EuPathDB Data Release Policies

The Eukaryotic Pathogen Genomics Database (http://EuPathDB.org) is a Bioinformatics Resource Center (BRC) funded under contract from the US National Institute of Allergy & Infectious Diseases (NIAID). EuPathDB is charged with ensuring that genomic (and other large-scale) datasets pertaining to supported pathogens are conveniently accessible to all researchers worldwide. This document provides a brief summary of policies associated with releasing datasets on EuPathDB and affiliated databases.

Some general principles:

- ***Providing data to EuPathDB does not in itself constitute commitment to immediate public release.*** While there is no point in depositing data that will never be made public, data is released only when both the data provider and EuPathDB staff agree that it is accurately represented and ready to go live. (Database staff are distinct from research students/postdocs; even within the groups responsible for EuPathDB, the latter see new datasets only when they become accessible to the general public.)
- ***Data providers know their data best.*** We expect to work with those who generate the underlying data to determine how best to analyze and represent new data types. This typically means taking in relatively raw data, and building an in-house analysis pipeline to ensure that all comparable datasets are handled similarly.
- ***The earlier we learn about new datasets, the easier it is to schedule timely release.*** The nature of database production, and competing demands from the many communities we support, means that several months' notice is often required to prepare for release. Note that it is often possible to use a preliminary dataset for planning, which can swapped for the final version before public release.
- ***Experience has shown that data not deposited prior to publication often fails to emerge at all!*** After publication, it may be difficult to focus on tracking down the raw data, associated metadata, analysis methods, etc. It is never too early to discuss planned datasets with the EuPathDB team!
- ***While not required, pre-publication data release often provides in favorable attention from colleagues (and reviewers).*** Note that all major scientific journals are now on record as agreeing that early release of genomic-scale datasets does not compromise publication.

Why submit my data to EuPathDB?

- Facilitates your own analysis of data, particularly in the context of other genomic scale experiments already available in EuPathDB.
- Permits others to analyze your data in greater depth than possible in print (and even in advance of publication, if desired).
- Keeps your data alive on a highly accessed genomics resource: EuPathDB is accessed by ~13,000 unique users each month.

How do I submit data to EuPathDB?

- Contact the EuPathDB Scientific Outreach Manager by clicking the 'Contact Us' link on any EuPathDB page, or emailing to help@EuPathDB.org.

- Tell us about your data as early as possible to allow ample time for scheduling into EuPathDB release cycles.
- Once you tell us about your data, we will provide instructions on how to transfer your data to us (formats may differ depending on the nature and scale of the data to be transferred).
- In order to avoid any confusion or errors, we follow strict Standard Operating Procedures (SOPs), as outlined below.

Standard Operating Procedures (SOPs) for EuPathDB and affiliated databases:

EuPathDB routinely handles both published and pre-publication datasets. In order to ensure timely and accurate data integration we strictly adhere to the following Standard Operating Procedures (SOPs):

1. ***Datasets come to our attention in several ways, including:***
   - Direct contact from researchers generating the data (during the earliest stages of project design, as data is being produced, in the course of data analysis, or in the context of manuscript preparation).
   - Information provided by our database advisers or other users of EuPathDB.
   - Information obtained by EuPathDB staff at meetings and conferences.
   - Published literature.
2. ***Decisions to include a dataset in EuPathDB are based on community needs.*** In prioritizing data for integration, we rely heavily on discussions with active researchers, including the scientific advisory committees established for each of the taxonomic groups supported by EuPathDB.
3. ***Regardless of how we first learn about a given dataset, communication is established with the original producer*** through email, teleconference, and/or face-to-face meetings to discuss the desirability and feasibility of integration into EuPathDB. In the course of these discussions, we consider what data is likely to be available, questions the larger community may wish to ask of this data, and ways to represent or display such information.
4. Depending on the type of data and publication status, we also discuss data formats and transfer protocols. ***Data provided to EuPathDB is housed on secure servers and never shared outside of EuPathDB staff without prior consent of the data provider.***
5. Datasets are assigned a provisional release date, in consultation with the data provider. ***Scheduling a dataset does not mean that it will ever be released without the data provider's examination and approval!*** We operate on the assumption that those who generate the data are best placed to evaluate its proper integration and representation in the database. Note that this 'golden rule' applies to both published and unpublished data.
6. Two to three months before the scheduled release date, the ***Data Loading*** team processes and integrates the data into our internal databases.
7. After data loading is complete, the ***Data Development*** team begins to analyze and develop searches against the data. At this point we will likely communicate with the data provider, if questions arise.
8. ***Once data development is underway, the data provider is granted access to a password-protected development version of the EuPathDB web site containing their data.*** This site is a mirror image of the current live (production) site, except that it also contains the new data.

Instructions are provided on how to search and view the data, including sample searches integrating the new data with what is currently available in the database. Important questions to consider include:

- Does the database accurately represent your data?
- Are the values and/or graphical displays provided appropriate?
- Are the questions that one can ask of your data appropriate?
- Are there additional questions that you would like to see implemented?

9. ***A series of 'back-and-forth' exchanges typically ensues,*** in which we work to iteratively address any concerns, with changes made on the password protected site allowing you to view them and continue to play with your data in the context of the rest of the database.
10. ***Public release is only considered after everyone is satisfied with how the data is represented.*** If the provider is not yet ready to authorize data public release, data is rescheduled for a future release, and removed from the development site before it goes live.
11. Once data is approved for public release, a description is included in the 'News' accompanying the next release, ***highlighting new datasets and functionality, and acknowledging all data providers.***
12. ***Post-release quality assurance*** provides the opportunity to modify displays and develop new queries if/as appropriate.

What species and data types are supported by EuPathDB?

In one form or another, EuPathDB currently represents sequence data (genomes, ESTs, RNA-seq, generated on various platforms), comparative genomic information, DNA polymorphism & population genetics data, information on field & clinical isolates (with geo-stratographic metadata), chromatin modification data (ChIP-chip & ChIP-seq), manually curated and automatically generated gene models and other annotation, transcript & proteomic profiling datasets (multiple platforms), interactome data, structural information, metabolic pathways & metabolomics data, phenotyping information, reagents (clones, antibodies, etc), publication references, image data, and more. Support for additional data types (including host response datasets, and inhibitor data) is under development. ***If you have data to provide that is not currently supported, please ask!***

NIAID provides support for the biosecurity pathogens *Babesia, Cryptosporidium, Entamoeba, Giardia,* Microsporidia (various genera), *Toxoplasma, Plasmodium,* and related taxa *(Acanthamoeba, Gregarina, Neospora, Theileria*). Support for kinetoplastid parasites (*Crithidia, Endotrypanum, Leishmania, Trypanosoma*) is provided by The Bill & Melinda Gates Foundation, and in partnership with Seattle Biomed and the Wellcome Trust Sanger Inst. The FungiDB project encompasses a large (and growing) number of species supported by the Burroughs Wellcome Fund, and in partnership with the USDA NIFA program on oomycete pathogens. *Trichomonas* is supported as a legacy of previous NIH funding. ***Please contact us if you have data from other species that should be supported!***