BERAC Unified Data Infrastructure Subcommittee

Kerstin Kleese van Dam

Charge

- Review the existing and anticipated BER data management and supporting infrastructure capabilities

- Recommend a strategy for the next generation data management and analysis within a unified framework

A Unified Data Infrastructure 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.

Unified Data Infrastructures

Unified Data Infrastructures are not new. They progressed from data sharing only, to including computing and now also experimental facilities directly. Europe has been driving their evolution until recently, but the US is now catching up with the latest concept.

Example Efforts:

Data

- NERC Data Catalog for environmental data (UK, 2002)
- ESGF (international, 2003)

Data and Compute

- ELIXIR for life sciences (EU, 2007)
- EUDAT for all science data and computing (EU, 2012)
- C-SCALE for Earth system science (EU, 2021)

Data, Compute and Experiments

- National Science Data Fabric (US, NSF, 2021)
- Integrated Research Infrastructure (US, DOE, 2023)

What is available today for BER Scientists?

BER Data Infrastructure

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

BER Data Infrastructure Status Today

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
- There are very limited connections between these data islands
- Data service interactions with other agencies is limited and cumbersome outside of DOE.

Are there Gaps?

Environmental Sciences Needs

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) – harmonization of metadata and data quality
- Accessible and sufficient server-side computing capability *integrated with up-to-date data archives*
- Support for shared community tools that integrate diverse data sources
- Access to all needed data types incl. e.g. access to data on human activities with sufficient privacy protection

Biological Sciences Needs

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
- Flexible infrastructure that adapts to science needs

Cross-cutting Science Needs

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 - Needs

- 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-use 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.

Sub-committee Observations and Recommendations

Observations

- BER research is increasingly complex, requiring the integration and study of processes across scales and modalities. However, current BER data infrastructure is not ready to support such efforts.
- More could be done to provide underserved communities and minorities with easy access to BER capabilities and encourage them to participate in BER research
- New infrastructure strategies could enhance workforce development and, in particular, support early career scientists better

Observations (2)

- Many unified data infrastructure efforts are underway worldwide. While none is ready for adoption yet, BER could learn a lot from these efforts, and useful collaborations could be formed
- Creating a unified data infrastructure requires not only technical developments but also the integration of researchers from different communities, allowing them to communicate and interact with ease
- A complete BER unified data infrastructure is not achievable in 5 years

Recommendations

- Pursue a project-driven collaboration strategy between infrastructure developers and researchers (adopt a “build it together” approach 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 integrate underserved researchers into the initial design phase.

Recommendations (2)

- Leverage existing BER facilities and data services to build an initial tightly integrated unified data infrastructure. Augment this infrastructure with a dedicated data facility (can be federated) that combines large-scale data and computing to alleviate the need for BER scientists to download data for integration and analysis.
- 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 and form new collaborations.
- Support targeted outreach and mentoring as data and tools come online to ensure, from the outset, a breadth of users and awareness of tools and data.

Recommendations (3)

- Support the integration of new technologies, such as artificial intelligence, quantum science, and digital twins, 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 (4)

- Co-develop a buildout plan, based on the requirements of early community adopters, that heavily leverages unified data infrastructures, such as (1) the DOE Advanced Scientific Computing Research program's Integrated Research Infrastructure High Performance Data Facility; (2) the National Science Foundation's National Scientific Data Fabric; and (3) efforts associated with the European Open Science Cloud, including the European Destination Earth project.
- Regularly review and amend the plan to incorporate the evolving requirements and priorities of communities as they work together in the new BER marketplace.

Recommendations (5)

- Selectively support integration and interaction with other agencies' data frameworks 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.
- Develop clear metrics of success for all stages and aspects of the unified data infrastructure program.

Comments

- There are many more detailed recommendations in the different report chapters, that combined create a much richer picture of the needs and for a strategy moving forward.
- AI will become an increasing driver for a unified data infrastructure, both for access to data and to scalable tools for the save, secure and trustworthy development and use of AI based solutions.
- The envisaged ASCR Integrated Research Infrastructure (IRI) could in time provide some of the base infrastructure capabilities, but a BER unified data infrastructure will only be successful if it includes science driven tools for data integration and analysis, currently not part of IRI.

Many Thanks to:
The Subcommittee Members
Workshop Participants
RFI Respondents