



BERAC Subcommittee Update – Unified Data Infrastructure

Kerstin Kleese van Dam

4/21/2022



Subcommittee Charge

The 2017 Grand Challenges report and the 2018 Scientific User Research Facilities report from BERAC recommended developing more consistent, integrated, and distributable data across the BER programs.

- (1) review the existing and anticipated capabilities in data management and supporting infrastructures that are relevant to the breadth of BER science
- (2) recommend a strategy for the next generation data management and analysis within a unified framework.

Subcommittee Charge (2)

- Identify new science opportunities that could be possible within and across BER programs if a unified data framework were to be developed;
- Assess recommendations from recent AI/ML reports that could potentially be incorporated into a future data framework for BER (e.g., with a component that includes training data);
- Consider data management strategies and investments in other agencies that could be leveraged in developing the BER unifying framework;

Subcommittee Charge (3)

- Provide a list and brief explanation of the components and specifications that would be needed in the development of a unified framework in service to BER science that is achievable in the next five years; and
- Examine the benefits of developing a unified data framework to the scientific research workforce, with particular attention to increased opportunities for enhancing career progression and which types of culture changes could help facilitate those benefits.

Additional Information

This new framework should allow for the **interoperability and compatibility of data, tools, and supporting information that span the biological, environmental, climate, and earth system sciences**, and it should facilitate the analysis and synthesis of data for complex and multi-disciplinary research efforts across BER.

For this assessment, **data** should include **laboratory and field observations** (e.g., ARM, AmeriFlux, ESS-DIVE, JGI, KBase, and NMDC), **model-generated** data (e.g., ESGF), **simulated and stochastic** data (e.g., ARM/LASSO), archives based on observations and models (e.g., Multisector Dynamics data and ILAMB), and **relevant metadata**, including **uncertainty characterization, data provenance**, and any tools used to generate the data.

Progress To Date

Convened a Subcommittee Steering Group:

- Kerstin Kleese van Dam (BNL, chair)
- Kjersten Fagnon (LBNL)
- Ann Fridlind (NASA/GISS)
- Susan Gregurick (NIH)
- Adam Schlosser (MIT, co-chair)
- Jeremy Schmutz (HudsonAlpha, co-chair)
- Daniel Segre (Boston University)
- Pamela Weisenhorn (ANL)
- Ben Bond-Lamberty (PNNL)
- Dev Niyogi (U of Texas, Austin)

Settled on General Strategy

- Identify priority science targets, that would be enabled or significantly accelerated by a unified data infrastructure
- Assess the current state of data services and interconnectivity between BER support DOE SC assets.
- Assess existing solutions for unified data infrastructures and support for AI applications within them.
- Collect input from BER scientists, user facilities, data repositories and major projects, the computer science community, relevant efforts in other agencies via RFI, workshops and interviews
- Review barriers to BER data usage

Establish Subcommittee Working Groups & Subcommittee Members*

Data Infrastructure	Integrated Data and AI	Biological Research	Environmental Sciences	Diversity, Equity, Inclusion, and Accessibility
Kjiersten Fagnan (LBNL), Casey Burleyson (PNNL)	Susan Gregurick (NIH), Kerstin Kleese van Dam (BNL)	Daniel Segre (Boston U), Pam Weisenhorn (ANL)	Ann Fridlind (NASA), Ben Bond-Lamberty (PNNL)	Dev Niyogi (U of Texas), Pam Weisenhorn (ANL)
Forrest Hoffman (ORNL)	Ben Blaiszik (U Chicago)	Steve Allison (UC-Irvine)	Charu Varadharajan (LBNL)	Sen Chaio (Howard U)
	Shantenu Jha (Rutgers)	Emily Eloie-Fadrosch (LBNL)	Gannet Hallar (Univ of Utah)	Lou Woodley (CSCCE)
	Ravi Madduri (ANL)	Chris Henry (ANL)	Jennie Rice (PNNL)	
	Carlos Soto (BNL)	Jeremy Schmutz (HudsonAlpha)	Luke van Roekel (LANL)	
		Shin-Han Shiu (MSU)		

* We anticipate ~15 total members per working group. This list only shows Subcommittee Members.

Planning

- The **Request for Information (RFI)** was issued on April 18 <https://www.govinfo.gov/content/pkg/FR-2023-04-17/pdf/2023-08029.pdf> – **Please respond as soon as possible, before June to have the most impact on our work!**
- **Workshop** dates have been set for July 18 and 25, general workshop structure defined, potential invitees identified
- **Report** background material review started, results will be presented at workshop
- We will present updates at the autumn BERAC meeting.

Early Insights

Biological Science

Gaps: what limits progress?

- Divide between experimental and computational languages/data/cultures
- Data formats at different scales. **Lack of a conceptual integration scheme**
- **Missing global map of available data**, their limitations, and what other data are most needed
- Hybrid mechanistic-AI models to **connect hierarchically across scales**: sequences \Rightarrow proteins \Rightarrow traits \Rightarrow functions \Rightarrow metabolic fluxes \Rightarrow processes \Rightarrow ecosystems

Research: what would be enabled by new bridges?

- Biology-inspired engineering solutions for current sustainability challenges. Harness genetic diversity of plants and microbes to advance sustainable crops/bioenergy production, and innovation in bioproducts and biomaterials
- Understanding the controls on and impacts of plant and microbial metabolism on the Earth system. Bidirectional: (1) Contribution of natural biological systems to biogeochemical cycles; (2) Long-term effects of climate change on biological systems

Environmental Science

- **First discussion yielded additions to WG scope**
 - group agreed that social science and hydrology data integration are emerging areas that should be included
 - relevant for multi-sector dynamics, urban environments, digital twins, etc.
- **Main foci of first RFI responses (from WG members)**
 - lowering barriers to combined use of BER/DOE data with NASA/NOAA/USDA/EPA/GAW/ACTRIS data sources
 - lowering barriers to carbon cycle meta-analyses and synthesis efforts
 - lack of human systems and high-resolution urban-scale field data
 - lack of central repositories for ocean data from ships/ARGO/NASA/other
 - lack of consistent metadata, data formats, data quality across databases
- **Completing workshop invitation list (aiming for 20% international)**

Diversity, Equity, Inclusion and Accessibility

Unified data infrastructure should improve access to BER data

Users with fewer resources rely upon public data and compute

Workforce development can be enhanced through community engagement, education, and training

PNAS

RESEARCH ARTICLE | SOCIAL SCIENCES | ✓

Research Article landing page

Improving data access democratizes and diversifies science

Abhishek Nagaraj , Esther Shears , and Mathijs de Vaan  [Authors Info & Affiliations](#)

Edited by Douglas S. Massey, Princeton University, Princeton, NJ, and approved July 28, 2020 (received for review January 30, 2020)

September 8, 2020 | 117 (38) 23490-23498 | <https://doi.org/10.1073/pnas.2001682117>

↗ 4,703 | 14



Significance

Data access is critical to empirical research, but past work on open access is largely restricted to the life sciences and has not directly analyzed the impact of data access restrictions. We analyze the impact of improved data access on the quantity, quality, and diversity of scientific research. We focus on the effects of a shift in the accessibility of

Diversity, Equity, Inclusion and Accessibility

Accessibility Challenges and Gaps:

- Data re-use often requires both the ability to find, combine, and integrate across heterogeneous data sets
- At least three types of users within Domain, Compute/Data ; Transdisciplinary with levels of expertise ranging from Foundational, Intermediate, to Advanced.
- Need sufficient data descriptions to allow non-advanced users to find, access, assess data quality and suitability for a purpose
- Need to learn how to interact with multiple distinct interfaces, many requiring some command line or coding experience
- Working with large assembled data can have associated compute needs

Diversity, Equity, Inclusion and Accessibility

Paths forward:

- Identify general and specific barriers to data access and re-use by diverse stakeholders from: non-research intensive universities, MSIs, and industry
- Identify best practices for low barrier to entry approaches to support broad user groups in both accessing and using the data
- Identify infrastructure and workforce training and development needs
- Identify mechanisms to support data usability efforts during project closing/continuation

Existing BER Data Management and Supporting Infrastructures

Quick Survey of existing Capabilities:



- Resource Size
- Number of Users
- Data Sources
- Data Types
- Data Sharing Policies

BER scientists generate large amounts of diverse data from genome sequencers to atmospheric sensors to high fidelity climate simulations.

- Discovering these datasets means navigating to the right resource
- Linking these datasets is challenging
- **Opportunity** – deploy common data and infrastructure APIs

Existing BER Infrastructure Serves the Global Scientific Community


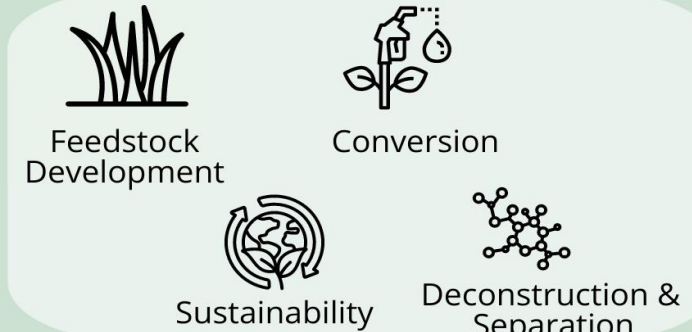
User Facility Raw & Value-Added Data Generation

20+ PB Data

>15,000 Users

Bioenergy Research Centers Support Biofuel & Bioproduct Innovation





Feedstock Development


Conversion

Sustainability



Deconstruction & Separation



Microbiome Data Made Discoverable via Sample Metadata Harmonization



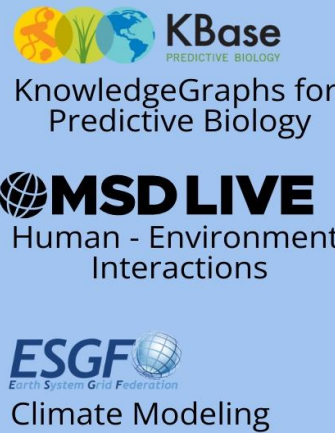

2,449 Samples



Unique Identifiers, Metadata & APIs Facilitate Data Exchange

Data and Computing Platforms Accessible Exploration & Interactive Analysis

KnowledgeGraphs for Predictive Biology

Human - Environment Interactions


Climate Modeling

9+PB Data

>7,000 Users

>10 Nodes

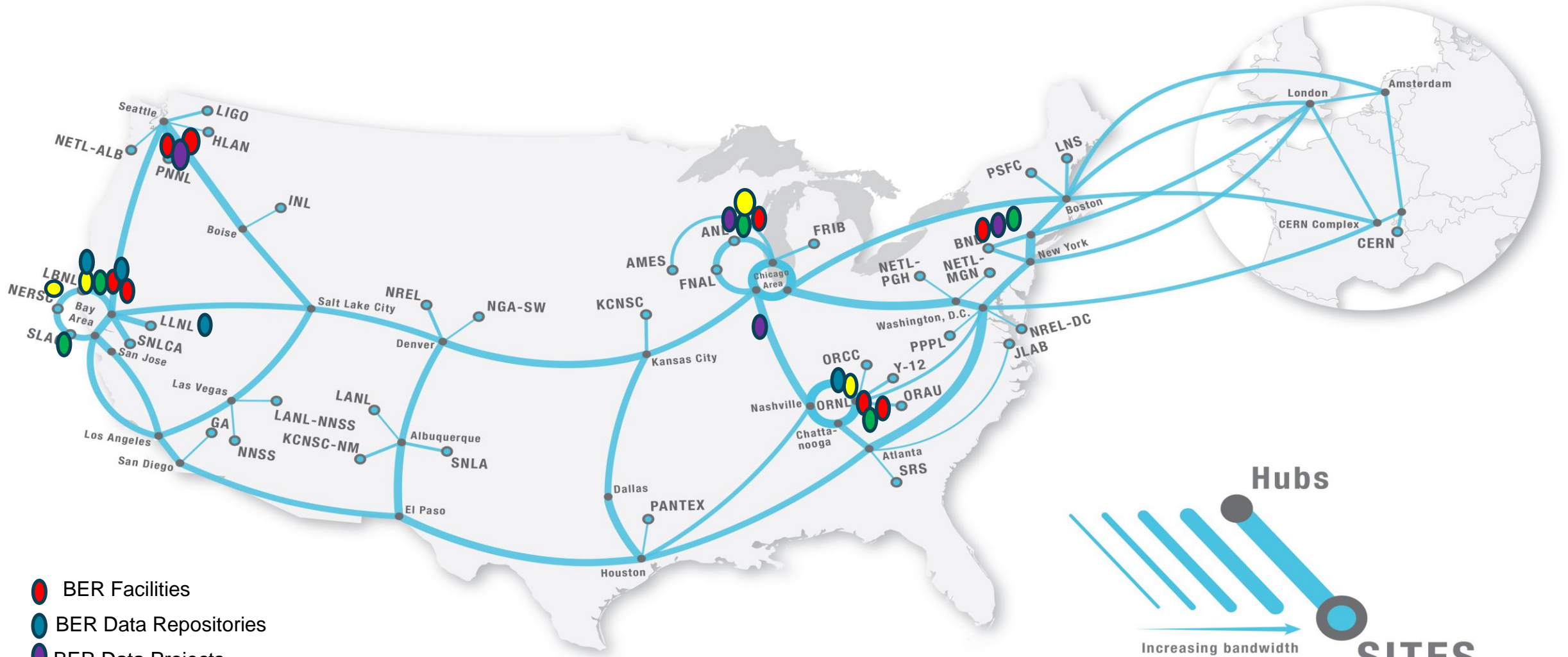
Archival Resource for User-provided Earth Sciences Data



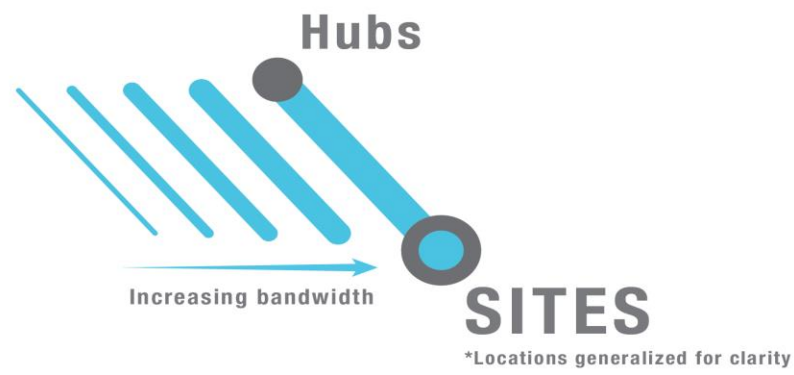

1.5 TB Data

*not all projects captured, but will be in the final report

ESnet6



- BER Facilities
- BER Data Repositories
- BER Data Projects
- BES
- ASCR



Unified Data Infrastructures

A Unified Data Infrastructure or Data Fabric is a federated, scalable architectural and operational approach to simplify self-service data governance, lifecycle management, access and analysis at scale across all connected data producers, data services and data users.

Oldest - 2002 UK NERC Data Catalog

Europe -

- 2006 ESFRI - asked for the creation of a data fabric.
- Many domain specific projects, now European Open Science Cloud.
- Existing community-based data fabrics of interest are C-Span and eliXir

Elsewhere -

- Australian Biocommons, African H3ABioNet

USA

- NSF NSDF
- NIH NCPI
- NNSA Tri-Labs
- **DOE SC IRI**
- BER NVBBCC

AI in a Unified Data Infrastructure

Tools that make it easy to use AI for users with different experience levels – both on integrated data and in a federated setting, sending the AI to the data.

Available Tools

- Tensorflow Federated (<https://www.tensorflow.org/federated>)
- Flower Framework (<https://flower.dev/docs/index.html>)
- Argonne Privacy Preserving Federated Learning (<https://appflx.link>)
- Google Colab with Tensorflow Federated

Today all require familiarity with command line, indepth knowledge of installation of software on HPC systems, and AI models.

Started Feature comparison, gap analysis

Unified Data Infrastructures - First Insights

- We are not alone, IRI vision aligns with BER interests
- **Timely subcommittee to define BER's priorities and requirements**
 - What exciting science could be enabled, what is priority
 - What data services are needed to transform science
 - What roles will existing data repositories play
- Many partial solutions available for Data Fabrics, but ease of use for users with different levels of experience still far from achieved
- **Tie in with existing BER data repositories and services requires engagement from day one**
- Use of AI in a federated environment presents a particular hurdle
- Existing BER data repositories are excellent, but not all data captured
- **Data Integration support biggest gap – semantic not physical!**

Questions?