BERAC Unified Data Infrastructure Subcommittee

Kerstin Kleese van Dam

Introduction

BER has facilities and projects that create significant volumes of data:

- BER supporting facilities : ARM, EMSL, JGI, ASCR Computing and BioImaging (incl. BES Light Sources, BER CryoEM facility)
- Large BER projects/programs: Bioenergy Centers, BRaVE, AmeriFlux, NGEE, E3SM, Urban IFLs

To preserve and share this data BER operates data services through: ARM, EMSL, JGI, ESS-DIVE, ESGF, KBase, NMDC and MSD-LIVE

Introduction

Today, BER operates a portfolio of data infrastructure resources that provide excellent support to the users of their data within a well defined scientific topic area. - Building small unified infrastructure islands.

- Each data island provides its own customized metadata, data management, access and analysis services
- Not all science data is accessible today through one of BER’s data services

Is what is available today enough to address BER science grand challenges?

Science Perspective

Scientific challenges in the biological and environmental sciences increasingly require the integration of multi-disciplinary and multi-scale datasets. Key challenges vary:

Environmental sciences:

- Exceedingly large data available, often need to integrate subsets of data
- Data size itself is prohibitive to repeatedly download/transfer/process

Biological sciences:

- Data size is more manageable, but data are sparse or incomplete
- Many challenges to integrating data across studies and projects

Important to capture multiple, diverse projects as science cases to prototype platform. While a rising tide raises all ships, coordination among these projects is required to ensure solutions are robust to disciplinary differences

Environmental Sciences

Earth system modeling

Opportunity

Robustly predict how the nonlinear Earth system will respond to 21st century climate change

Barriers

- Lack of efficient simultaneous access to massive and growing multi-agency data sources (e.g., satellite, ground-based and oceanic networks)
- Inconsistent metadata conventions
- Lack of computationally-intensive data processing capabilities that can operate centrally using shared tools
- Global datasets require bespoke efforts to integrate (e.g., oceanic field campaigns, international data sources)

Environmental Sciences

Multi-sector dynamics

Opportunity

Incorporate state-of-the-art knowledge of human-Earth system dynamics in societal decision-making

Barriers

Earth System Modeling barriers in addition to

- Lack of data on human activities such as household energy use or agricultural water use
- Lack of access to industry data subject to privacy and other protections

Environmental Sciences

Field site and campaign data

Opportunity

Fill gaps in process-level understanding needed to improve multi-scale ESM physics

Barriers

- Lack of data integration and storage coordination across US agency and international participants (e.g., TRACER and MOSAIC campaigns)
- Inconsistent methodological, metadata, and other standards
- Inconsistent data deposition by individual PIs and labs
- No coordinated searchability across what is available

Environmental Sciences

Recommendations:

- Unified, consistent access to data archives maintained by differing US agencies at the funded project level (projects address this on a one-off local basis currently)
- Accessible and sufficient server-side computing capability *integrated with up-to-date data archives*
- Support for shared community tools that integrate diverse data sources

Biological Sciences

Microbial integration

Opportunity

Understand how microbial genomes and the environment determine ecosystem-level structure and functioning

Barriers

- Incompatibility of data types (e.g. 16S, shotgun)
- Non-standard data scattered across studies
- Lack of environmental meta-data
- Lack of framework for data integration across scales (genes, proteins, pathways, cells, communities)

Biological Sciences

Plants for sustainability

Opportunity

Efficiently grow plants for production of bioproducts / biofuels under a varying climate

Barriers

- Curation of plant and microbial genetic datasets and integration with field collected data
- Climate data not easily accessible to plant/microbe researchers
- Efforts often limited to short-term collaborations in specific crops. Long term continuity needed to build links across different groups
- Lack of multi-scale modeling approach to connect genetic variation to production in future scenarios

Biological Sciences

Bioengineering for negative C

Opportunity

Engineer plants, microbes, and ecosystems capable of absorbing and retaining C

Barriers

- Integration of mechanistic/ML models for predictive modeling need multiomics data infrastructure
- Lack of unified ontology for synthetic biology experiments
- Barriers in modeling vs. experimental data comparison

Biological Sciences

Recommendations:

- Development of an effective, scalable, and federated search engine to help researchers find relevant datasets
- Secure repository to make available all relevant biological and contextual data (including experimental design details, imaging, e.g. BES light sources)
- Support for improved cross-walks, ontologies, and standards for data and metadata that go beyond individual disciplinary boundaries

Cross-Cutting Science

Putting data in context

Opportunity

Understand biological system functions under realistic field conditions at multiple scales

Barriers

- Laborious, repetitive manipulation of diverse datasets needed to synthesize/use field data
- Field data collection is heterogeneous and lacks mechanism and culture of sharing
- Many groups forced to re-invent tools that could be universally adopted
- Few incentives for integration owing to diverse agency data sources and objectives

Cross-Cutting Science

Multi-scale models from genes to Earth

Opportunity

Incorporate variables at multiple biological scales in high-resolution regional and Earth system predictive models

Barriers

- Specialized knowledge needed for accessing/interpreting regional and ESMs vs. biological data
- Disparate location, ontology, format between bio vs env data
- Challenges in large variation in spatial and temporal scales
- Lack of organized datasets to test models

Cross-Cutting Science

Taking research from useful to usable

Opportunity

Make BER data, especially those that have high societal linkages, accessible, inclusive, and usable by the broader community

Barriers

- New area especially as it pertains to work with communities
- Weak incentives, processes, and unclear timelines for making data available and reusable
- Lack of opportunities for data users and decision makers to interface
- Lack of access or knowledge about how to access datasets and models to test decisions or usability

Cross-cutting Science

Cross-cutting opportunities can bring great gains in understanding and manipulating natural and agriculture systems to meet DOE goals.

Recommendations:

- Encourage shareable, coordinated data collection with standards for field data
- Develop curated, standardized and open datasets that can address multiscale modeling
- Place a focus on making field, environmental, variation, and climate data accessible to expand inclusions across all of these BER areas

Inclusion and Accessibility - Barriers

- Awareness of data, analytical tools, and models are a prerequisite to their findability, accessibility, and reuse. This digital divide impacts user experience and abilities.
- Effective collaboration between experts working in different disciplines and at different scales requires development of interpersonal relationships and development of a common language, which is slow and puts some groups at a disadvantage.
- New research avenues include a greater societal impact component. This brings in both new data types and a wider range of potential data users, including decision makers and community members who are non-experts.

Inclusion and Accessibility - Needs

- Codevelopment should be prioritized.
- Data products, and test cases/prototypes can benefit from, and need to be available to, non-domain experts.
- Documentation and *readily available training designed to meet people where they are* (experience, culture, etc) are essential.

Inclusion and Accessibility - Recommendations

- Targeted outreach to ensure meaningful inclusion of diverse stakeholders from the initial design phase.
- A single user interface that is easy to find and makes data and tools from across facilities accessible. This includes ready-to-access documentation and training to promote data and tool use among a broader community of researchers.
- Provide accessible compute that can handle data volumes and tool requirements without input from non-expert users.
- Available funds, compute, tools, and data provide an incentive for broad community participation and engagement.

Workforce Development - Data Literacy

- Developing *the ability to explore, understand, and communicate with data in a meaningful way* is critical to the future workforce and needs broad support
- Current efforts to broadly advance data literacy are underway, but the depth of adoption within any one agency is still limited.

Establishing and using data and metadata conventions and standards can enable increased data literacy, enabling researchers to focus on analyzing and understanding the significance of the information.

Workforce Development - AI

- AI is projected to change analytical approaches, data management, and data wrangling processes.
- Workforce training needs to include:
  - leveraging large, established models efficiently (e.g. Foundation Models)
  - equipping researchers with the know-how to adapt these models through prompt engineering and various tuning techniques
  - mastering a systematic approach to verification and validation becomes imperative.
- Another aspect to consider is the impact of potentially-disruptive Large Language Models (LLMs) on the data and research lifecycle.

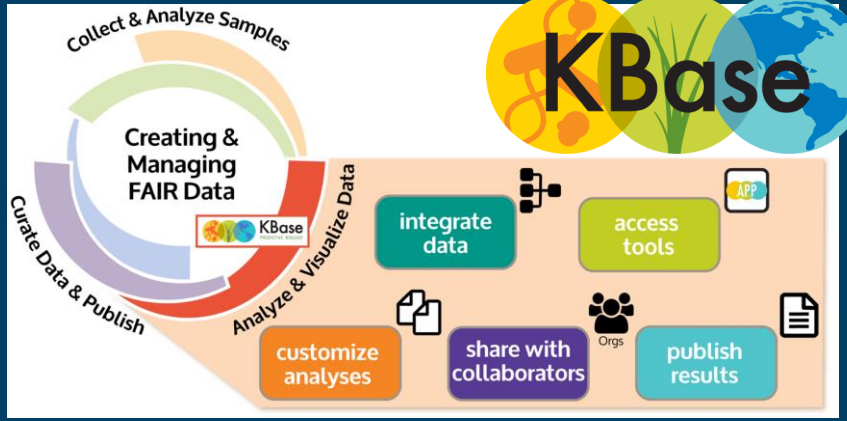
Existing Facilities

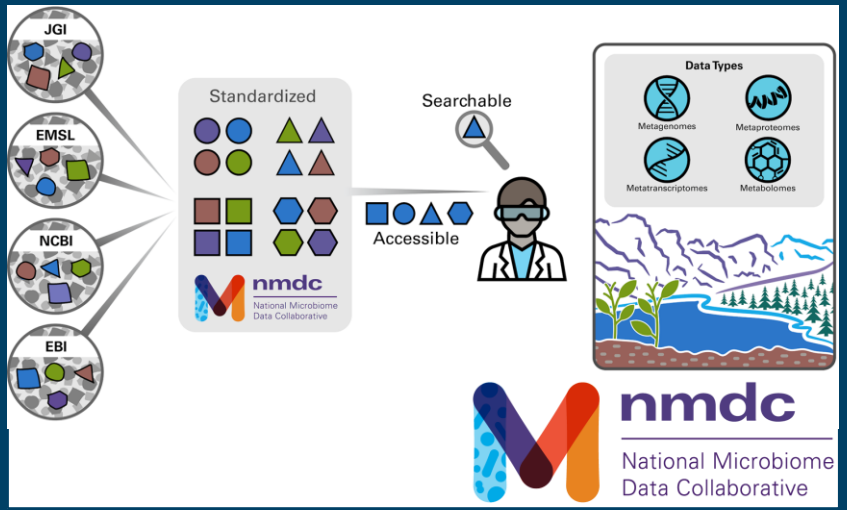
Many existing BER resources serve their respective communities well

Opportunities exist to network resources together to lower barriers to data integration and collaboration

KBase allows for **bespoke analyses** of public data.

NMDC provides powerful search for **multiomics data collection and discovery** for microbiomes.

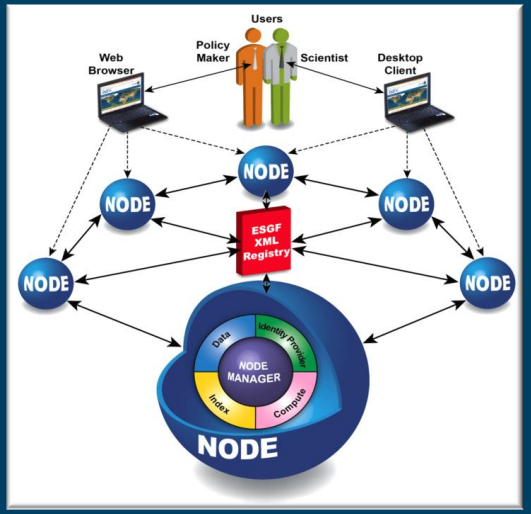


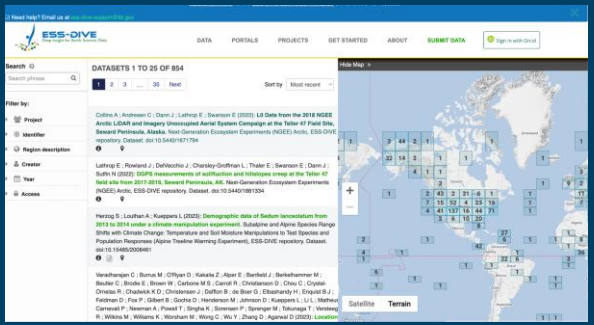


Existing Facilities

Earth Systems Grid Federation supports an international community in data integration efforts by providing collocated data and computing.

ESS-DIVE preserves data and metadata for discovery and reuse. They are leading efforts to establish common biological and environmental sample metadata standards.





Integrating User Facility Proposal Calls

FICUS
Facilities Integrating Collaborations for User Science

Scientists utilize the broad range of the capabilities of each facility, and generate datasets beyond what each of these facilities could do alone, within Focused Topic Areas of:

– Biofuels, biomaterials and bioproducts

– Hydro-biogeochemistry

– Inter-organismal interactions

– Novel applications of molecular techniques

– Ecosystem-scale research using samples from the NEON Biorepository

JGI JOINT GENOME INSTITUTE
- DNA&RNA sequencing
- DNA Synthesis
- Metabolomics

EMSL
- 'Omics
- Advanced imaging
- Biogeochemistry
- Isotope analysis

NSLS-II National Synchrotron Light Source II
Brookhaven National Laboratory
- Structural biology

Center for Structural and Molecular Biology
- Structural biology

Argonne National Laboratory
Advanced Photon Source
- Structural biology

117 Accepted Proposals

150+ Publications

6 National User Facilities

Unified Data Infrastructure

Identified Barriers:

- Working Practices - Download and work on data locally
- Unified Search Capability - Across all BER data services
- Data Integration - metadata and data format misalignment, labor intensive manual integration effort
- Localized Solutions - lack of interoperability, reusability and scalability.
- Sufficient, accessible data and compute capabilities
- Support for new technologies (incl. AI)

Unified Data Infrastructure - R&D Needs

Policy driven efforts:

- Inclusive, BER wide governance structure
- Harmonization of User IDs, Authentication and Authorization
- Harmonization of metadata - working with communities to enable search across data services
- Standardization of data formats were needed
- Interoperability of services
- New infrastructure investments have to integrate by design

Unified Data Infrastructure R&D

Development Efforts:

- **Scalable search engine** that can discover data, workflows, tools and potential collaborators across all BER science areas
- **Attribution of contribution** mechanism to reward all that participate
- **Common, distributed data fabric** that ties all BER resources together
- **Marketplace** in which participants can find and use available data and tools, share results and most importantly find new collaborators

Unified Data Infrastructure

Research Required:

- AI support for search, recommendation, workflow composition, content (data and tools) evaluation
- AI training ground - high quality data sets, validation and UQ frameworks directed in particular at large scale foundation models or Digital Twins

Training, Support and Documentation:

- Essential for users at all levels to make the most of the new capabilities and provide valuable feedback to developers

Response - First Charge Question

Review the existing and anticipated capabilities in data management and supporting infrastructures that are relevant to the breadth of BER science:

- BER has a sophisticated set of data infrastructure capabilities both in support of specific programs and to facilitate the integration across dedicated user communities
- However, the subcommittee identified gaps in the data services support, some data sets collected through BER programs and projects are not easily accessible. Furthermore, cross community integration across different data services has currently only limited support.

Recommendations - Second Charge Question

Pursue a project-driven collaboration strategy between infrastructure developers and researchers ("build it together" rather than "build it and they will come"):

- Identify a select number of high impact science goals that require a unified data infrastructure to empower early adopters, and ultimately affect a culture change across the BER research space.
- Explicitly include targeted outreach in early science demonstrators to reach diverse stakeholders and achieve integration of underserved researchers from the initial design phase.
- Establish a BER ‘marketplace’ where BER scientists can discover and use data, tools, services, and resources across all BER programs, as well as interact with each other forming new collaborations.
- Support for targeted outreach/mentoring as data and tools come online to have a breadth of users from the outset as well as awareness of tools and data.

Recommendations - Second Charge Question

- The integration of new technologies such as AI, Quantum and Digital Twins needs to be supported through dedicated training, validations and verification frameworks.
- Support the incubation of a community-based unified data infrastructure, through policies to harmonize user IDs, authentication, and authorization across BER facilities and data services.
- Integrate all new infrastructure into the unified data infrastructure and incentivize participation, which likely requires long-term commitment to host data and access.

Recommendations - Second Charge Question

- Based on the requirements of the early community adopters, co-develop a buildout plan that heavily leverages unified data infrastructures such as the DOE ASCR Integrated Research Infrastructure HPDF, NSF's National Scientific Data Fabric and efforts associated with the European Open Science Cloud including the European Destination Earth project.
- Regularly review and amend the plan. As communities work together in the new BER marketplace, their requirements and priorities will evolve.
- Selectively support integration and interaction with data frameworks of other agencies important to BER science. Given the effort that such connections require, target only core partners on a project-driven basis in the first 5 years.