Integration of server-side computing into ESGF
  - Tool development utilizing this API will empower a broad spectrum of environmental data inclusion.
- Implement a visualization/analysis platform (UV-CDAT)

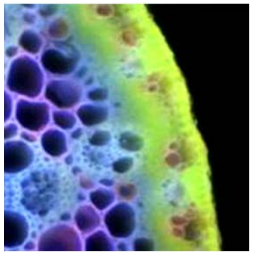
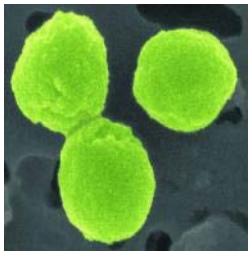
Disc Space

- Currently enhancing the volume of spinning disc space at ESGF node sites.

Timeline

- 2016 – Finalize strategic implementation plan
- 2016 - 2017 – efforts will emphasize the methods by which data is published into ESGF.
- 2016 - 2017 – work with the community to establish metadata connections for QA/QC, with emphasis on CDIAC and ARM data, within ESGF.
- 2017 – Continue infrastructure storage build-out at appropriate ESGF nodes.



Thank you

http://science.energy.gov/~/media/ber/berac/pdf/20130221/BERACVirtualLaboratory_Feb-18-2013.pdf

http://science.energy.gov/~/media/ber/pdf/workshop%20reports/Virtual_Data_Integration_workshop_report.pdf

justin.hnilo@science.doe.gov