A Strategic Data Roadmap for CESD/BER

• Justin Jay Hnilo
• Data Informatics Program Manager CESD/BER/SC
• May 2014
What’s this all about?
It’s about how we do research, today..

Data formats (or standards) are like toothbrushes, everyone has one but no one wants to use yours.

Data are like toothbrushes, everyone has some but, no one wants to share.

*Toothbrushes.. Who knew?*

*Scientific advancement is limited by this approach*
Scientific Drivers/Grand Challenges: BER

Today’s predictive capabilities are fragmented across many disciplines and system components. Such segmentation leads to uncertainties in how these coupled systems interact across many space and time scales. These uncertainties act to complicate our ability to sustainably manage and mitigate energy and environmental challenges.

We must understand complex systems science across scales, leveraging and expanding multidisciplinary research, advancing dynamically adaptable computing and mathematical capabilities and assessing human impacts on the Earth system. What we need to develop are efficient capabilities that enable a predictive understanding of complex, multi-scale, coupled and biologically based environmental systems behavior.

We envision a transformative knowledge based predictive modeling environment (a virtual lab). It is an environment that describes a continuum of events occurring from the molecular to the regional and global scales. The development of such would entail the unprecedented integration of in-situ hypothesis-driven experimentation and observations; technological advances across multiple scales of space, time and system organization; and major advances in theory, mathematics and computation associated with multi-scale models.

BER Virtual Laboratory:
Innovative Framework for Biological and Environmental Grand Challenges (2013)
Grand Challenges for Biological and Environmental Research: A Long-Term Vision (2010)
Scientific Drivers/Grand Challenges: Data

Rapid advances in experimental, sensor, computational technologies, and techniques are driving exponential growth in the volume, acquisition rate, variety, and complexity of scientific data. This new wealth of scientifically meaningful data has tremendous potential for scientific discovery.

However, to achieve scientific breakthroughs, these data must be exploitable—they must be analyzed effectively and efficiently, and the results shared and communicated easily within the wider Department of Energy’s Biological and Environmental Research (BER) Climate and Environmental Sciences Division (CESD) community.

The explosion in data complexity and scale makes these tasks exceedingly difficult to achieve—particularly given that an increasing number of disciplines are working across techniques, integrating simulation and experimental or observational results. Consequently, we need new approaches to data management, analysis, and visualization that provide research teams with easy-to-use and scalable end-to-end solutions.

These solutions must facilitate (and where feasible, automate) every stage in the data lifecycle, from collection to management, annotation, sharing, discovery, analysis, and visualization. In addition, core functionalities required are the same between different climate science communities but require customization to adapt to their specific needs and fit into their research and analysis workflows.

The mission of CESD’s Data and Informatics Program is to integrate all existing and future distributed CESD data holdings in a seamless and unified environment for the acceleration of Earth System Science.
Where do we start?

First step is to define and build a federated system of specialized nodes and ontologies that will allow for data integration.
How are we going to do this?

Leverage the success of the ESGF distributed data archival and retrieval system

- Distributed and federated architecture
- Support discipline specific portals
- Support browser-based and direct client access
- Single Sign-on to access all data
- Automated script and GUI-based publication tools
- Full support for data aggregations
  - A collection of files, usually ordered by simulation time, that can be treated as a single file for purposes of data access, computation, and visualization
- User notification service
  - Users can choose to be notified when a data set has been modified
Integrated Data Ecosystem

Collection and Data Management
- Sensors, field and lab experiments
- Data Models
- Transport and Communications
- Data Quality and uncertainty
- Storage, provenance and discovery

Data-intensive Computing
- Architectures – persistent data to streams
- Programming environments
- Human Computer Interface

Pattern Discovery
- Descriptive statistics
- Graph analytics
- Machine Learning
- Signal and image processing

Predictive Models
- Statistical prediction, classification, and anomaly detection
- Steering discrete-events and continuous simulations

Decisions and control
Design optimization
Policy making (Humans)
Enabling Integrated Earth System Research

DOE SC Data Infrastructure

Global Climate Modeling • Atmospheric Research • Land Surface Modeling • Watershed Modeling • Terrestrial Ecosystem • Subsurface Biogeo-Chemistry

1.0 Data Center • 1.1 Data Center • 1.2 Data Center • 1.n Data Center

API’s and Services

External Data Centers

Virtual Laboratory Infrastructure

An Integrated Cyber-infrastructure leveraging core Office of Science resources to enable discovery, analytics, simulation, and knowledge innovation

Data Center & Interoperable Services
ARM and CDIAC Data and Metadata Publication into the Earth Systems Grid Federation and Other Data Gateways

- **Science and Research User Community**
  - Search and access to data, plots, and quality information

- **Other Data Gateways**

- **ARM and CDIAC Data Centers**
  - Standards Based Science Metadata
  - Standards Based Data Access for Extraction, Visualization, and Analysis

- **Earth System Grid Federation Gateway**

- **Integrated Data Analysis and Visualization**

- **Custom Module**

- **Metadata Harvesting**
  - Data Access


justin.hnilo@science.doe.gov
Thank You