

NEW TRIPAL MODULES: ELASTICSEARCH AND EXPRESSION

Margaret Staton



Open source content management system (CMS) for biological data

- Specializing in genetic, genomic, breeding, etc.

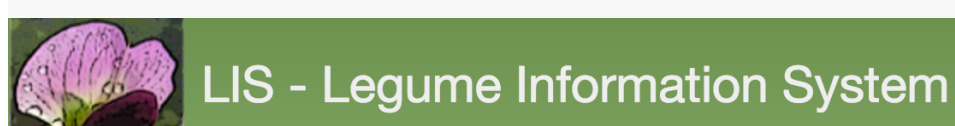
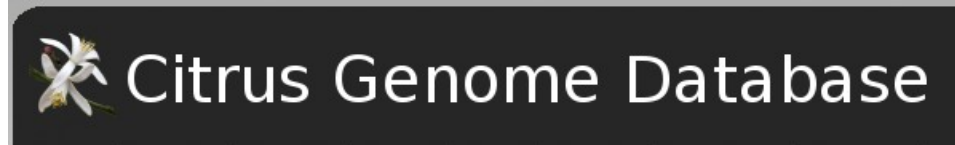
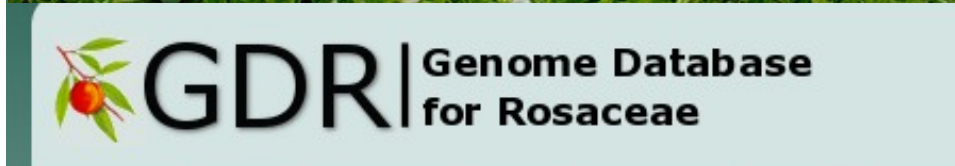
97 sites report using Tripal!

Benefits:

- Reduces IT costs
- Publishes simple genome sites out-of-the-box
- Provides an API for complete customization
- Uses Chado and community ontologies for standardization
- Allows for sharing of extensions between sites

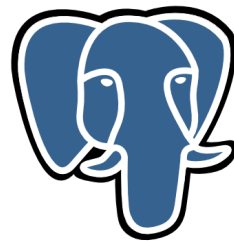
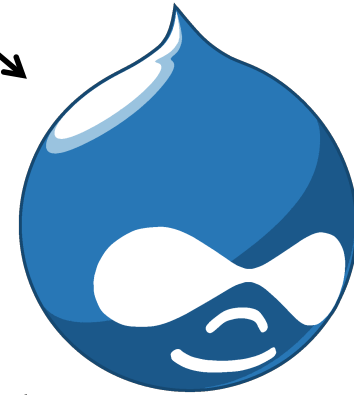


The Hardwood Genomics Project

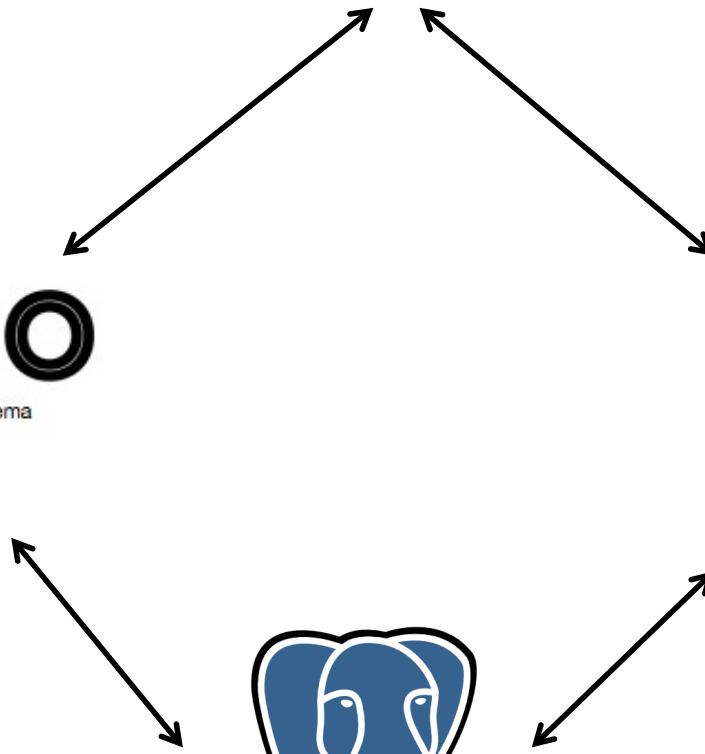




CHADO
Chado: Biological database schema



PostgreSQL



Extension Module System

- Core modules
 - Developed by core team or vetted by core team
 - Likely to be needed/appreciated by all Tripal sites
 - Well tested
- Extension modules
 - Anyone can contribute extra functionality
 - Take 'em or leave 'em - If you don't need it, it doesn't clutter your server

<http://tripal.info>

Extension modules can eventually be integrated with core.

Tripal Extensions

The following is a list of extensions that have been provided by Tripal site developers and which can be imported into another Tripal site. If you have custom extensions you would like to share, please let us know by joining [the Tripal mailing list](#) and sending a request. Consider adding your extensions even if only in development state as this encourages collaboration and reduces duplication of effort. Use the fields below to filter.

Extension Type	Tripal	Chado	Module Status	Categories
- Any -	v1.0 v1.1 v2.0	v1.1x v1.2x v1.3x	No longer supported In Development Released: Needs debugging Released: Needs Review Released: Ready for Use	- Any -

ND Genotypes (create marker/stock) Bulk Loader Template

Bulk Loader Template

This template provides a means of loading marker by germplasm genotype matrices (See GenotypeSampleData.txt for an example) into Chado by creating Natural Diversity Experiments for each Genotype (ie: each element in the matrix) and linking it to the corresponding stock and marker. This module requires the sequence and relationship ontologies.

Breeding API Extension Module

Extension Module

Breeding API (<https://github.com/plantbreeding>) implementation for Tripal. Currently not released and in heavy development. This project has been sponsored by the Bill and Melinda Gates Foundation which funded the breeding API hackathon in June 2015 in Seattle.

<http://github.com/tripal>

HARDWOOD GENOMICS PROJECT



The Hardwood Genomics Project

[Home](#) [About](#) [Trees](#) [Genomic Data](#) [Tools](#) [Blast](#) [Search](#) [Contact](#)

[View](#) [Edit](#) [Outline](#) [Track](#)

Welcome to the Hardwood Genomics Project

We house transcriptome and genome resources for hardwood trees.

Interested in contributing data? [Please contact us!](#)

This website was originally developed for the NSF grant **Comparative Genomics of Environmental Stress Responses in North American Hardwoods**, which has now ended. We are expanding to a larger set of tree genomic resources with the support of the NSF grant **CIF21 DIBBS: Tripal Gateway, a Platform for Next-Generation Data Analysis and Sharing** (PI Stephen Ficklin, Washington State University). Species and resources currently available:

- [Chinese Chestnut \(draft reference genome, physical map, genetic map\)](#)
- [American Beech \(transcriptome sequencing\)](#)
- [American Chestnut \(transcriptome sequencing\)](#)
- [Black Cherry \(low coverage genome sequence, SSRs\)](#)
- [Black Walnut \(transcriptome sequencing, SSRs, BAC, QTL Map, Reference Populations, BAC Sequencing\)](#)
- [Blackgum \(transcriptome sequencing, SSRs\)](#)
- [European Chestnut \(transcriptome sequencing\)](#)



Upcoming Meetings

Aug 8th-11th, 2016

[American Society for Horticultural Science](#)
Atlanta, GA

Oct 4th-6th, 2016

Training Course: [Molecular phylogeography of forest tree species using newly developed genomic resources \(EST-SSRs and SNPs\)](#)
Transilvania University of Brasov, Romania

June 26th-29th, 2017

[Forest Genetics 2017: Health and Productivity under Changing Environments](#)
Hosted by [Western Forest Genetics Association \(WFGA\)](#)
Edmonton, Alberta

September 19th-27th, 2017

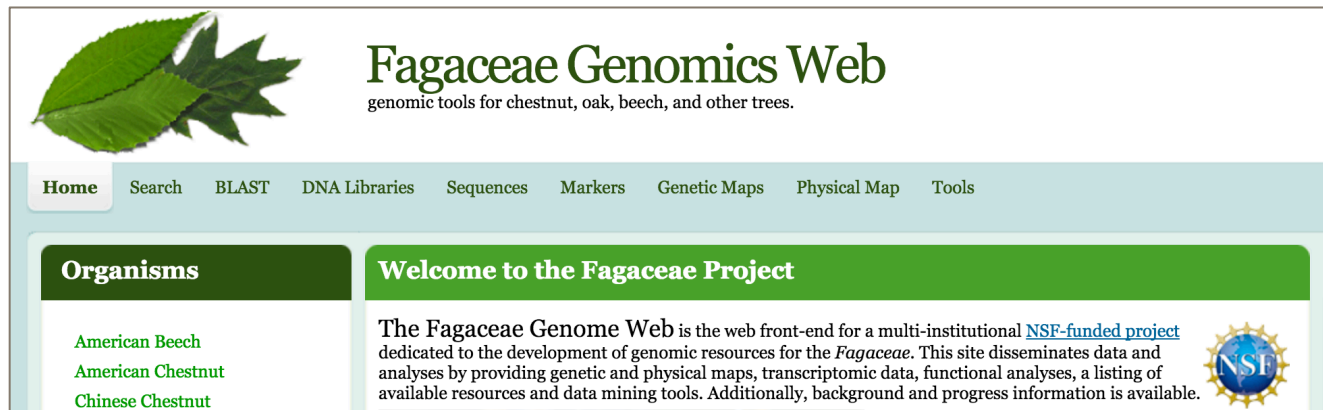
[IUFRO 125th Anniversary Congress](#)
Freidberg, Germany

Hardwood Genomics - Data

- Built with a grant for hardwood tree genomics (2010, PI Carlson)
- Seedling stress testing to mimic climate change:
 - Ozone, heat, cold, drought, wounding
- Transcriptome sequencing for 8 species
 - Libraries from diversity of tissue types
 - Libraries from abiotic stress treatments
- Genetic mapping populations for 6 species
- Molecular marker development for 12 species
 - (ranging from in silico only to laboratory confirmed)
- Genetic mapping for 4 species
- QTL mapping for 2 species

Hardwood Genomics - History

- Incorporates data from the original Fagaceae Genomics Web (built to house data from NSF grant to develop Fagaceae family genomic resources, 2007, PI Sederoff)



The screenshot shows the homepage of the Fagaceae Genomics Web. At the top left is a logo featuring a green leaf and a white flower. To the right of the logo, the text reads "Fagaceae Genomics Web" in a large green font, with the subtitle "genomic tools for chestnut, oak, beech, and other trees." below it. A navigation bar contains links for "Home", "Search", "BLAST", "DNA Libraries", "Sequences", "Markers", "Genetic Maps", "Physical Map", and "Tools". Below the navigation bar, there are two main sections. The left section is titled "Organisms" and lists "American Beech", "American Chestnut", and "Chinese Chestnut". The right section is titled "Welcome to the Fagaceae Project" and contains a paragraph of text: "The Fagaceae Genome Web is the web front-end for a multi-institutional [NSF-funded project](#) dedicated to the development of genomic resources for the *Fagaceae*. This site disseminates data and analyses by providing genetic and physical maps, transcriptomic data, functional analyses, a listing of available resources and data mining tools. Additionally, background and progress information is available." To the right of this text is the NSF logo.

- Continuing interest in Chestnut genomics (Forest Health Initiative, The American Chestnut Foundation, USDA)

Hardwood Genomics - Data

- Chestnut
- Genome
 - Genetic map
 - Physical map

Common Name	Transcriptome	RNASeq Biomaterial records	Expression Analysis	Predicted SSRs (transcriptomic)	Predicted SSRs (genomic)	Polymorphic SSRs	Population Description	BAC Libraries
Sugar Maple	x	x	x	x	x	x		
White Alder	x	o		x				
Red Alder	x	o		x				
Japanese Chestnut	x	o		o				
American Chestnut	x	o		x		x		x
Chinese Chestnut		o				x		x
European Chestnut	x	o		o				
American Beech	x	o		x				
White Ash					x	x		
Green Ash	x	x	x	x	x	x	x	
Honeylocust	x	x	x	x	x	x	x	
Black Walnut	x	x	x	x	x		x	x
English Walnut								
American Sweetgur	x	x	x	x	x	x		
Tulip Poplar	x	x	x	x		x	x	x
Blackgum	x	x	x	x	x			
Redbay					x			
Black Cherry					x	x		
White Oak	x	o		x	x	x	x	
Pedunculate Oak								x
Northern Red Oak	x	x	x	x		x	x	x

Hardwood Genomics - Tools

- Jbrowse
- Apollo
- CMap
- BLAST

- Symap – need to replace, any ideas?



Ongoing work - DIBBS

Tripal Gateway Project

NSF Data Information Building Blocks (DIBBs) grant

- Award #1443040
- PI Stephen Ficklin, Washington State University
- 3 years (1.5 years in)

Three components:

- RESTful Web services for Tripal sites
 - Allow sites to exchange data
- Integration with Galaxy
 - Allow sites to provide next-generation sequence analysis tools
- Improve data transfer
 - Big Data Smart Socket Client (BDSS) available
 - Explore Software Defined Networking (SDN)



Upcoming work - PGRP

- Dorrie Main, WSU (PI)
- Ontologies
 - Structure, trait, phenotypic quality and environment
 - Curation of current data
 - Standardize data collection in the future
 - Standardize data submission for users
- Communication between sites
 - Web services
 - Tripal extension module for cross-site querying - enabling a user to collate or view data from multiple Tripal sites
- Better querying and visualization of complex phenotype, genotype, and environment data
- Online educational modules, training courses, and developer/user support for Tripal

ELASTICSEARCH

What problem is being solved?

- Drupal internal search
 - Easy to set up and customize (for normal Drupal data types)
 - No native support for external DBs
 - Slow to index, slow to return results
- Need a solution that will:
 - Access chado database
 - Provide flexible and customizable indexing – index only what is needed, not everything
 - Scales to very large biological data sets

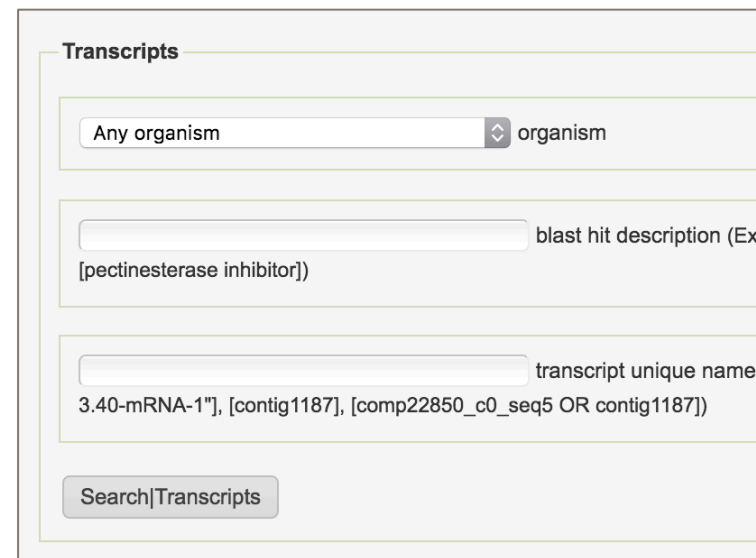
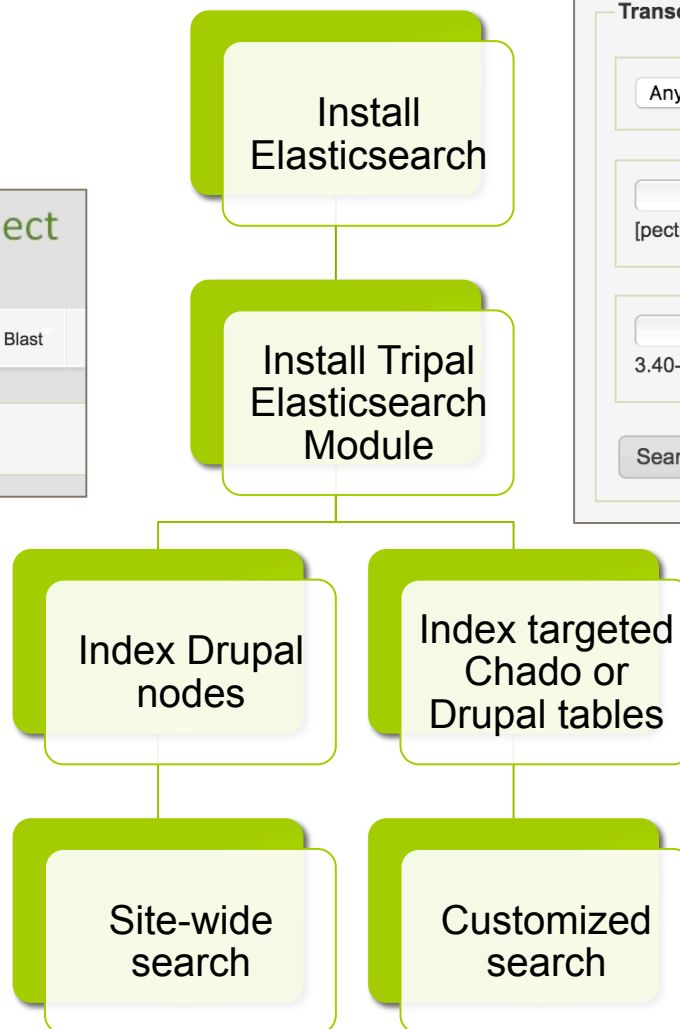
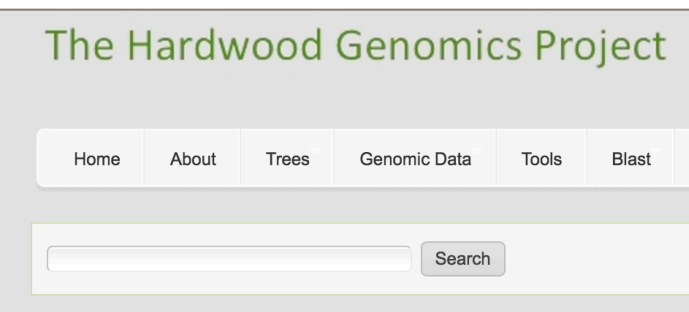
Elasticsearch Software



elastic

- distributed, open source search and analytics engine
- Massively distributed – can scale horizontally
- Multitenancy – a search cluster can manage many individual indices that can be queried individually or as a group
- Built on Apache lucene -> autocomplete, fuzzy searching, “did you mean” suggestions
- Document oriented – export database tables as JSON
- RESTful API can be leveraged with JSON over HTTP
- Open source

Elasticsearch Module



Elasticsearch Module - Example

Create a view with the materialized view table.

View Name *

Please enter the name for this materialized view.

MView Description

This view joins feature uniquenames to BLAST hit information (description, e-value, and hit score) and organism information (genus, species, common_name).

Optional. Please provide a description of the purpose for this materialized view.

Elasticsearch Module - Example

Create a view with the materialized view table.

Query *

```
hit_best_eval, b.hit_best_score AS hit_best_score, o.common_name AS common_name, o.genus AS  
genus, o.species AS species  
FROM  
chado.feature f  
INNER JOIN chado.blast_hit_data b ON b.feature_id = f.feature_id  
INNER JOIN chado.organism o ON f.organism_id = o.organism_id
```

Please enter the SQL statement used to populate the table.

Elasticsearch Module - Example

Schema Array

```
array (  
  'description' => 'This view joins feature uniquenames to BLAST hit information (description, e-value, and hit score) and  
organism information (genus, species, common_name).',  
  'table' => 'search_features_all',  
  'fields' => array (  
    'uniquename' => array (  
      'type' => 'text',  
      'not null' => true,  
    ),  
    'hit_description' => array (  
      'type' => 'text',  
      'not null' => true,  
    ),  
    'hit_best_eval' => array (  
      'type' => 'text',  
      'not null' => true,  
    ),  
    'hit_best_score' => array (  
      'type' => 'text',  
      'not null' => true,  
    ),  
    'common_name' => array (  
      'type' => 'text',  
      'not null' => true,  
    ),  
    'genus' => array (  
      'type' => 'text',  
      'not null' => true,  
    ),  
    'species' => array (  
      'type' => 'text',  
      'not null' => true,  
    ),  
  ),  
)
```

After
describing,
populate.

Elasticsearch Module - Example

Now to Elasticsearch admin.

Select the materialized view to index (or any other table).

Select the fields.

TRIPAL_ELASTICSEARCH INDEXING

Select a table to index

search_features_all

SELECT FIELDS TO INDEX

Fields available to index


- uniqueness
- hit_description
- hit_best_eval
- hit_best_score
- common_name
- genus
- species


Index

Elasticsearch Module - Example

Queue UI and ultimate cron are dependencies. You can check on cron jobs and run them in parallel. This is convenient if you have many processors and are on a dev server.

[Home](#) » [Administration](#) » [Configuration](#) » [System](#)

Queue manager 



▼ **SYSTEMQUEUE**

<input type="checkbox"/>	NAME	TITLE	NUMBER OF ITEMS	CLASS	INSPECT
<input type="checkbox"/>	update_fetch_tasks		8	SystemQueue	Inspect

Batch process

Remove leases

Cron process

Delete queues

Elasticsearch Module - Example

Move to demo...

Elasticsearch Module

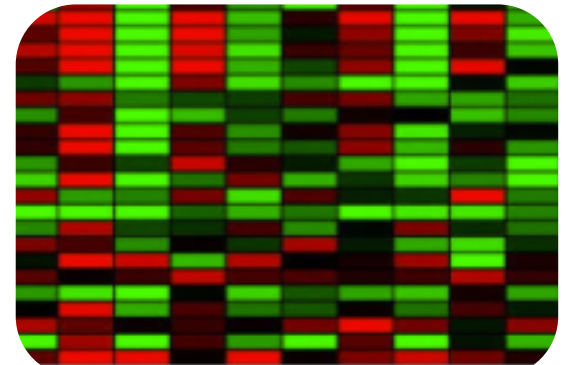
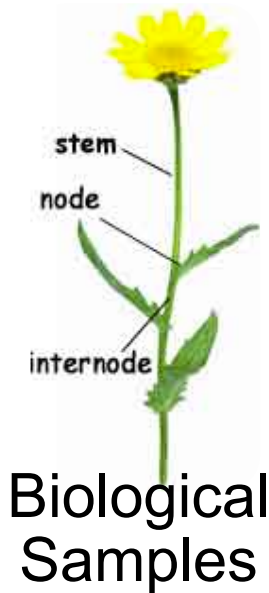
Future Development

- Edit the search form – change field labels, type of search field (dropdown, checkboxes), order of fields
- Paths - Are all fields easily accessed by URLs? Automate discovery of URL links for datatypes?
- Fasta file download for feature table (include script)
- Multisite installs – use the flexibility of elasticsearch
- Scale with bigger data and different types of data
- Port to Tripal 3.0 and compare to new internal searching

EXPRESSION MODULE

What problem is being solved?

Need a better way to store and visualize RNASeq differential gene expression experiments.



Expression Module

Module to display expression data collected from RNASeq

We have left open the possibility for microarray expression data sources as well, currently untested

Chado Tables/Modules used:

- MAGE
- Organism
- Contact
- Sequence
- Companalysis modules.

Content Types

- Tripal content types are created for these tables:
 - Biomaterial
 - Similar to NCBI BioSample and SRA
 - We do not differentiate between samples and libraries
 - Array design
 - Can be used for microarray data, but not used for NGS projects
 - Protocol
 - Define protocols for the experimental analysis
- New Chado analysis content type:
 - Analysis: Expression.

Loading Data

- Import biomaterial
 - BioSample data downloaded from NCBI (xml)
 - Flat file format (based on NCBI biomaterial bulk load form)
- Import expression values
 - (assumed to be normalized, features must already exist)
 - Individual file per sample
 - Tab delimited file with gene rows, sample columns

Visualization

- Demo...

Future Work on Expression Module

- Biomaterials

- Upload SRA records from NCBI automatically via web services
- Link the properties to ontologies
- Link to individual analyses (currently only displays as associated with an organism)
 - IE – A transcriptome is built from a subset of biomaterials

- Expression

- Allow user to provide a list of genes (cart system) and generate heatmap for all
- Add significance/p-values from differential gene expression test results
 - Important functional data
 - Aid searching – limit results only to genes that respond to cold stress

Acknowledgements



Washington State University

- Stephen Ficklin

University of Tennessee

- Ming Chen
- Nathan Henry



University of Saskatchewan

- Lacey Anne Sanderson



All the developers of

