

BETYdb Data Entry Workflow

David LeBauer, Moein Azimi, David Bettinardi, Rachel Bonet, Emily Cheng, Michael Dietze, Patrick Mulrooney, Andy Tu

Abstract

This is the userguide for entering data into the BETYdb database. The goal of this guide is to provide a consistent method of data entry that is transparent, reproducible, and well documented. The steps here generally accomplish one of two goals. The first goal is to provide data that is associated with the experimental methods, species, site, and other factors associated with the original study. The second goal is to provide a record of all the transformations, assumptions, and data extraction steps used to migrate data from the primary literature to the standardized framework of the database.

1 Table Of Contents

- Getting Started 1
- Preparing Publications 2
- Adding Data
 - Citations 3
 - Site 4
 - Treatments 5
 - Managements 6
 - Traits 7
 - Yields 8
- Bulk uploads 9
- QA/QC 10

2 Getting Started

You will need to create the following accounts:

- BETYdb (<https://www.betydb.org/signup>) (To use the database; request "creator" access during signup to enter data; request "manager" to perform QA/QC)
- Mendeley (<https://www.mendeley.com/>) is used to track and annotate citations
- Google Docs (<https://drive.google.com>) is used to prepare and transform data prior to entry.
- Redmine (<https://ebi-forecast.igb.illinois.edu/redmine/account/register>) is used to track data that need to be checked and/or corrected.

3 Preparing Publications for Data Entry

3.1 Mendeley

Mendeley provides a central location for the collection, annotation, and tracking of the journal articles that we use. Features of Mendeley that are useful to us include:

- Collaborative annotation & notes sharing
 - Text highlighter
 - Sticky notes for comments in the text
 - Notes field for text notes in the reference documentation
- Read/ unread & favorites: Papers can be marked as **read** or **unread**, and may be **starred**.
- Groups
- Tagging

Each project has two groups: "projectname" and "projectname_out" for the papers with data to be entered and for the papers with data that has been entered, respectively. Papers in the _out group may contain data for future entry (for example, traits that are not listed in Table 7).

Each project manager may have one or more projects and each project should have one group. Group names should refer to plant species, plant functional types, or another project specific name. Please make sure that David LeBauer is invited to join each project folder.

1. Open Mendeley desktop
2. Click **Edit** → **New Group** OR **Ctrl+Shift+M**
3. Create group name following instructions above
4. Enter group name
5. Set **Privacy Settings** → **Private**
6. Click **Create Group**
7. Click **Edit Settings**
8. Under **File Synchronization**, check **Download attached files to group**

3.1.1 Adding and Annotating Papers

When naming a group, tag folders so that instructions for a technician would include the folder and the tag to look for, e.g. "please enter data from projectx" or "please enter data from papers tagged y from project x". To access the full text and PDF of papers from off campus, use the UIUC VPN (<http://www.cites.illinois.edu/vpn/download-install.html>) service. If you are managing a Mendeley folder that undergraduates are actively entering data from, please plan to spend between 15 min and 1 hour per week maintaining it - enough to keep up with the work that the undergraduates are doing.

3.1.2 Adding a reference

- If the DOI number is available (most articles since 2000)
 1. Select project folder
 2. Right click and select **Add entry manually...**
 3. Paste DOI number in *DOI* field
 4. Select the search spyglass icon
 5. Drag and drop PDF onto the record.

- If DOI not available:
 1. Download the paper and save as `citation_key.pdf`
 2. Add using the *Files* field
 3. The citation key should be in `authorYYYYabc` where `YYYY` is the four digit year and `abc` is the acronym for the first three words excluding articles (the, a, an), prepositions (on, in, from, for, to, etc...), and the conjunctions (for, and, nor, but, or, yet, so) with less than three letters.

3.1.3 Annotating a Reference

Each week, please identify and prepare papers that you would like to be entered next by completing the following steps:

1. Use the star label to identify the papers that you want the student to focus on next.
 - Start by keeping a minimum of 2 and a maximum of 5 highlighted at once so that students can focus on the ones that you want. Students have been entering 1-3 papers per week, once we get closer to 3-5, the min/max should change.
 - Choose papers that are the most data rich.
2. For each paper, use comment bubbles, notes field, and highlighter to indicate:
 - Name(s) of traits to be collected
 - Methods:
 - Site name
 - Location
 - Number of replicates
 - Statistics to collect
 - Identify treatment(s) and control
 - Indicate if study was conducted in greenhouse, pot, or growth chamber
 - Data to collect
 - Identify figures number and the symbols to extract data from.
 - Table number and columns with data to collect
 - Covariates
 - Management data (for yields)
 - Units in 'to' and 'from' fields used to convert data
 - Esoteric information that other scientists or technicians might not catch and that is not otherwise recorded in the database
 - Any data that may be useful at a later date but that can be skipped for now.

Comment or Highlight the following information

- Sample size
- Covariates (see table 9)
- Treatments
- Managements
- Other information entered into the database, e.g. experimental details

3.1.4 Finding a citation in Mendeley

To find a citation in Mendeley, go to the project folder. By default, data entry technicians should enter data from papers which have been indicated by a yellow star and in the order that they were added to the list. Information and data to be collected from a paper can be found under the 'Notes' tab and in highlighted sections of the paper.

3.2 Recording extracted data and transformations

Google Spreadsheets are used to keep a record of any data that is not entered directly from the original publication. Please share all spreadsheets with the user betydb@gmail.com (mailto:betydb@gmail.com) in addition to any collaborators.

- Any raw data that is not directly entered into the database but that is used to derive data or stats using equations in Tables 3 or 6.
- Any data extracted from figures, along with the figure number
- Any calculations that were made. These calculations should be included in the cells.

Each project has a Google document spreadsheet with the title "project_data". In this spreadsheet, each reference should have a separate worksheet labeled with the citation key (authorYYYabc format). Do not enter data into excel first as this is prone to errors and information such as equations may be lost when uploading or copy-pasting.

4 Data Entry Overview

Before entering data, it is first necessary to add and select the citation that is the source of the data. It is also necessary for each data point to be associated with a Site, Treatment, and Species. Cultivar information is also required when available, but it is only relevant for domesticated species. Fields with an asterisk (*) are required.

5 Adding a Citation (<https://www.betydb.org/citations/new>)

Citation provides information regarding the source of the data. A PDF copy of each paper should be available through Mendeley.

1. Select one of the starred papers from your project's Mendeley folder.
2. The data to be entered should be specified in the notes associated with the paper in Mendeley
3. Identify (highlight or underline) the data (means and statistics) that you will enter
4. Enter citation information
 - Data entry form (<https://www.betydb.org/citations/new>) for a new site: BETYdb → Citations → new
 - **Author:** Input the first author's last name only
 - **Year:** Input the year the paper was published, not submitted, reviewed, or anything else
 - Fill out Title, Journal, Vol, & Pg. For unknown information, input 'NA'
 - **DOI:** The 'digital object identifier'. If DOI is available, PDF and URL are optional. This can be located in the article or in the article website. Use Ctrl+F 'DOI' to find it. Some older articles do not have a DOI.

- **URL:** Web address of the article, preferably from publisher's website
- **PDF:** URL of the PDF of the article

Figure 1. Adding a new citation

6 Adding a Site (<https://www.betydb.org/sites/new>)

Each experiment is conducted at a unique site. In the context of BETY, the term 'site' refers to a specific location and it is common for many sites to be located within the same experimental station. By creating distinct records for multiple sites, it is possible to differentiate among independent studies.

1. Before adding a site, search to make sure that site is not already entered in the database.
2. Search for the site given latitude and longitude
 - If an institution name or city and state are given, try to locate the site on Google Maps
 - If a site name is given, try to locate the site using a combination of Google and Google Maps
 - If latitude and longitude are given in the paper, search by lat and lon, which will return all sites within ± 1 degree lat and long.
 - If an existing site is plausibly the same site as the one mentioned in the paper, it will be necessary to check other papers linked to the existing site.
 - Use the same site if the previous study uses the *exact same location* and experimental setup.
 - Create a new site if the study was conducted in a different field (i.e., not the exact same location).
 - Create a new site if one study was conducted in a greenhouse and another was conducted in a field.
 - Do not use distinct sites for seed source in a common garden experiment (see 'When not to enter a new site' below)

3. To use an existing site, click Edit for the site, and then select current citation under Add Citation Relationships
4. If site does not exist, add a new site.

Table 1: Attributes of a site record

Descriptors	Additional Notes
Site Name	Site identifier, sufficient to uniquely identify the site within the paper
City	Nearest city
State	State, if in the US
Country	Country
Longitude	Decimal Form. For conversion see the equation in table 9
Latitude	Decimal Form. For conversion see the equation in table 9
Greenhouse	TRUE if plants were grown in a greenhouse, growth chamber or pots.
Soil	By percent clay, sand, and silt if given
SOM	Soil organic matter (% by weight)
MAT	Mean Annual Temperature (°C)
MAP	Mean Annual Precipitation (mm)
MASL	Elevation (meters above sea level, m)
Notes	Site Details not included above
Soil Notes	Soil details not included above
Rooting Zone Depth	Measured in Meters (m)
Depth of Water	Measured in Meters (m)
Table	

1. Do **not** enter a new site When plants (or seeds) are collected from multiple locations and then grown in the same location, this is called 'common garden experiment'. In this case, the location of the study is included as site information. Information about the seed source can be entered as a distinct cultivar.

6.1 Site Location

If latitude and longitude coordinates are not available, it is often possible to determine the site location based on the site name, city, and other information. One way to do this would be to look up a location name in Google Maps (<http://maps.google.com>) and then locate it on the embedded map. Google Maps can provide decimal degrees if the LatLng feature is enabled, which can be done here (<http://maps.google.com/maps?showlabs=1>). Google Earth can be particularly useful in locating sites, along with their coordinates and elevation. Alternatively, the site website or address might be found through an internet search (e.g. Google).

Use Table 2 to determine the number of significant digits to indicate the level of precision with which a study location is known.

Table 2: Table 2 Level of accuracy to record in lat and lon fields.

Location Detail	Degree Accuracy
City	0.1
Mile	0.01
Acre	0.001
10 Meters	0.0001

The screenshot shows a web browser window with the URL <https://www.betydb.org/sites/new>. The page title is "New Site". The form contains the following fields:

- Site name:
- Elevation (m): Mean Annual Precipitation (mm/yr): Mean Annual Temperature (°C):
- City: State: Country:
- Lat: Lon:
- Soil:
- % Clay: % Sand: SOM: Greenhouse:
- Notes:
- Soil Notes:

A map of the United States is displayed on the right side of the form, with a "Map" button and a "Satellite" button. The map data is attributed to Google, 2014.

Figure 2. Adding a new site

7 Adding Treatments and Managements

7.1 Treatments (<https://www.betydb.org/treatments/new>)

Treatments provide a description of a study's treatments. Any specific information such as rate of fertilizer application should be recorded in the managements table. In general, managements are recorded when Yield data is collected, but not when only Trait data is collected.

When not to use treatment: predictor variables that are not based on distinct managements, or that are distinguished by information already contained in the trait (e.g. site, cultivar, date fields) should not be given distinct treatments. For example, a study that compares two different species, cultivars or genotypes can be assigned the same control treatment; these categories will be distinguished by the species or cultivar field. Another example is when the observation is made at two sites: the site field will include this information.

- A treatment name is used as a categorical (rather than continuous) variable: it should be easy to find the treatment in the paper based on the name in the database. The treatment name does not have to indicate the level of treatment used in a particular

treatment - this information will be included in the management table.

- It is essential that a control group is identified with each study. If there is no experimental manipulation, then there is only one treatment. In this case, the treatment should be named 'observational' and listed as control. To determine the control when it is not explicitly stated, first determine if one of the treatments is most like a background condition or how a system would be in its non-experimental state. In the case of crops, this could be how a farmer would be most likely to treat a crop.

Name: indicates type of treatment; it should be easy for anyone with the original paper to be able to identify the treatment from its name.

Control: make sure to indicate if the treatment is the study 'control' by selecting true or false

Definition: indicates the specifics of the treatment. It is useful for identification purposes to use a quantified description of the treatment even though this information can only be used for analysis when entered as a management.

7.2 Managements (<https://www.betydb.org/managements/new>)

Managements refers to something that occurs at a specific time and has a quantity. Managements include actions that are done to a plant or ecosystem, such as the planting density or rate of fertilization, for example. Managements are distinct from treatments in that a treatment is used to categorically identify an experimental treatment, whereas a management is used to describe what has been done. Managements are the way a treatment becomes quantified. Each treatment is often associated with multiple managements. The combination of managements associated with a particular treatment will distinguish it from other treatments. The management types that can be entered into BETY are described in Table 4. Each management may be associated with one or more treatments. For example, in a fertilization experiment, planting, irrigation, and herbicide managements would be applied to all plots but the fertilization will be specific to a treatment. For a multi-year experiment, there may be multiple entries for the same type of management, reflecting, for example, repeated applications of herbicide or fertilizer.

note: By default managements are recorded for Yields but not for Traits, unless specifically required by the data or project manager.

To associate a management with multiple treatments, first create the management, then edit the management and add treatment relationships.

Dateloc: date level of confidence, explained in Section 7 and defined in Table 7.

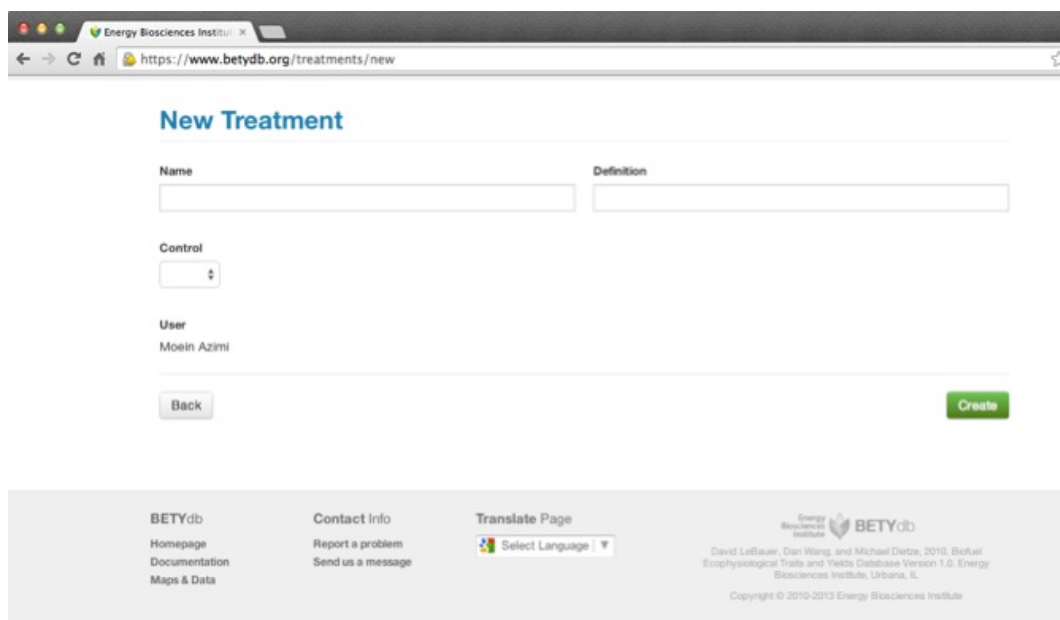
Mgmttype: the name of the management being used. A list of standardized management types can be found in Table 4

Level: a quantification of mgmttype

Units: refers to the units of the level. Units should be converted to those in Table 4

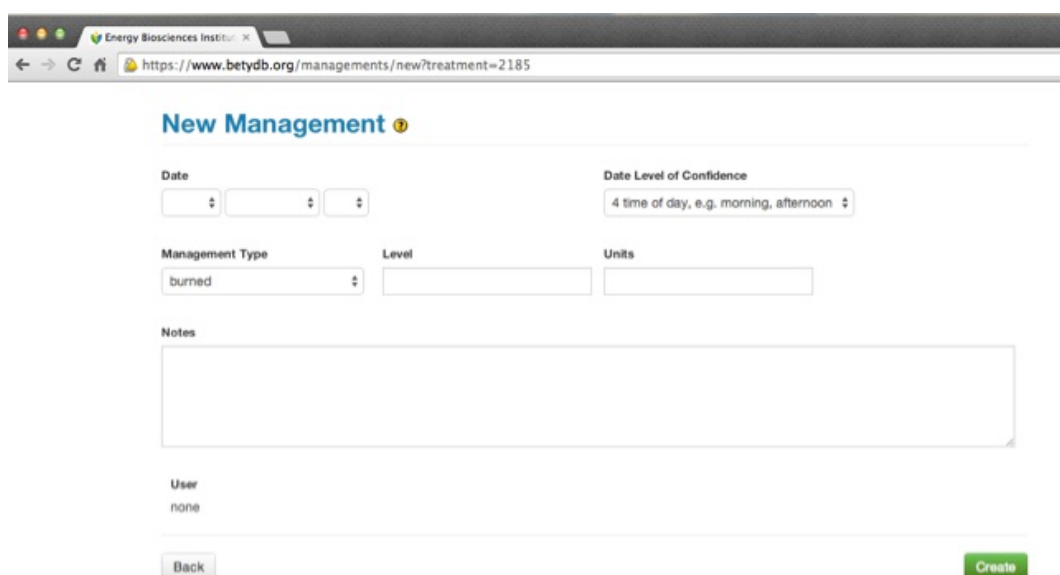
7.3 Editing Management-Treatment Relationships (<https://www.betydb.org/treatments>)

Under Construction for Fall 2014



The screenshot shows a web browser window with the URL <https://www.betydb.org/treatments/new>. The page title is "New Treatment". It contains two input fields for "Name" and "Definition". Below these is a "Control" dropdown menu. The "User" field is populated with "Moein Azimi". At the bottom left is a "Back" button, and at the bottom right is a green "Create" button. A footer section contains links for "BETYdb" (Homepage, Documentation, Maps & Data), "Contact Info" (Report a problem, Send us a message), a "Translate Page" button with a language selector, and the BETYdb logo with a copyright notice: "Copyright © 2010-2013 Energy Biosciences Institute".

Figure 3. Adding a new treatment



The screenshot shows a web browser window with the URL <https://www.betydb.org/managements/new?treatment=2185>. The page title is "New Management". It contains a "Date" field with a calendar icon, a "Date Level of Confidence" dropdown menu (set to "4 time of day, e.g. morning, afternoon"), a "Management Type" dropdown menu (set to "burned"), a "Level" input field, and a "Units" input field. Below these is a large "Notes" text area. The "User" field is populated with "none". At the bottom left is a "Back" button, and at the bottom right is a green "Create" button.

Figure 4. Adding a new management

8 Adding a Trait (<https://www.betydb.org/traits/new>)

In general, a 'trait' is a phenotype (a characteristic that the plant exhibits). The traits that we are primarily interested in collecting data for are listed in Table 7. Before adding trait data, it is necessary to have the citation, treatments, and site information already entered. If the correct citation is not identified at the top of the page Figure 8. To add a new Trait, go to the new trait (<http://www.betydb.org/traits/new>) page: Trait → new.

Presently, we are also using the Trait table to record ecosystem level measurements other than Yield. Such ecosystem level measurements can include leaf area index or net primary productivity, but are only collected when required for a particular project. Most of the fields

in the Traits table are also used in the Yields table. Here is a list of the fields with a brief description, followed by more thorough explanations:

- **Species:** Search for species in the database using the search box; if species is not found, then the new species should be added to the database.
- **Cultivar:** primarily used for crops; If the cultivar being used is not found in drop-down box
- **DateLOC:** Date Level of confidence. See for values.
- **TimeLOC:** Time Level of confidence. See for values.
- **Mean:** For yield, mean is in units of tons per hectare per year (t/ha)
- **Stat name:** is the name of the statistical method used (usually one of SE, SD, MSE, CI, LSD, HSD, MSD). See for more details.
- **Statistic:** is the value of the statistic associated with Stat name.
- **N:** Always record N if provided. N is the number of experimental replicates, often referred to as the sample size; N represents the number of independent units within each treatment: in a field setting, this is often the number of plots in each treatment, but in a greenhouse, growth chamber, or pot-study this may be the number of chambers, pots, or individual plants. Sometimes this value is not clearly stated.

8.1 dateLOC

The date level of confidence (DateLOC, Table 5) provides an indication of how accurately the date associated with the trait or yield observation is known. It provides the values that should be entered in this field. If the event occurred at a level of precision not defined by an integer in this table, then use fractions. For example, we commonly use 5.5 to indicate a one week level of precision. If the exact year is not known, but the time of year is, then use 91 to 97, with the second digit to indicate the information known within the year.

8.2 TimeLOC

The time level of confidence (TimeLOC, Table 5) provides an indication of how accurately the time associated with the trait or yield observation is known. It provides the values that should be entered in this field.

8.3 Statistics

Our goal is to record statistics that can be used to estimate standard deviation or standard error (<https://www.authorea.com/users/5574/articles/68111/> (<https://www.authorea.com/users/5574/articles/68111/>)). Many different methods can be used to summarize data, and this is reflected in the diversity of statistics that are reported. An overview of these methods is given in a description below.

Where available, direct estimates of variance are preferred, including Standard Error (SE), sample Standard Deviation (SD), or Mean Squared Error (MSE). SE is usually presented in the format of mean (\pm SE). MSE is usually presented in a table. When extracting SE or

SD from a figure, measure from the mean to the upper or lower bound. This is different than confidence intervals and range statistics (described below), for which the entire range is collected.

If MSE, SD, or SE are not provided, it is possible that LSD, MSD, HSD, or CI will be provided. These are range statistics and the most frequently found range statistics include a Confidence Interval (95%CI), Fisher's Least Significant Difference (LSD), Tukey's Honestly Significant Difference (HSD), and Minimum Significant Difference (MSD). Fundamentally, these methods calculate a range that indicates whether two means are different or not, and this range uses different approaches to penalize multiple comparisons. The important point is that these are ranges and that we record the entire range.

Another type of statistic is a "test statistic"; most frequently there will be an F-value that can be useful, but this should not be recorded if MSE is available. Only if there is no other information available should you record the P-value.

9 Adding a Yield (<http://www.betydb.org/yields/new>)

The protocol for entering yield data is identical to entering data for a trait, with a few exceptions:

1. There are no covariates associated with yield data
2. Yield data is always the dry harvestable biomass; if necessary, moisture content can be added as a trait

Yield is equivalent to aboveground biomass on a per-area basis, and has units of $\text{Mg ha}^{-1} \text{y}^{-1}$

10 Adding a Covariate

(<http://www.betydb.org/covariates/new>)

Covariates are required for many of the traits. Covariates generally indicate the environmental conditions under which a measurement was made. Without covariate information, the trait data will have limited value.

A complete list of required covariates can be found in Table 9. For all respiration rates and photosynthetic parameters, temperature is recorded as a covariate. Soil moisture, humidity, and other such variables that were measured at the time of the measurement may be required in order to standardize across studies.

When root data is recorded, the root size class needs to be entered as a covariate. The term 'fine root' often refers to the $<2\text{mm}$ size class, and in this case, the covariate `root_maximum_diameter` would be set to 2. If the size class is a range, then the `root_minimum_diameter` can also be used.

The screenshot shows the 'New Trait' form on the Betydb website. The form is organized into several sections:

- Header:** 'New Trait' title.
- Basic Info:** Fields for Mean, Std, Method, Statname, Date, Date Level of Confidence, Time, and Time Level of Confidence.
- Location:** Site (1133: Confluence of Casiquiare River and Rio Negro - San Carlos).
- Species:** Search by Symbol or Scientific Name (No species associated with this record).
- Cultivar:** Cultivar (2180: Succession : growth following fire - UN 1987) and Treatment (2180: Succession : growth following fire - UN 1987).
- Trait:** Trait (540:) and R (2.058 Researchers).
- Access level:** Access level (2.058 Researchers).
- Notes:** A large text area for notes.
- Covariate:** A section titled 'Add a covariate to this trait' with a 'Variable' field and a 'Level' field.
- Buttons:** 'Back' and 'Create' buttons at the bottom.

Figure 5. Adding a new trait & new covariate

11 Adding a PFT, Species, or Cultivar

Plant functional types (PFTs) are used to group plants for statistical modeling and analysis. PFTs are associated with both a specific set of priors, and a subset species for which the traits and yields data will be queried. In many cases, it is appropriate to use default PFTs (e.g. `tempdecid` is temperate deciduous trees)

In other cases, it is necessary to define PFTs for a specific project. For example, to query a specific set of priors or a subset of a species, a new PFT may be defined. For example, Xiaohui Feng defined PFTs for the species found at the EBI Farm prairie. Such project-specific PFTs can be defined as ``projectname`.`pft`` (i.e. `ebifarm.c4grass` instead of `c4grass`).

Species that are found or cultivated in the United States should be in the Plants table. Look it up there first.

To add a new Cultivar, go to the new cultivar (<https://www.betydb.org/cultivars/new>) page:

`Cultivar` → `new`.

The screenshot shows a web browser window with the URL <https://www.betydb.org/cultivars/new>. The page title is 'New Cultivar'. It features a 'Species' section with a search bar labeled 'Search by Symbol or Scientific Name' and a message 'No species associated with this record.' Below this are three input fields: 'Name', 'Ecotype', and 'Notes'. At the bottom of the form are two buttons: 'Back' and 'Create'.

Figure 6. Adding a new cultivar

12 BETYdb: Bulk Data Upload

Currently the web interface does not support bulk data upload, although this is a planned feature for BETY 2.0.

There are three phases for a basic bulk upload of data :

1. Use the web interface
 - to enter metadata (new sites, citations, treatments, managements)
 - obtain a template appropriate for your data set.
2. Fill in the template with your data.
 - traits.csv
(<https://docs.google.com/spreadsheets/d/1lans4FMJ8avn34dcKkMzkEavqyZu6I4WuPP9oqcformat=csv&gid=0>) and can be downloaded in .xls format
 - yields.csv
(<https://docs.google.com/spreadsheets/d/1maK1uKr6i9KERaYdU5zSiXcBndQoiG4Vgn2DTformat=csv&gid=0>).
3. Use the web interface to upload your data set and insert it into the database.

For now, only the steps needed to upload yields data will be outlined in order to make this case simpler.

For clarity, in what follows, the term "field" will be used to refer to the heading used in the uploaded CSV file and the term "column" or "attribute" to refer to an attribute of a yield datum in the yields table of the database.

12.1 Required fields:

1. Citation
 - If only one citation for the entire dataset exists, this may be specified interactively by choosing a citation on the citations page
 - otherwise, specify in the CSV table using either (doi) or (author, year, first n characters of (title)) for some number n; or perhaps (author, year, first 3 words of

(title))

- For citations, if citation_ doi is available, then the rest of fields pertaining to the citation may be left blank. However, if the citation_ doi is not available, then citation_ author, citation_ year, and citation_ title must all be entered.

2. site: use sitename
3. species: use scientificname
4. treatment: use name date: require one of the forms "2003-07-25", "2003-07", or "2003".

Of these, the citation, site, species, treatment, and access_level may be specified interactively when uploading the dataset (if they are uniform for the whole set) rather than appearing as a field or set of fields of the CSV file. As noted above, for citations, this is done outside of the upload wizard by choosing a citation on the citations page.

12.2 Data for Yields

1. mean: must be one of the fields of the CSV table, though for uploads of yields data, we will by default call this field "yield" in the provided templates
2. n: required if and only if an SE column is given
3. SE: required if and only if an n column is given; this datum will be inserted into the stat column, and the statname will be set to "SE" access_level

12.3 Data for Traits

1. <variable_1>, <variable_2>, ... <variable_n>: These column names should be replaced with values from 7 (or see the variables table (<https://www.betydb.org>) for a comprehensive listing).
2. <covariate_1>, <covariate_2>, ... <covariate_n> these columns should be replaced with values from 9.
3. To enter n and SE, add additional columns "<variable_1> n" and "<variable_1> SE" as needed.

12.4 Optional fields:

1. n and SE: as noted above, if one of these is present, the other must be as well; if SE is given, the value will go into the stat column of the yields table, and the statname column will be set to "SE"
2. cultivar: use name; defaults to NULL (for the cultivar_id column) if not provided
3. notes: defaults to the empty string if not provided If a uniform value for the species is provided interactively when uploading the data set, the cultivar may be specified this way as well provided that it also has a uniform value for the whole data set. If n and SE are not given fields of the uploaded CSV file, the value of the n column of the yields table will default to 1 and the stat and statname column values will default to NULL.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	citation_doi	citation_author	citation_year	citations_site	cultivar	species	site	treatment	date	dateloc	mean	n	SE	notes	access_level
2															
3															
4															
5															
6															
7															
8															
9															
10															
11															
12															
13															
14															
15															
16															
17															
18															
19															
20															
21															
22															
23															
24															
25															
26															
27															
28															
29															
30															
31															

Figure 7. Sample template for bulk upload of yield data

13 QA/QC with the Web Interface

Quality assurance and quality control (QA/QC) is a critical step that is used to ensure the validity of data in the database and of the analyses that use these data. When conducting QA/QC, your data access level needs to be elevated to “manager”.

1. Open citation in Mendeley
2. Locate citation in BETYdb
 - Select
 - Select
 - Check that author, year, title, journal, volume, and page information is correct
 - Check that links to URL and PDF are correct, using DOI if available
 - If any information is incorrect, click 'edit' to correct
3. Check that site(s) at bottom of citation record match site(s) in paper
 - Check that latitude and longitude are consistent with manuscript, are in decimals not degrees, and have appropriate level of precision
 - Click on site name to verify any additional information site information that is present
 - Enter any additional site level information that is found
4. Select treatments (<http://www.betydb.org/treatments/>) from menu bar
 - Check that there is a control treatment
 - Ensure that treatment name and definition are consistent with information in the manuscript
 - Under “treatments from all citations associated with associated sites”, ensure that there is no redundancy (i.e. if another citation uses the same treatments, it should not be listed separately)
 - If managements are listed, make sure that managment-treatment associations are correct
5. Check managements (<http://www.betydb.org/managements/>) if there are any listed on the treatments page.

- If yield data has been collected, ensure that required managements have been entered
 - If managements have been entered, ensure that they are associated with the correct treatments
6. Click Yields (<http://www.betydb.org/yields/>) or Traits (<http://www.betydb.org/traits/>) to check data.
- Check that means, sample size, and statistics have been entered correctly
 - If data has been transformed, check that transformation was correct in the associated google spreadsheet (or create a new google spreadsheet following instructions)
 - For any trait data that requires a covariate

14 Extracting information from tables and graphs

To extract information from a figure, the general method is:

1. upload an image
2. set the x and y scales by indicating the values at two points on each axis
3. indicate if the scale is linear, log, etc,
4. click on the points.

Some software programs automatically recognize lines or points. However, since points are usually sought after, the results are often too inconsistent to be helpful even with 100s of points. Also, no program has yet been found to be able to distinguish different symbols. This feature could be worth the trouble for digitizing lines, but this is not commonly required.

The program returns each point as an x-y matrix.

Often it helps selecting points if the image is zoomed, either by uploading a zoomed version of the image or using the zooming feature available in some of the programs.

14.0.1 List of available Programs

I have experience with the following programs. All of these work well fine. Except in contexts where measurement error is very small, error from graph scraping is insignificant (e.g. error from digitization << size of error bars or uncertainty in the estimate). I have not tested the accuracy of any of these programs, but it would be interesting to compare among users, among programs, and against the results of reproduced statistical analyses.

- Digitizer (<http://digitizer.sourceforge.net/>) (shareware) auto point / line recognition. Available in Ubuntu repository (engage-digitizer)
- Get Data (<http://www.getdata-graph-digitizer.com/>) (shareware) has zoom window, auto point / line recognition
- Digitizeit (<http://www.digitizeit.de/>) (shareware) auto point / line recognition
- ImageJ (<http://rsbweb.nih.gov/ij/>) (open source, most extensible after R digitize)
- R digitize (<http://cran.r-project.org/web/packages/digitize/index.html>) (free, open source), because it simplifies the process of getting data from the graph into an analysis by keeping all of the steps in R. See the tutorial in R-Journal (http://journal.r-project.org/archive/2011-1/RJournal_2011-1_Poisot.pdf)

- GrabIt! (<http://www.datatrendsoftware.com/home.html>) (free demo, \$69) Excel plug-in

I have not used these:

- WebPlotDigitizer (<http://arohatgi.info/WebPlotDigitizer/app/>) (free, online). Extracts data from images. Demo here (<http://blog.plot.ly/post/70293893434/automatically-grab-data-from-an-image-with/>).
- GraphClick (<http://www.arizona-software.ch/graphclick/>) (Mac, \$8)
- g3data (<http://www.frantz.fi/software/g3data.php>) (open source - GNU GPL) Has zoom window, no auto-recognition. Available in Ubuntu repository.

See related question on Stats.stackexchange (<http://stats.stackexchange.com/a/14440/1381>)

1. Identify the data that is associated with each treatment *note*: If the experiment has many factors, the paper may not report the mean and statistics for each treatment. Often, the reported data will reflect the results of more than one treatment (for example, if there was no effect of the treatment on the quantity of interest). In some cases it will be possible to obtain the values for each treatment, e.g. if there are $n-1$ values and n treatments. If this is not the case, the treatment names and definitions should be changed to indicate the data reflect the results of more than one experimental treatment.
2. Enter the mean value of the trait
3. Enter the `statname`, `stat`, and number of replicates, `n` associated with the mean
 - `stat` is the value of the `statname` (i.e. `statname` might be 'standard deviation' (SD) and the `stat` is the numerical value of the statistic)
 - Always measure size of error bar from the mean to the end of an error bar. This is the value when presented as ($X \pm SE$) or $X(SE)$ and may be found in a table or on a graph.
 - Sometimes CI and LSD are presented as the entire range from the lower to the upper end of the confidence interval. In this case, take 1/2 of the interval representing the distance from the mean to the upper or lower bound.

14.0.2 Extracting Data using R

To extract data from a jpg file in R using the digitize package:

1. Save image as a `*.jpg` file
2. Open R
3. Change the directory that R is using to the one where the image is
4. Use R code below to extract data, display it, and save it in a `csv` file (steps below)
5. Upload csv to the project file in google spreadsheet, or open as excel/openoffice and copy/paste to google spreadsheet

14.0.3 Extracting Data From a Figure using GetData

1. Open PDF in Adobe Reader.
2. Zoom in on the figure
3. Choose `Tools` → `Select and Zoom`
4. Open Paint

5. Paste Picture
6. Save as `authorYYYYabc_figX.jpg`
7. Open Get Data
8. `File` → `open` open figure
9. Select button with two arrows (fourth from left)
10. Follow instructions to select x min, x max, y min and y max. If the x-axis has a categorical variable, it does not matter what values you use for x min and x max.
11. Make sure to set the correct values for the max and min of each axis, and indicate if the axis is log-scaled
12. Select the target button (seven from left)
13. Click over center of desired data points and error bars
14. Copy data to a Google spreadsheet. See Google Spreadsheets.
15. Calculate SE as the distance between the error bar upper bound and the mean (absolute value of difference between the two points)

How to convert statistics from P , LSD , or MSD to SE

Many statistical transformations are implemented in the `transformstats` (<https://github.com/PecanProject/pecan/blob/master/utils/R/transformstats.R>) function within the `PEcAn.utils` package. However, these transformations make conservative (variance inflating) assumptions about study-specific experimental design (especially degrees of freedom) that is not captured in the `BETYdb` schema, for example HSD , LSD , P . More accurate estimates of SE can be obtained at time of data entry using the formulas in "Transforming ANOVA and Regression statistics for Meta-analysis" (<https://www.authorea.com/users/5574/articles/68111/>).

14.1 Converting Units and Adjustment to Temperature

For many transformations, particularly when automated, please use the `udunits2` software where possible. For example, in R, you can use

```
library(udunits2)
## transform meters to mm
ud.convert(10, "m", "mm")
## equivalently, via the udunits synonym database
ud.convert(10, "meters", "millimeters")
## it can also handle more complex units
ud.convert(10, "m/s", "mm/d")
```

NB: Many of these conversions have been automated within `PEcAn` (<https://github.com/PecanProject/pecan>).

Table 3: Useful conversions for entering site, management, yield, and trait data

From (X)	to (Y)	Conversion	Notes
	$X_1 = \text{root}$		

X_2 = root production	biomass & root turnover rate	$Y = X_2/X_1$	
DD° MMSS	XX.ZZZZ	$XX.ZZZZ = XX + MM/60 + SS/60$	to convert latitude or longitude to minutes, seconds
lb	kg	$Y = X \times 2.2$	
mm/s	$\mu \text{ mol CO}_2 \text{ m}^{-2} \text{ s}^{-1}$	$Y = X \times 0.04$	
m^2	ha	$Y = X/10^6$	
g/m^2	kg/ha	$Y = X \times 10$	
US ton/acre	Mg/ha	$Y = X \times 2.24$	
m^3/ha	cm	$Y = X/100$	units used for irrigation
% roots	root:shoot (q)	$Y = \frac{X}{1-X}$	% root
$\mu \text{ mol cm}^{-2} \text{ s}^{-1}$	$\text{mmol m}^{-2} \text{ s}^{-1}$	$Y = X/10$	
$\text{mol m}^{-2} \text{ s}^{-1}$	$\text{mmol m}^{-2} \text{ s}^{-1}$	$Y = X/10^6$	
$\text{mol m}^{-2} \text{ s}^{-1}$	$\mu \text{ mol cm}^{-2} \text{ s}^{-1}$	$Y = X/10^5$	
mm s^{-2}	$\text{mmol m}^{-3} \text{ s}^{-1}$	$Y = X/41$	
$\text{mg CO}_2 \text{ g}^{-1} \text{ h}^{-1}$	$\mu \text{ mol kg}^{-1} \text{ s}^{-1}$	$Y = X \times 6.31$	used for respiration
$\mu \text{ mol}$	mol	$Y = X \times 10^6$	
julian day (1--365)	date		http://disc.gsfc.nasa.gov/julian (NASA)
spacing (m)	density (plants m^{-2})	$Y = \frac{1}{\text{row spacing} \times \text{plant spacing}}$	
$\text{kg ha}^{-1} \text{ y}^{-1}$	$\text{Mg ha}^{-1} \text{ y}^{-1}$	$Y = X/1000$	
$\text{g m}^{-2} \text{ y}^{-1}$	$\text{Mg ha}^{-1} \text{ y}^{-1}$	$Y = X/100$	
kg	mg	$Y = X \times 10^6$	
cm^2	m^2	$Y = X \times 10^4$	

15 Reference Tables

Table 4 Managements This is a list of managements to enter, with the most common management types in bold. It is more important to have management records for Yields than for traits. For greenhouse experiments, it is not necessary to include informaton on fertilizaton, lighting, or greenhouse temperature.

Management Type	Units	Definition	Notes
Burned	aboveground biomass burned		
CO2 fumigation	ppm		
Fertilization_X	kg x ha ⁻¹	fertilization rate, element X	
Fungicide	kg x ha ⁻¹		add type of fungicide to notes
Grazed	years	livestock grazing	pre-experiment land use
Harvest			no units, just date, equivalent to coppice, aboveground biomass removal
Herbicide	kg x ha ⁻¹		add type of herbicide to notes: glyphosate, atrazine, many others
Irrigation	cm		convert volume \ area to depth as required
Light	W m ⁻²		
O3 fumigation	ppm		
Pesticide	kg x ha ⁻¹		add type of pesticide to notes
Planting	plants m ⁻²		Convert row spacing to planting density if possible
Seeding	kg seeds x ha ⁻¹		
Tillage			no units, maybe depth; <i>tillage</i> is equivalent to <i>cultivate</i>

Table 5: Table 4: Table 5: Date level of confidence (DateLOC) field Numbering convention for the DateLOC (Date level of confidence) and TimeLOC (Time level of confidence) field, used in managements, traits, and yields table. .

Dateloc Definition

9

no data

8	year
7	season
6	month
5	day
95	unknown year, known day
96	unknown year, known month
...etc	

Timeloc Definition

9	no data
4	time of day i.e. morning, afternoon
3	hour
2	minute
1	second

Table 6: Table 6: List of statistical summaries List of the statistics that can be entered into the statname field of traits and yields tables. Please see David (or Mike) if you have questions about statistics that do not appear in this list. If you have P, or LSD in a study with $n \neq b$ (e.g. not a RCBD, see Table 8), please convert these values prior to entering the data, and add a note that stat was transformed to the table. Note: These are listed in order of preference, e.g., if SD, SE, or MSE are provided then use these values.

Statname	Name	Definition	Notes
SD	Standard Deviation	$\sqrt{\frac{1}{N} \sum (x_i - \bar{x})^2}$	\bar{x} is the mean
SE	Standard Error	$\frac{s}{\sqrt{n}}$ &	
MSE	Mean Squared Error		
95%CI	95% Confidence Interval	$t_{1-\alpha/2, n} * s$	measure the 95% CI from the mean, this is actually $1/2$ of the CI
LSD	Least Significant Difference	$t_{1-\frac{\alpha}{2}, n} \sqrt{2\text{MSE}/b}$	b is the number of blocks (Rosenberg 2004)
MSD	Minimum Significant Difference		

Table 7: Table 7 Key Trait Variables

Variable	Units	Median (90%CI) or Range	Definition
----------	-------	-------------------------	------------

V _c max	$\mu \text{ mol CO}_2 \text{ m}^2 \text{ s}^{-1}$	44(12, 125)	maximum rubisco carboxylation capacity
SLA	$\text{m}^2 \text{ kg}^{-1}$	15(4, 27)	Specific Leaf Area area of leaf per unit mass of leaf
LMA	kg m^{-2}	0.09(0.03, 0.33)	Leaf Mass Area (LMA = SLM = 1/SLA) mass of leaf per unit area of leaf
leafN	%	2.2(0.8, 17)	leaf percent nitrogen
c2n leaf	leaf C:N ratio	39(21, 79)	use only if leafN not provided
leaf turnover rate	1/year	0.28(0.03, 1.0)	
J _{max}	$\mu \text{ mol photons m}^{-2} \text{ s}^{-1}$	121(30, 262)	maximum rate of electron transport
stomatal slope		9(1, 20)	
GS			stomatal conductance (= gs _{max})
q*		0.2--5	ratio of fine root to leaf biomass
*grasses	ratio of root:leaf = below:above ground biomass		
aboveground biomass	g m^{-2} or g plant^{-1}		
root biomass	g m^{-2} or g plant^{-1}		
*trees	ratio of fine root:leaf biomass		
leaf biomass	g m^{-2} or g plant^{-1}		
fine root biomass (<2mm)	g m^{-2} or g plant^{-1}		
root turnover rate	1/year	0.1--10	rate of fine root loss (temperature dependent) year ⁻¹
leaf width	mm	22(5, 102)	
growth respiration factor	%	0--1	proportion of daily carbon gain lost to growth respiration
R _{dark}		$\mu \text{ mol CO}_2 \text{ m}^{-2} \text{ s}^{-1}$	dark respiration

quantum efficiency	%	0--1	efficiency of light conversion to carbon fixation, see Farquhar model
dark respiration factor	%	0--1	converts Vm to leaf respiration
seedling mortality	%	0--1	proportion of seedlings that die
r fraction	%	0--1	fraction of storage to seed reproduction
root respiration rate*	CO ₂ kg ⁻¹ fine roots s ⁻¹	1--100	rate of fine root respiration at reference soil temperature
f labile	%	0--1	fraction of litter that goes into the labile carbon pool
water conductance			

Table 9: Table 8: Table 9: Traits with required covariates A list of traits and the covariates that must be recorded along with the trait value in order to be converted to a constant scale from across studies. *notes:* stomatal conductance (gs) is only useful when reported in conjunction with other photosynthetic data, such as Amax. Specifically, if we have Amax and gs, then estimation of Vcmax only covaries with dark_respiration_factor and atmospheric CO2 concentration.

We also now have information to help constrain stomatal_slope. If we have Amax but not gs, then our estimate of Vcmax will covary with: darkrespirationfactor, CO2, stomatalslope, cuticularconductance, and vapor-pressure deficit VPD (which is more difficult to estimate than CO2, but still possible given lat, lon, and date). Most important, there will be a strong covariance between Vcmax and stomatal_slope.

Variable	Required Covariates	Optional Covariates
vcmax	irradiance and temperature (leaf or air)	
any leaf measurement		canopy height
root_respiration_rate	temperature (root or soil)	soil moisture
	root_diameter_max	root size class (usually 2mm)
any respiration	temperature	
root biomass		min. size cutoff, max. size cutoff

root, soil	depth (cm)	used for max and min depths of soil, if only one value, assume min depth = 0; negative values indicate above ground
gs (stomatal conductance)	A_{max}	see notes in caption
stomatal_slope (m)	humidity, temperature	specific humidity, assume leaf T = air T

All public data in BETYdb is made available under the Open Data Commons Attribution License (ODC-By) v1.0 (<http://opendatacommons.org/licenses/by/1-0/>). You are free to share, create, and adapt its contents. Data with an access_level field and value <= 2 is not covered by this license, but may be available for use with consent.

Please cite the source of data as:

LeBauer, David; Dietze, Michael; Kooper, Rob; Long, Steven; Mulrooney, Patrick; Rohde, Gareth Scott; Wang, Dan; (2010): Biofuel Ecophysiological Traits and Yields Database (BETYdb); Energy Biosciences Institute, University of Illinois at Urbana-Champaign. <http://dx.doi.org/10.13012/J8H41PB9> (<http://dx.doi.org/10.13012/J8H41PB9>)

16 Acknowledgements

Patrick Mulroony originally implemented the data entry interface, and it is currently maintained by Scott Rohde. Rob Kooper, Andrew Shirk, and Carl Crott have contributed (see visualization on GitHub (<https://github.com/PecanProject/bety/graphs/contributors>)).

Many data entry technicians (undergrads) have contributed to the implementation and development of the interface and documentation. These include Moein Azimi, David Bettinardi, Nick Brady, Emily Cheng, Anjali Patel, along with other members of the EBI Feedstock Productivity and Ecosystem Services modeling group.