Analyzed Phenotypes Documentation

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

Apr 20, 2023

Contents:

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-choosen indexes.

CHAPTER 1

Administrative Guide

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

This module requires the following system be setup prior to installation:

- Drupal 7
- Tripal 3.x
- PostgresSQL 9.3

Additionally, the following extension modules and libraries are pre-requisites: Unpack the following in your `sites/all/modules` directory:

- Tripal Download API
- Tripal D3.js

and unpack these libraries in the `sites//all/libraries` directory:

- PHP Excel Writer Libraries
- D3 JavaScript Library

1.1.1 Quickstart

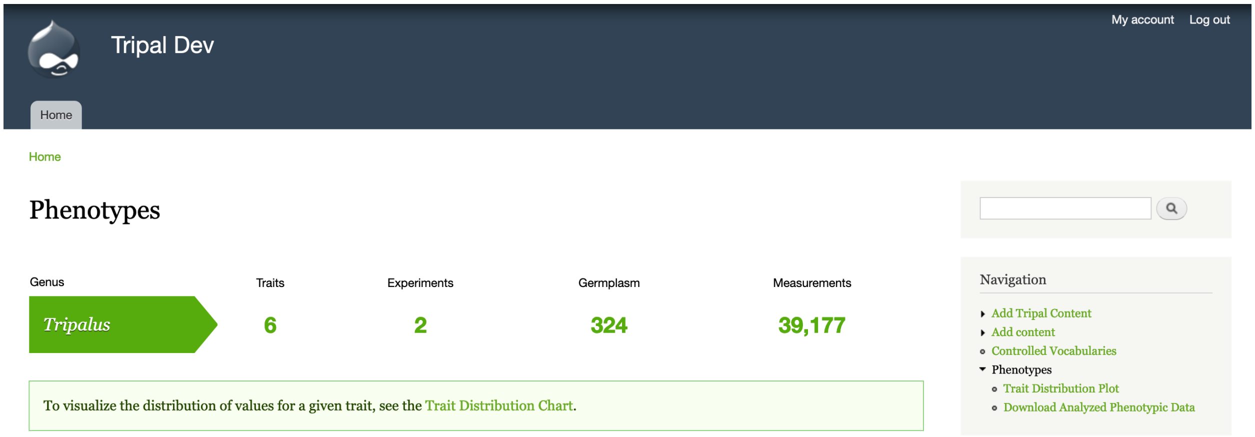
1. Install pre-requisites (see above).
2. Download this module into your `sites/all/modules` directory and enable it.
3. Set-up ontology terms at Admin > Tripal > Extensions > Analyzed Phenotypes > Setup Ontologies.
4. Upload data at Admin > Tripal > Data Loaders > Phenotypic Data Importer.

1.2 General Usage

> **Warning:** Before you can use this module, you need to set it up the module by going to Admin > Tripal > Extensions> Analyzed Phenotypes > Set-up Ontologies. For more detail on how to do this, see the "Setup" documentation in the Administrators Guide.

Phenotypic data should be uploaded by trusted users, preferably the biologists who generated the data. This is done by going to Admin > Tripal > Data Loaders > Phenotypic Data Importer. Trusted users should be given permission to this page via Admin > People > Permissions and checking `Upload Analyzed Phenotypic Data` for the desired role. I suggest creating a Specific Role for users who will upload data and assigning users to that role via Admin > People.

Once you have phenotypic data uploaded, you should see it summarized at Navigation > Phenotypes (`https://[yoursite]/phenotypes`). This summary lets users know how many traits and experiments are available for a given genus. It also summarizes the number of germplasm assayed and the number of measurements stored to give an indication of magnitude.

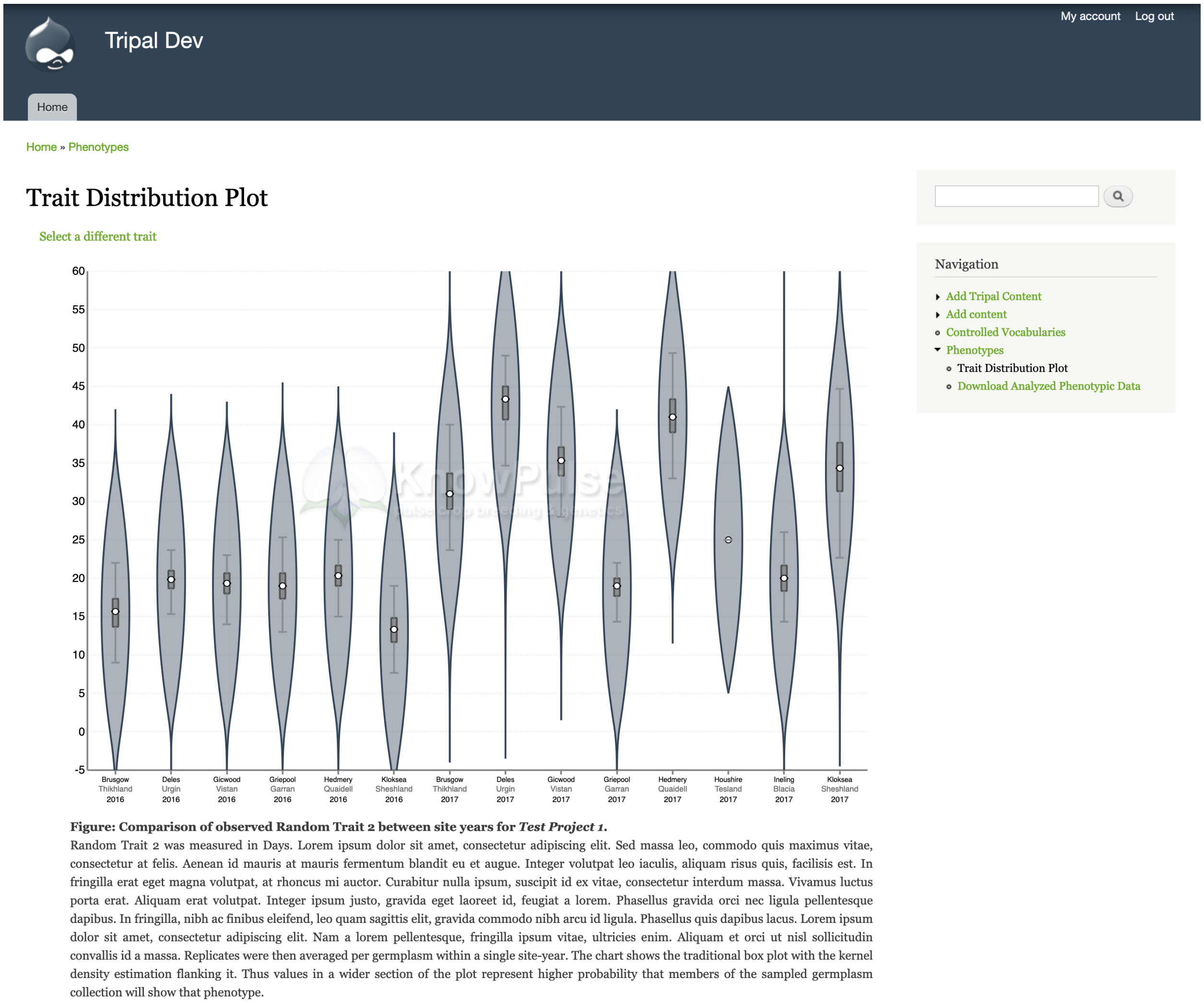


Phenotypes

Genus	Traits	Experiments	Germplasm	Measurements
Tripalus	6	2	324	39,177

To visualize the distribution of values for a given trait, see the Trait Distribution Chart.

Users can then see a summary of specific traits within a project through the **Trait Distribution Chart**. When you click on the link for the Trait Distribution chart at the bottom of the summary, you will be taken to a form where you choose the specific experiment, trait, method, unit combination you want to visualize. This level of granularity ensures that data is not combined when it is not statistically correct to do so. Also, quantitative data will be visualized with a violin plot and qualitative data will be visualized by a multi-series bar chart. Additionally, you can configure a watermark (shown in the screenshot below) for these charts to protect unpublished data.



Trait Distribution Plot

Figure: Comparison of observed Random Trait 2 between site years for *Test Project 1*.
Random Trait 2 was measured in Days. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed massa leo, commodo quis maximus vitae, consectetur at felis. Aenean id mauris at mauris fermentum blandit eu et augue. Integer volutpat leo iaculis, aliquam risus quis, facilisis est. In fringilla erat eget magna volutpat, at rhoncus mi auctor. Curabitur nulla ipsum, suscipit id ex vitae, consectetur interdum massa. Vivamus luctus porta erat. Aliquam erat volutpat. Integer ipsum justo, gravida eget laoreet id, feugiat a lorem. Phasellus gravida orci nec ligula pellentesque dapibus. In fringilla, nibh ac finibus eleifend, leo quam sagittis elit, gravida commodo nibh arcu id ligula. Phasellus quis dapibus lacus. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nam a lorem pellentesque, fringilla ipsum vitae, ultricies enim. Aliquam et orci ut nisl sollicitudin convallis id a massa. Replicates were then averaged per germplasm within a single site-year. The chart shows the traditional box plot with the kernel density estimation flanking it. Thus values in a wider section of the plot represent higher probability that members of the sampled germplasm collection will show that phenotype.

Users can also download the data for further analysis. Currently, this module offers download of the data with replicates averaged. This protects the original filtered data while still facilitating genome-wide association studies. We suggest pairing this module with the ND Genotypes module to allow your users to easily download genotypic and phenotypic data associated with the same germplasm.

1.3 Set-up Ontologies

1.3.1 Trait Ontologies

When collecting phenotypic data, it is very important to ensure you use the same methodology including units across replicates.

- the trait it was measuring (e.g. plant height)
- the method used to measure it (e.g. stretching the plant)
- the unit it was measured with (e.g. centimetres)

The trait, method and unit are all stored as controlled vocabulary terms with each in it's own ontology. The Trait Ontologies fieldset in the Set-up Ontologies form allows you to set these ontologies (see configuration instructions below).

There are two schools of thought when it comes to storing phenotypic data:

1. Use published ontologies directly
2. Create custom controlled vocabularies which link to published ontologies.

This module supports both methodologies although it caters more to the second one. This is due to my experience working directly with both breeder's and researchers who are very particular in how traits are named. By using the same names they use, we ensure we are accurately capturing the trait which was measured and by linking that trait to the correct published ontologies (preferably with input from the data collector) we facilitate sharing of data. In my opinion this is the "best of both worlds" so to speak.

Method #1

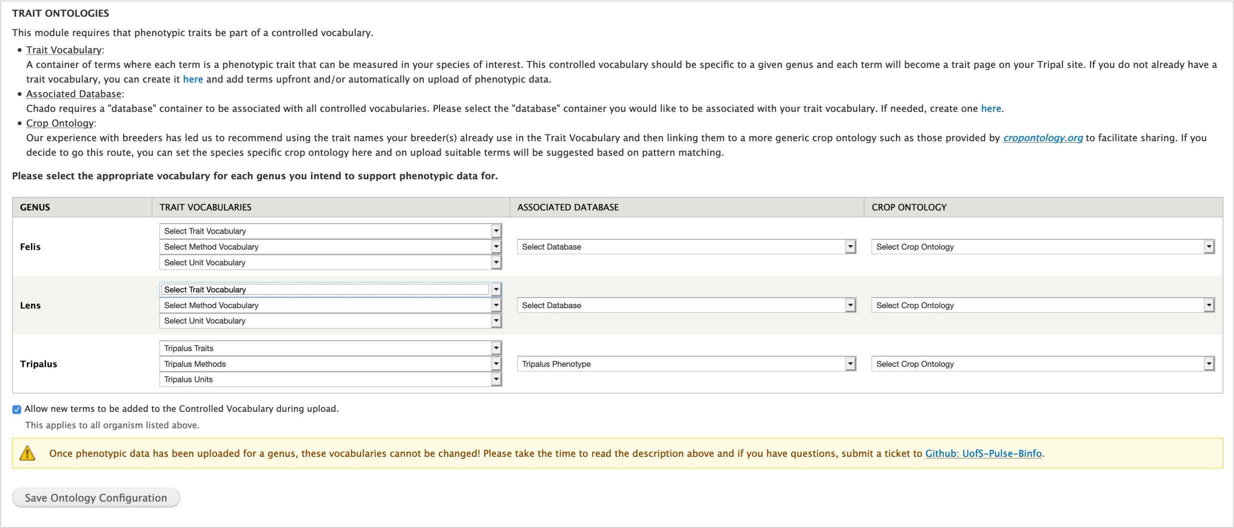
1. Use the ontology loader available with Tripal (Admin > Tripal > Data Loaders > Chado Vocabularies > OBO Vocabulary Loader) to load an associated organism-specific ontology (e.g. http://www.cropontology.org/).
2. (OPTIONAL) Load a more generic trait ontology (e.g. http://www.obofoundry.org/ontology/to.html) to associate your traits with. This allows you to utilize two public ontologies with the link between them being specified by the trusted researcher loading the phenotypic data.
3. Configure each genus with the controlled vocabularies you loaded in step #1 and the database you created in step #3. To work with this module the controlled vocabulary must have three "namespace": one for trait, method and unit. If you loaded an additional public ontology (e.g. plant trait ontology), select it under "Crop Ontology".
4. Uncheck the "Allow new terms to be added to the Controlled Vocabulary during upload" checkbox to ensure new terms are not added to the public ontology during upload. Click the "Save Ontology Configuration" button.

Method #2 (Recommended)

1. Create a new controlled vocabulary for the following by using the "Add Vocabulary" form available at Admin > Tripal > Data Loaders > Chado Vocabularies > Manage Chado CVs. You will need one of the following for each genus you would like to manage phenotypic data for.

- Trait (e.g. "Phenotypic Traits: Tripalus")
- Method (e.g. "Phenotypic Methods: Tripalus")
- Unit (e.g. "Phenotypic Units: Tripalus")

2. Create a new Database Reference container (i.e. chado.db) per genus by using the "Add Database" form available at Admin > Tripal > Data Loaders > Chado Databases. Chado requires database references for all controlled vocabulary terms; the one created here will be used for all three controlled vocabularies created in step #1. Leave the URL and URL prefix blank.
3. (OPTIONAL) Use the ontology loader available with Tripal (Admin > Tripal > Data Loaders > Chado Vocabularies > OBO Vocabulary Loader) to load an associated organism-specific ontology (e.g. http://www.cropontology.org/).
4. Configure each genus with the controlled vocabularies you created in step #1 and the database you created in step #3. If you loaded a genus-specific ontology, select it under "Crop Ontology". Click the "Save Ontology Configuration" button.



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 VOCABULARIES	ASSOCIATED DATABASE	CROP ONTOLOGY
Felis	Select Trait Vocabulary Select Method Vocabulary Select Unit Vocabulary	Select Database	Select Crop Ontology
Lens	Select Trait Vocabulary Select Method Vocabulary Select Unit Vocabulary	Select Database	Select Crop Ontology
Tripalus	Tripalus Traits Tripalus Methods Tripalus Units	Tripalus Phenotype	Select Crop Ontology

☑ Allow new terms to be added to the Controlled Vocabulary during upload.

This applies to all organism listed above.

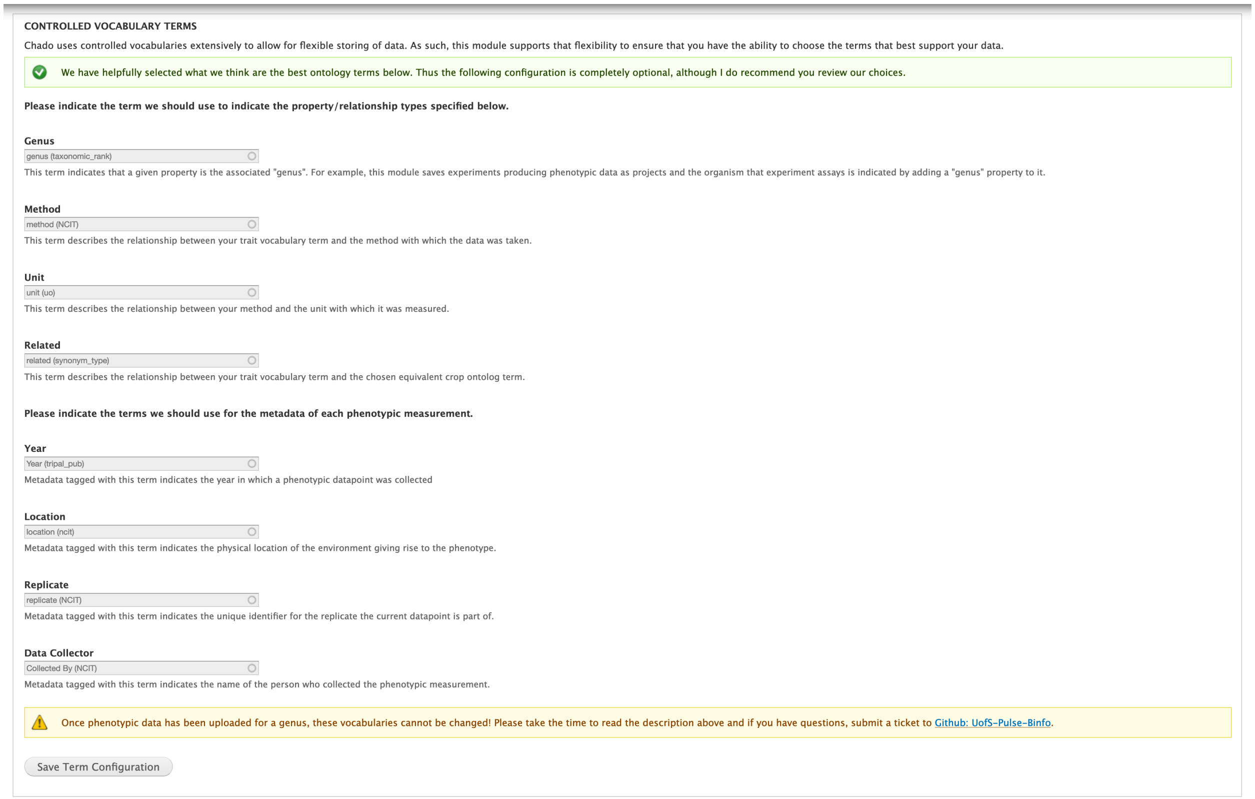
⚠ Once phenotypic data has been uploaded for a genus, these vocabularies cannot be changed! Please take the time to read the description above and if you have questions, submit a ticket to Github: UofS-Pulse-Binfo.

Note: You can also configure whether you would like new controlled vocabulary terms to be added during upload of phenotypic data using the checkbox directly above the "Save Ontology Configuration" button. If you want to control the names of traits, I suggest un-clicking this checkbox and loading your trait dictionary using Admin > Tripal > Data Loaders > Phenotypic Trait Importer. If you do not have an admin with enough time or expertise to do this then I suggest leaving it checked and allowing traits to be created as needed.

1.3.2 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 regardless of the types used for your data, this module will still be able to navigate the necessary relationships and interpret the data.

To provide ease of use, we have already chosen a set of controlled vocabulary terms and inserted them by default. This makes this portion of the set-up optional.



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.

Genus
genus (taxonomic_rank)
This term indicates that a given property is the associated "genus". For example, this module saves experiments producing phenotypic data as projects and the organism that experiment assays is indicated by adding a "genus" property to it.

Method
method (NCIT)
This term describes the relationship between your trait vocabulary term and the method with which the data was taken.

Unit
unit (uo)
This term describes the relationship between your method and the unit with which it was measured.

Related
related (synonym_type)
This term describes the relationship between your trait vocabulary term and the chosen equivalent crop ontolog term.

Please indicate the terms we should use for the metadata of each phenotypic measurement.

Year
Year (tripal_pub)
Metadata tagged with this term indicates the year in which a phenotypic datapoint was collected

Location
location (ncit)
Metadata tagged with this term indicates the physical location of the environment giving rise to the phenotype.

Replicate
replicate (NCIT)
Metadata tagged with this term indicates the unique identifier for the replicate the current datapoint is part of.

Data Collector
Collected By (NCIT)
Metadata tagged with this term indicates the name of the person who collected the phenotypic measurement.

⚠ 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.

If you would like to change the above controlled vocabulary terms simply type the term you would like to use instead in the autocomplete box and then select it from the list. Once you have done this for all terms you would like to change, click the "Save Term Configuration" button.

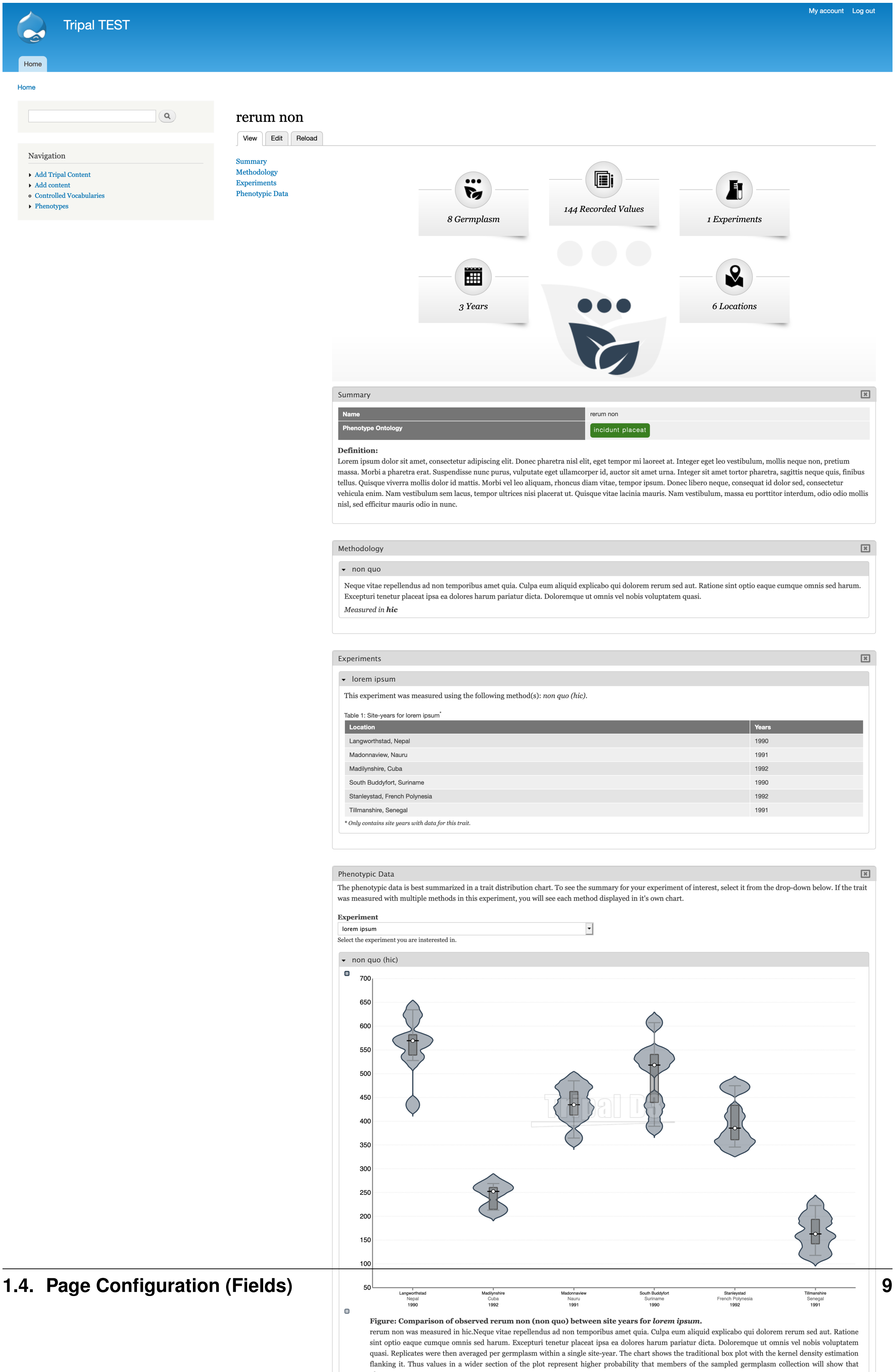
> **Warning:** Once you upload data, you can no longer change these terms.

1.4 Page Configuration (Fields)

This module provides a large number of fields for displaying phenotypic data on various Tripal Content Pages. This guide is going to show you how to configure all of them to produce the following pages. However, every field is optional so pick and choose to your hearts content.

1.4.1 Trait Pages

The following shows the full trait page configured on a default Tripal site with Antonelli theme and the following field configuration.



rerum non

Summary

Name	rerum non
Phenotype Ontology	incidunt placeat

Definition:
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec pharetra nisl elit, eget tempor mi laoreet at. Integer eget leo vestibulum, mollis neque non, pretium massa. Morbi a pharetra erat. Suspendisse nunc purus, vulputate eget ullamcorper id, auctor sit amet urna. Integer sit amet tortor pharetra, sagittis neque quis, finibus tellus. Quisque viverra mollis dolor id mattis. Morbi vel leo aliquam, rhoncus diam vitae, tempor ipsum. Donec libero neque, consequat id dolor sed, consectetur vehicula enim. Nam vestibulum sem lacus, tempor ultrices nisi placerat ut. Quisque vitae lacinia mauris. Nam vestibulum, massa eu porttitor interdum, odio odio mollis nisl, sed efficitur mauris odio in nunc.

Methodology

non quo

Neque vitae repellendus ad non temporibus amet quia. Culpa eum aliquid explicabo qui dolorem rerum sed aut. Ratione sint optio eaque cumque omnis sed harum. Excepturi tenetur placeat ipsa ea dolores harum pariatur dicta. Doloremque ut omnis vel nobis voluptatem quasi.

Measured in ***hic***

Experiments

lorem ipsum

This experiment was measured using the following method(s): *non quo (hic)*.

Table 1: Site-years for lorem ipsum*

Location	Years
Langworthstad, Nepal	1990
Madonnaview, Nauru	1991
Madilynshire, Cuba	1992
South Buddyfort, Suriname	1990
Stanleystad, French Polynesia	1992
Tillmanshire, Senegal	1991

* *Only contains site years with data for this trait.*

Phenotypic Data

The phenotypic data is best summarized in a trait distribution chart. To see the summary for your experiment of interest, select it from the drop-down below. If the trait was measured with multiple methods in this experiment, you will see each method displayed in it's own chart.

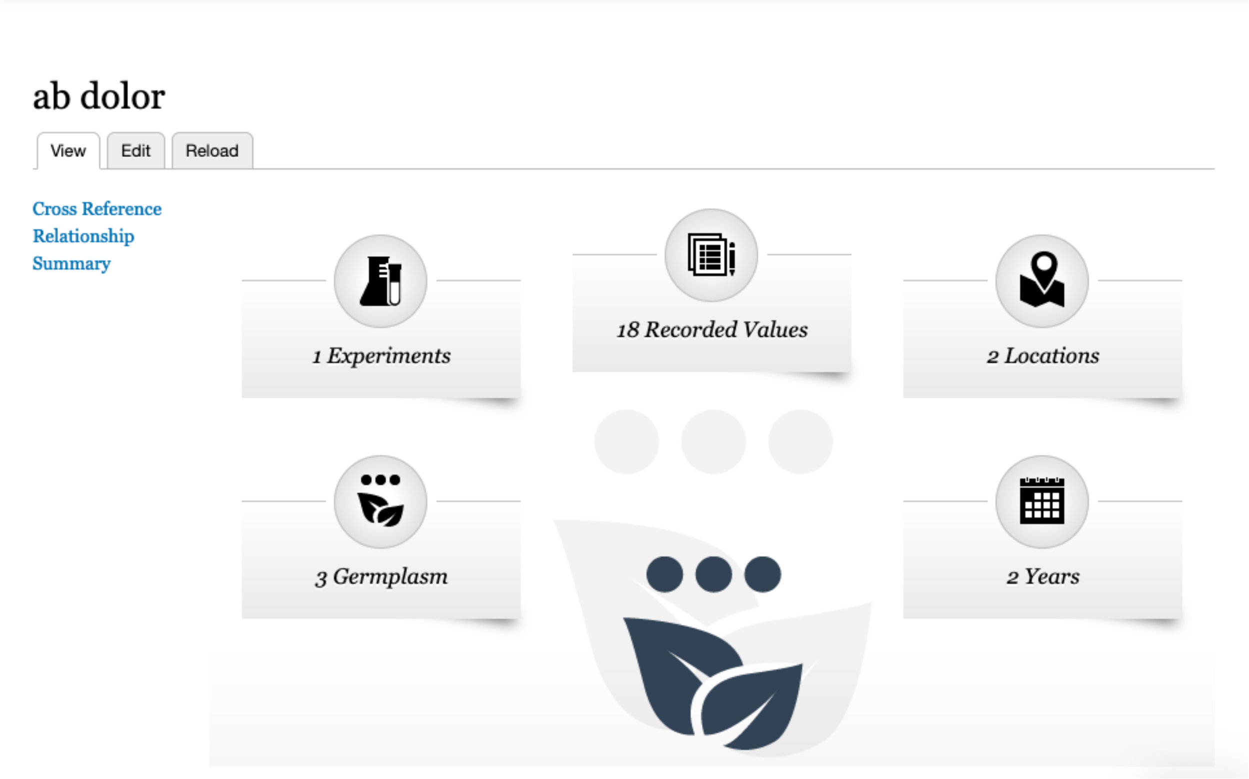
Experiment
lorem ipsum
Select the experiment you are interested in.

non quo (hic)

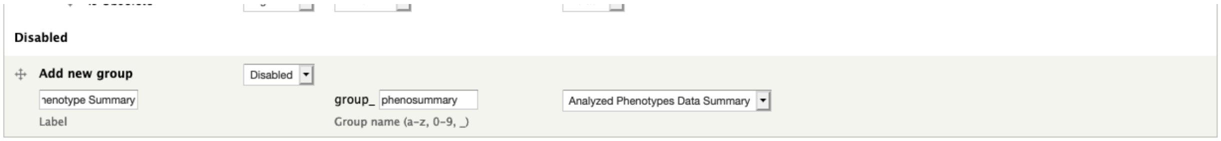
Figure: Comparison of observed rerum non (non quo) between site years for *lorem ipsum*.
rerum non was measured in hic.Neque vitae repellendus ad non temporibus amet quia. Culpa eum aliquid explicabo qui dolorem rerum sed aut. Ratione sint optio eaque cumque omnis sed harum. Excepturi tenetur placeat ipsa ea dolores harum pariatur dicta. Doloremque ut omnis vel nobis voluptatem quasi. Replicates were then averaged per germplasm within a single site-year. The chart shows the traditional box plot with the kernel density estimation flanking it. Thus values in a wider section of the plot represent higher probability that members of the sampled germplasm collection will show that

Note: First off, we recommend installing `tripal_ds` and applying the default layout. This will unfortunately pile all our new Analyzed Phenotypes fields into the summary table but it gives us a good place to start.

Phenotype Summary



1. Go to "Manage Fields" and check that a series of "Phenotype Summary: Number of *" fields are available. If not, click "Find new fields" and they should appear.
   - while not necessary, I suggest grouping these fields in order at the bottom of the page for better readability. These fields will not add widgets to your add/edit form.
2. Scroll to the bottom of the "Manage Display" page and add a new group named "Phenotype Summary" with a machine name of "phenosummary" and the type being "Analyzed Phenotypes Data Summary". Once filling it out click save to create it.



Disabled

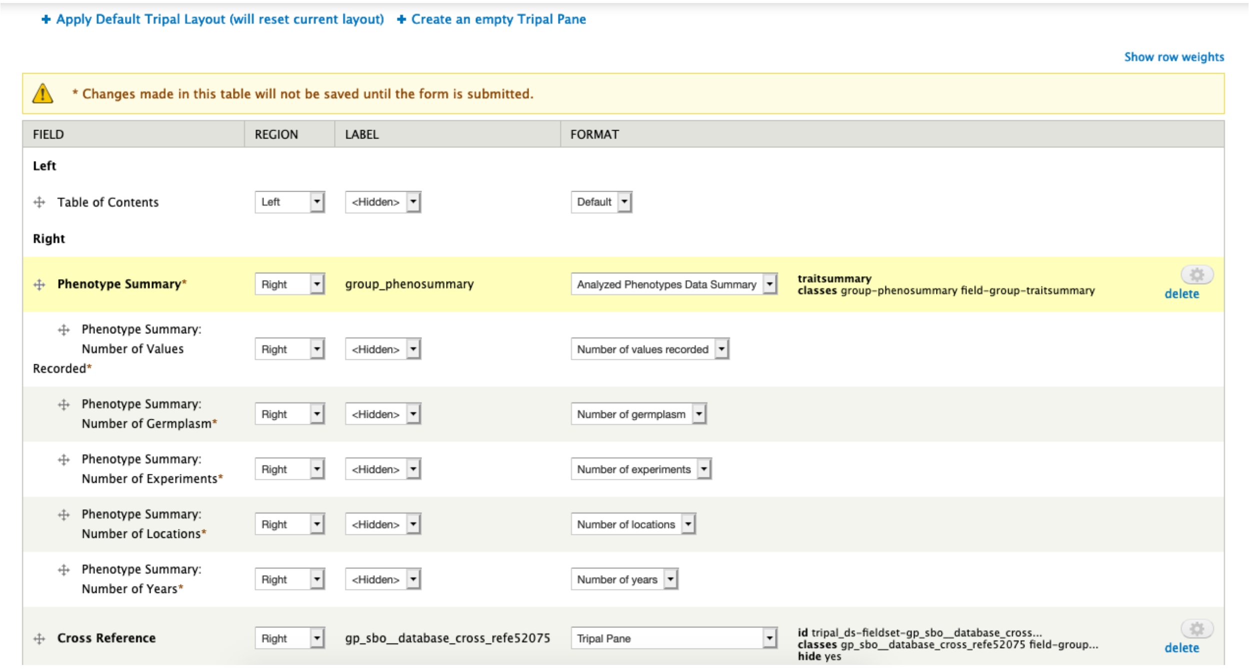
Add new group: Disabled

Label: henotype Summary

Group name (a-z, 0-9, _): group_ phenosummary

Analyzed Phenotypes Data Summary

3. Arrange the new group near the top of the page. It can be within or outside a Tripal Pane based on your preference. Then arrange each "Phenotype Summary: Number of *" nested underneath the group as shown in the following image. Click "Save" at the bottom of the page to save the order.



✚ Apply Default Tripal Layout (will reset current layout) ✚ Create an empty Tripal Pane

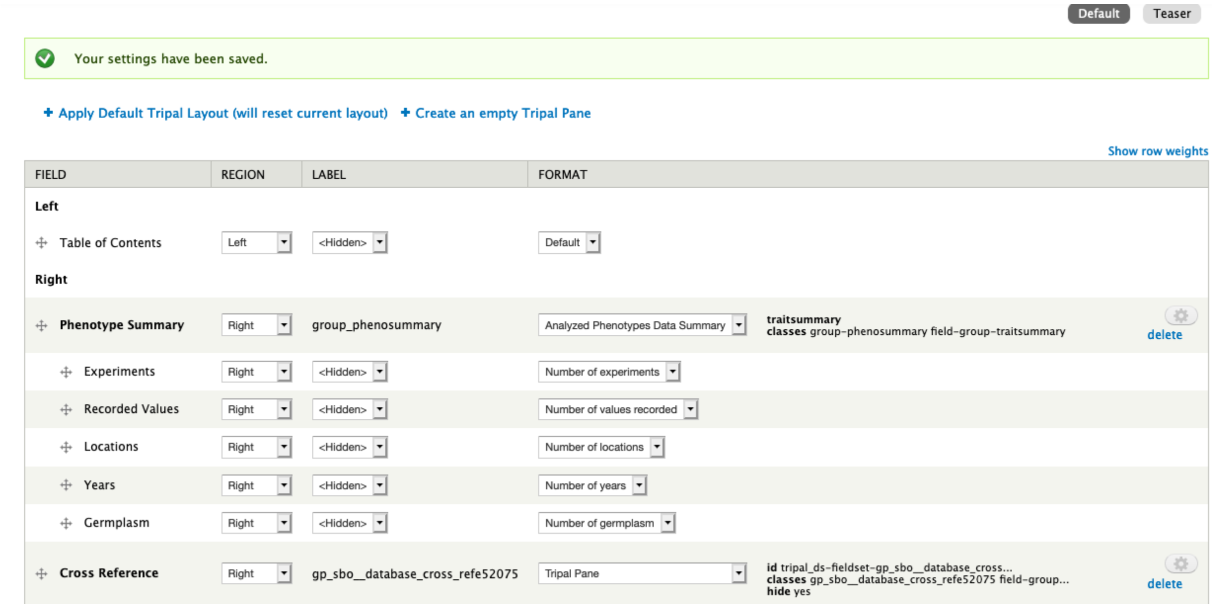
Show row weights

⚠ * Changes made in this table will not be saved until the form is submitted.

FIELD	REGION	LABEL	FORMAT		
Left					
Table of Contents	Left	<Hidden>	Default		
Right					
Phenotype Summary*	Right	group_phenosummary	Analyzed Phenotypes Data Summary	**traitsummary** **classes** group-phenosummary field-group-traitsummary	delete
Phenotype Summary: Number of Values Recorded*	Right	<Hidden>	Number of values recorded		
Phenotype Summary: Number of Germplasm*	Right	<Hidden>	Number of germplasm		
Phenotype Summary: Number of Experiments*	Right	<Hidden>	Number of experiments		
Phenotype Summary: Number of Locations*	Right	<Hidden>	Number of locations		
Phenotype Summary: Number of Years*	Right	<Hidden>	Number of years		
Cross Reference	Right	gp_sbo__database_cross_refe52075	Tripal Pane	**id** tripal_ds-fieldset-gp_sbo__database_cross... **classes** gp_sbo__database_cross_refe52075 field-group... **hide** yes	delete

Note: The order of the fields on this page controls the order of the numbers in the summary graphic.

4. Go back to the "Manage Fields" page and change the name of each "Phenotype Summary: Number of *" field to be the label you would like displayed on the page. For example, you may change "Phenotype Summary: Number of Germplasm" to "Germplasm".



Default Teaser

✓ Your settings have been saved.

✚ Apply Default Tripal Layout (will reset current layout) ✚ Create an empty Tripal Pane

Show row weights

FIELD	REGION	LABEL	FORMAT		
Left					
Table of Contents	Left	<Hidden>	Default		
Right					
Phenotype Summary	Right	group_phenosummary	Analyzed Phenotypes Data Summary	**traitsummary** **classes** group-phenosummary field-group-traitsummary	delete
Experiments	Right	<Hidden>	Number of experiments		
Recorded Values	Right	<Hidden>	Number of values recorded		
Locations	Right	<Hidden>	Number of locations		
Years	Right	<Hidden>	Number of years		
Germplasm	Right	<Hidden>	Number of germplasm		
Cross Reference	Right	gp_sbo__database_cross_refe52075	Tripal Pane	**id** tripal_ds-fieldset-gp_sbo__database_cross... **classes** gp_sbo__database_cross_refe52075 field-group... **hide** yes	delete

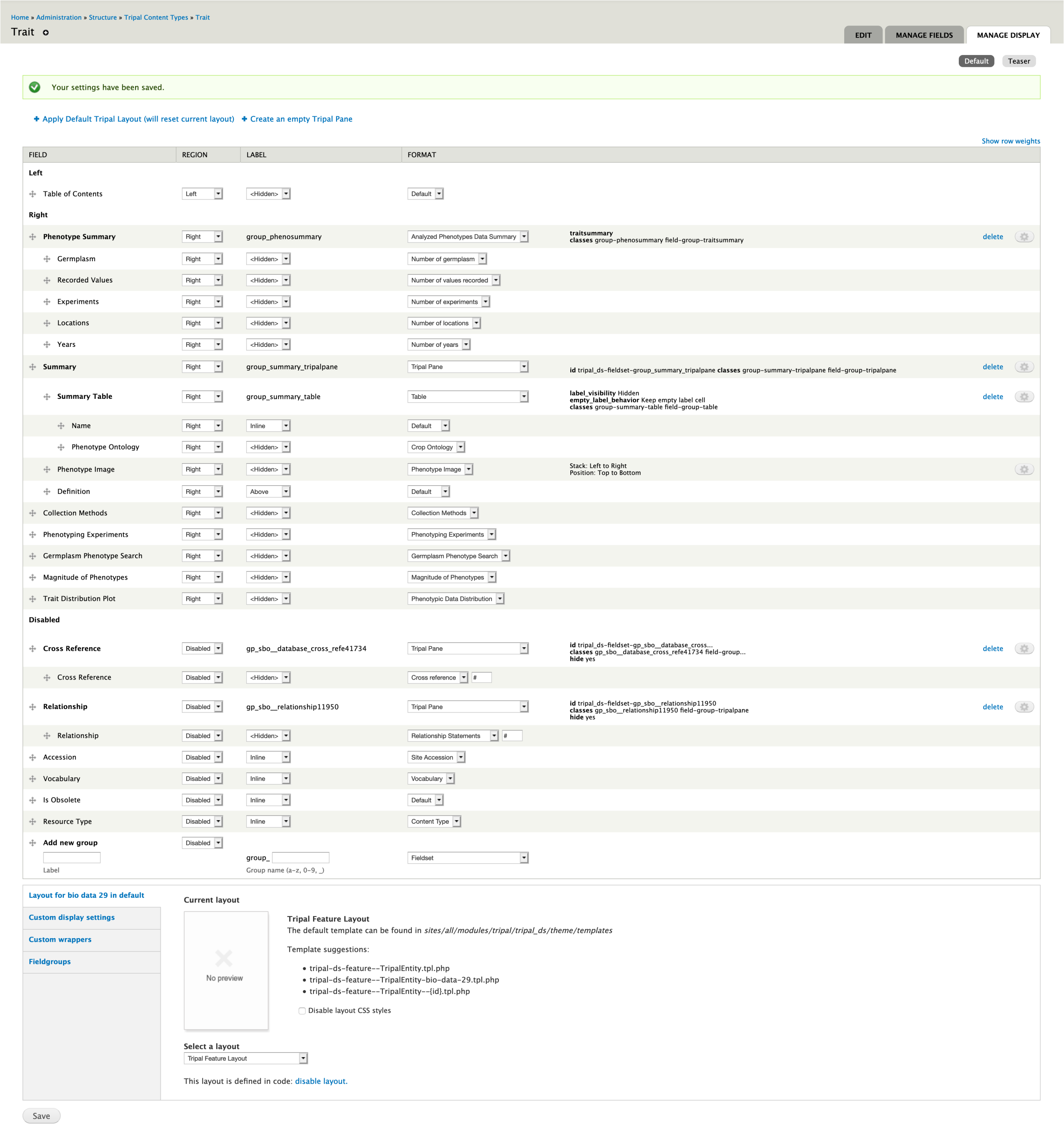
This will produce a summary at the top of your page!

Core Trait Details

This sections shows you which fields contain the core trait information such as definition, method and units.

> **Warning:** Remember to click save after each step!

1. Disable the following fields: Accession, Cross Reference, Relationship, Is Obsolete, Resource Type, and Vocabulary. The information provided by these fields is provided by the remaining fields in a more intuitive way.
2. Move the following fields out of the summary table (still in the summary): Phenotype Image and Definition.
3. Move the following fields out of the summary all together: Collection Methods, Phenotyping Experiments, Germplasm Phenotype Search, Magnitude of Phenotypes, Trait Distribution Plot.



Home » Administration » Structure » Tripal Content Types » Trait

Trait

Your settings have been saved.

✚ Apply Default Tripal Layout (will reset current layout) ✚ Create an empty Tripal Pane

Show row weights

FIELD	REGION	LABEL	FORMAT		
Left					
Table of Contents	Left	<Hidden>	Default		
Right					
Phenotype Summary	Right	group_phenosummary	Analyzed Phenotypes Data Summary	**traitsummary** **classes** group-phenosummary field-group-traitsummary	delete
Germplasm	Right	<Hidden>	Number of germplasm		
Recorded Values	Right	<Hidden>	Number of values recorded		
Experiments	Right	<Hidden>	Number of experiments		
Locations	Right	<Hidden>	Number of locations		
Years	Right	<Hidden>	Number of years		
Summary	Right	group_summary_tripalpane	Tripal Pane	**id** tripal_ds-fieldset-group_summary_tripalpane **classes** group-summary-tripalpane field-group-tripalpane	delete
Summary Table	Right	group_summary_table	Table	**label_visibility** Hidden **empty_label_behavior** Keep empty label cell **classes** group-summary-table field-group-table	delete
Name	Right	Inline	Default		
Phenotype Ontology	Right	<Hidden>	Crop Ontology		
Phenotype Image	Right	<Hidden>	Phenotype Image	Stack: Left to Right Position: Top to Bottom	
Definition	Right	Above	Default		
Collection Methods	Right	<Hidden>	Collection Methods		
Phenotyping Experiments	Right	<Hidden>	Phenotyping Experiments		
Germplasm Phenotype Search	Right	<Hidden>	Germplasm Phenotype Search		
Magnitude of Phenotypes	Right	<Hidden>	Magnitude of Phenotypes		
Trait Distribution Plot	Right	<Hidden>	Phenotypic Data Distribution		
Disabled					
Cross Reference	Disabled	gp_sbo__database_cross_refe41734	Tripal Pane	**id** tripal_ds-fieldset-gp_sbo__database_cross... **classes** gp_sbo__database_cross_refe41734 field-group... **hide** yes	delete
Cross Reference	Disabled	<Hidden>	Cross reference #		
Relationship	Disabled	gp_sbo__relationship11950	Tripal Pane	**id** tripal_ds-fieldset-gp_sbo__relationship11950 **classes** gp_sbo__relationship11950 field-group-tripalpane **hide** yes	delete
Relationship	Disabled	<Hidden>	Relationship Statements #		
Accession	Disabled	Inline	Site Accession		
Vocabulary	Disabled	Inline	Vocabulary		
Is Obsolete	Disabled	Inline	Default		
Resource Type	Disabled	Inline	Content Type		
Add new group Label	Disabled	group_ Group name (a-z, 0-9, _)	Fieldset		

Current layout

Tripal Feature Layout
The default template can be found in *sites/all/modules/tripal/tripal_ds/theme/templates*

Template suggestions:

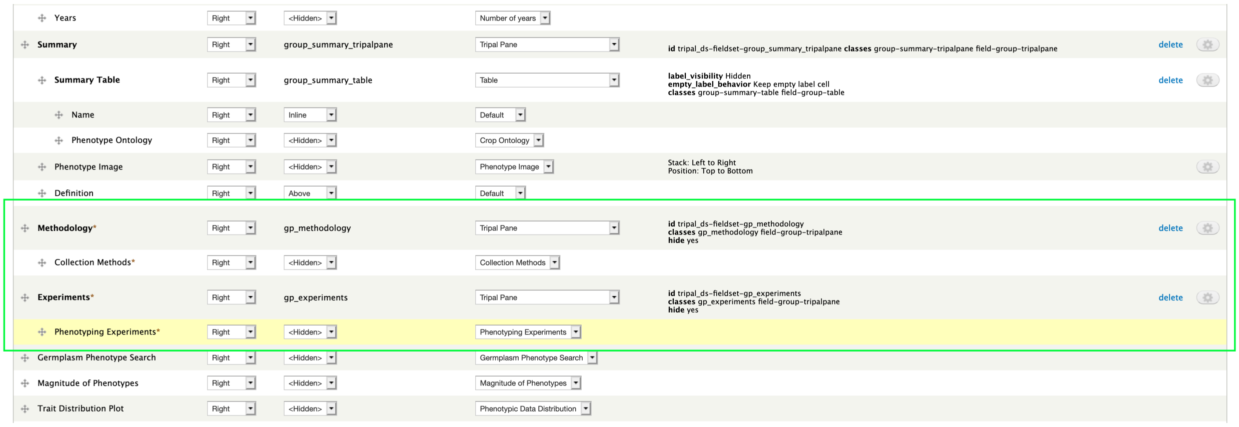
- tripal-ds-feature--TripalEntity.tpl.php
- tripal-ds-feature--TripalEntity-bio-data-29.tpl.php
- tripal-ds-feature--TripalEntity--{id}.tpl.php

☐ Disable layout CSS styles

Select a layout
Tripal Feature Layout

This layout is defined in code: disable layout.

4. Create a Tripal Pane for Collection Methods and another or Experiments by clicking "Create an empty Tripal Pane" at the top, entering the title and clicking save. Then nest the field within it.

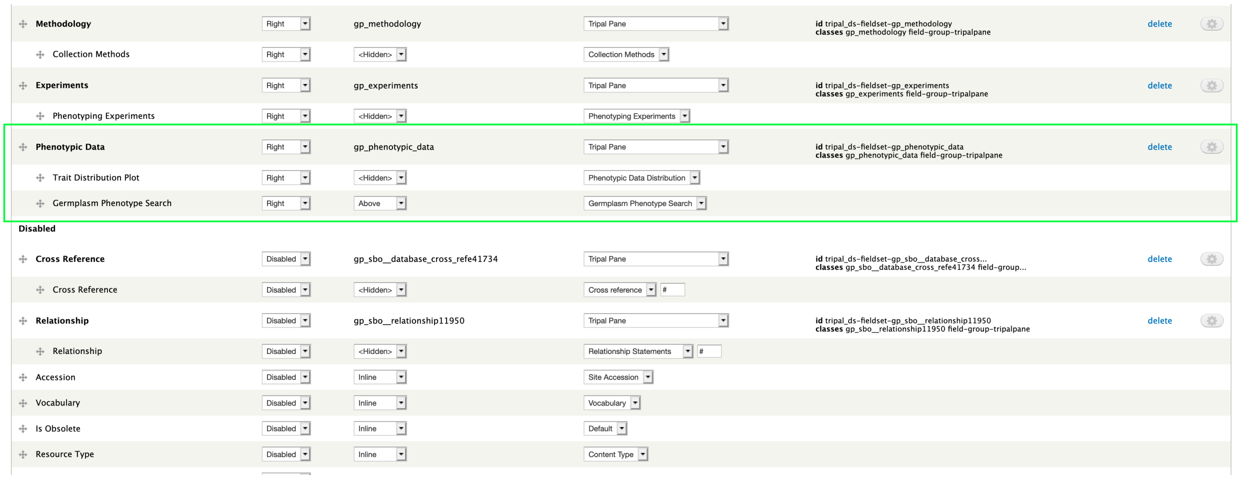


Years	Right	<Hidden>	Number of years			
Summary	Right	group_summary_tripalpane	Tripal Pane	**id** tripal_ds-fieldset-group_summary_tripalpane **classes** group-summary-tripalpane field-group-tripalpane	delete	⚙
Summary Table	Right	group_summary_table	Table	**label_visibility** Hidden **empty_label_behavior** Keep empty label cell **classes** group-summary-table field-group-table	delete	⚙
Name	Right	Inline	Default			
Phenotype Ontology	Right	<Hidden>	Crop Ontology			
Phenotype Image	Right	<Hidden>	Phenotype Image	Stack: Left to Right Position: Top to Bottom		⚙
Definition	Right	Above	Default			
Methodology*	Right	gp_methodology	Tripal Pane	**id** tripal_ds-fieldset-gp_methodology **classes** gp_methodology field-group-tripalpane **hide** yes	delete	⚙
Collection Methods*	Right	<Hidden>	Collection Methods			
Experiments*	Right	gp_experiments	Tripal Pane	**id** tripal_ds-fieldset-gp_experiments **classes** gp_experiments field-group-tripalpane **hide** yes	delete	⚙
Phenotyping Experiments*	Right	<Hidden>	Phenotyping Experiments			
Germplasm Phenotype Search	Right	<Hidden>	Germplasm Phenotype Search			
Magnitude of Phenotypes	Right	<Hidden>	Magnitude of Phenotypes			
Trait Distribution Plot	Right	<Hidden>	Phenotypic Data Distribution			

Note: If you do not want a Tripal Pane hidden on page load, click on the gear on the right hand side of the "Manage Display" for that Tripal Pane and uncheck the "Hide panel on page load" checkbox. Click both "update" and "save" and changing these settings.

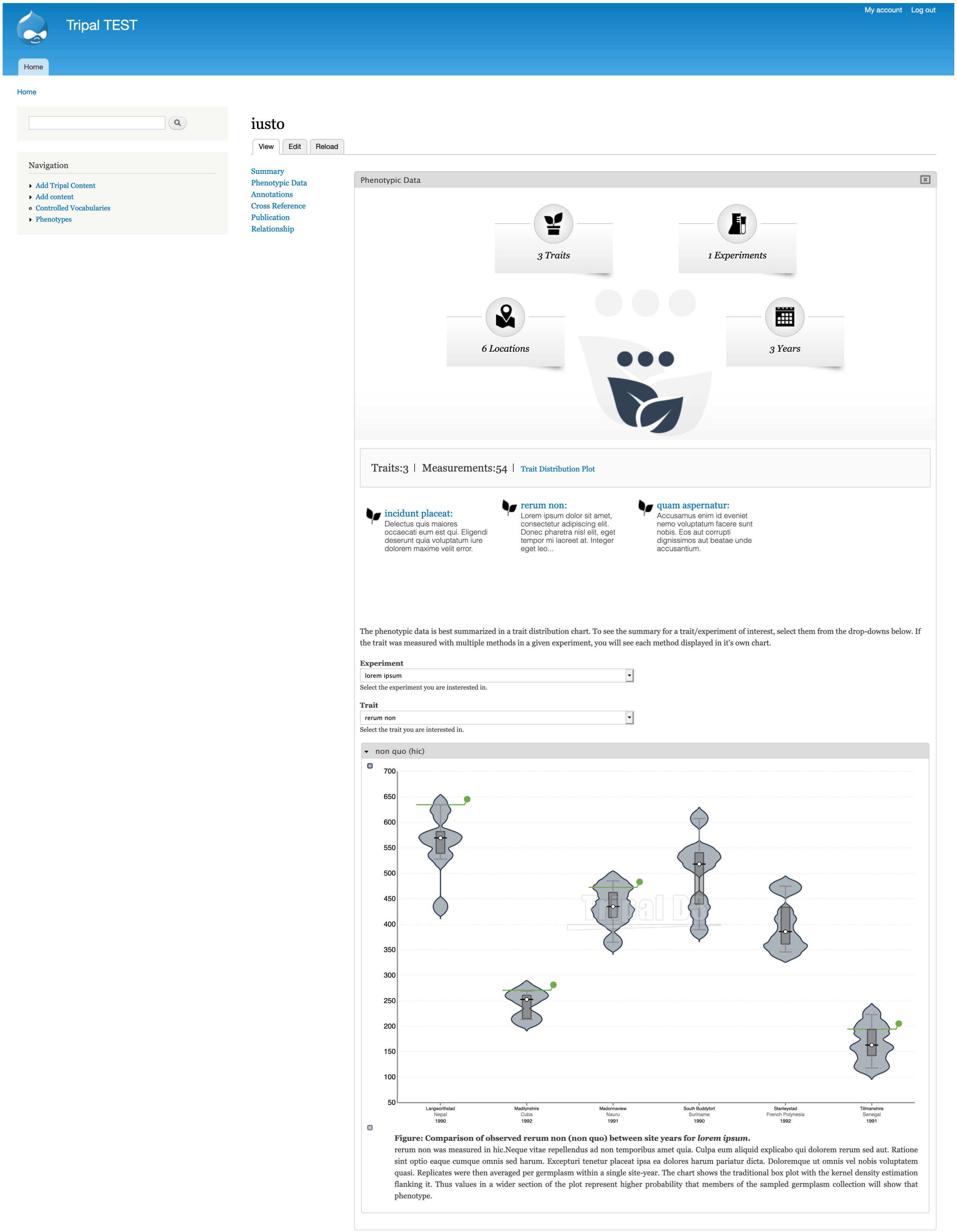
Trait Distribution Plot

The trait distribution plot tool is also provided for embedding on trait pages via the `Trait Distribution Plot` field. We recommend placing this field in a Tripal Pane with the "Germplasm Phenotype Search" field beneath it.



Methodology	Right	gp_methodology	Tripal Pane	**id** tripal_ds-fieldset-gp_methodology **classes** gp_methodology field-group-tripalpane	delete	⚙
Collection Methods	Right	<Hidden>	Collection Methods			
Experiments	Right	gp_experiments	Tripal Pane	**id** tripal_ds-fieldset-gp_experiments **classes** gp_experiments field-group-tripalpane	delete	⚙
Phenotyping Experiments	Right	<Hidden>	Phenotyping Experiments			
Phenotypic Data	Right	gp_phenotypic_data	Tripal Pane	**id** tripal_ds-fieldset-gp_phenotypic_data **classes** gp_phenotypic_data field-group-tripalpane	delete	⚙
Trait Distribution Plot	Right	<Hidden>	Phenotypic Data Distribution			
Germplasm Phenotype Search	Right	Above	Germplasm Phenotype Search			
Disabled						
Cross Reference	Disabled	gp_sbo__database_cross_refe41734	Tripal Pane	**id** tripal_ds-fieldset-gp_sbo__database_cross... **classes** gp_sbo__database_cross_refe41734 field-group...	delete	⚙
Cross Reference	Disabled	<Hidden>	Cross reference #			
Relationship	Disabled	gp_sbo__relationship11950	Tripal Pane	**id** tripal_ds-fieldset-gp_sbo__relationship11950 **classes** gp_sbo__relationship11950 field-group-tripalpane	delete	⚙
Relationship	Disabled	<Hidden>	Relationship Statements #			
Accession	Disabled	Inline	Site Accession			
Vocabulary	Disabled	Inline	Vocabulary			
Is Obsolete	Disabled	Inline	Default			
Resource Type	Disabled	Inline	Content Type			

1.4.2 Germplasm Page



iusto

Traits:3 | Measurements:54 | Trait Distribution Plot

incidunt placeat: Delectus quis maiores occaecati eum est qui. Eligendi deserunt quia voluptatum iure dolorem maxime velit error.

rerum non: Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec pharetra nisl elit, eget tempor mi laoreet at. Integer eget leo...

quam aspernatur: Accusamus enim id eveniet nemo voluptatum facere sunt nobis. Eos aut corrupti dignissimos aut beatae unde accusantium.

The phenotypic data is best summarized in a trait distribution chart. To see the summary for a trait/experiment of interest, select them from the drop-downs below. If the trait was measured with multiple methods in a given experiment, you will see each method displayed in it's own chart.

Experiment

lorem ipsum

Select the experiment you are insterested in.

Trait

rerum non

Select the trait you are interested in.

non quo (hic)

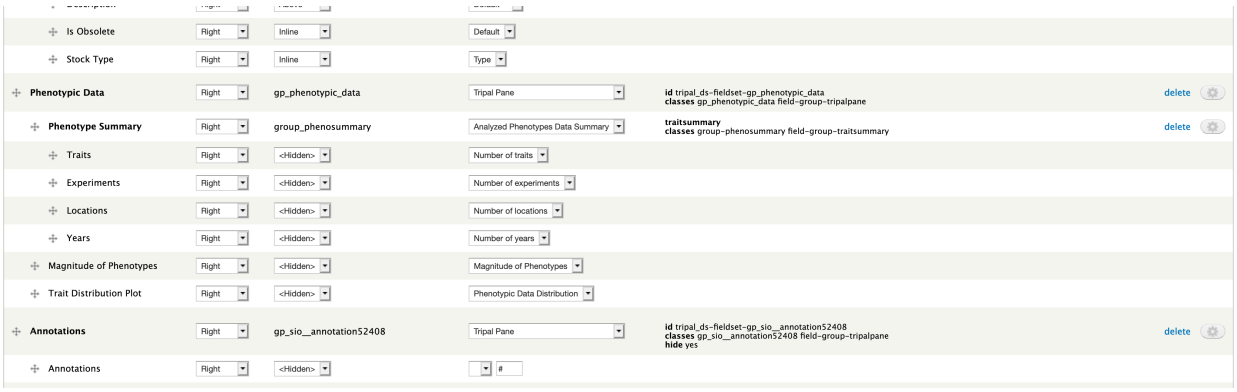
Figure: Comparison of observed rerum non (non quo) between site years for *lorem ipsum*.
rerum non was measured in hic.Neque vitae repellendus ad non temporibus amet quia. Culpa eum aliquid explicabo qui dolorem rerum sed aut. Ratione sint optio eaque cumque omnis sed harum. Excepturi tenetur placeat ipsa ea dolores harum pariatur dicta. Doloremque ut omnis vel nobis voluptatem quasi. Replicates were then averaged per germplasm within a single site-year. The chart shows the traditional box plot with the kernel density estimation flanking it. Thus values in a wider section of the plot represent higher probability that members of the sampled germplasm collection will show that phenotype.

This configuration applies to any stock-based content type. For example, "cultivar" or "Germplasm Accession" are content types created by default Tripal using the Chado stock table for storage.

1. Create a Tripal pane to contain all phenotypic data fields.
2. Arrange the Phenotype summary

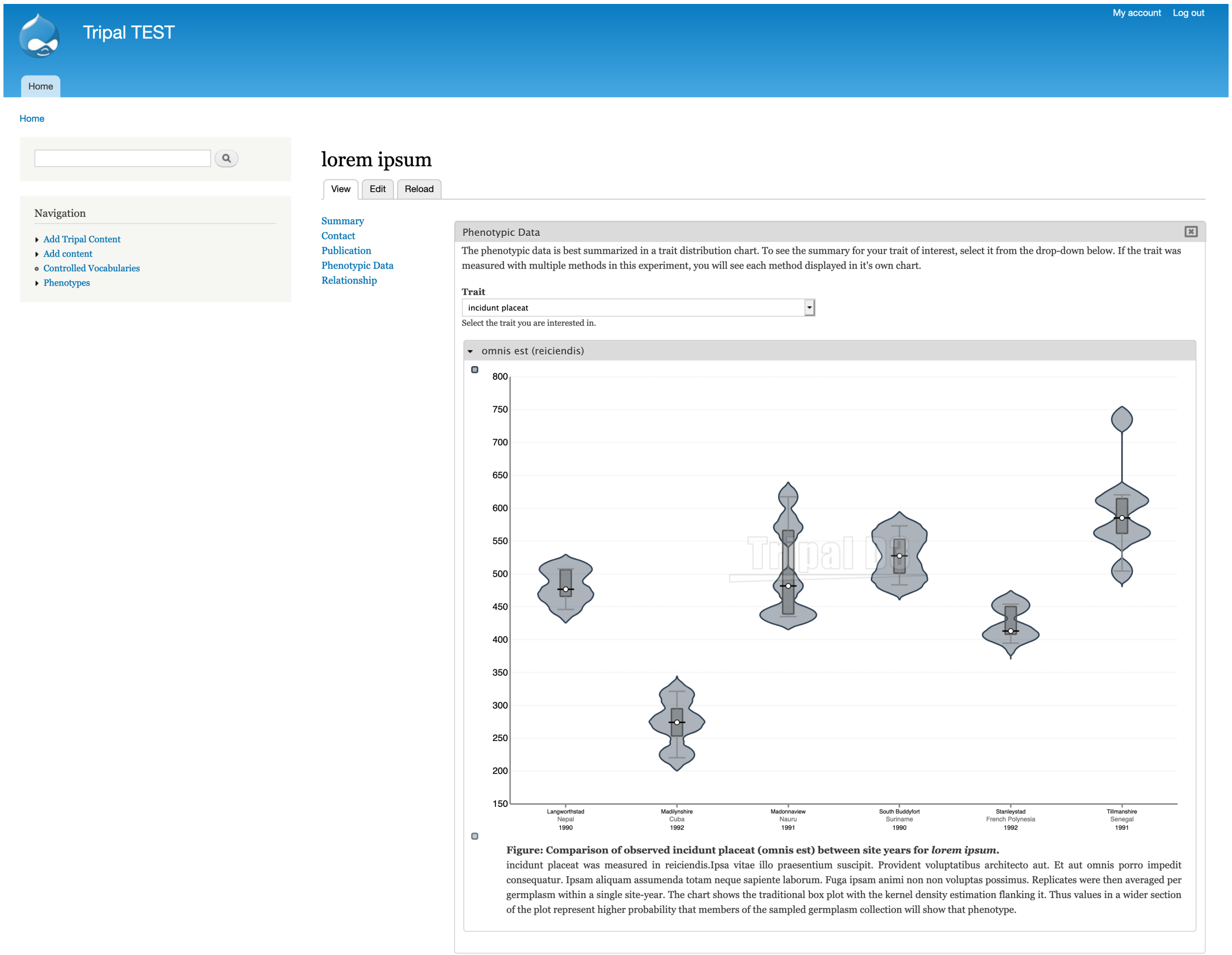
- Create a new group of type "Analyzed Phenotypes Data Summary" to contain the summary statistics just as we did on the trait page.
- Move all "Phenotype Summary" fields into this group.
- Rename each "Phenotype Summary" field in the "Manage Fields" page to match what you would like as a label.
- Move the group into the Phenotypic Data Tripal Pane

3. Move the "Magnitude of Phenotypes" field into the Phenotypic Data Tripal Pane, underneath the summary group but not nested within it. This provides a listing of traits that there is data for the current germplasm and gives an idea of the magnitude.
4. Move the "Trait Distribution plot" field beneath the "Magnitude of Phenotypes" field. Notice that the trait plot on germplasm pages has green lines showing where the germplasm falls within each site-year.



Is Obsolete	Right	Inline	Default		
Stock Type	Right	Inline	Type		
Phenotypic Data	Right	gp_phenotypic_data	Tripal Pane	**id** tripal_ds-fieldset-gp_phenotypic_data **classes** gp_phenotypic_data field-group-tripalpane	delete
Phenotype Summary	Right	group_phenosummary	Analyzed Phenotypes Data Summary	**traitsummary** **classes** group-phenosummary field-group-traitsummary	delete
Traits	Right	<Hidden>	Number of traits		
Experiments	Right	<Hidden>	Number of experiments		
Locations	Right	<Hidden>	Number of locations		
Years	Right	<Hidden>	Number of years		
Magnitude of Phenotypes	Right	<Hidden>	Magnitude of Phenotypes		
Trait Distribution Plot	Right	<Hidden>	Phenotypic Data Distribution		
Annotations	Right	gp_sio__annotation52408	Tripal Pane	**id** tripal_ds-fieldset-gp_sio__annotation52408 **classes** gp_sio__annotation52408 field-group-tripalpane **hide** yes	delete
Annotations	Right	<Hidden>	#		

1.4.3 Project Page



lorem ipsum

Phenotypic Data

The phenotypic data is best summarized in a trait distribution chart. To see the summary for your trait of interest, select it from the drop-down below. If the trait was measured with multiple methods in this experiment, you will see each method displayed in it's own chart.

Trait

incidunt placeat

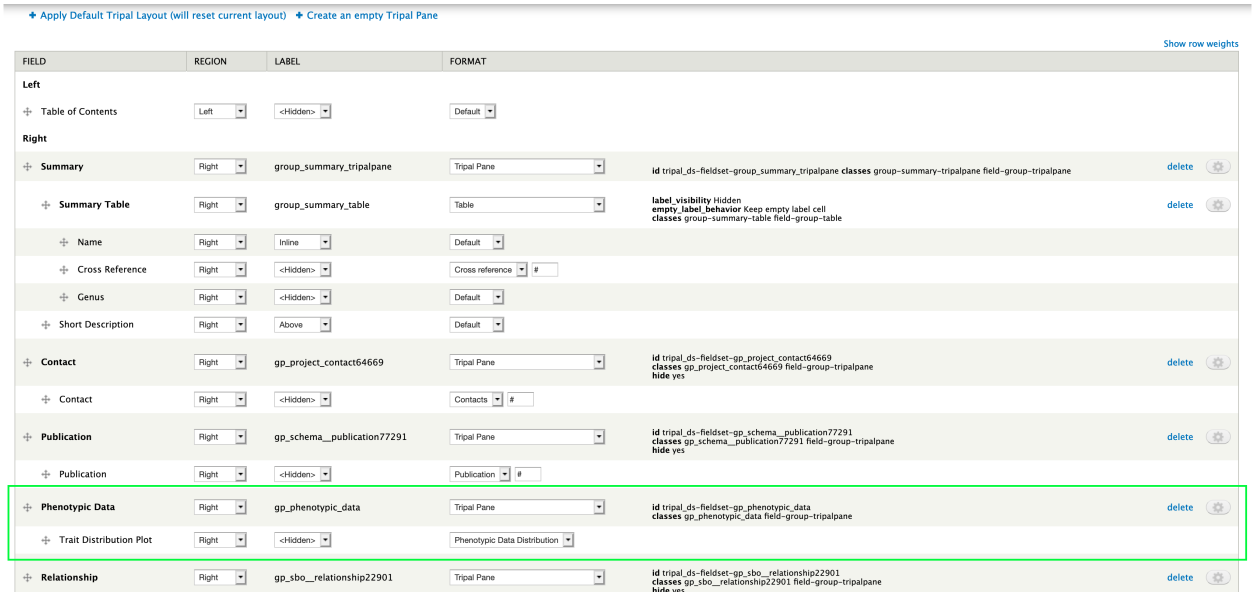
Select the trait you are interested in.

omnis est (reiciendis)

Figure: Comparison of observed incidunt placeat (omnis est) between site years for *lorem ipsum*. incidunt placeat was measured in reiciendis.Ipsa vitae illo praesentium suscipit. Provident voluptatibus architecto aut. Et aut omnis porro impedit consequatur. Ipsam aliquam assumenda totam neque sapiente laborum. Fuga ipsam animi non non voluptas possimus. Replicates were then averaged per germplasm within a single site-year. The chart shows the traditional box plot with the kernel density estimation flanking it. Thus values in a wider section of the plot represent higher probability that members of the sampled germplasm collection will show that phenotype.

This configuration applies to any project-based content type.

1. Create a Tripal pane to contain all phenotypic data fields.
2. Move the “Trait Distribution plot” field into the Phenotypic Data Tripal Pane.



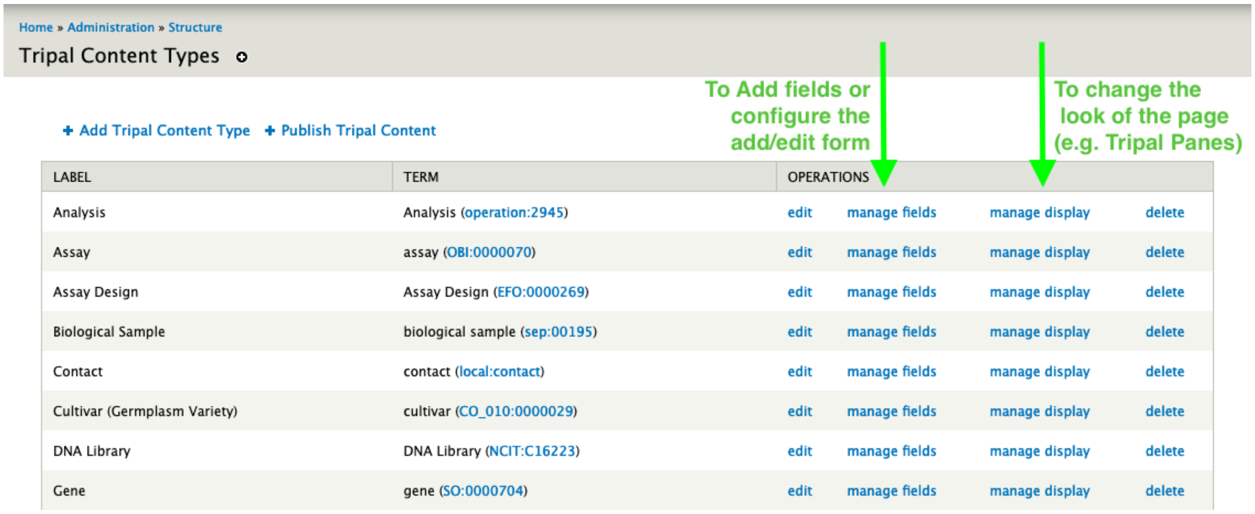
✚ Apply Default Tripal Layout (will reset current layout) ✚ Create an empty Tripal Pane

Show row weights

FIELD	REGION	LABEL	FORMAT		
Left					
Table of Contents	Left	<Hidden>	Default		
Right					
Summary	Right	group_summary_tripalpane	Tripal Pane	**id** tripal_ds-fieldset-group_summary_tripalpane **classes** group-summary-tripalpane field-group-tripalpane	delete
Summary Table	Right	group_summary_table	Table	**label_visibility** Hidden **empty_label_behavior** Keep empty label cell **classes** group-summary-table field-group-table	delete
Name	Right	Inline	Default		
Cross Reference	Right	<Hidden>	Cross reference #		
Genus	Right	<Hidden>	Default		
Short Description	Right	Above	Default		
Contact	Right	gp_project_contact64669	Tripal Pane	**id** tripal_ds-fieldset-gp_project_contact64669 **classes** gp_project_contact64669 field-group-tripalpane **hide** yes	delete
Contact	Right	<Hidden>	Contacts #		
Publication	Right	gp_schema__publication77291	Tripal Pane	**id** tripal_ds-fieldset-gp_schema__publication77291 **classes** gp_schema__publication77291 field-group-tripalpane **hide** yes	delete
Publication	Right	<Hidden>	Publication #		
Phenotypic Data	Right	gp_phenotypic_data	Tripal Pane	**id** tripal_ds-fieldset-gp_phenotypic_data **classes** gp_phenotypic_data field-group-tripalpane	delete
Trait Distribution Plot	Right	<Hidden>	Phenotypic Data Distribution		
Relationship	Right	gp_sbo__relationship22901	Tripal Pane	**id** tripal_ds-fieldset-gp_sbo__relationship22901 **classes** gp_sbo__relationship22901 field-group-tripalpane **hide** yes	delete

1.4.4 Where is the Configuration?

All of the following field configuration assumes you are on the "Manage Fields" or "Manage Display" pages for the tripal content type you are configuring. Specifically, this means that for Traits you would go to Administration » Structure » Tripal Content Types and then click on either the "Manage Fields" or "Manage Display" links on the same line as "Trait".



Home » Administration » Structure

Tripal Content Types

✚ Add Tripal Content Type ✚ Publish Tripal Content

To Add fields or configure the add/edit form

To change the look of the page (e.g. Tripal Panes)

LABEL	TERM	OPERATIONS			
Analysis	Analysis (operation:2945)	edit	manage fields	manage display	delete
Assay	assay (OBI:0000070)	edit	manage fields	manage display	delete
Assay Design	Assay Design (EFO:0000269)	edit	manage fields	manage display	delete
Biological Sample	biological sample (sep:00195)	edit	manage fields	manage display	delete
Contact	contact (local:contact)	edit	manage fields	manage display	delete
Cultivar (Germplasm Variety)	cultivar (CO_010:0000029)	edit	manage fields	manage display	delete
DNA Library	DNA Library (NCIT:C16223)	edit	manage fields	manage display	delete
Gene	gene (SO:0000704)	edit	manage fields	manage display	delete

1.5 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 :-).

1.5.1 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.

1.5.2 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.

1.5.3 Datasets

The queries were tested on two phenotypic datasets with different composition. Both datasets were generated using the Generate Tripal Data Drush module; 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 is 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

1.5.4 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,
  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;
```

1.5.5 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

1.6 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 then 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.

CHAPTER 2

User Guide

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 Phenotypic Data

2.1.1 Upload File Format

The upload data page imports analyzed phenotypic data provided in Tab-Delimited Values (TSV) format. 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.

This file is expected to contain the following columns:

- **Trait Name**: The full name of the trait as you would like it to appear on a trait page. This should not be abbreviated. (e.g. Days till one open flower)
- **Method Name**: A short (<4 words) name describing the method. This should uniquely identify the method while being very succinct. (e.g. 10% Plot at R1)
- **Unit**: The unit the trait was measured with. In the case of a scale this column should defined the scale. (e.g. days)
- **Germplasm Accession**: The stock.uniquename for the germplasm whose phenotype was measured. (e.g. ID:1234)
- **Germplasm Name**: The stock.name for the germplasm whose phenotype was measured. (e.g. Variety ABC)
- **Year**: The 4-digit year in which the measurement was taken. (e.g. 2020)
- **Location**: The full name of the location either using "location name, country" or GPS coordinates (e.g. Saskatoon, Canada)
- **Replicate**: The number for the replicate the current measurement is in. (e.g. 3)

- **Value**: The measured phenotypic value. (e.g. 34)
- **Data Collector**: The name of the person or organization which measured the phenotype.

The following is a short example:

```
Trait Name    Method Name     Unit    Germplasm Accession     Germplasm Name  Year    ␣
↪Location        Replicate       Value   Data Collector
Lorem ipsum   dolor sit amet  metris  ID:1    GERM1   2015    "Neither up, nor Down" ␣
↪1        5.3     Lacey Sanderson
Lorem ipsum   dolor sit amet  metris  ID:2    GERM2   2015    "Neither up, nor Down" ␣
↪1        2.2     Lacey Sanderson
Lorem ipsum   dolor sit amet  metris  ID:3    GERM3   2015    "Neither up, nor Down" ␣
↪1        4.9     Lacey Sanderson
Lorem ipsum   dolor sit amet  metris  ID:1    GERM1   2015    There   1       5.1     ␣
↪Lacey Sanderson
Lorem ipsum   dolor sit amet  metris  ID:2    GERM2   2015    There   1       3.6     ␣
↪Lacey Sanderson
Lorem ipsum   dolor sit amet  metris  ID:3    GERM3   2015    There   1       4       ␣
↪Lacey Sanderson
Lorem ipsum   dolor sit amet  metris  ID:1    GERM1   2015    Here    1       5.1     ␣
↪Lacey Sanderson
Lorem ipsum   dolor sit amet  metris  ID:2    GERM2   2015    Here    1       3.3     ␣
↪Lacey Sanderson
Lorem ipsum   dolor sit amet  metris  ID:3    GERM3   2015    Here    1       4.5     ␣
↪Lacey Sanderson
```

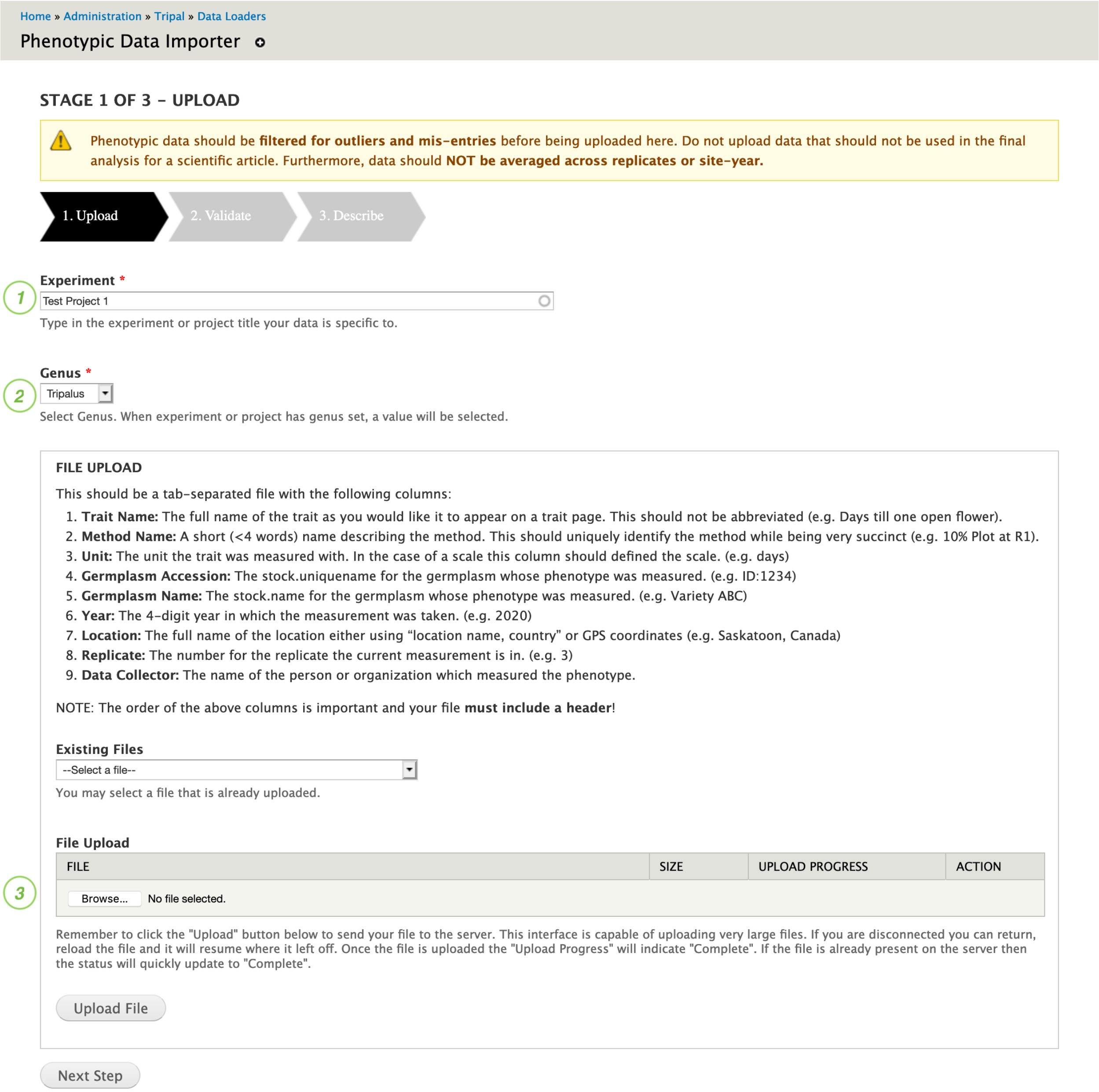
Note: You can see a full example of this file distributed with the module: `tests/example_files/AnalyzedPhenotypes-TestData-1trait3loc2yr3rep.txt`. The screenshots in this tutorial were taken importing this particular example file.

2.1.2 Stage 1: Upload File

The importer is available through the Administrative Toolbar > Tripal > Data Loaders > Phenotypic Data Importer.

1. Enter the name of the experiment your phenotypic data was taken for. The experiment should already exist. If your experiment does not exist, first go to Admin Toolbar > Content > Tripal Content > Add Tripal Content to add a "Research Experiment" describing your experiment.
2. Choose the genus of the organism of the germplasm samples your data was taken on. This system supports data from multiple species but requires each file only contain a single genus. If the genus is not available, contact your administrator to configure ontologies for your genus.
3. Next, under file upload, use the "Browse" button to select the file containing your phenotypic data. Navigate to the file and click open, then click "Upload File".

Warning: The trait, method and unit names in your file will be used in the describe phase but you will be unable to change you. Please ensure the trait name (e.g. Days till Plants have one open flower) is generic to allow data to be combined with the method name (e.g. Days till 10% of plants/plot have one open flower) very specific.



Phenotypic Data Importer

STAGE 1 OF 3 – UPLOAD

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.**

Experiment *

Test Project 1

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

Genus *

Tripalus

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

FILE UPLOAD

This should be a tab-separated file with the following columns:

1. **Trait Name:** The full name of the trait as you would like it to appear on a trait page. This should not be abbreviated (e.g. Days till one open flower).
2. **Method Name:** A short (<4 words) name describing the method. This should uniquely identify the method while being very succinct (e.g. 10% Plot at R1).
3. **Unit:** The unit the trait was measured with. In the case of a scale this column should defined the scale. (e.g. days)
4. **Germplasm Accession:** The stock.uniquename for the germplasm whose phenotype was measured. (e.g. ID:1234)
5. **Germplasm Name:** The stock.name for the germplasm whose phenotype was measured. (e.g. Variety ABC)
6. **Year:** The 4-digit year in which the measurement was taken. (e.g. 2020)
7. **Location:** The full name of the location either using "location name, country" or GPS coordinates (e.g. Saskatoon, Canada)
8. **Replicate:** The number for the replicate the current measurement is in. (e.g. 3)
9. **Data Collector:** The name of the person or organization which measured the phenotype.

NOTE: The order of the above columns is important and your file **must include a header**!

Existing Files

--Select a file--

You may select a file that is already uploaded.

File Upload

FILE	SIZE	UPLOAD PROGRESS	ACTION
Browse... No file selected.			

Remember to click the "Upload" button below to send your file to the server. This interface is capable of uploading very large files. If you are disconnected you can return, reload the file and it will resume where it left off. Once the file is uploaded the "Upload Progress" will indicate "Complete". If the file is already present on the server then the status will quickly update to "Complete".

Basic compliance tests at the file level are performed to ensure that requirements outlined are met. For instance, the file must be a valid tab-delimited file following the format outlined above, the genus must be configured for phenotypic data and the experiment must be selected. You will only be notified about errors so you can assume the file validated properly if you see the screen shown in Stage 2.

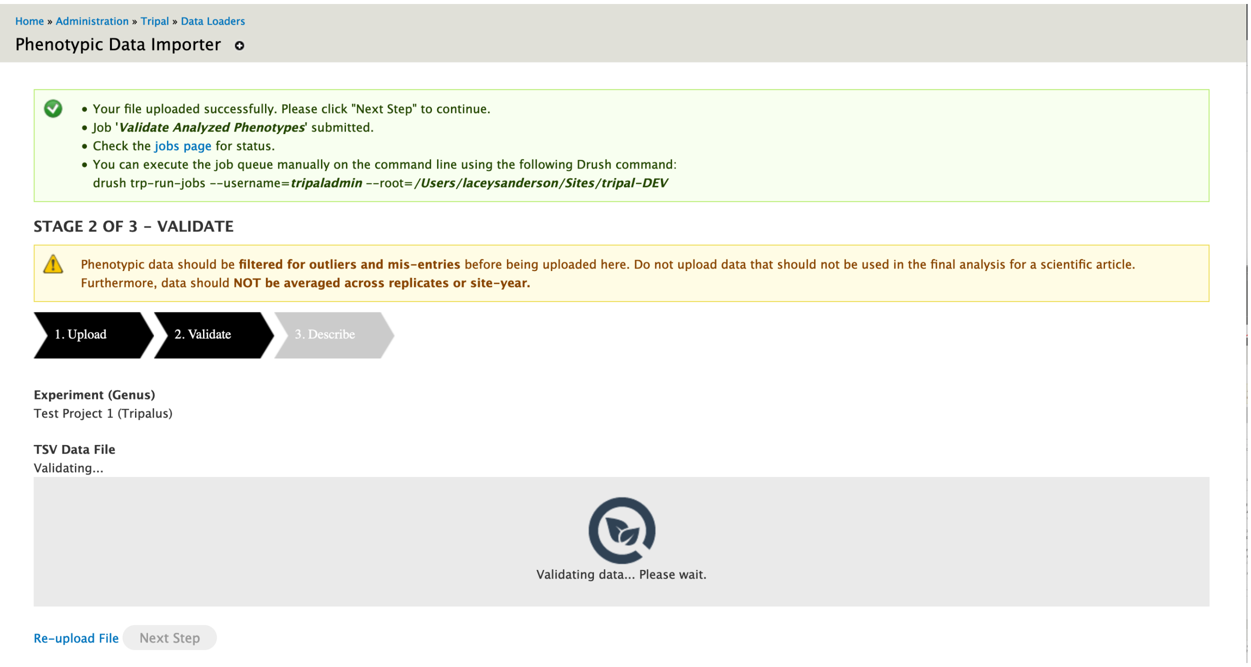
2.1.3 Stage 2: Validate File

In this stage, the file undergoes a data level validation. The entire file is tested against a set of validation rules to ensure that the

- File complies with the format specified above
- If "Allow new traits…" has been unchecked in configuration, ensure all traits already exist.

- All columns have a value
- All metadata columns have the correct data type (e.g. integer for replicates)
- Each value matches the expected data type, based on the unit
- All germplasm must already exist in the site and matches to the file must be exact
- Trait, method, unit, germplasm, location, year, replicate combination must be unique in the file

While validation is ongoing, the user is presented with spinner.



Home » Administration » Tripal » Data Loaders

Phenotypic Data Importer

- Your file uploaded successfully. Please click "Next Step" to continue.
- Job '***Validate Analyzed Phenotypes***' submitted.
- Check the jobs page for status.
- You can execute the job queue manually on the command line using the following Drush command:
  drush trp-run-jobs --username=***tripaladmin*** --root=***/Users/laceysanderson/Sites/tripal-DEV***

STAGE 2 OF 3 - VALIDATE

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.**

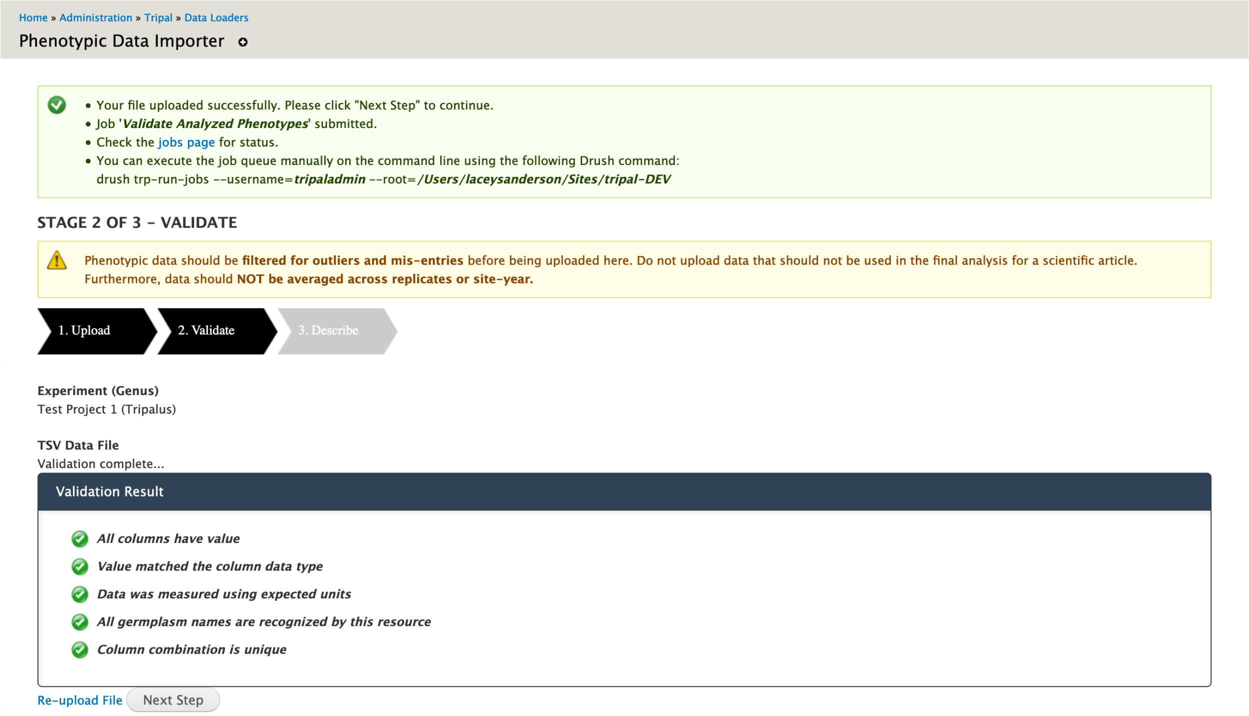
Experiment (Genus)
Test Project 1 (Tripalus)

TSV Data File
Validating...

Validating data... Please wait.

Warning: Administrators are urged to setup the Tripal Daemon since validation occurs during a Tripal Job.

Once validation is complete, you will be shown a list of green checkmarks detailing which criteria passed. If any criteria were not met, a red stop symbol will be shown with helpful information on how to fix the problem.



Home » Administration » Tripal » Data Loaders

Phenotypic Data Importer

- Your file uploaded successfully. Please click "Next Step" to continue.
- Job '*Validate Analyzed Phenotypes*' submitted.
- Check the jobs page for status.
- You can execute the job queue manually on the command line using the following Drush command:
  drush trp-run-jobs --username=*tripaladmin* --root=*/Users/laceysanderson/Sites/tripal-DEV*

STAGE 2 OF 3 - VALIDATE

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.**

Experiment (Genus)
Test Project 1 (Tripalus)

TSV Data File
Validation complete...

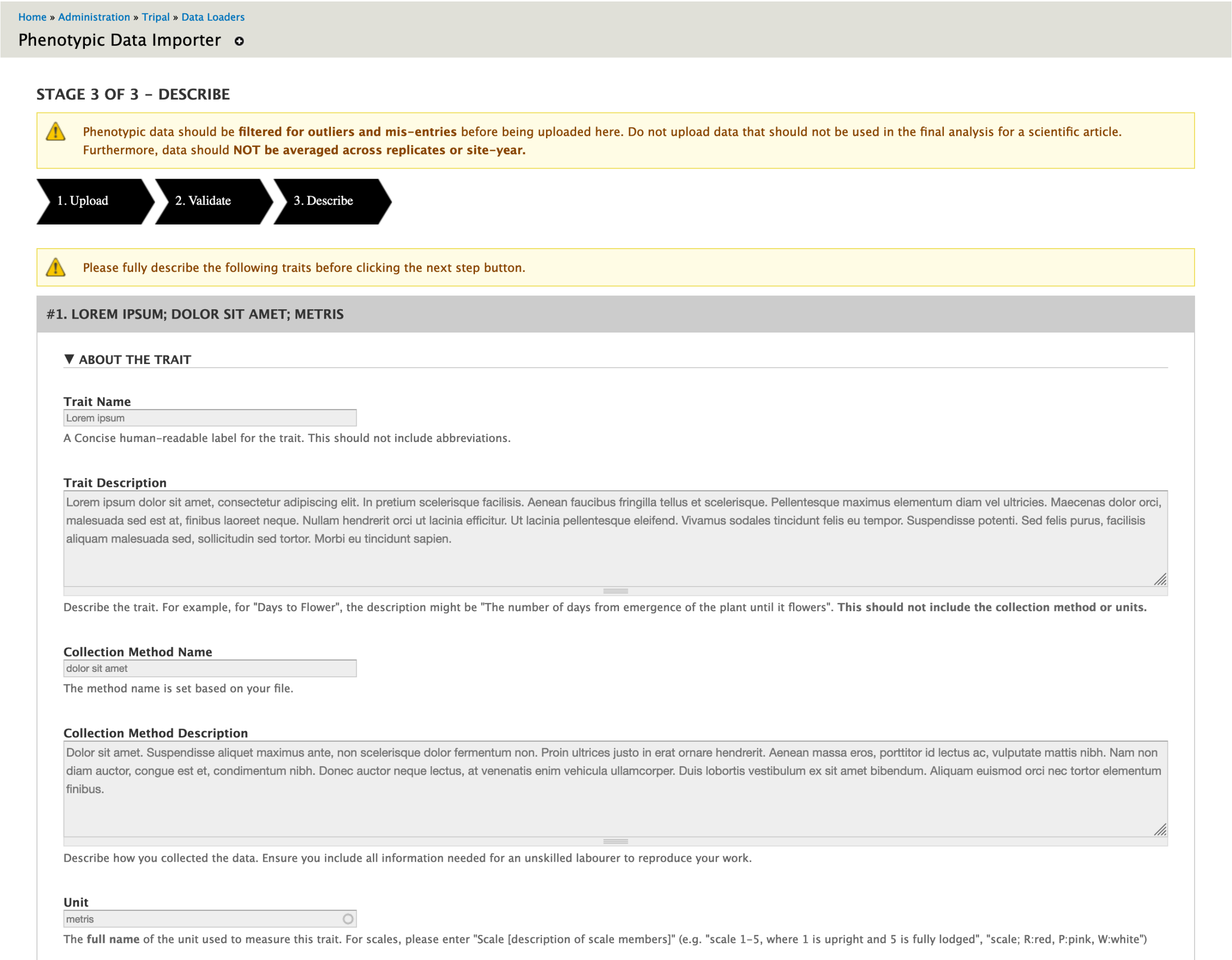
Validation Result

- ✓ *All columns have value*
- ✓ *Value matched the column data type*
- ✓ *Data was measured using expected units*
- ✓ *All germplasm names are recognized by this resource*
- ✓ *Column combination is unique*

2.1.4 Stage 3: Describe Traits

Each unique trait in the file is described in this step.

If the trait already exists, you will be shown the trait name, method and unit details existing in the system. This allows you to confirm you have chosen the right values for your file. If the trait does not already exist and the system has been configured to allow entry of traits on upload, you will be asked to describe your trait, data collection method and units.



Phenotypic Data Importer

STAGE 3 OF 3 – DESCRIBE

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.**

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

#1. LOREM IPSUM; DOLOR SIT AMET; METRIS

▼ ABOUT THE TRAIT

Trait Name
Lorem ipsum
A Concise human-readable label for the trait. This should not include abbreviations.

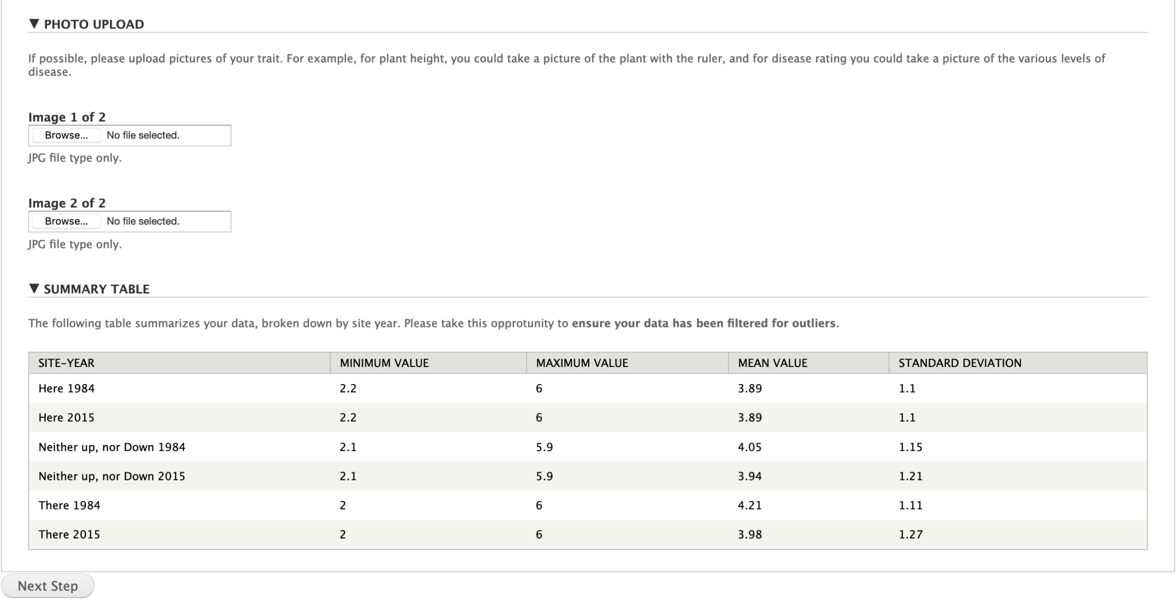
Trait Description
Lorem ipsum dolor sit amet, consectetur adipiscing elit. In pretium scelerisque facilisis. Aenean faucibus fringilla tellus et scelerisque. Pellentesque maximus elementum diam vel ultricies. Maecenas dolor orci, malesuada sed est at, finibus laoreet neque. Nullam hendrerit orci ut lacinia efficitur. Ut lacinia pellentesque eleifend. Vivamus sodales tincidunt felis eu tempor. Suspendisse potenti. Sed felis purus, facilisis aliquam malesuada sed, sollicitudin sed tortor. Morbi eu tincidunt sapien.
Describe the trait. For example, for "Days to Flower", the description might be "The number of days from emergence of the plant until it flowers". **This should not include the collection method or units.**

Collection Method Name
dolor sit amet
The method name is set based on your file.

Collection Method Description
Dolor sit amet. Suspendisse aliquet maximus ante, non scelerisque dolor fermentum non. Proin ultrices justo in erat ornare hendrerit. Aenean massa eros, porttitor id lectus ac, vulputate mattis nibh. Nam non diam auctor, congue est et, condimentum nibh. Donec auctor neque lectus, at venenatis enim vehicula ullamcorper. Duis lobortis vestibulum ex sit amet bibendum. Aliquam euismod orci nec tortor elementum finibus.
Describe how you collected the data. Ensure you include all information needed for an unskilled labourer to reproduce your work.

Unit
metris
The **full name** of the unit used to measure this trait. For scales, please enter "Scale [description of scale members]" (e.g. "scale 1-5, where 1 is upright and 5 is fully lodged", "scale; R:red, P:pink, W:white")

You also have the ability to upload images describing a trait. For example, an image showing the scale is particularly helpful!



▼ PHOTO UPLOAD

If possible, please upload pictures of your trait. For example, for plant height, you could take a picture of the plant with the ruler, and for disease rating you could take a picture of the various levels of disease.

Image 1 of 2
Browse... No file selected.
JPG file type only.

Image 2 of 2
Browse... No file selected.
JPG file type only.

▼ SUMMARY TABLE

The following table summarizes your data, broken down by site year. Please take this opprotunity to **ensure your data has been filtered for outliers.**

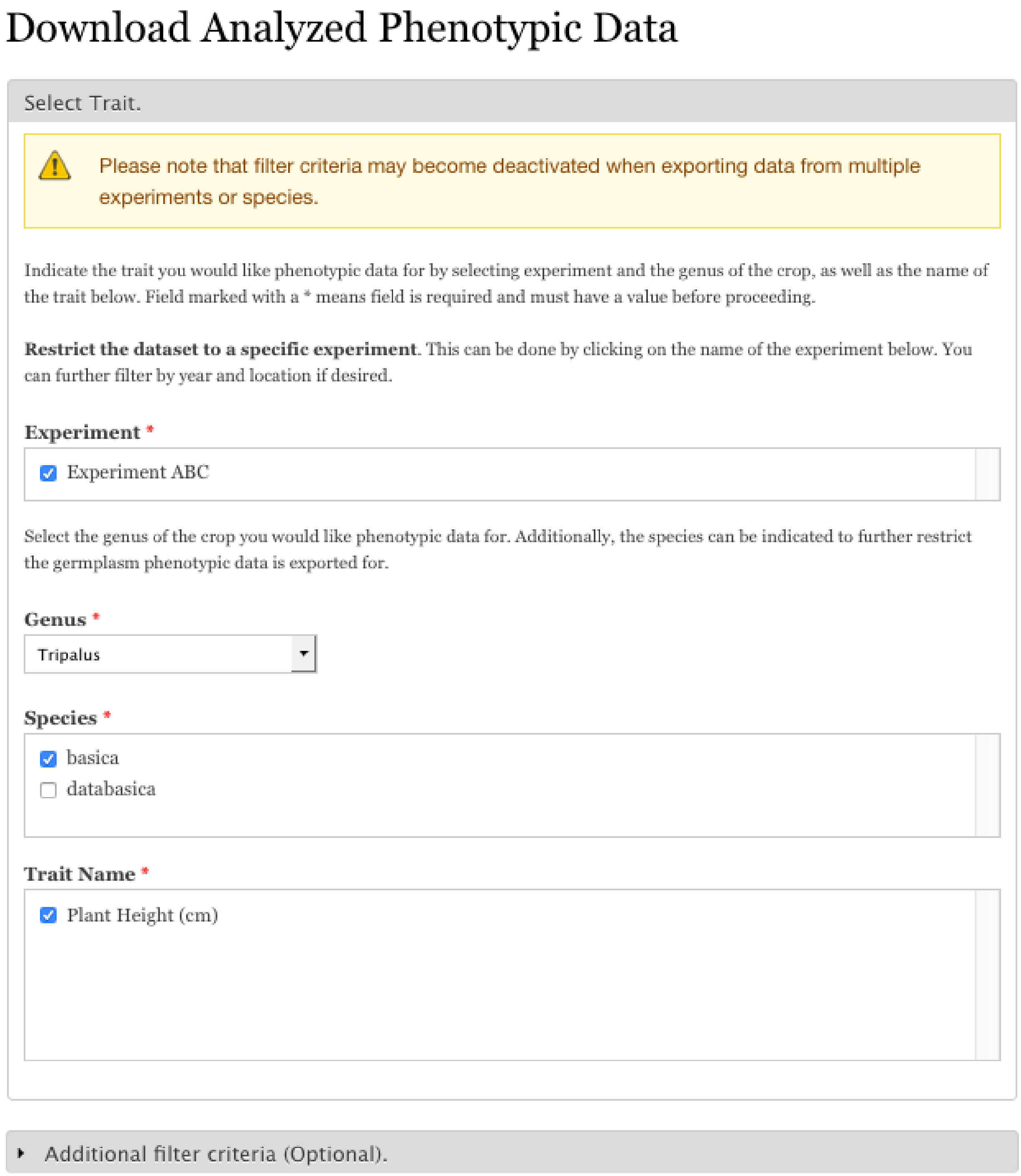
SITE-YEAR	MINIMUM VALUE	MAXIMUM VALUE	MEAN VALUE	STANDARD DEVIATION
Here 1984	2.2	6	3.89	1.1
Here 2015	2.2	6	3.89	1.1
Neither up, nor Down 1984	2.1	5.9	4.05	1.15
Neither up, nor Down 2015	2.1	5.9	3.94	1.21
There 1984	2	6	4.21	1.11
There 2015	2	6	3.98	1.27

Finally, the data in the file for a given trait is summarized. This can help you pick out problems such as outliers with the data before you upload it. Furthermore, it can be used to confirm the correct file was uploaded.

Warning: Once you click “Import File”, the form submits a Tripal job to complete the importing process. The form will reset and the Tripal Job will be completed in the background.

2.2 Downloading 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 xlxs, 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 *

Tripalus

Species *

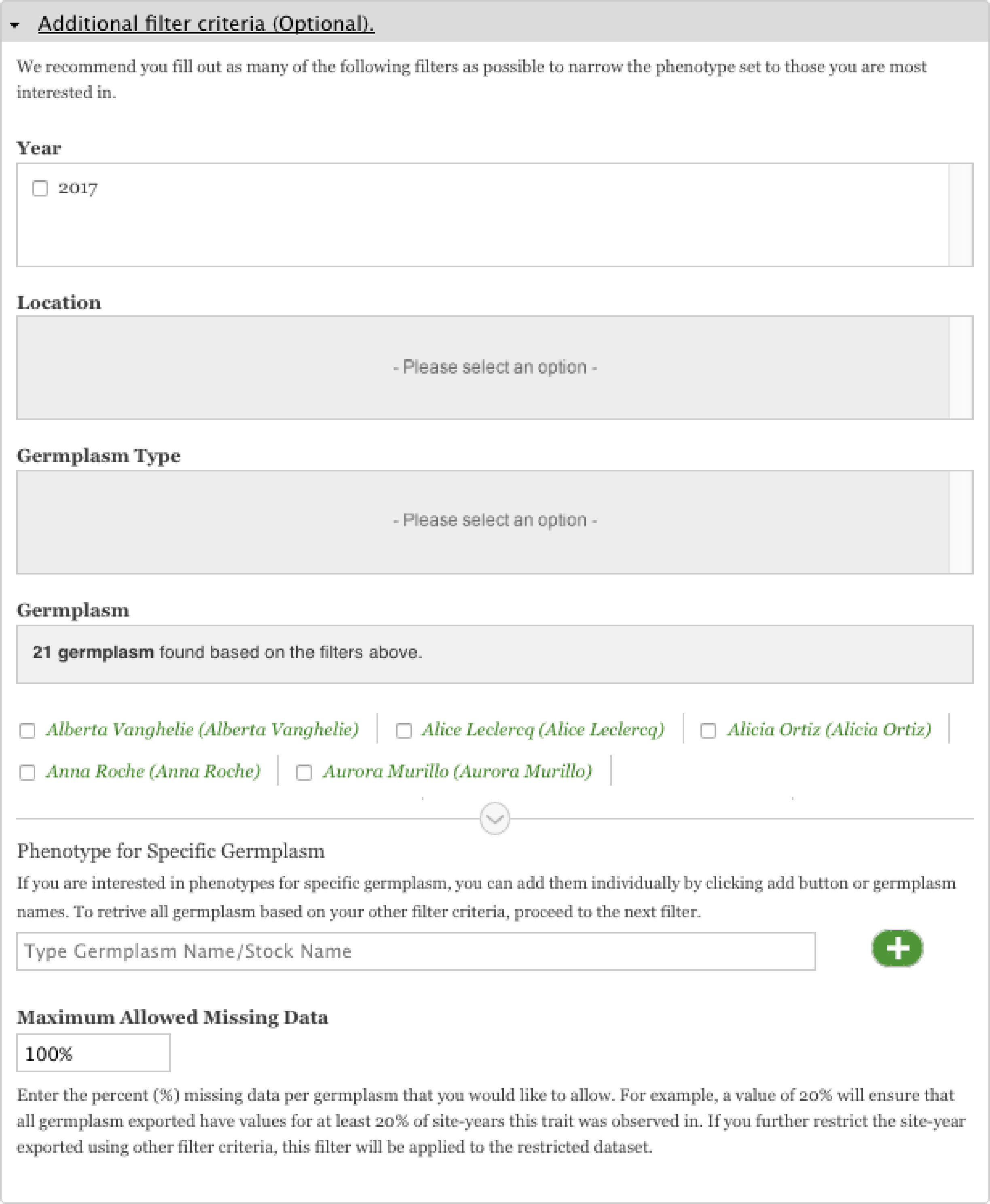
☑ basica
☐ databasica

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 retrive all germplasm based on your other filter criteria, proceed to the next filter.

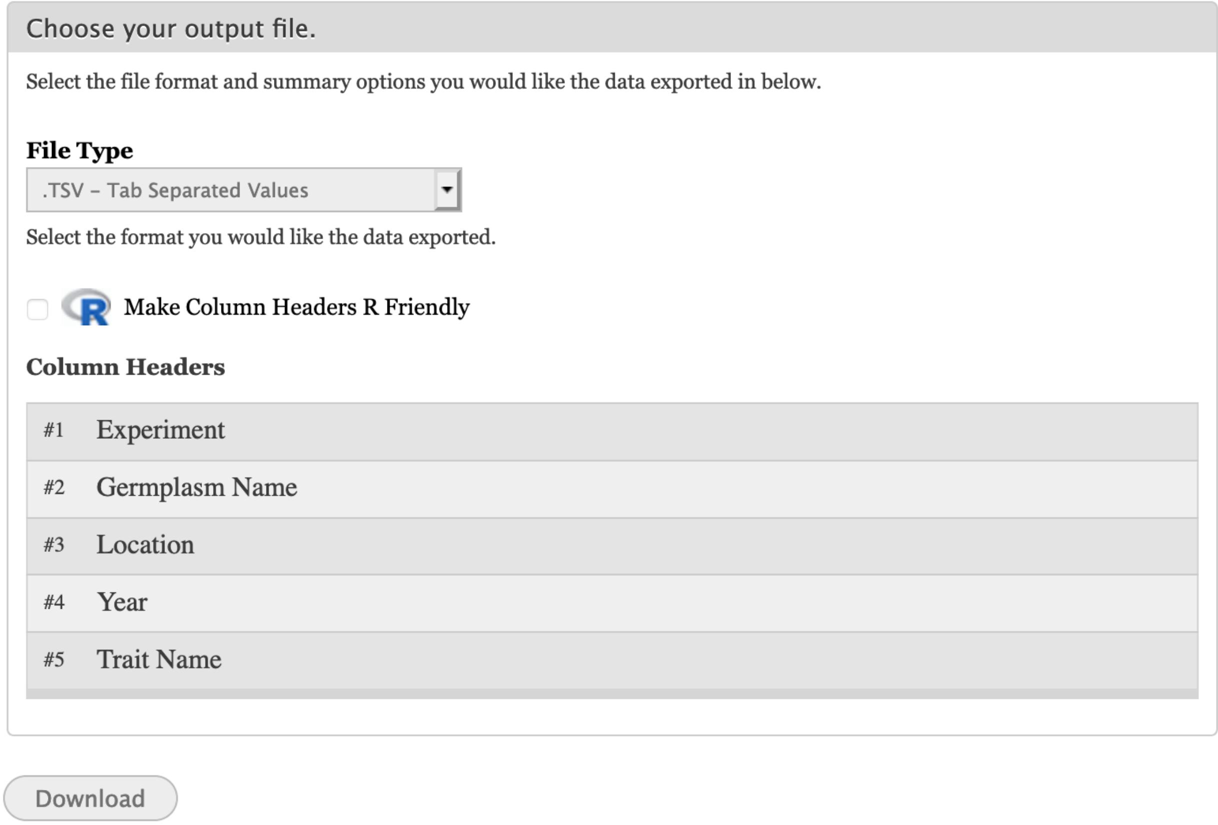
Type Germplasm Name/Stock Name

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 and summary options you would like the data exported in below.

File Type

.TSV – Tab Separated Values

Select the format you would like the data exported.

□ Make Column Headers R Friendly

Column Headers

#1 Experiment

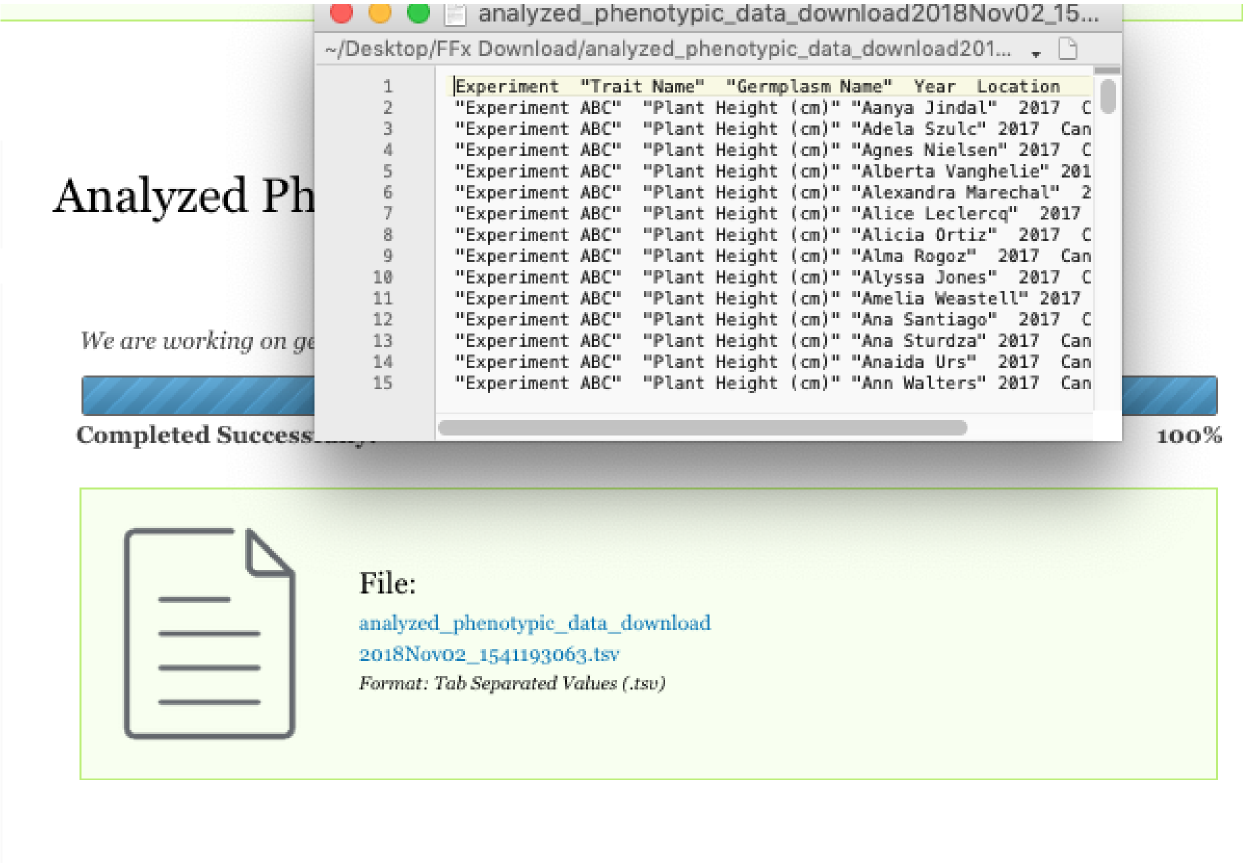
#2 Germplasm Name

#3 Location

#4 Year

#5 Trait Name

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



analyzed_phenotypic_data_download2018Nov02_15...

~/Desktop/FFx Download/analyzed_phenotypic_data_download201...

```
Experiment  "Trait Name"  "Germplasm Name"  Year  Location
"Experiment ABC"  "Plant Height (cm)" "Aanya Jindal"  2017  C
"Experiment ABC"  "Plant Height (cm)" "Adela Szulc" 2017  Can
"Experiment ABC"  "Plant Height (cm)" "Agnes Nielsen" 2017  C
"Experiment ABC"  "Plant Height (cm)" "Alberta Vanghelie" 201
"Experiment ABC"  "Plant Height (cm)" "Alexandra Marechal"  2
"Experiment ABC"  "Plant Height (cm)" "Alice Leclercq"  2017
"Experiment ABC"  "Plant Height (cm)" "Alicia Ortiz"  2017  C
"Experiment ABC"  "Plant Height (cm)" "Alma Rogoz"  2017  Can
"Experiment ABC"  "Plant Height (cm)" "Alyssa Jones"  2017  C
"Experiment ABC"  "Plant Height (cm)" "Amelia Weastell" 2017
"Experiment ABC"  "Plant Height (cm)" "Ana Santiago"  2017  C
"Experiment ABC"  "Plant Height (cm)" "Ana Sturdza" 2017  Can
"Experiment ABC"  "Plant Height (cm)" "Anaida Urs"  2017  Can
"Experiment ABC"  "Plant Height (cm)" "Ann Walters" 2017  Can
```

Analyzed Ph

We are working on ge

Completed Success...

100%

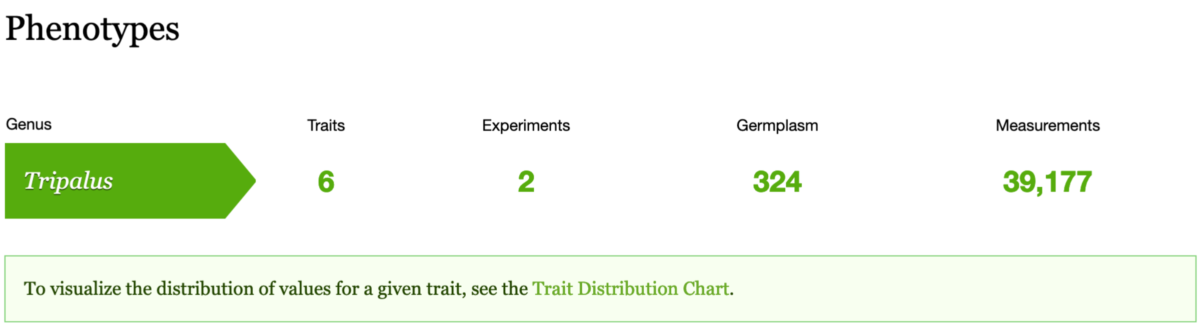
File:

analyzed_phenotypic_data_download2018Nov02_1541193063.tsv

Format: Tab Separated Values (.tsv)

2.3 Visualize Phenotypic Data

The data you have loaded is summarized on the phenotypic summary page. The following example summarizes two experiments where a total of six traits were measured. The number of germplasm represents the unique set assayed among all traits and experiments; it does not imply that any one trait was measured by 324 germplasm although this may be the case. The number of measurements indicates the number of phenotypic data points available.

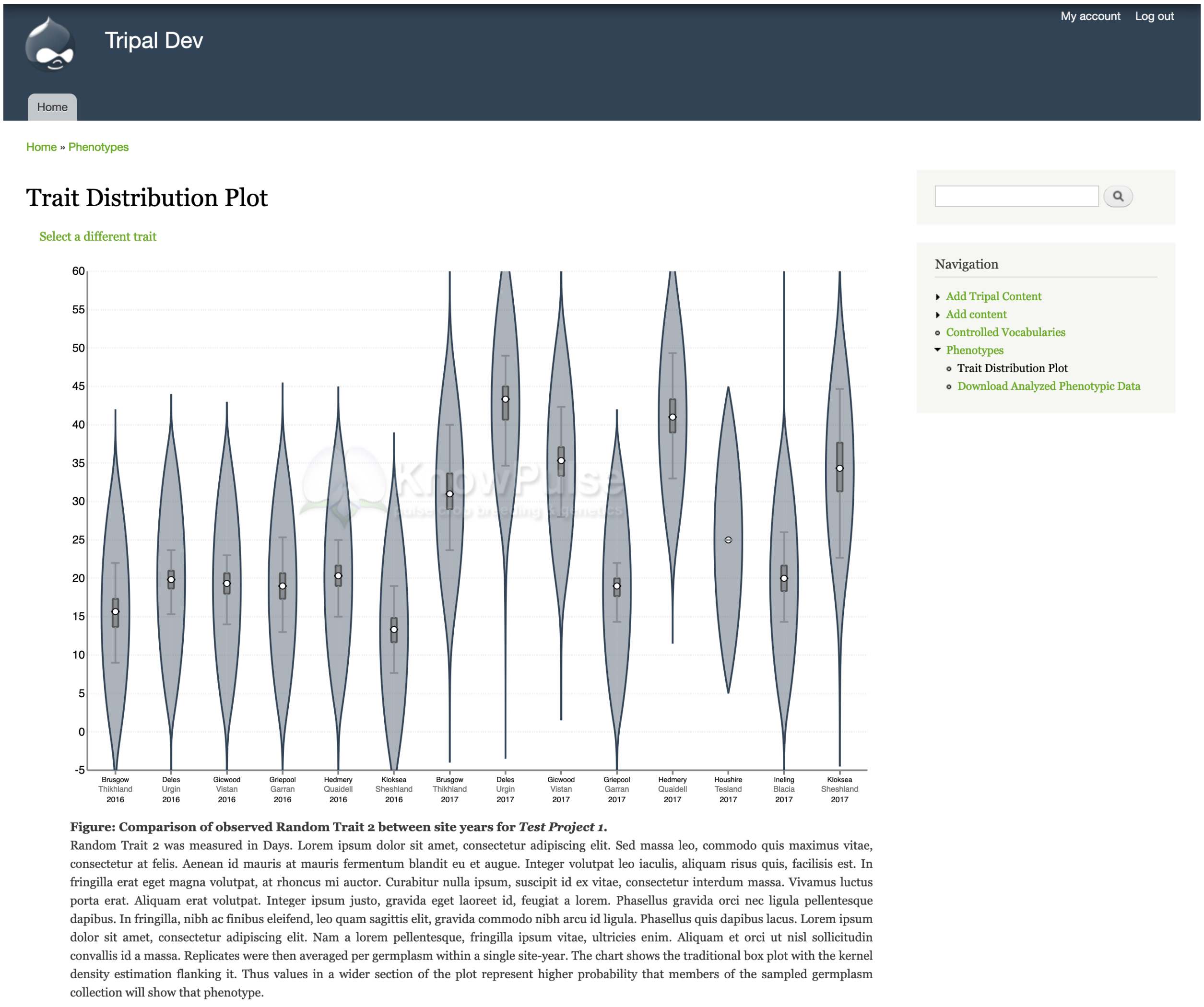


Phenotypes

Genus	Traits	Experiments	Germplasm	Measurements
Tripalus	6	2	324	39,177

To visualize the distribution of values for a given trait, see the Trait Distribution Chart.

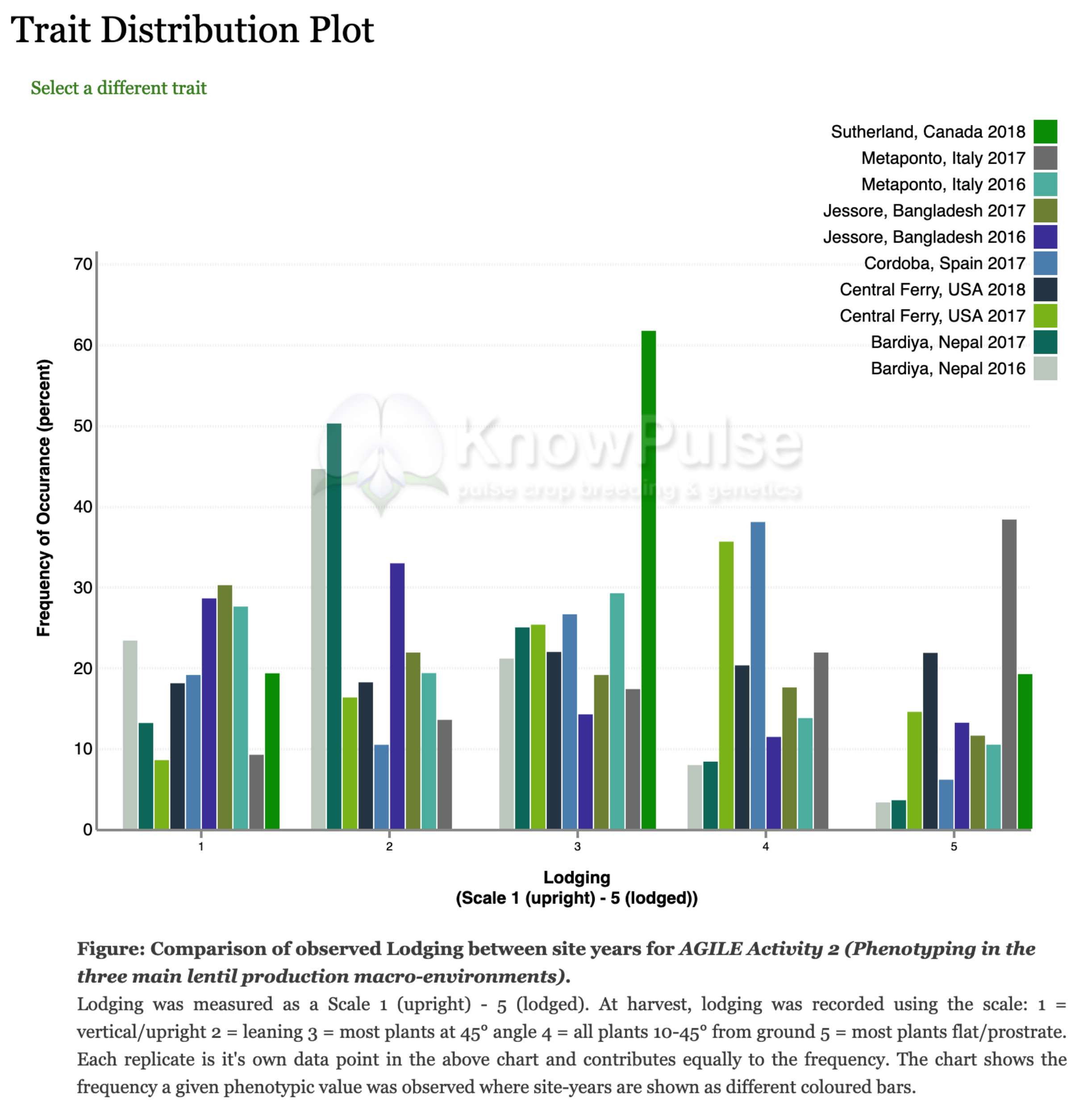
To visualize the distribution of values for a single trait, see the Trait Distribution chart for that trait. This chart can be accessed from the summary page above and will summarize the data for a single trait within a single experiment. Data is averaged across replicates but not across site-years. This allows you to compare the trait distribution between site-years for consistency and/or environmental effect. For quantitative traits a violin plot is shown and for qualitative traits a multi-bar chart is shown.



Trait Distribution Plot

Figure: Comparison of observed Random Trait 2 between site years for *Test Project 1*.

Random Trait 2 was measured in Days. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed massa leo, commodo quis maximus vitae, consectetur at felis. Aenean id mauris at mauris fermentum blandit eu et augue. Integer volutpat leo iaculis, aliquam risus quis, facilisis est. In fringilla erat eget magna volutpat, at rhoncus mi auctor. Curabitur nulla ipsum, suscipit id ex vitae, consectetur interdum massa. Vivamus luctus porta erat. Aliquam erat volutpat. Integer ipsum justo, gravida eget laoreet id, feugiat a lorem. Phasellus gravida orci nec ligula pellentesque dapibus. In fringilla, nibh ac finibus eleifend, leo quam sagittis elit, gravida commodo nibh arcu id ligula. Phasellus quis dapibus lacus. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nam a lorem pellentesque, fringilla ipsum vitae, ultricies enim. Aliquam et orci ut nisl sollicitudin convallis id a massa. Replicates were then averaged per germplasm within a single site-year. The chart shows the traditional box plot with the kernel density estimation flanking it. Thus values in a wider section of the plot represent higher probability that members of the sampled germplasm collection will show that phenotype.



Trait Distribution Plot

Select a different trait

Figure: Comparison of observed Lodging between site years for *AGILE Activity 2 (Phenotyping in the three main lentil production macro-environments)*.

Lodging was measured as a Scale 1 (upright) - 5 (lodged). At harvest, lodging was recorded using the scale: 1 = vertical/upright 2 = leaning 3 = most plants at 45° angle 4 = all plants 10-45° from ground 5 = most plants flat/prostrate. Each replicate is it's own data point in the above chart and contributes equally to the frequency. The chart shows the frequency a given phenotypic value was observed where site-years are shown as different coloured bars.

2.4 Tripal Fields

This module offers a number of Tripal Fields for enhancing trait, project, and germplasm pages. Fields are available to content types with specific chado base tables as indicated below. In addition to the default Tripal content types listed, these fields will also be available to any custom content types based on the listed chado tables.

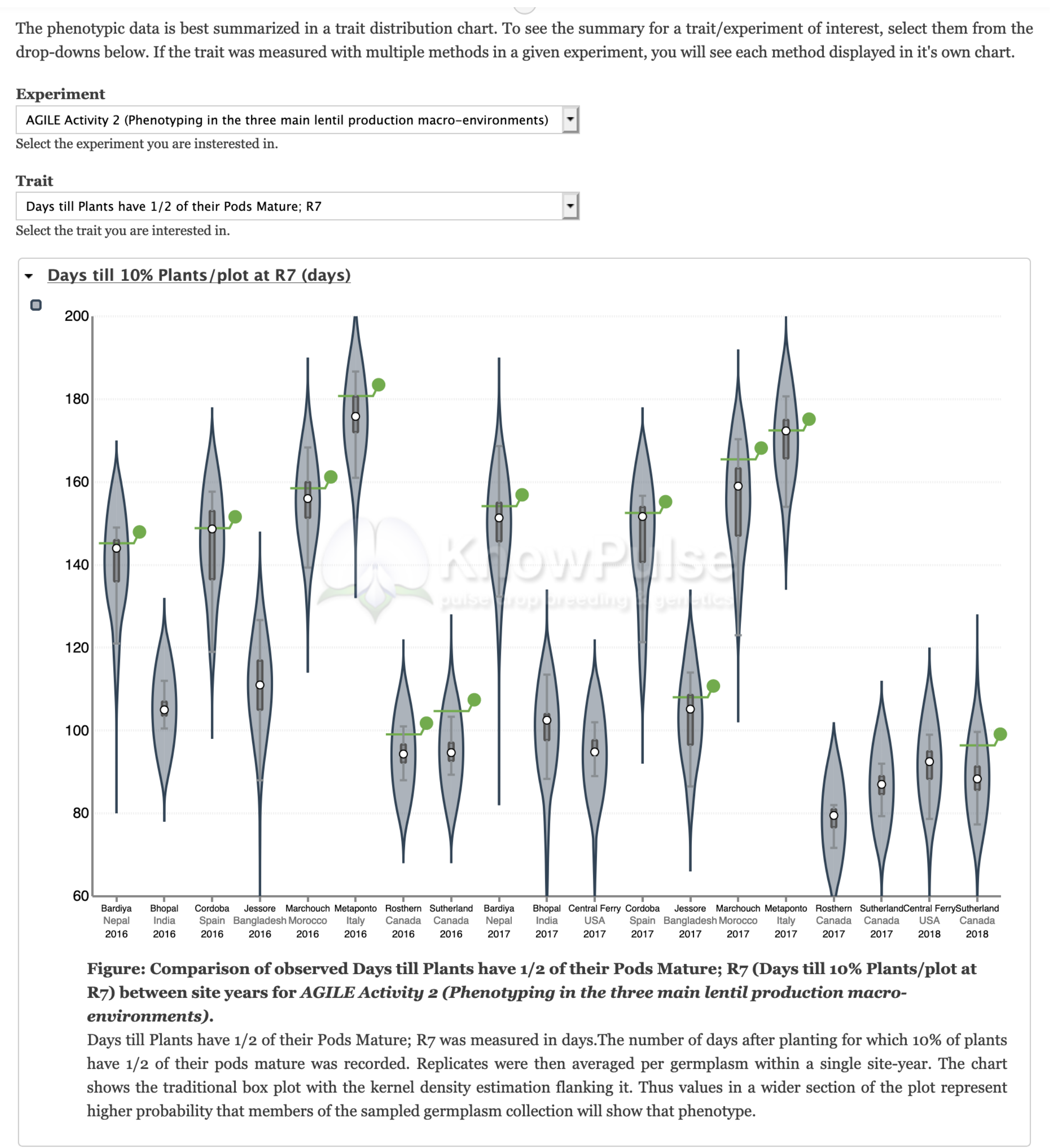
2.4.1 Trait Distribution Chart (hp__phenotypic_variability)

Base Chado table	project, stock, cvterm (traits)
Default Tripal Content Types	Project, Study, Cultivar (germplasm Variety), Generated Germplasm (breeding Cross), Germplasm Accession, Recombinant Inbred Line

In all cases, this field will embed the Trait Distribution Chart in your Tripal Content page. However, there is also specific forms and functionality depending on the content type it is attached to.

Germplasm

On stock-based pages, the user can select an experiment (project) and a trait after which they will see the trait distribution plot specific to that combination. If the trait was measured using multiple methods (e.g. if the method changed through the project) then a chart per method will be shown. Additionally, the current germplasm is indicated in each site-year by a green line intersecting the violin. The user can hover-over the attached circular handle to see the exact value recorded for the current germplasm in a particular site-year.



The phenotypic data is best summarized in a trait distribution chart. To see the summary for a trait/experiment of interest, select them from the drop-downs below. If the trait was measured with multiple methods in a given experiment, you will see each method displayed in it's own chart.

Experiment

AGILE Activity 2 (Phenotyping in the three main lentil production macro-environments) ▾

Select the experiment you are insterested in.

Trait

Days till Plants have 1/2 of their Pods Mature; R7 ▾

Select the trait you are interested in.

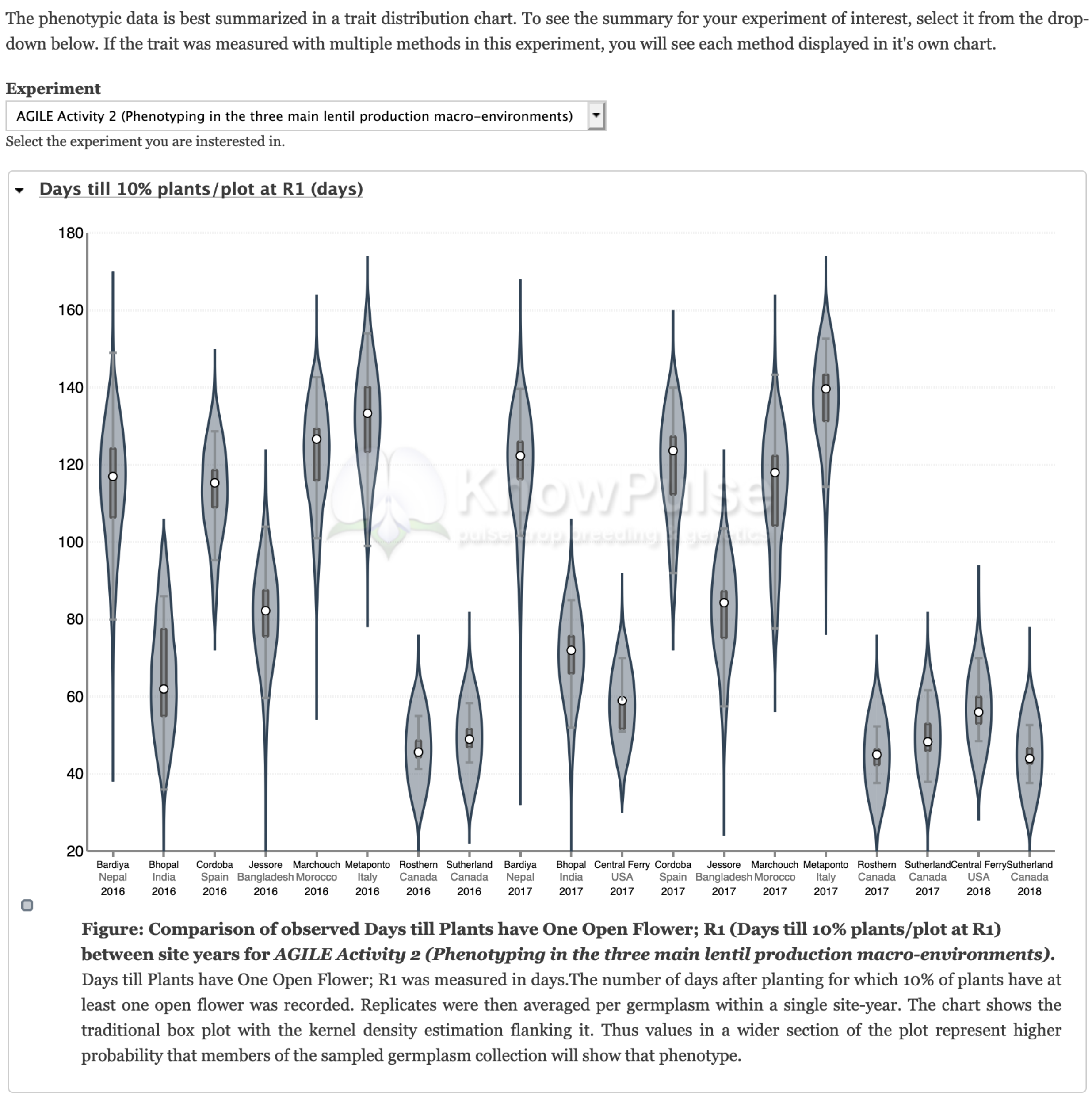
▾ Days till 10% Plants/plot at R7 (days)

Figure: Comparison of observed Days till Plants have 1/2 of their Pods Mature; R7 (Days till 10% Plants/plot at R7) between site years for *AGILE Activity 2 (Phenotyping in the three main lentil production macro-environments)*.

Days till Plants have 1/2 of their Pods Mature; R7 was measured in days.The number of days after planting for which 10% of plants have 1/2 of their pods mature was recorded. Replicates were then averaged per germplasm within a single site-year. The chart shows the traditional box plot with the kernel density estimation flanking it. Thus values in a wider section of the plot represent higher probability that members of the sampled germplasm collection will show that phenotype.

Traits

On cvterm-based trait pages, the user can select the specific experiment after which they will see the trait distribution plot for the current trait specific to the experiment chosen. If the trait was measured using multiple methods (e.g. if the method changed through the project) then a chart per method will be shown.



The phenotypic data is best summarized in a trait distribution chart. To see the summary for your experiment of interest, select it from the drop-down below. If the trait was measured with multiple methods in this experiment, you will see each method displayed in it's own chart.

Experiment

AGILE Activity 2 (Phenotyping in the three main lentil production macro-environments) ▾

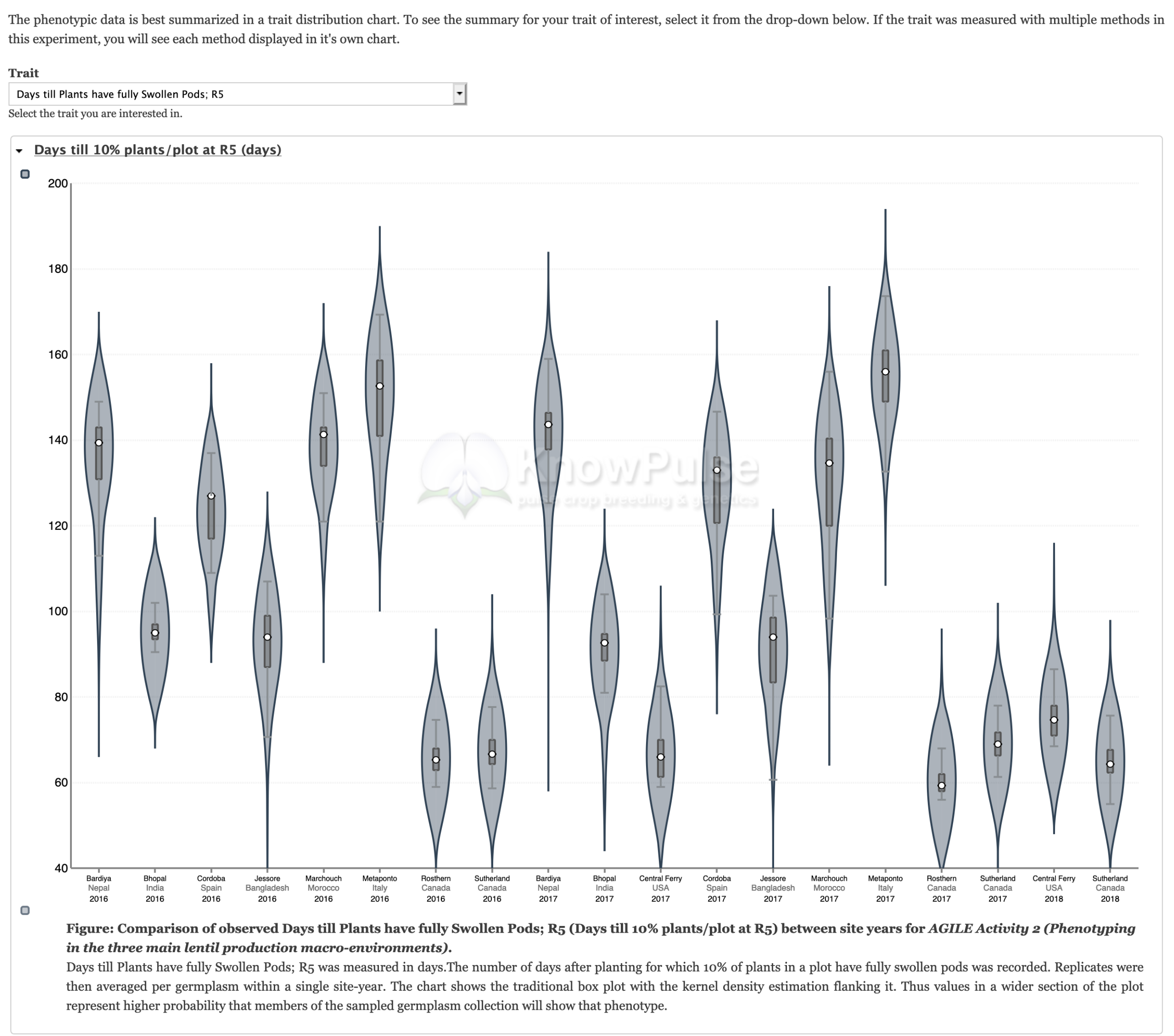
Select the experiment you are insterested in.

▾ Days till 10% plants/plot at R1 (days)

Figure: Comparison of observed Days till Plants have One Open Flower; R1 (Days till 10% plants/plot at R1) between site years for *AGILE Activity 2 (Phenotyping in the three main lentil production macro-environments)*. Days till Plants have One Open Flower; R1 was measured in days.The number of days after planting for which 10% of plants have at least one open flower was recorded. Replicates were then averaged per germplasm within a single site-year. The chart shows the traditional box plot with the kernel density estimation flanking it. Thus values in a wider section of the plot represent higher probability that members of the sampled germplasm collection will show that phenotype.

Project

On project-based pages, the user can choose a trait to visualize after which they will see the trait distribution plot specific to the current project. If the trait was measured using multiple methods (e.g. if the method changed through the project) then a chart per method will be shown.



The phenotypic data is best summarized in a trait distribution chart. To see the summary for your trait of interest, select it from the drop-down below. If the trait was measured with multiple methods in this experiment, you will see each method displayed in it's own chart.

Trait

Days till Plants have fully Swollen Pods; R5

Select the trait you are interested in.

▾ Days till 10% plants/plot at R5 (days)

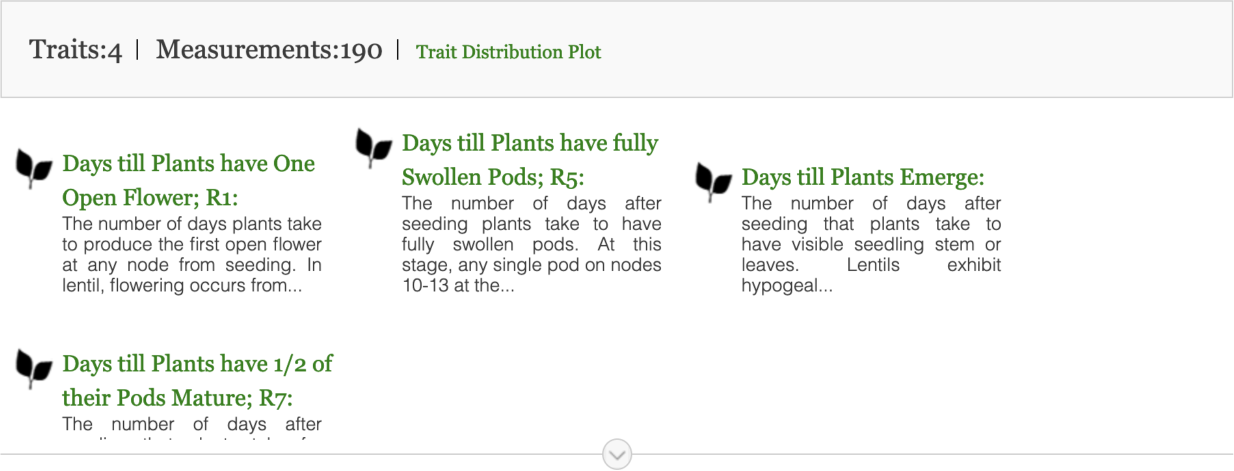
Figure: Comparison of observed Days till Plants have fully Swollen Pods; R5 (Days till 10% plants/plot at R5) between site years for *AGILE Activity 2 (Phenotyping in the three main lentil production macro-environments)*.

Days till Plants have fully Swollen Pods; R5 was measured in days.The number of days after planting for which 10% of plants in a plot have fully swollen pods was recorded. Replicates were then averaged per germplasm within a single site-year. The chart shows the traditional box plot with the kernel density estimation flanking it. Thus values in a wider section of the plot represent higher probability that members of the sampled germplasm collection will show that phenotype.

2.4.2 Magnitude of Phenotypes (local__magnitude_of_phenotypes)

Base Chado table	stock
Default Tripal Content Types	Cultivar (germplasm Variety), Generated Germplasm (breeding Cross), Germplasm Accession, Recombinant Inbred Line

This field is meant to give an indication for how much phenotypic data is available for a given germplasm. It indicates the number of traits measured and the total number of phenotypic measurements collected, as well as linking to both the trait distribution chart and individual trait pages for more information. The first few traits are shown by default with an expandable arrow giving access to the rest.



Traits:4 | Measurements:190 | Trait Distribution Plot

Days till Plants have One Open Flower; R1:
The number of days plants take to produce the first open flower at any node from seeding. In lentil, flowering occurs from...

Days till Plants have fully Swollen Pods; R5:
The number of days after seeding plants take to have fully swollen pods. At this stage, any single pod on nodes 10-13 at the...

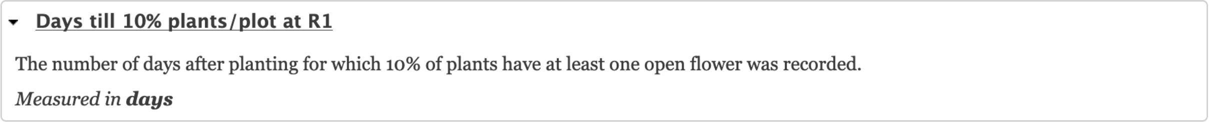
Days till Plants Emerge:
The number of days after seeding that plants take to have visible seedling stem or leaves. Lentils exhibit hypogeal...

Days till Plants have 1/2 of their Pods Mature; R7:
The number of days after

2.4.3 Methodology (ncit__method)

Base Chado table	cvterm (traits)
Default Tripal Content Types	None available. Please create a content type through the Administrative UI.

This field simply lists the methods used to measure a given trait. This is helpful for understanding the differences you may see between data from different experiments, as well as, to inform possible techniques for future experiments.



▾ Days till 10% plants/plot at R1

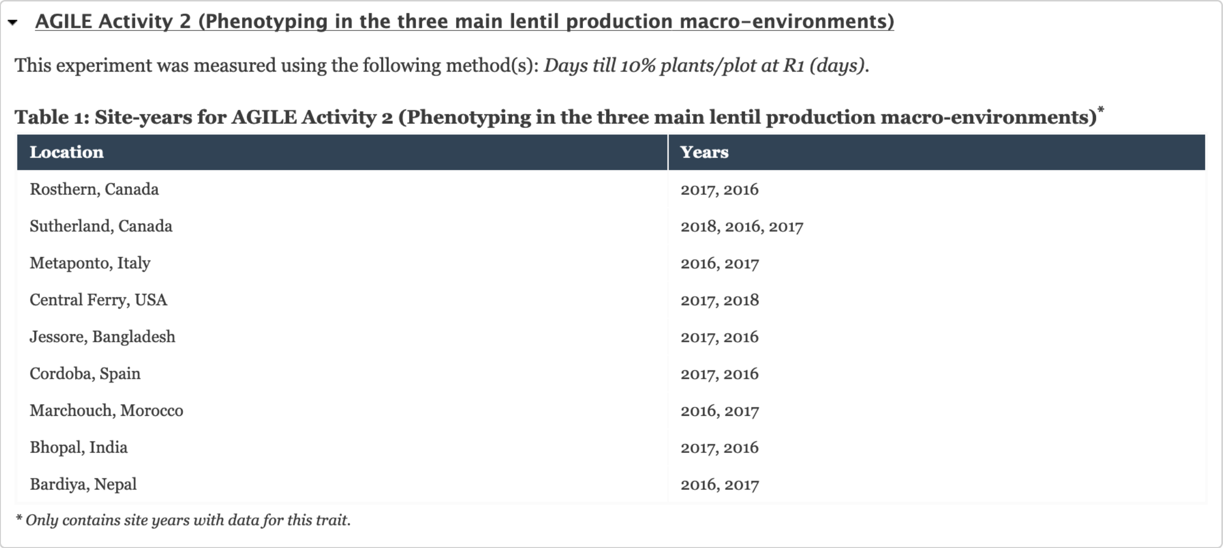
The number of days after planting for which 10% of plants have at least one open flower was recorded.

*Measured in **days***

2.4.4 Experiments (sio__study)

Base Chado table	cvterm (traits)
Default Tripal Content Types	None available. Please create a content type through the Administrative UI.

This field lists each experiment assaying the current trait. For each experiment the method used and site-years are listed.



▾ AGILE Activity 2 (Phenotyping in the three main lentil production macro-environments)

This experiment was measured using the following method(s): *Days till 10% plants/plot at R1 (days)*.

Table 1: Site-years for AGILE Activity 2 (Phenotyping in the three main lentil production macro-environments)*

Location	Years
Rosthern, Canada	2017, 2016
Sutherland, Canada	2018, 2016, 2017
Metaponto, Italy	2016, 2017
Central Ferry, USA	2017, 2018
Jessore, Bangladesh	2017, 2016
Cordoba, Spain	2017, 2016
Marchouch, Morocco	2016, 2017
Bhopal, India	2017, 2016
Bardiya, Nepal	2016, 2017

** Only contains site years with data for this trait.*

CHAPTER 3

Contributing

We're excited to work with you! Post in the issues queue with any questions, feature requests, or proposals.

3.1 Automated Testing

This module uses Tripal Test Suite. To run tests locally:

```
cd MODULE_ROOT
composer up
./vendor/bin/phpunit
```

This will run all tests associated with the Analyzed Phenotypes extension module. If you are running into issues, this is a good way to rule out a system incompatibility.

> **Warning:** It is highly suggested you ONLY RUN TESTS ON DEVELOPMENT SITES. We have done our best to ensure that our tests clean up after themselves; however, we do not guarantee there will be no changes to your database.

3.2 Manual Testing (Demonstration)

We have provided a Tripal Test Suite Database Seeder to make development and demonstration of functionality easier. To populate your development database with fake phenotypic data:

1. Install this module according to the instructions in the administration guide.
2. Create an organism (genus: Tripalus; species: databasica)
3. Configure the module terms by Navigating to Administration » Tripal » Extensions » Analyzed Phenotypes » Set-up Ontologies and click "Save term configuration" at the bottom of the page.
4. Run the database seeder to populate the database using the following commands:

```
cd MODULE_ROOT
composer up
./vendor/bin/tripaltest db:seed PhenotypeSeeder
```

4. Populate the materialized views by going to Administration » Tripal » Data Storage » Chado » Materialized Views and clicking “Populate” beside `mview_phenotype` and `mview_phenotype_summary`. Finally run the Tripal jobs submitted.
5. Create the trait content type by going to Administration » Structure » Tripal Content Types » Add Tripal Content Type. We suggest the following values:

- Chado Base Table: cvterm
- Use a Parent Chado cvterm? No. All records belong to a single controlled vocabulary.
- Restrict to Vocabulary: [get value from set-up ontologies page]

6. To play with trait pages go to Administration » Structure » Tripal Content Types » Publish Tripal Content and select the term used in step 5 to create pages.

Warning: NEVER run database seeders on production sites. They will insert fictitious data into Chado.

3.3 Stress Testing

We have also provided a Tripal Test Suite Database Seeder to be used for stress testing this module. It inserts 3 billion phenotype records including associated metadata. To populate your development database with this fake phenotypic dataset:

1. Install this module according to the instructions in the administration guide.
2. Run the database seeder to populate the database using the following commands:

```
cd MODULE_ROOT
composer up
./vendor/bin/tripaltest db:seed MassivePhenotypeSeeder
```

3. Populate the materialized views by going to Administration » Tripal » Data Storage » Chado » Materialized Views and clicking “Populate” beside `mview_phenotype` and `mview_phenotype_summary` and run the Tripal jobs submitted.
4. Edit tests/massivePhenotypesTimings.php to include a trait and project ID which exist.
5. Run the timings script to determine how specific queries in the module may respond to the current dataset.

```
cd MODULE_ROOT
drush php-script tests/massivePhenotypesTimings.php
```

Warning: This script will take at least 4 hours to run due to 9 spaced replicates. Additionally, the execution time will increase depending on how your system handles these queries.