
Analyzed Phenotypes Documentation

Reynold Tan et al., University of Saskatchewan, Pulse Bioinforma

Feb 21, 2019

Contents:

1	Administrative Guide	3
1.1	Installation	3
1.2	General Usage	3
1.3	Configuration	8
1.4	Benchmarking	9
1.5	Data Storage	12
2	User Guide	13
2.1	Uploading of Analyzed Phenotypic Data	13
2.2	Download of Analyzed Phenotypic Data	18
2.3	Visualize the data	22

This module provides support and visualization for partially analyzed data stored in a modified GMOD Chado schema. It is meant to support large scale phenotypic data through backwards compatible improvements to the Chado schema including the addition of a project and stock foreign key to the existing phenotype table, optimized queries and well-chosen indexes.

This guide is meant for administrators of a Tripal site. It will show you how to install, configure and provide basic usage orientation.

1.1 Installation

Install this module as you would any other Drupal module after ensuring you have the following dependencies:

1. Drupal 7 (<https://www.drupal.org/>)
2. Tripal 3.x (<http://tripal.info/>)
3. Tripal Download API (https://github.com/tripal/trpdownload_api)
4. PostgreSQL 9.3 (<https://www.postgresql.org/>)
5. Drag and Drag Upload module (https://www.drupal.org/project/dragndrop_upload)
6. PHP Excel Writer Libraries (https://github.com/SystemDevil/PHP_XLSXWriter_plus)
7. D3 JavaScript Library (<https://github.com/d3/d3/releases/download/v3.5.14/d3.zip>)

1.2 General Usage

1. Configure this module by accessing Configuration page: `[your site]/admin/tripal/extension/analyzedphenotypes/configuration`.

Configuration

TRAIT ONTOLOGIES

This module requires that phenotypic traits be part of a controlled vocabulary.

- **Trait Vocabulary:**
A container of terms where each term is a phenotypic trait that can be measured in your species of interest. This controlled vocabulary should be specific to a given genus and each term will become a trait page on your Tripal site. If you do not already have a trait vocabulary, you can create it [here](#) and add terms upfront and/or automatically on upload of phenotypic data.
- **Associated Database:**
Chado requires a "database" container to be associated with all controlled vocabularies. Please select the "database" container you would like to be associated with your trait vocabulary. If needed, create one [here](#).
- **Crop Ontology:**
Our experience with breeders has led us to recommend using the trait names your breeder(s) already use in the Trait Vocabulary and then linking them to a more generic crop ontology such as those provided by cropontology.org to facilitate sharing. If you decide to go this route, you can set the species specific crop ontology here and on upload suitable terms will be suggested based on pattern matching.

Please select the appropriate vocabulary for each genus you intend to support phenotypic data for.

GENUS	TRAIT VOCABULARY	ASSOCIATED DATABASE	CROP ONTOLOGY
Citrus	<input type="text" value="feature_property"/>	<input type="text" value="AGL"/>	<input type="text" value="Select Crop Ontology"/>
Tripalus	<input type="text" value="local"/>	<input type="text" value="data"/>	<input type="text" value="Select Crop Ontology"/>

Allow new terms to be added to the Controlled Vocabulary during upload.
This applies to all organism listed above.

 Once phenotypic data has been uploaded for a genus, these vocabularies cannot be changed! Please take the time to read the description above and if you have questions, submit a ticket to [Github: UofS-Pulse-Binfo](#).

CONTROLLED VOCABULARY TERMS

Chado uses controlled vocabularies extensively to allow for flexible storing of data. As such, this module supports that flexibility to ensure that you have the ability to choose the terms that best support your data.

 We have helpfully selected what we think are the best ontology terms below. Thus the following configuration is completely optional, although I do recommend you review our choices.

Please indicate the term we should use to indicate the property/relationship types specified below:

Figure 1. Configuration page.

2. Load or upload analyzed phenotypic data using the Upload page. To access upload page use the following link: [your site]/admin/tripal/extension/analyzedphenotypes/upload.

Home » Administration » Tripal » Extensions » Analyzed Phenotypes

Upload Phenotypic Data



Phenotypic data should be **filtered for outliers and mis-entries** before being uploaded here. Do not upload data that should not be used in the final analysis for a scientific article. Furthermore, data should **NOT be averaged across replicates or site-year**.

STAGE 1 OF 4 - UPLOAD



Experiment *

Type in the experiment or project title your data is specific to.

Genus *

Select Genus. When experiment or project has genus set, a value will be selected.

TSV Data File

- Drag and drop your file here -
or [choose a file](#)

Tab Separated Values (tsv), Text File (txt) only

Figure 2. Upload Page.

3. You can export a fullset or subset of data using data download page. To access download page use the following link: [your site]/phenotype/download

Download Analyzed Phenotypic Data

Select Trait.

 Please note that filter criteria may become deactivated when exporting data from multiple experiments or species.

Indicate the trait you would like phenotypic data for by selecting experiment and the genus of the crop, as well as the name of the trait below. Field marked with a * means field is required and must have a value before proceeding.

Restrict the dataset to a specific experiment. This can be done by clicking on the name of the experiment below. You can further filter by year and location if desired.

Experiment *

Experiment ABC

Select the genus of the crop you would like phenotypic data for. Additionally, the species can be indicated to further restrict the germplasm phenotypic data is exported for.

Genus *

- Select -

Species *

- Please select an option -

Trait Name *

- Please select an option -

▼ Additional filter criteria (Optional).

Figure 3. Data Download Page.

4. Summary of analyzed phenotypic data and data visualization can be viewed using the following link [your site]/phenotypes.

Phenotypes

Genus	Traits	Experiments	Germplasm	Measurements
<i>Tripalus</i>	1	1	132	132

To visualize the distribution of values for a given trait, see the [Trait Distribution Chart](#).

Trait Distribution Chart

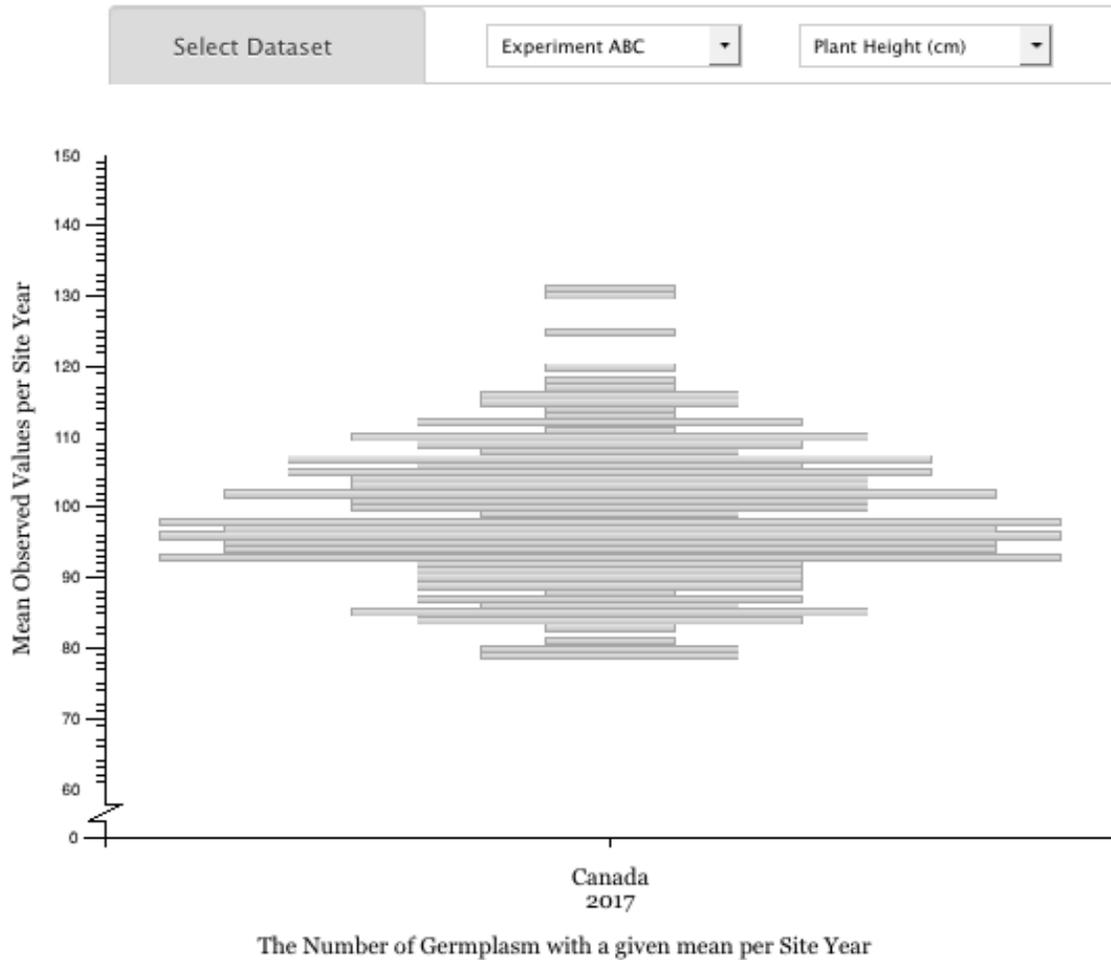


Figure: The Distribution of mean *Plant Height (cm)* per site year for *Experiment ABC*. This chart shows the distribution of means observed (y-axis) per site year (x-axis) with the width of each bar representing the number of germplasm a given mean was observed/measured in. Only the data for the *Experiment ABC* phenotyping experiment is being shown.

[Go back to Summary Table](#)

Figure 4: Summary Page and Trait Distribution Chart

1.3 Configuration

Chado uses controlled vocabulary extensively to allow for flexible storing of data. As such this module supports that flexibility to ensure that regardless of the type used for your data, this module will still be able to navigate the necessary relationships and interpret your types.

1.3.1 Controlled Vocabulary

A container of terms where each term is a phenotypic trait that can be measured in your species of interest. This controlled vocabulary should be specific to a given genus and each term will become a trait page on your Tripal site. If you do not already have a trait vocabulary, you can create it and add terms upfront and/or automatically on upload of phenotypic data (add link to create cv).

1.3.2 Database

Chado requires a “database” container to be associated with all controlled vocabularies. Please select the “database” container you would like to be associated with your trait vocabulary. If needed, create one by clicking this link (add link to phenotype page)

1.3.3 Ontology

Our experience with breeders has led us to recommend using the trait names your breeder(s) already use in the Trait Vocabulary and then linking them to a more generic crop ontology such as those provided by cropontology.org to facilitate sharing. If you decide to go this route, you can set the species specific crop ontology here and on upload suitable terms will be suggested based on pattern matching.

1.3.4 Allow New Traits

Allow new terms to be added to the Controlled Vocabulary during upload option will allow and prevent new records to be inserted into the Controlled Vocabulary configuration. This option can be used to limit the number of traits that user can upload data to a given experiment.

1.3.5 Controlled Vocabulary Terms

Chado uses controlled vocabularies extensively to allow for flexible storing of data. As such, this module supports that flexibility to ensure that you have the ability to choose the terms that best support your data.

1.4 Benchmarking

We decided to do more formal benchmarking on two of our modules for the ISMB 2017 Conference. The details of such are included here for the benefit of the community :-).

1.4.1 Caveats

1. All timings were done on the same hardware (see specification below).
2. Queries were timed at the database level using PostgreSQL 9.4.10 EXPLAIN ANALYZE [query] and as such don't include rendering time in Tripal. Note: the addition of the analyze keyword ensures the query is actually run and the actual total time was reported.
3. The system the tests were run on includes a production Tripal site with small and uneven load. The tests were run 3 times on the same day over the span of at least 4 hours to help mitigate the differences in load.
4. Datasets are computationally derived with no missing data points.

1.4.2 Timings

Timings were done on July 18,2017

Dataset	Query	Rep1	Rep2	Rep3	Average
#1	Quantitative Mview	32.709 ms	25.628 ms	25.981 ms	28.106 ms
#1	Quantitative Directly	1167.909 ms	1159.963 ms	1158.73 ms	1162.2 ms
#1	Summary	0.011 ms	0.004 ms	0.003 ms	0.006 ms

- See “Datasets” for a description of the datasets the tests were run on and how they were generated.
- See “Queries” section below for the exact queries executed.
- See “Hardware” section for the specification of the database server all tests were run on.

1.4.3 Datasets

The queries were tested on two phenotypic datasets with different composition. Both datasets were generated using the [Generate Tripal Data Drush module](https://github.com/UofS-Pulse-Binfo/generate_trpdata); specifically, the drush generate-phenotypes command. While the data is computationally derived, it does attempt to simulate real data by choosing the range of values for each trait and then generating quantitative values along a normal distribution. Furthermore, it ensures that replicate values are within 3 units of each other.

Name	Trait	SiteYears	Germplasm	Measurements (Averaged across reps)
Dataset #1	100	100	4500	135 million
Dataset #2	100	10,000	45	135 million

1.4.4 Queries

The queries executed represent those used to summarize phenotypic data results. Keep in mind that the results from the queries may be further processed before display and that times reported here do not include render times as stated in the caveats section above.

Quantitative Measurement Distribution

This is the query executed to extract the quantitative data collected for a single trait within a single experiment. The data retrieved represents pre-computed means per germplasm and site-year combination for a given trait (denoted :trait_id) and experiment (denoted :project_id).

```
SELECT location, year, stock_name, mean
FROM chado.mview_phenotype
WHERE experiment_id=:project_id AND trait_id=:trait_id
```

This query is made much simpler thanks to the use of a materialized view. For context, the following query is used to generate the materialized view:

```
SELECT
  o.genus as organism_genus,
  trait.cvterm_id as trait_id,
  trait.name as trait_name,
  proj.project_id as project_id,
  proj.name as project_name,
```

(continues on next page)

(continued from previous page)

```

loc.value as location,
yr.value as year,
s.stock_id as germplasm_id,
s.name as germplasm_name,
avg( CAST(p.value as FLOAT) ) as mean
FROM chado.phenotype p
LEFT JOIN chado.cvterm trait ON trait.cvterm_id=p.attr_id
LEFT JOIN chado.project proj USING(project_id)
LEFT JOIN chado.stock s USING(stock_id)
LEFT JOIN chado.organism o ON o.organism_id=s.organism_id
LEFT JOIN chado.phenotypeprop loc ON loc.phenotype_id=p.phenotype_id
AND loc.type_id IN (SELECT cvterm_id FROM chado.cvterm WHERE name='Location')
LEFT JOIN chado.phenotypeprop yr ON yr.phenotype_id=p.phenotype_id
AND yr.type_id IN (SELECT cvterm_id FROM chado.cvterm WHERE name='Year')
GROUP BY trait.cvterm_id, trait.name, proj.project_id, proj.name, loc.value, yr.value,
→ s.stock_id, s.name, o.genus;

```

Experiment Summary

This is the query executed on the main phenotype page which summarizes how many traits, experiments, unique site-years and measurements (averaged across reps) in the current Tripal site broken down by crop/organism. This query is greatly improved by the use of a materialized view.

```
SELECT * FROM chado.mview_phenotype_summary;
```

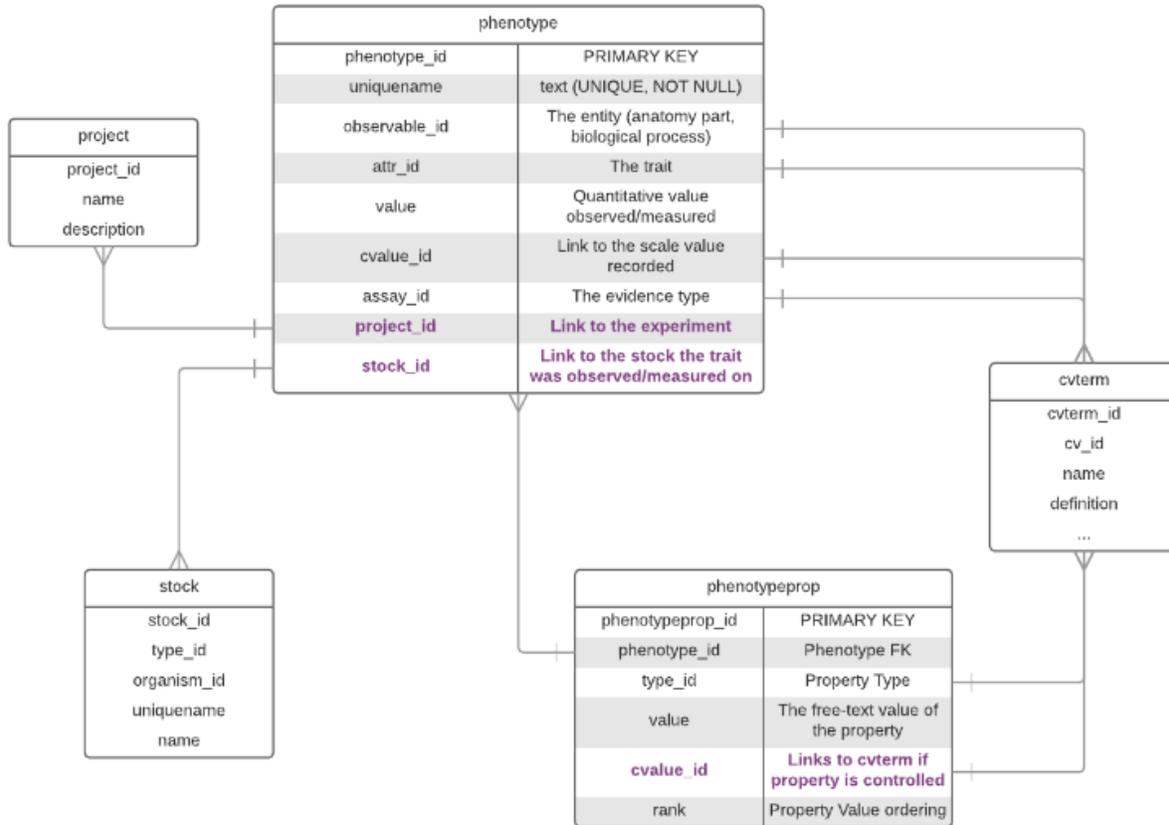
1.4.5 System Specification

Our Production Tripal site is setup on a dedicated two-box system (webserver + database server) with Apache + PHP installed on the first box and PostgreSQL installed on the second box. All testing for this benchmarking was done on a clean Tripal v3 site setup on the same two boxes in order to show queries time on a Production Server versus a less powerful Development server.

- RAID 10 configuration
- Debian GNU/Linux 8.7 (jessie)
- PostgreSQL 9.4.10
- Minimal PostgreSQL configuration tuning
- Hardware Specification (Database Server only)
 - Lenovo X3650 M5 2U Rackmount
 - Server 2x Xeon 6C E52643 V3 3.4GHz
 - 128GB RAM (8x 16GB TruDDR4 Memory (2Rx4, 1.2V) LP RDIMM) 1x ServeRAID M5210 Controller w/ 1GB Flash/RAID 5 Upgrade
 - 8x 600GB 15K 6Gbps SAS 2.5in G3HS HDD
 - Redundant Power Supplies
 - 4x 1GbE Onboard Ethernet

1.5 Data Storage

Phenotypic data is stored in the existing Chado phenotype table with the addition of a project and stock foreign key. This allows phenotypic data measurements to be linked directly to the germplasm they were taken from rather than through the Chado nd_experiment tables providing a huge efficiency boost.



This allows the trait (attr_id), measurement (value or cvalue_id), germplasm (stock_id) combination for a given project (project_id) to be stored as a single record. The location, year, data collector, etc for that data point are then stored in the phenotypeprop table.

This guide is meant for data curators and users of your Tripal site. It demonstrates some of the functionality and provides tutorials for data import and export.

2.1 Uploading of Analyzed Phenotypic Data

The upload data page handles and processes analyzed phenotypic data translated in the form of a Tabbed Delimited Values or .TSV file. The process is divided into 4 stages.

STAGE 1 OF 4 - UPLOAD



Figure 1: Stage Indicator in Upload Page.

2.1.1 Stage 1: Upload

Basic compliance tests on the file level are performed to ensure that requirements outlined are met. For instance, file must be a valid .tsv file and experiment has been selected.

STAGE 1 OF 4 - UPLOAD



Experiment *

Type in the experiment or project title your data is specific to.

Genus *

Select Genus. When experiment or project has genus set, a value will be selected.

TSV Data File

- Drag and drop your file here -
or [choose a file](#)

Tab Separated Values (tsv), Text File (txt) only

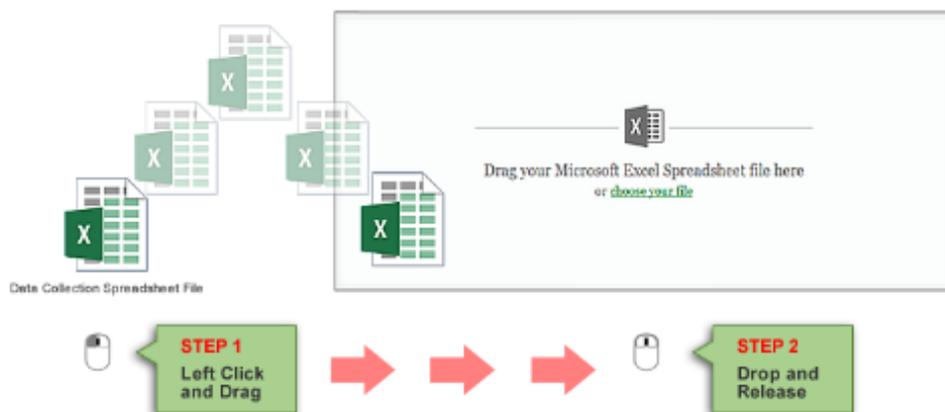


Figure 2: Upload Page Supports Drag and Drop File Upload, as well as Manual Upload.

2.1.2 Stage 2: Validate

In this stage, the file undergoes a data level validation where data in rows and columns are tested against a set of validation rules to ensure that a value meets a set of conditions and requirements.

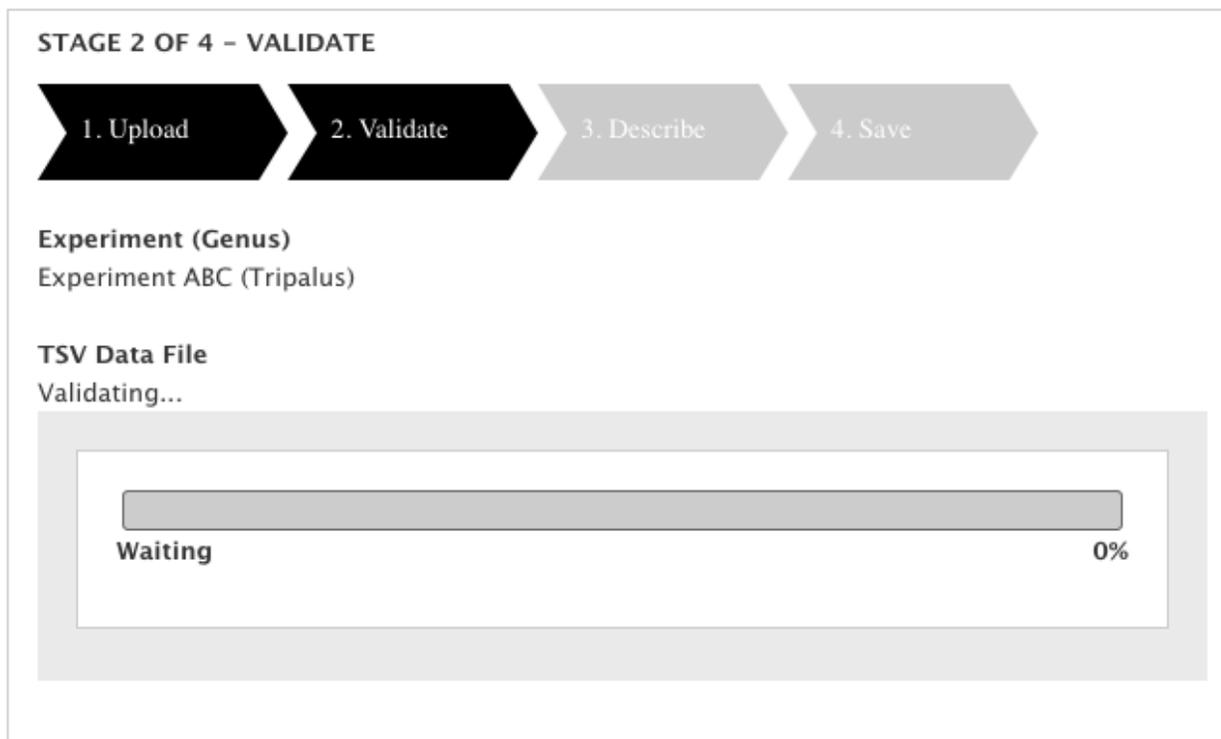


Figure 3: Validate Stage shows progress bar to show user the status of data validation process.

2.1.3 Stage 3: Describe

The file is further examined for all the unique traits in the Trait Name column. Information is then requested from the user for each trait detected.

STAGE 3 OF 4 – DESCRIBE

1. Upload 2. Validate 3. Describe 4. Save

Experiment (Genus)
Experiment ABC (Tripalus)

! Please fully describe the following traits before clicking the next step button.

#1. PLANT HEIGHT (CM)

Did you mean?
None of these apply

The system has detected a similar trait in the database. It is recommended that you select a trait from the select box that best describes your data. If trait is not listed, please select None of these apply option and use the form below to describe the trait.

▼ ABOUT THE TRAIT

Name
Plant Height

A Concise human-readable name or label for the trait.

Unit
cm

Unit used to measure this trait.

Description of Trait Including Collection Method

Text definition or description of trait.

▼ PHOTO UPLOAD

Image 1 of 2
Browse... No file selected.

JPG file type only.

Image 2 of 2
Browse... No file selected.

Figure 4: Describe Stage showing data form requesting user to provide information about the trait detected in data file.

2.1.4 Stage 4: Save

File and data are stored.

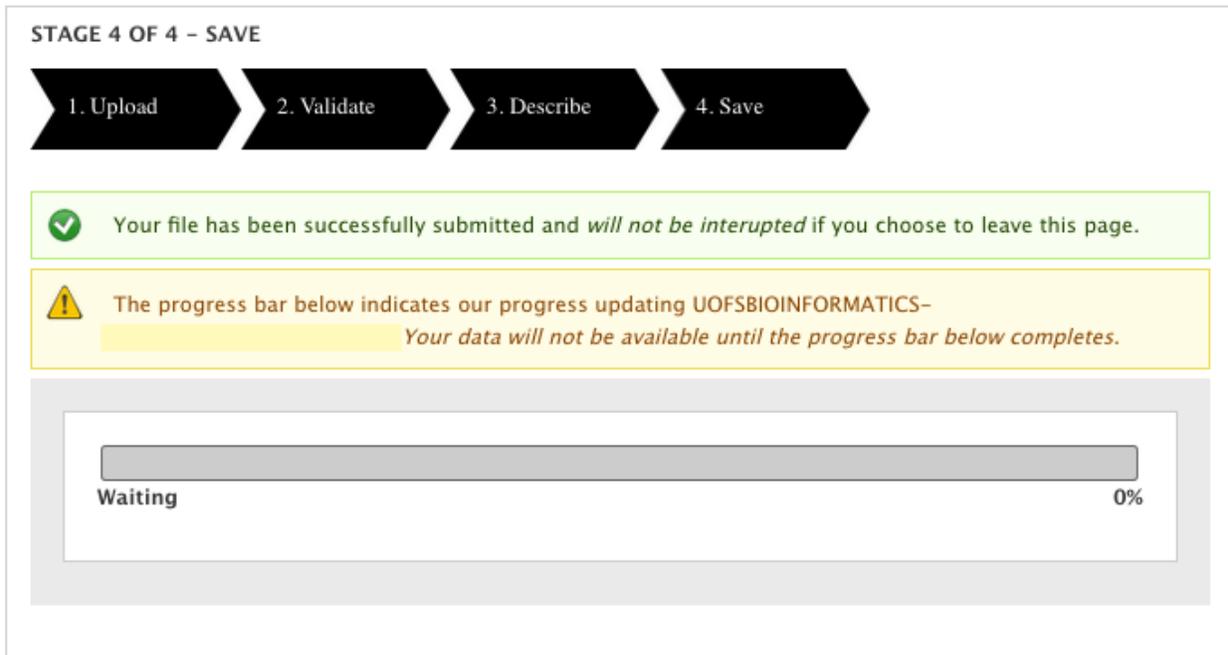


Figure 5: Show the final stage of the upload process. Similar to validate stage, a progress bar shows the status of saving process.

2.2 Download of Analyzed Phenotypic Data

Download data page is where user can download a subset of analyzed phenotypes data, as well as, the full set. Data generated in this page can be xlsx, tsv and csv file format.

Download Analyzed Phenotypic Data

Select Trait.

 Please note that filter criteria may become deactivated when exporting data from multiple experiments or species.

Indicate the trait you would like phenotypic data for by selecting experiment and the genus of the crop, as well as the name of the trait below. Field marked with a * means field is required and must have a value before proceeding.

Restrict the dataset to a specific experiment. This can be done by clicking on the name of the experiment below. You can further filter by year and location if desired.

Experiment *

Experiment ABC

Select the genus of the crop you would like phenotypic data for. Additionally, the species can be indicated to further restrict the germplasm phenotypic data is exported for.

Genus *

Species *

basica
 databasica

Trait Name *

Plant Height (cm)

▸ Additional filter criteria (Optional).

To further filter data for desired set, additional filter options are available to user.

Additional filter criteria (Optional).

We recommend you fill out as many of the following filters as possible to narrow the phenotype set to those you are most interested in.

Year

2017

Location

- Please select an option -

Germplasm Type

- Please select an option -

Germplasm

21 germplasm found based on the filters above.

Alberta Vanghelie (Alberta Vanghelie) | *Alice Leclercq (Alice Leclercq)* | *Alicia Ortiz (Alicia Ortiz)* |
 Anna Roche (Anna Roche) | *Aurora Murillo (Aurora Murillo)* |



Phenotype for Specific Germplasm

If you are interested in phenotypes for specific germplasm, you can add them individually by clicking add button or germplasm names. To retrieve all germplasm based on your other filter criteria, proceed to the next filter.

Type Germplasm Name/Stock Name



Maximum Allowed Missing Data

100%

Enter the percent (%) missing data per germplasm that you would like to allow. For example, a value of 20% will ensure that all germplasm exported have values for at least 20% of site-years this trait was observed in. If you further restrict the site-year exported using other filter criteria, this filter will be applied to the restricted dataset.

To organize result set, options are provided to ensure that exported data meet users requirements in terms of file format, header ordering and average values.

Choose your output file.

Select the file format, column headers and summary options you would like the data exported in below.

File Type

Select the format you would like the data exported.

Average Replicates per Site-Year

By default, all replicates in a single site-year will be averaged. Please specify if you would like all replicates to be included. To export the value of each replicate, uncheck this option.

 Make Column Headers R Friendly

Column Headers

#1	<input checked="" type="checkbox"/> (default) - Plant Height (cm)	↓
#2	<input checked="" type="checkbox"/> (default) - Germplasm Name	↑ ↓
#3	<input checked="" type="checkbox"/> (default) - Year	↑ ↓
#4	<input checked="" type="checkbox"/> (default) - Location	↑

[Preview headers](#)

Optional Headers

+ - Data Collector

+ - Country of Origin

+ - Experiment

Check the column headers you want to include in the output file. Use the up and down arrow buttons to change the order they appear in the file. Click Preview headers to preview selected column headers.

Download

File is generated instantly based on the filter and format options selected by user.

analyzed_phenotypic_data_download2018Nov02_15...

```

1 | Experiment | Trait Name | Germplasm Name | Year | Location
2 | "Experiment ABC" | "Plant Height (cm)" | "Aanya Jindal" | 2017 | C
3 | "Experiment ABC" | "Plant Height (cm)" | "Adela Szulc" | 2017 | Can
4 | "Experiment ABC" | "Plant Height (cm)" | "Agnes Nielsen" | 2017 | C
5 | "Experiment ABC" | "Plant Height (cm)" | "Alberta Vanghelie" | 2017 | C
6 | "Experiment ABC" | "Plant Height (cm)" | "Alexandra Marechal" | 2017 | C
7 | "Experiment ABC" | "Plant Height (cm)" | "Alice Leclercq" | 2017 | C
8 | "Experiment ABC" | "Plant Height (cm)" | "Alicia Ortiz" | 2017 | C
9 | "Experiment ABC" | "Plant Height (cm)" | "Alma Rogoz" | 2017 | Can
10 | "Experiment ABC" | "Plant Height (cm)" | "Alyssa Jones" | 2017 | C
11 | "Experiment ABC" | "Plant Height (cm)" | "Amelia Weastell" | 2017 | C
12 | "Experiment ABC" | "Plant Height (cm)" | "Ana Santiago" | 2017 | C
13 | "Experiment ABC" | "Plant Height (cm)" | "Ana Sturdza" | 2017 | Can
14 | "Experiment ABC" | "Plant Height (cm)" | "Anaida Urs" | 2017 | Can
15 | "Experiment ABC" | "Plant Height (cm)" | "Ann Walters" | 2017 | Can
    
```

Completed Successfully 100%

File:
[analyzed_phenotypic_data_download2018Nov02_1541193063.tsv](#)
 Format: Tab Separated Values (.tsv)

2.3 Visualize the data

The data you have loaded is summarized on the phenotypic summary page. The following example summarizes a single experiment where only one trait was measured for 132 germplasm.

Phenotypes

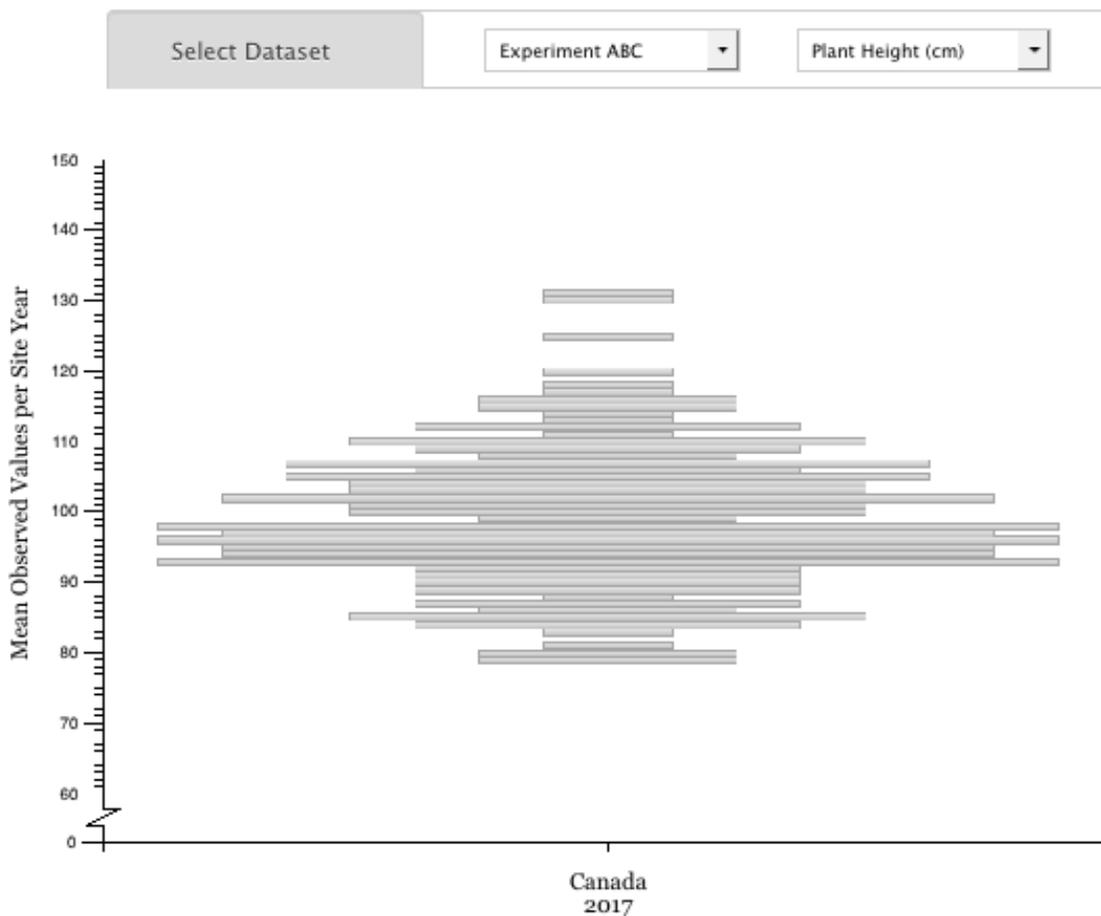
Genus	Traits	Experiments	Germplasm	Measurements
Tripalus	1	1	132	132

To visualize the distribution of values for a given trait, see the [Trait Distribution Chart](#).

To visualize the distribution of values for a single trait, see the Trait Distribution chart for that trait. This chart can be accessed from the summary page above and will summarize the data for a single trait within a single experiment.

Data is averaged across replicates but not across site-years. This allows you to compare the trait distribution between site-years for consistency and/or environmental effect.

Trait Distribution Chart



The Number of Germplasm with a given mean per Site Year

Figure: The Distribution of mean *Plant Height (cm)* per site year for *Experiment ABC*.

This chart shows the distribution of means observed (y-axis) per site year (x-axis) with the width of each bar representing the number of germplasm a given mean was observed/measured in. Only the data for the *Experiment ABC* phenotyping experiment is being shown.

[Go back to Summary Table](#)