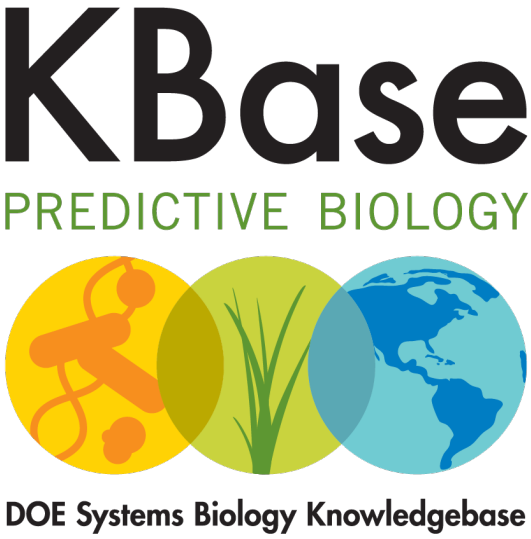

DOE BSSD Performance Management Metrics Report FY 2024 Q1: New Challenges in Data Operability Being Addressed in KBase

Authors: Elisha-Wood Charlson[1] (elishawc@lbl.gov), Chris Henry[2] (chenry@mcs.anl.gov),Gazi Mahmud[1] (GaziMahmud@lbl.gov), Paramvir Dehal[1] (psdehal@lbl.gov), Roy Kamimura[1] (royk@lbl.gov), Adam Arkin[1] (aparkin@lbl.gov)

[1]Lawrence Berkeley National Laboratory, Berkeley, CA 94720, [2]Argonne National Laboratory, Argonne IL 60439,

KBase and the challenges of data-driven systems biology

The Department of Energy's Systems Biology Knowledgebase (KBase), was conceived as a collaborative platform to enable researchers across the world to share data and analyses with the goal of building models and making predictions from complex, multiscale biological data. At the time KBase launched, the community had just begun to grapple with increased rates of production and diversity of data, challenges of doing domain-specific data interoperability at scale, and early hints as to the role machine learning (ML) and artificial intelligence (AI) might play in aiding analyses. Advancements, both technical and sociological, have improved how we find and access data, as laid out by the FAIR data principles (Findable, Accessible, Interoperable, and Reusable; 1). As a community, however, we have not addressed the domain-specific challenges of interoperability and reuse of data at scale. The continued increase in data volume and diversity, and the requirements for ML/AI to have well-curated training data to build robust models, have made it obvious that establishing ***what, where, and how diverse data can be made comparable across (or even within) programs is truly the next big challenge in data interoperability.***

In the following series of quarterly highlights, we address four topics:

1. The new challenges and approaches to data interoperability we are addressing within KBase (this highlight).
2. The strategies we are developing to allow users to better collaborate in team-oriented science efforts (Q2 highlight).
3. How we are engaging with other data-oriented institutions to ensure access to data and analysis resources across BER (Q3 highlight).
4. How we are incorporating new AI and/or ML capabilities that will enhance KBase capabilities and empower our users to accelerate BER research (Q4 highlight).

For this Q1 highlight, we focus on the challenges surrounding data interoperability and the approaches KBase is applying to address these challenges. This specifically draws on what we have learned about FAIR data and what needs to come next. We call it "enabling researchers to COPE with complex data" (making data Comparable, Organized, Predictive, and Engaging).

Challenge: Increased rates of production and diversity of data

Over the last decade, massive advancements in both experimental data collection technologies and information management systems have resulted in substantial growth in the scale and complexity of biological science performed today. It is now routine to produce a terabase (1 trillion DNA base pairs) of genomic sequencing data within a single dataset, with commensurate growth in all other sequencing approaches, including targeted regions within a genome (amplicons), expressed genes (RNA transcripts), and large-scale screens of mutant strains of cultured microbes. Mass spectrometry-based approaches, which are used to identify small molecules generated by cellular metabolism (metabolomics) and sequence short fragments of proteins (proteomics) have undergone similar improvements. Additionally, new methods are

emerging that can make these data inherently more complex. For example, quantitative stable isotope probing, or qSIP, labels biological molecules with isotopes to measure activity, vs measuring DNA alone, which could be present in the environment after cell rupture or in genomes of dormant or inactive cells. Thus, qSIP enables quantitative measurements of metabolic activity within a given environment. With modern-day sequencing capacity, amplicon-based methods that target specific regions of a genome with diagnostic benefits (i.e., species identification), can now be applied at scale across large temporal and spatial experiments collected by large collaboratives (2) (Figure 1).

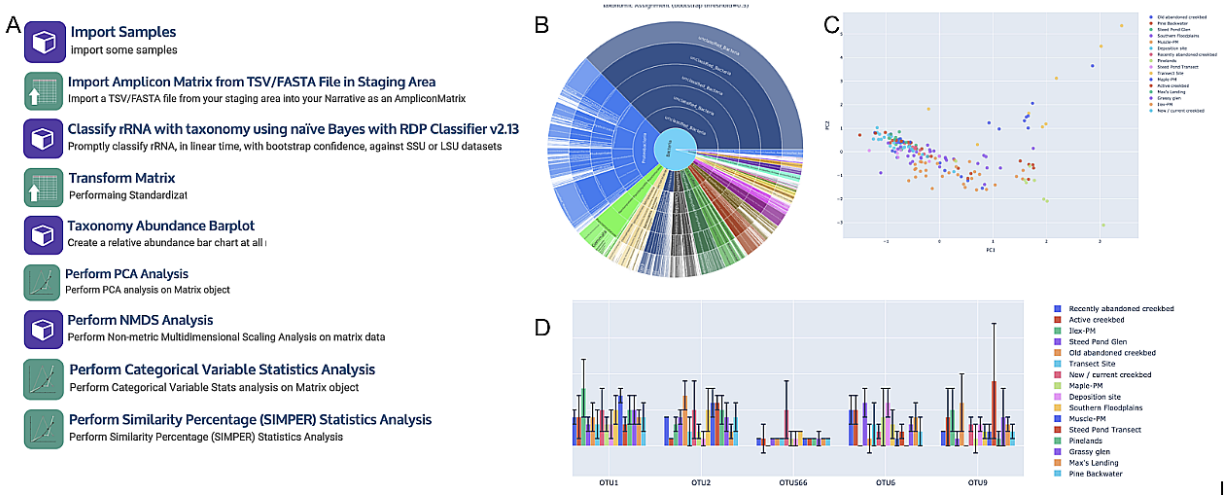

Figure 1. Amplicon Analysis Workflow Supported on KBase Platform. KBase provides a common workflow for microbial community focused analysis and links results with existing data in KBase (A). After importing Samples with environment attributes, the amplicons are taxonomically classified with RDP (B). The abundance matrix can be normalized, filtered and subsetted in preparation for statistical reports, bar plots, PCA (C) and NMDS analysis. Finally the SIMPER app (D) returns plots of the most influential OTU for a given attribute of the Samples. These are screenshots of app output in KBase. Text in legends is not meant to be readable.

Finally, gradual reduction in cost for genome sequencing (3) through improved technologies has helped democratize the use of biological sequence data. More research labs and even classrooms are able to ask novel questions, generate large amounts of relevant data, and participate in the collaborative process of solving complex, multiscale biological research problems (Figure 2a, 4).

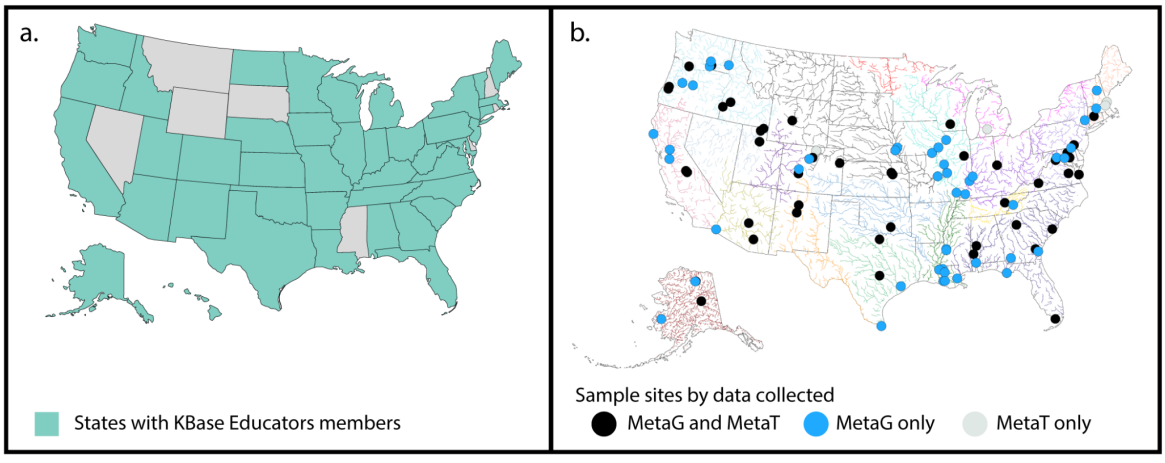

Figure 2. KBase supports (a) professional development training at universities in 42 of 50 US states (so far), and (b) samples and data from every US watershed.

Now, more than ever, there is a critical need for these data to be placed in a broader, environmental context to better understand biological questions from a systems perspective. This includes pairing biological sequence data with the chemical and physical parameters that govern the conditions the organisms live in (Figure 2b, MetaG = metagenomics, MetaT = metatranscriptomics, 5). As well as recording related data concerning their host (for animal or plant associated microbes), soil properties, or water features, and the extended ecosystem they reside in (e.g., tundra vs tropics).

Challenge: Domain specific data interoperability at scale

For individual programs within the Department of Energy's (DOE) Biological and Environmental Research (BER) program, data are collected with the general purpose of inferring causal mechanisms by which organisms transform and are transformed by their environments and each other. Our collective goal is to predict how future interventions, including bioengineering and biosystem design, may modify existing features of an organism, or establish new ones, to create a desired change in the system. For example, modification of microbes relevant to crop yield that increases soil nutrient levels and decreases the need for commercial fertilizer. However, establishing sufficient understanding at a systems-level requires large and diverse data sets to be made interoperable, including an evaluation of the data context and quality, normalization of units and scale, and determination of interrelationships between data sets. With the ultimate goal of modeling a system so that each specific BER program can explore and tailor components of the model to achieve their goals. ***Data interoperability at this level and scale is still a technical and sociological challenge.*** Even within a single project, different experimental questions may require intentional standardization to be made interoperable (6). Conducting that level of coordination across collaborating labs in complex programs, like the DOE Scientific Focus Areas (SFAs) and Bioenergy Research Centers (BRCs), has not been a priority because of the immense effort required to pivot projects with established research practices. However, as BER's research questions continue to expand beyond individual programs, and even across biological and ecosystem divisions, we now see the value and importance of standardized data reporting (a technical challenge) and data sharing (a technical and social challenge). With the recent Office of Scientific and Technology Policy (OSTP) memos around open data (7), the agency-wide pledge for open science (8), and the need to ensure scientific integrity and reproducibility - where researchers have persistent access to all stages of data generation, we have an arduous, tedious, and expensive task before us. One that often has little financial, cultural, or technical support in the budget.

KBase's current approach to data interoperability

When KBase was conceived, it came with the fundamental goal of accelerating sophisticated analysis and modeling of biological systems within the context of their environment(s). KBase is program and research question agnostic, with a focus on collaborative, integrative biology. Our premise was that by providing access to 1) high-performance computing resources; 2) advanced suites of analytical tools that could be effectively and easily chained together; and 3) systems that make data *and analyses* FAIR for our users. Also important to the KBase mission

were mechanisms to ensure data could be “integrated” and made comparable between programs, and organized such that the *relationships* among the data could be identified. For example, users want to rapidly identify genomic data from similar environments or organisms that have similar metabolic potential, and generate novel data sets for analysis within KBase. *We have made a great deal of progress on these goals with users seeing value in what we provide.* KBase now serves >42,000 users working in >102,000 KBase Narratives (open, collaborative Jupyter notebook-like analysis workflows), with access to >22 TB of public data and hundreds of tools that operate on diverse biological data types (Figure 3). KBase has also been a leading example of how BER data programs integrate all research products (samples, data, and analyses) into the broader publishing infrastructure (9). By generating FAIR Narratives, with DOIs that reference any literature, data, and software used in the analysis, KBase produces fully citable research products that 1) directly connect to the existing publishing infrastructure, 2) indexed by web search engines, and 3) can be cited in research publications, linking reproducible data analysis to scientific conclusions. Beyond that, KBase also tracks all Narrative/data views, copies, and workflow/data reuse, which aims to transform how we demonstrate the impact of research products (such as analysis tools and experimental data, not just publications). One of our goals for object provenance tracking and credit is to support the success of early career researchers that want to openly share data and results, while also making it easier for funding agencies to quantitatively demonstrate success of their funded initiatives. For example, the use of software tools is much higher than the research citations would indicate as not everyone publishes for proprietary or other reasons, such as education or routine processing.

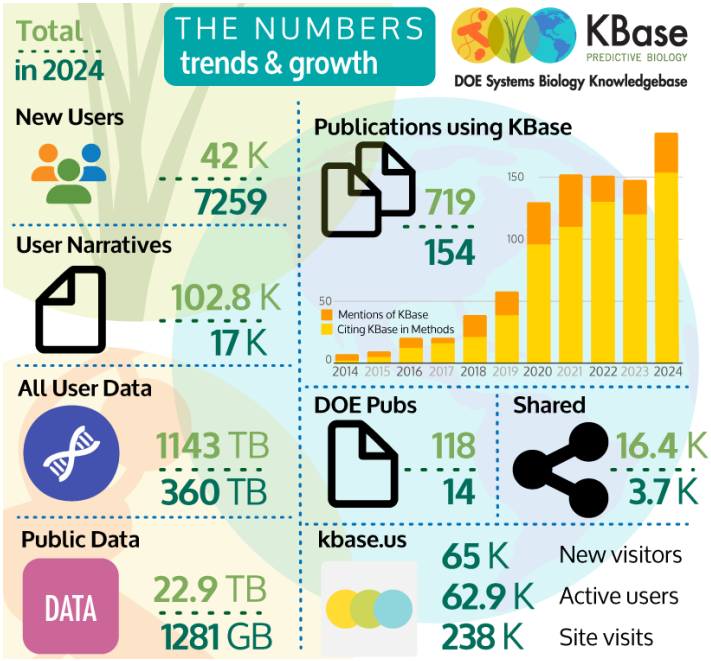

Figure 3. Total KBase metrics, with growth in 2024

KBase supports analyses ranging from taxonomic annotation of organisms or functional annotation of proteins, to assembly of large metagenome data sets and generation of community-level metabolic models (Figure 4). The system provides democratized access to public data sets and community reference data, access to DOE computational infrastructure - both cluster and high performance-based, and enables sharing of sample metadata, data, and analyses. KBase has also pioneered methods for ‘abstracting’ the complexities of data interoperability. By developing extensible object-oriented data models that accurately represent biological systems, from the perspective of how we conduct research, and standardized data

formats for inputs and outputs of analysis tools, we have ensured that any data in KBase can be rapidly analyzed in any user-defined workflow. However, the rapid increase in the rate and diversity of data (discussed previously) has rendered our existing data models and data formats insufficient and fragile to new features requests from our users. In addition, analysis of KBase user behavior, based on publications and workshop feedback, has concluded that both the data complexity and the sheer number of options available to analyze the data are overwhelming, even in a system designed to make both easier. For many researchers, as well as students just starting out in science, not having cyberinfrastructure options like KBase would make these analyses impossible.

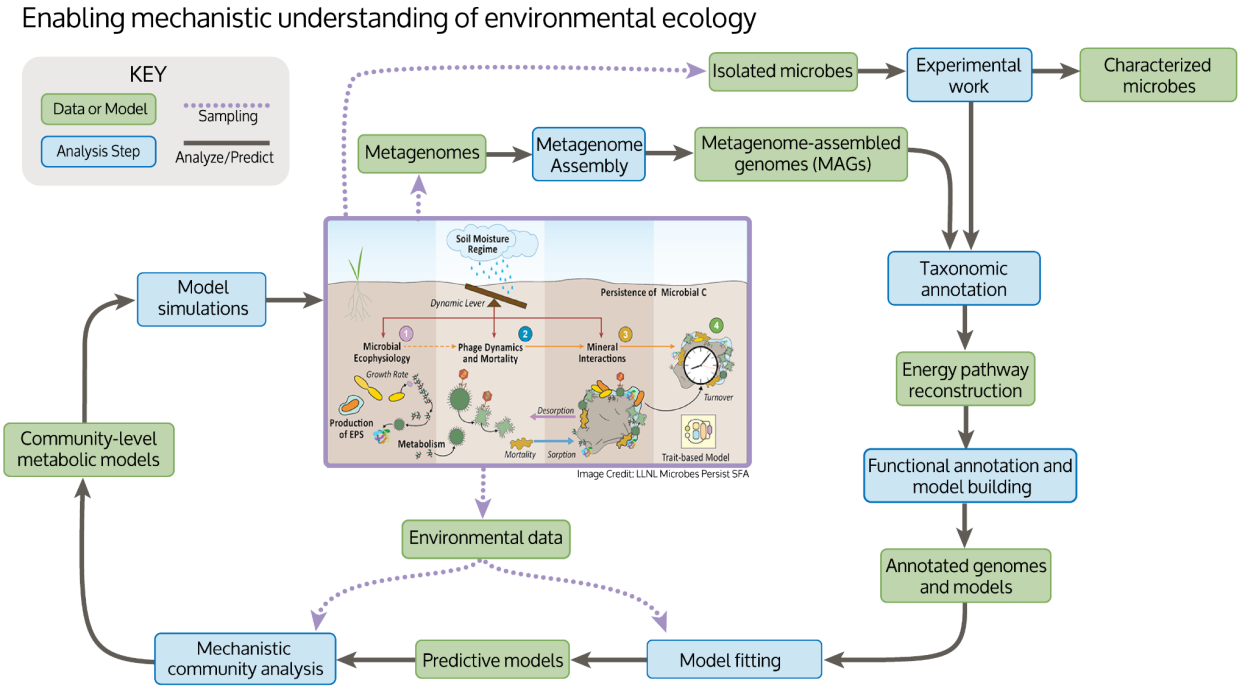

Figure 4. KBase now supports rich workflows so building a mechanistic understanding of environmental ecology, including metagenome analysis, genome annotation, modeling, and phenotype prediction is within reach of all.

KBase's plan approach to data interoperability at scale

To address scalable data interoperability, KBase is making a significant pivot. In October 2024, we began our next four year performance period, and we have already begun to prototype how technological innovations, like GPUs and AI, can support a significantly expanded, modernized, and more comprehensive biological data model.

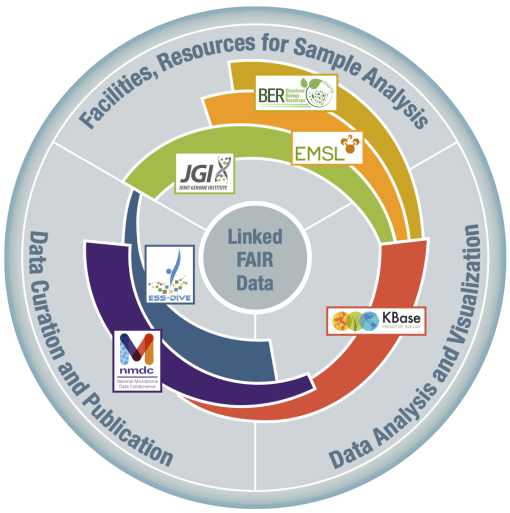

Figure 5. BER data and analysis ecosystem supports integrative computational and data science to facilitate community access, analysis, and sharing. Figure from 2021 BSSD Strategic Plan (https://www.genomicscience.energy.gov/doe-ber-biological-systems-science-division-strategic-plan/)

Advancing the KBase Central Data Model (CDM) will greatly improve our ability to make diverse data comparable across programs. KBase aims to provide provenanced data (tracking who and how it was generated), that is well-modeled and labeled with standardized conceptual and ontological frameworks, and made technologically accessible for sophisticated data science and systems-level modeling. Because of our existing object-based architecture, KBase is well positioned to rapidly advance into these new, innovative technological frameworks.

All KBase advancements in data interoperability are being done in close coordination with the main DOE data programs, including the 2 BER user facilities: the Joint Genome Institute (JGI) and the Environmental Molecular Sciences Laboratory (EMSL), as well as the National Microbiome Data Collaborative (NMDC) and the Environmental System Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE). Collectively, we are discovering what will be required to enable seamless discovery, integration, and analysis of complex data generated by multiple programs for a diversity of researchers, institutions, and research questions (Figure 5, 10).

FAIR data principles improved the research landscape, to a point

The concept of the FAIR data principles (1) can be summarized as making data:

- Findable, typically through a globally unique persistent identifier (PID), such as a digital object identifier (DOI) that contains rich metadata describing origin, data, creator, etc.,
- Accessible via a standard web or application-based protocol;
- Interoperable file/data format(s) for analysis, storage, processing; and
- Reusable with sufficient provenance, ideally using community-standardized terms, and a clear usage license.

The global support and effort to adopt and adhere to the FAIR data principles has been highly encouraging, a demonstration of what the scientific community can accomplish when provided with a logical, achievable foundation (11, 12). This momentum has helped establish FAIR principles for more than just data, including software (13) and facilities and instruments (14). By making all components of the research process FAIR, data platforms and repositories that support that various components (data, protocols, software, projects, and publications) can be linked. This enables better tracking of research accomplishments, and improves our ability to navigate and query the connected components (9), not just individual FAIR data sets. KBase is at the forefront of linking and tracking access and reuse of all research products in our platform, and is working to ensure all contributors to the scientific process are acknowledged.

On the other hand, the FAIR data principles have also been effective at highlighting the technical and sociological challenges that have slowed our ability to fully render biological research in a computer-based reality. Even though the “bio- / natural science” domains have been strong adopters of the FAIR data principles (11, 12)., much of our research is still conducted “offline” - as fieldwork: slogging through wetlands, braving freezing conditions at the pole, or on research vessels far from land; or labwork: trying to cultivate novel organisms, extracting metabolic compounds from cultures or environmental samples, or running samples from monitoring

programs to detect pollutants or contaminants. In addition, the FAIR data principles were not specifically designed to capture complexities inherent in life science research questions. Finally, the biological sciences do not exist in isolation; we often leverage discoveries made in Earth and health sciences (12), and even computer science (13). The FAIR data principles need to be applicable to a variety of scientific disciplines, and therefore lack domain-specific guidance that truly addresses data interoperability within individual domains. This challenge - what to do when you get the data files and try to make sense of their contents - has yet to be addressed properly. Lamprecht *et al.* (13) describes how to take computer science "beyond FAIR", discussing the need to distinguish between "form" (how the code is provided) and "function" (what the code actually does). We need to do the same for biological research, and KBase has begun to tackle this challenge. But first, let us provide additional context to better understand the factors that have previously hindered success in this area.

Need for domain-specific data interoperability at scale

With the adoption of the FAIR data principles, and increased generation of FAIR data, the next biggest hurdle to accelerating discovery lies in evaluating domain-specific data interoperability. Domain-specific standards are necessary to assess interoperability of the data described within the files, especially if we aim to automate integration or training of any new ML/AI models. This is another area where the biological sciences are primed to act. There are several groups establishing ontologies and controlled vocabularies to describe biological components in standardized ways (15, 16), and several groups leveraging those to train AI models to assist in this process (17). The largest hurdle to the application of standards in the biological sciences is both technical and sociological. Many researchers aren't aware of standards, and if they are aware, they are often overwhelmed by the options and the still very manual application of terminology without clear guidance or training on how to decide between terms/ontologies that are similar/overlapping. We need to both increase awareness *and* lower the effort required for adoption, likely by 1) providing tools for rapid term selection, and 2) auto-assigning terms based on properties inherent to the data (e.g., assign biome descriptors based on latitude(lat)/longitude(lon)). Another component of ensuring data can be integrated appropriately is the curation effort required to evaluate if the data products are measuring the same thing, because units of measurement matter. Some measurements are directly convertible, such as lat/lon (decimal-degrees vs degrees-minutes-seconds) and time (local vs coordinated universal time, or UTC). Others require specific knowledge of what and how the data were collected or the type of standard label(s) applied. For example, absolute abundance of an organism may not be comparable across studies if one study sampled for a day and another for a week. Or, organismal labels could be difficult to compare without the source data if one data set applied National Center for Biotechnology Information (NCBI) taxon labels and another applied Genome Taxonomy Database (GTDB) labels.

This problem is compounded when the science community needs to leverage each others' data to answer questions that the original data collection was not necessarily designed for. For example, to find microbes that are diagnostic of high levels of metals in the environment,

researchers would need to combine metagenomic and environmental data sets across a large set of studies to look for statistical increases in abundance of certain metal-related genes (e.g., transporters, chelating agents, etc.) in environments known to have higher concentrations of specific metals. However, the reported abundance of relevant genes across those metagenomes may differ, with differences in lab-based protocols or analysis pipelines introducing variance that compromises the statistics. And the same is likely true for measured metal concentration data from the environment. Finally, the nature of the environment matters, as the mechanisms and effect levels of different metals may be different in marine, freshwater, sediment, or human gut environments. Data must be evaluated for comparability, and represented in ways that make the relationships clear (e.g., from the same organism, chemical measurement, or environment), and with as much knowledge of their causal relationships as possible, both biological (e.g., DNA sequence denotes genomic potential while expressed RNA sequences and ribosome profiles indicate proteins are being actively expressed), and experimental design (e.g., three replicates sample were taken at the same time, with paired measurements of temperature and pH).

Enabling researchers to COPE (comparable, organized, predictive, and engaged) with complex data

FAIR data enables researchers to leverage existing data for their research, but it does not solve the challenge of data interoperability across data sets. By applying standardized labels and units, data can be rapidly assessed for **Comparability**. *Are two data sets describing the same aspect of biology, and from the same perspective?* Can numbers/measurements from two datasets be combined in a principled manner? Once comparability has been confirmed, the data can quickly be **Organized** such that *different data types have meaningful relationships between them, based on the experimental design that generated the data and a Central Data Model (CDM) that links components of biological systems to each other through evolutionary and physical relationships*. The KBase CDM is being updated to be more comprehensive, inclusive of work done by other DOE data programs, and designed to be interoperable with other agencies (joint publication with JGI, EMSL, NMDC, ESS-DIVE, NASA, NIH, and academic partners, in prep). Linked data that are comparable and organized by experimental design and/or biological concept becomes a

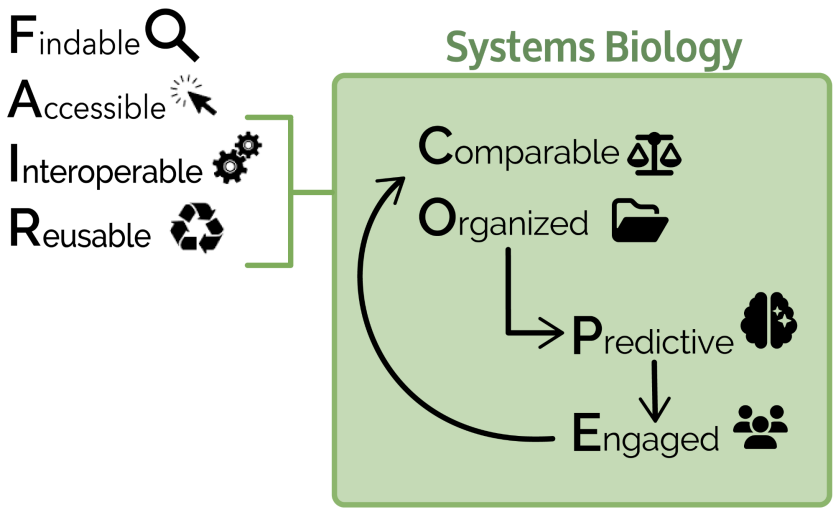

Figure 6. Connecting FAIR to domain-specific best practices to help us COPE with complex data.

framework for **Predictive** science. For example, taxonomic labels depend on the database used to apply them (e.g., NCBI vs GTDB taxonomy) and sequences attached to those labels are used to predict the taxonomy of new organisms. As more diverse and novel organisms are added to the databases, those labels may need to change. Consequently, the predictions of the classifications of organisms need to be updated. The quality or confidence of each prediction may also vary, depending on the underlying data and assumptions made by the algorithms. For example, protein function can be assigned by: primary sequence homology, structural homology, gene neighborhood information, inference from genetic experiment, or direct biochemical measurement. It is difficult to assess the quality of each prediction, which requires tracking how they were made, but the best assessment of a prediction are those that have been experimentally validated. Which is why it is critical to have an **Engaged** community invested in assessing and validating predictions, which can be used to update the CDM. In return, as discussed previously, those contributions must be both recognized and the broader research community needs to be made aware *and* enabled to leverage these advancements.

The KBase Central Data Model Platform

At the heart of the KBase CDM is a comprehensive biological data model comprised of the entities and relationships that define the behavior and interactions of the principal biological actors operating within the DOE BER mission space: microbes, plants, fungi, and associated viruses (Figure 7). Entities include molecular constructs like genomes, genes, proteins, and molecules, which are linked to conceptual constructs that convey our knowledge and understanding of biological systems, like reactions, functions, taxonomy, and classifications. We also integrate environmental data including sample metadata, geophysical measurements, and microbiome compositions and we represent the processes that govern their interactions such as representation of chemical transformation (reactions). All public DOE data is being integrated within this extensible framework, which improves interoperability by first and foremost imposing rigorous data standards on all integrated data (e.g., all annotations need evidence; all genomes must meet quality standards), curating comparability (e.g. by normalizing to common units, and linking to protocols).

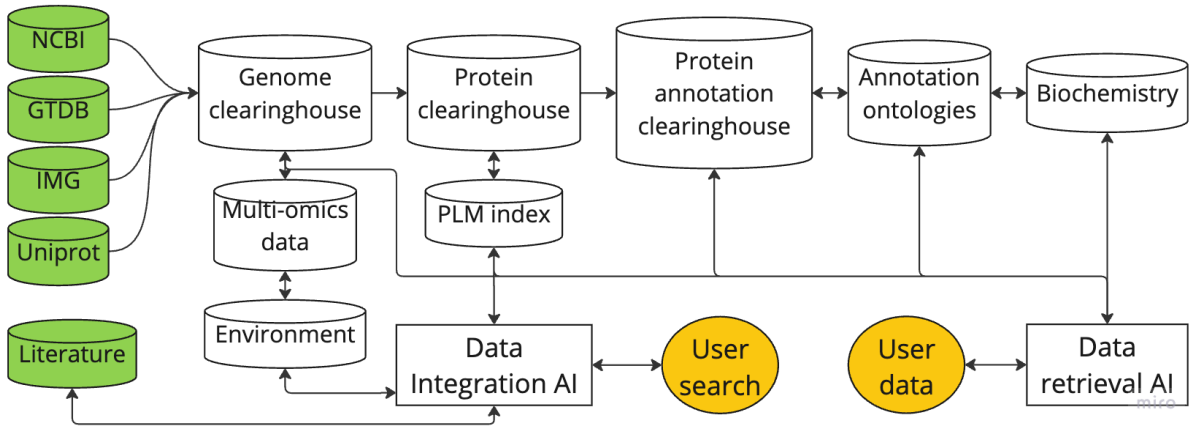

Figure 7. KBase Central Data Model combines data from major genomics resources, mapping IDs, intermapping genomes, and proteins, and consolidating annotations to provide a genomic rossetta stone database to support interoperability and maximize data synergy by merging data from all databases and offering scalable mapping of user data to this framework. New AI tools power scalable sequence comparison (PLM) and powerful query through agentic AI. Consolidated ontologies enable annotations to be rapidly compared and reconciled. Integration of all IDs also maximizes potential for mapping literature data to entities within the database and indirectly to user data as well.

However, the CDM is going far beyond the imposition of data standards. Data within the CDM is also being rigorously interrelated. For example, in support of an evolutionary relational model within the CDM, all genomes and proteins within the CDM are being scalably compared so that similar genomes and proteins may be clustered together. Based on these relationships, pangenomes are being computed, capturing all the functional capabilities a group of closely related genomes encodes, as well as the likelihood that any given function will occur within a given set of closely related organisms. Pangenomes are crucial for providing a robust means to fill in missing information and COPE with often sparse data that occurs in many biological systems being studied by DOE. Studying biological systems on the level of pangenomes makes them more scalable. While there will be 1 million genomes and 1 billion genes in the CDM, the core is likely to be only 100K pangenomes with 100 million protein families. More importantly, this sort of interrelation of all DOE generated genomes and proteins effectively links all DOE funded projects together in this body of overlapping shared pangenomes. This provides all DOE funded projects with a sophisticated understanding of the overlap that exists in the biological entities these projects are studying. This empowers these research teams to collaborate more effectively and pool their efforts to understand these shared entities, avoiding generation of conflicting knowledge products that need to be reconciled later. Finally, calculation of pangenomes is a prediction based on current genomes and classification methods. By providing an integrated, updating resource we reduce the need for redundant calculation by multiple groups of the similar objects and increase the speed with which similarities among biological entities and processes across projects can be navigated.

Our ability to construct and maintain the CDM platform is being driven in part by new advances in AI, including agentic AI and protein language models (PLMs). Agentic AI enables users to intuitively craft complex powerful data queries from a simple English language prompt. PLMs offer a means of scalably indexing protein sequence data, which greatly improves the speed and efficiency with which protein sequences may be compared and interrelated. KBase has constructed a PLM-based index of proteins stored within the CDM, enabling new emerging data produced by projects within and outside of DOE to be rapidly mapped into the CDM to facilitate the rapid discovery of novel genes and proteins and the rapid integration of new data for existing genomes and proteins.

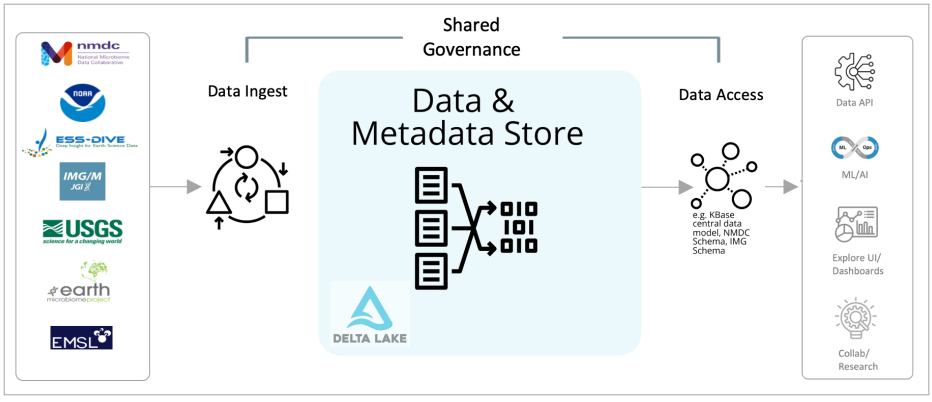

Figure 8. KBase Central Data Model Platform - unified data lakehouse platform with shared governance

Of course, the volume of data that is being aggregated within the CDM is enormous, with an anticipated 10^6 genomes and 10^9 genes, and the CDM will be expected to serve a scientific community of ≥ 100K users. Performance is particularly important because interoperability is a core goal in developing the CDM, meaning the CDM infrastructure must serve not only KBase users, but all other DOE facilities, facilities outside the DOE, and users of those facilities. To meet these needs performantly, the CDM is being implemented within a sophisticated cyberinfrastructure. This cutting-edge platform is designed to manage and utilize the rapidly growing volumes of genomic and biosciences data critical to advancing research and innovation. By integrating advanced data storage and computational capabilities, the platform creates a unified ecosystem that supports efficient data management, seamless analysis, and collaboration across diverse scientific domains.

The KBase CDM Data Lakehouse Platform is a cornerstone of this system, providing reliable, scalable, and well-governed access to data. Leveraging lessons learned in industry from decades of handling large and diverse data types and databases,this platform combines the scalability and agility of a data lake with the structured performance and transactional integrity of a traditional data warehouse. This approach ensures that all data is stored in a format that not only supports diverse and complex research workflows but is also adaptable to evolving scientific needs and standards.

One of the platform's core strengths lies in its ability to enable data interoperability across diverse biosciences research domains. By facilitating seamless integration and standardization of data from multiple sources, the platform fosters collaboration among researchers across various DOE facilities and beyond (Figure 8). This data interoperability ensures that insights and discoveries are no longer siloed, but instead shared and built upon an interoperable data storage ecosystem that fosters acceleration of scientific progress.

Furthermore, the platform simplifies cross-facility computing and infrastructure challenges by abstracting complexities and enabling the seamless integration of computational resources. Researchers can perform sophisticated data processing and analysis without being burdened by the technical details of resource allocation, ensuring efficient and cost-effective operations. The underlying cyberinfrastructure is designed to handle the exponential growth of data while maintaining high performance for data-intensive tasks.

A critical component of the platform is its shared governance framework, which unifies data quality standards, compliance policies, and data access controls across all domains. This shared governance not only improves the reliability and trustworthiness of the data but also ensures that the data is FAIR (Findable, Accessible, Interoperable, and Reusable). Such governance empowers researchers to confidently work with high-quality data, paving the way for AI-readiness by ensuring that data is interoperable and actionable for downstream ML/AI workflows.

The platform's user-centric, auto-scalable, and flexible architecture prioritizes ease of access, streamlined workflows, and enhanced data interoperability. This allows researchers to focus on scientific discovery rather than navigating technical challenges. The CDM Platform serves as a unified hub where Data and AI assets converge under a common governance framework, making cross-domain data and AI assets interoperable and actionable. This enables researchers to generate AI-driven insights, advancing breakthroughs in biosciences and fostering transformative, data-driven solutions to pressing global challenges.

Conclusions

Overall, we see how the KBase platform has taken enormous strides in driving interoperability through data standards, through interoperable tools chains, through provenance, and more recently, through the development of the KBase Central Data Model, an exciting new data infrastructure that maps data from all DOE facilities beyond, offers a standardized scalable interface to that data, and produces a synergistic output by facilitating the combination of all data with user data. The Data Lakehouse platform, along with its shared governance frameworks, is a demonstration of an open adaptable system that, if adopted widely, could facilitate improved interoperability across programs. This infrastructure is already empowering scientists and educators today with rich reference datasets, a rich ability to share and compare user data, and a rich set of hundreds of interoperable scientific applications implementing the latest important scientific workflows. The CDM elevates this capability by integrating reference data from primary repositories, enabling scalable mapping of user data to these repositories, empowering rich user queries, and allowing for advanced AI capabilities.

References

1) https://www.go-fair.org/fair-principles/
2) Thompson, L., Sanders, J., McDonald, D. et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551, 457–463 (2017). https://doi.org/10.1038/nature24621
3) https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data
4) Dow EG, Wood-Charlson EM, Biller SJ, Paustian T, Schirmer A, Sheik CS, Whitham JM, Krebs R, Goller CC, Allen B, Crockett Z and Arkin AP (2021) Bioinformatic Teaching Resources – For Educators, by Educators – Using KBase, a Free, User-Friendly, Open Source Platform. Front. Educ. 6:711535. doi: 10.3389/feduc.2021.711535
5) Borton, M.A., McGivern, B.B., Willi, K.R. et al. A functional microbiome catalogue crowdsourced from North American rivers. Nature 637, 103–112 (2025). https://doi.org/10.1038/s41586-024-08240-z
6) Kosmopoulos, J.C., Klier, K.M., Langwig, M.V. et al. Viromes vs. mixed community metagenomes: choice of method dictates interpretation of viral community ecology. Microbiome 12, 195 (2024). https://doi.org/10.1186/s40168-024-01905-x

7) https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-access-Memo.pdf
8) https://www.energy.gov/doe-public-access-plan
9) Wood-Charlson EM, Crockett Z, Erdmann C, Arkin AP, Robinson CB (2022) Ten simple rules for getting and giving credit for data. PLoS Comput Biol 18(9): e1010476. https://doi.org/10.1371/journal.pcbi.1010476
10) https://www.genomicscience.energy.gov/doe-ber-biological-systems-science-division-strategic-plan/
11) Mirjam van Reisen, Mia Stokmans, Mariam Basajja, Antony Otieno Ong'ayo, Christine Kirkpatrick, Barend Mons; Towards the Tipping Point for FAIR Implementation. Data Intelligence 2020; 2 (1-2): 264–275. doi: https://doi.org/10.1162/dint_a_00049
12) Kalinin, N.A., Skvortsov, N.A. Difficulties of FAIR Principles Implementation in Cross-Domain Research Infrastructures. Lobachevskii J Math 44, 147–156 (2023). https://doi.org/10.1134/S199508022301016X
13) Lamprecht, Anna-Lena et al. 'Towards FAIR Principles for Research Software'. 1 Jan. 2020 : 37 – 59.
14) Johnson, A., Julian, R., Mayernik, M., Mundoma, C., Murray, M., Ranganath, A., & Stossmeister, G. J. (2024). FAIR Facilities and Instruments Workshop #1 Report: Exploring Persistent Identifier Needs, Barriers and Incentives. https://doi.org/10.5065/zgsx-2d06 (Original work published 2024)
15) Yilmaz P, Gilbert JA,Knight R, Amaral-Zettler L, Karsch-Mizrachi I, Cochrane G, Nakamura Y, Sansone SA, Glöckner FO, Field D, The genomic standards consortium: bringing standards to life for microbial ecology, The ISME Journal, Volume 5, Issue 10, October 2011, Pages 1565–1567, https://doi.org/10.1038/ismej.2011.39
16) Jackson R, Matentzoglu N, Overton JA, Vita R, Balhoff JP, Buttigieg PL, Carbon S, Courtot M , Diehl AD, Dooley DM, Duncan WD, Harris NL, Haendel MA, Lewis SE, Natale DA, Osumi-Sutherland D, Ruttenberg A, Schriml LM, Smith B, Stoeckert Jr. CJ, Vasilevsky NA, Walls RL, Zheng J, Mungall CJ,Peters B, OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies, *Database*, Volume 2021, 2021, baab069, https://doi.org/10.1093/database/baab069
17) Toro, S., Anagnostopoulos, A.V., Bello, S.M. et al. Dynamic Retrieval Augmented Generation of Ontologies using Artificial Intelligence (DRAGON-AI). J Biomed Semant 15, 19 (2024). https://doi.org/10.1186/s13326-024-00320-3