Plan Overview

A Data Management Plan created using DMP Tool

DMP ID: https://doi.org/10.48321/D13BA96b4d

Title: DOE Systems Biology Knowledgebase (KBase)

Creator: Elisha Wood-charlson - **ORCID:** 0000-0001-9557-7715

Affiliation: Lawrence Berkeley National Laboratory (lbl.gov)

Principal Investigator: Adam Arkin, Chris Henry, Robert Cottingham

Data Manager: Elisha Wood-Charlson, Gazi Mahmud, Paramvir Dehal

Project Administrator: Roy Kamimura

Contributor: Paramvir Dehal

Funder: United States Department of Energy (DOE) (energy.gov)

Template: Department of Energy (DOE): Office of Science

Project abstract:

The DOE Systems Biology Knowledgebase (KBase) is an ambitious systems biology software development and operations project led by Lawrence Berkeley National Laboratory in close collaboration with Argonne, Oak Ridge, and Brookhaven National Laboratories. Established in 2011, KBase is developing a community-extensible computational environment to meet the key challenges of systems biology: predicting and ultimately designing biological function. We aim to accelerate research into how plants, microbes, and their communities transform and are transformed by the environment; drive the Earth’s biogeochemical cycles; affect the fates of contaminants in soil and water; and can be harnessed to improve the environment and provide sustainable routes for energy production and security. This requires a multiscale understanding of biological function from molecular to ecological.

KBase enables secure sharing of data, tools, methods, and conclusions in a unified, extensible system where researchers collaboratively generate, test, and share hypotheses about biological functions; perform large-scale analyses on scalable computing infrastructure; combine multiple lines of evidence to accurately model plant and microbial physiology and community dynamics, and ultimately ‘publish’ their work in FAIR (Findable, Accessible, Interoperable and Reusable) ways.

Start date: 10-01-2024

End date: 09-30-2028

Last modified: 07-08-2024

Copyright information:

DOE Systems Biology Knowledgebase (KBase)

1. Data sharing and preservation

Data management plans should describe whether and how data generated in the course of the proposed research will be shared and preserved. If the plan is not to share and/or preserve certain data, then the plan must explain the basis of the decision (for example, cost/benefit considerations, other parameters of feasibility, scientific appropriateness, or limitations discussed in #4). At a minimum, DMPs must describe how data sharing and preservation will enable validation of results, or how results could be validated if data are not shared or preserved.

While not a formal repository, KBase is explicitly constructed to enable users to capture, share, preserve, and publish their data. KBase's Narrative interface, based on Jupyter Notebooks, enables reproducible analyses accompanied with rich documentation, provenance tracking, and versioning. The data sharing, preservation, and publishing mechanisms of KBase support Findable, Accessible, Interoperable, and Reusable (FAIR) data principles, while also enabling rapid validation and exploration of results, without having to download, organize, or provide compute resources.

KBase provides mechanisms for users to keep their data private, share it with collaborators or inside a KBase Organization (invite only), or to share it publicly on the platform with all KBase users. KBase also supports sharing outside the KBase login, with a Narrative "snapshot" shared as a static HTML page that is indexed and discoverable by search engines. These "static Narrative" can be made FAIR Narrative upon request from the user, which prompts the DOE Office of Scientific and Technical Information (OSTI) to issue a digital object identifier (DOI) for the static Narrative that includes appropriate metadata, persistent identifiers, and documentation for a FAIR Narrative.

KBase users are encouraged to make their data FAIR and globally accessible, but are not required to do so. Full description of KBase's data policies are available at: https://www.kbase.us/about/terms-and-conditions-v2/

Most of the data that KBase generates falls into five categories:

- **Data generated as part of KBase services and reference data for the community**: All data integrated and developed by KBase staff are open and available on the KBase platform and may be downloaded directly. Notes: any user uploaded data or data generated on the KBase platform is also available for download, once they share it or make it public.
    - Types of data uploaded to and generated by KBase include: genomes, annotations, metagenomes, expression, protein-protein interactions, models of organismal and community metabolism, gene regulation, and sample metadata.
- **Data generated by KBase on the usage of the system; the effectiveness of tools, methods, and services; and derived data from usage, such as provenance networks**: Data around usage of the platform, services, and tools are captured and made reported in aggregate, protecting the privacy of users. Details on user statistics are available to KBase leadership and DOE management, only.
- **Data generated as part of experiments by KBase partners or collaborators**: Besides supporting the archiving and sharing of user data, KBase staff are directly involved in generating scientific data through research collaborations. Any scientific data generated by KBase-supported projects will be made publicly available on KBase as soon as possible, but no later than at the time of publication.

- **KBase source code**: All code and software used to operate KBase are available under open source licenses on GitHub (https://github.com/kbase).
- **Back up of KBase data:** KBase data stores are regularly backed up in several locations, locally at ANL or LBNL and to cloud storage (Google S3). Backups are done daily or weekly (depending on the data store).

KBase’s operational target is to support 99% uptime over a yearly interval, providing users reliable access to their data.

2. Data used in publications

Data management plans should provide a plan for making all research data displayed in publications resulting from the proposed research open, machine-readable, and digitally accessible to the public at the time of publication. This includes data that are displayed in charts, figures, images, etc. In addition, the underlying digital research data used to generate the displayed data should be made as accessible as possible to the public in accordance with the Principles published in the DOE Policy for Digital Research Data Management. The published article should indicate how these data can be accessed.

All KBase data and analyses shared publicly by users and cited/mentioned in a publication are immediately available at the time they chose to share it. FAIR Narratives with DOIs are registered at OSTI and DataCite, and linked by citation to any sample, data, software, instruments, funders/proposal persistent identifiers (PIDs) as supplied by the user at the time of DOI request.

Any publication citing/mentioning KBase is made available at the time of publication, linked to the KBase website (https://www.kbase.us/research/). All raw data in KBase is preserved and made available via the KBase system, though we recommend users back up their data in classic domain repositories as well.

Finally, KBase has partnered with the California Digital Library (CDL) to ensure the long term preservation of data associated with a FAIR Narrative. In an event where KBase is unable to support published workflows, all data will be stored in the CDL's generalist repository under a CC0 license.

3. Data management resources

Data management plans should consult and reference available information about data management resources to be used in the course of the proposed research. In particular, DMPs that explicitly or implicitly commit data management resources at a facility beyond what is conventionally made available to approved users should be accompanied by written approval from that facility. In determining the resources available for data management at DOE Scientific User Facilities, researchers should consult the published description of data management resources and practices at that facility and reference it in the DMP. Information about other Office of Science facilities can be found in the additional guidance from the sponsoring program.

KBase is intimately linked to the DOE Office of Science data management resources and leverages those connections to ensure that data owners (e.g., research scientists/engineers), data producers (e.g., DOE User Facilities), data curators, data managers and data management projects, and data publishers are all linked through the use of appropriate PIDs, metadata, and relationships. We collaborated with OSTI to publish a data management best practices article in 2022, "Ten simple rules for getting and giving credit for data" (https://doi.org/10.1371/journal.pcbi.1010476).

4. Confidentiality, security and rights

Data management plans must protect confidentiality, personal privacy, Personally Identifiable Information and U.S. national, homeland, and economic security; recognize proprietary interests, business confidential information, and intellectual property rights; avoid significant negative impact on innovation and U.S. competitiveness; and otherwise be consistent with all applicable laws, regulations, agreement terms and conditions, and DOE orders and policies. There is no requirement to share proprietary data.

KBase terms and conditions are available at: https://www.kbase.us/about/terms-and-conditions-v2. KBase privacy policy is in compliance with Lawrence Berkeley National Lab’s privacy policy (https://www.lbl.gov/terms-and-conditions/)

KBase protects user confidentiality and privacy by requiring users to only provide the minimum amount of information sufficient to validate their credentials and enable KBase to contact them if necessary. All other data that users may make available via the social networking features of KBase is voluntary and under control of the user. In addition, for data brought into KBase from reference sources (https://www.kbase.us/data-policy-and-sources/), such as JGI, KBase will remind users to honor the data policies of the original source.

Users have control over their own data and can choose when and how to make that data available to others. This includes datasets they have uploaded to the system, Narratives they have created, and the results of their analyses. KBase also has a user policy that requires users to confirm that they have the rights to data they enter into the system and/or make available to others.

KBase data is stored in dedicated data centers at ANL and LBL, with fire suppression, climate control and secured access via card keys only by operational staff. KBase data is not physically accessible by anyone other than operational staff at ANL and LBL, and enjoys the physical and cyber security of their hosting institutions which includes 24x7 physical perimeter security, constant network based intrusion monitoring and proactive detection and remediation of network vulnerabilities. All production data and computation occurs within these facilities. In rare instances when cloud based compute services are needed for spikes in activity, access to these services are protected by standard KBase network security mechanisms that include encryption, authentication and network firewalls. Backups of KBase data are kept in confidential storage only accessible with carefully managed credentials.