Data Sharing: Examples from the Tripal Community

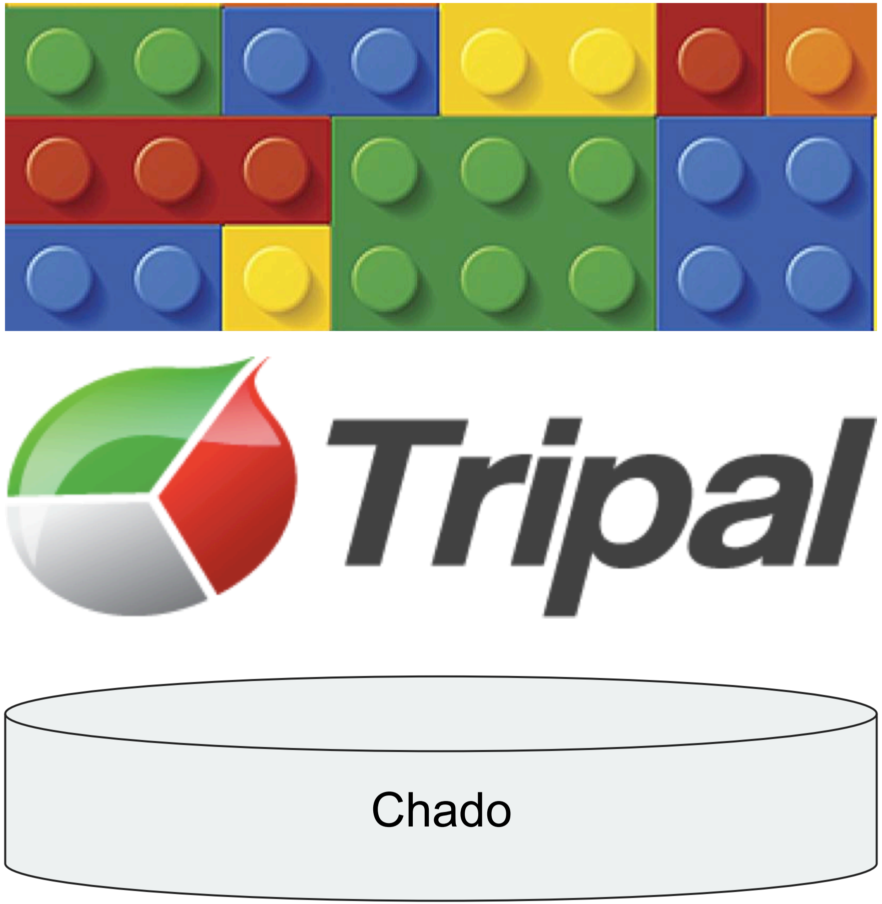
Meg Staton
University of Tennessee, Institute of Agriculture
mstaton1@utk.edu
@HardwoodGenomic

Tripal

A web framework for genetic and genomic data

Goals:

- Simplify construction of websites that have biological data
- Encourage high-quality, standards-based websites for data sharing and collaboration
- Expand and reuse code



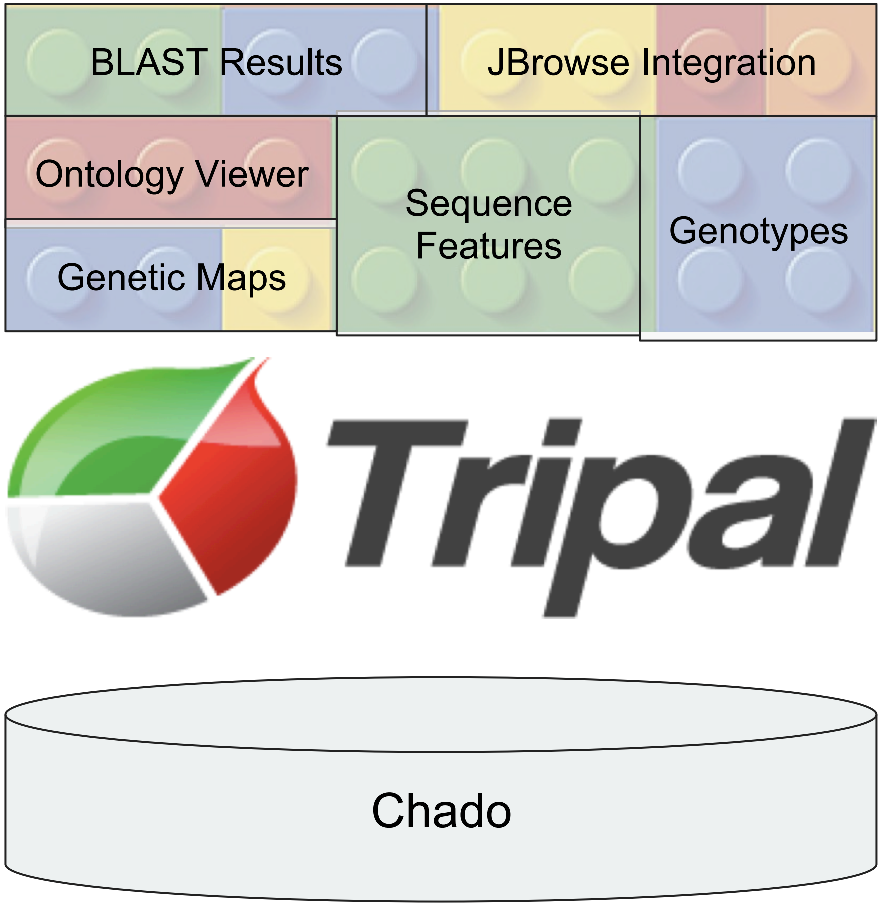
Chado

Tripal

A web framework for genetic and genomic data

Goals:

- Simplify construction of a websites that have biological data
- Encourage high-quality, standards-based websites for data sharing and collaboration
- Expand and reuse code



BLAST Results

JBrowse Integration

Ontology Viewer

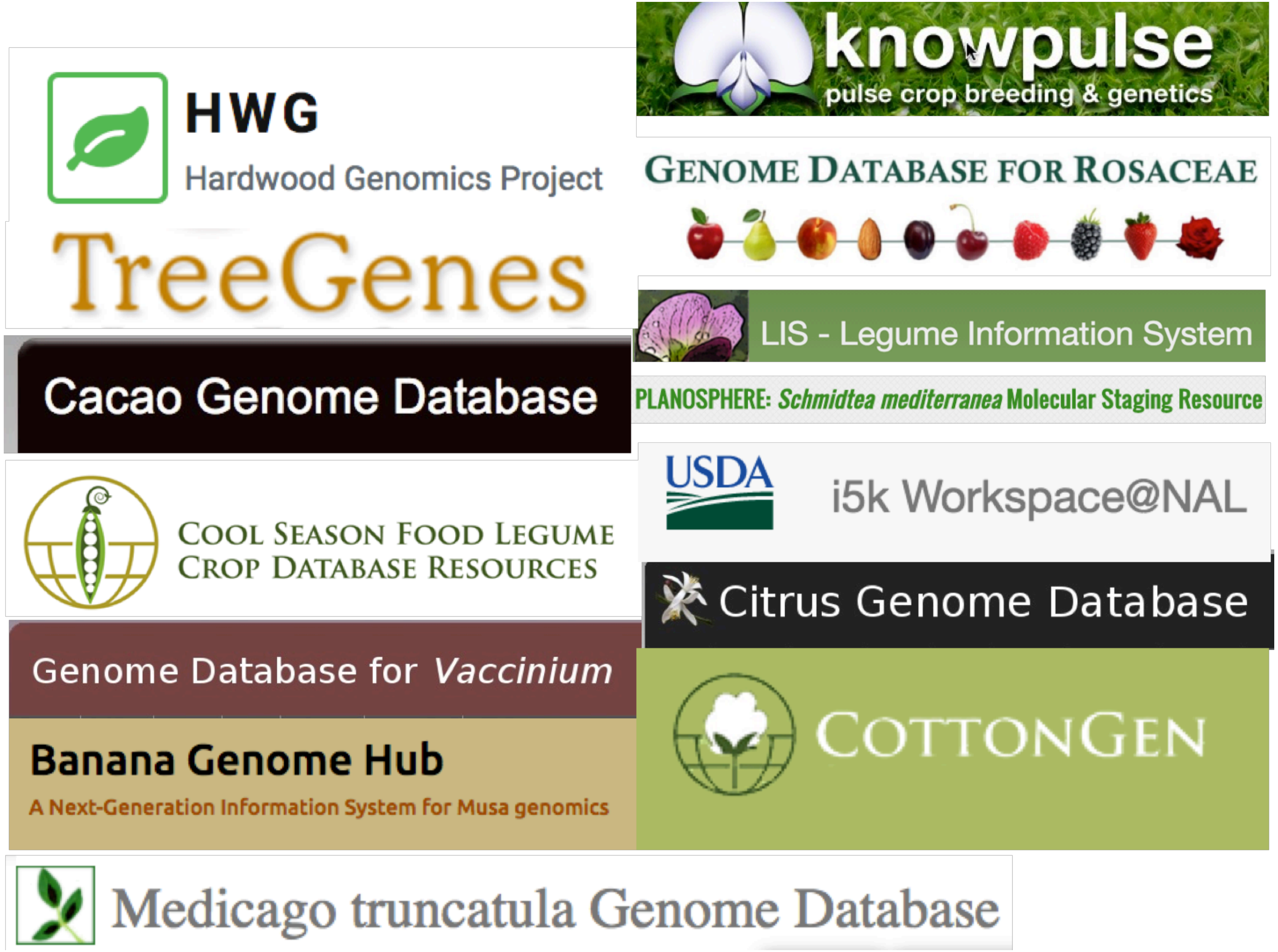
Sequence Features

Genotypes

Genetic Maps

Tripal

Chado



HWG

Hardwood Genomics Project

knowpulse

pulse crop breeding & genetics

GENOME DATABASE FOR ROSACEAE

TreeGenes

LIS - Legume Information System

Cacao Genome Database

PLANOSPHERE: *Schmidtea mediterranea* Molecular Staging Resource

USDA

i5k Workspace@NAL

COOL SEASON FOOD LEGUME CROP DATABASE RESOURCES

Citrus Genome Database

Genome Database for *Vaccinium*

COTTONGEN

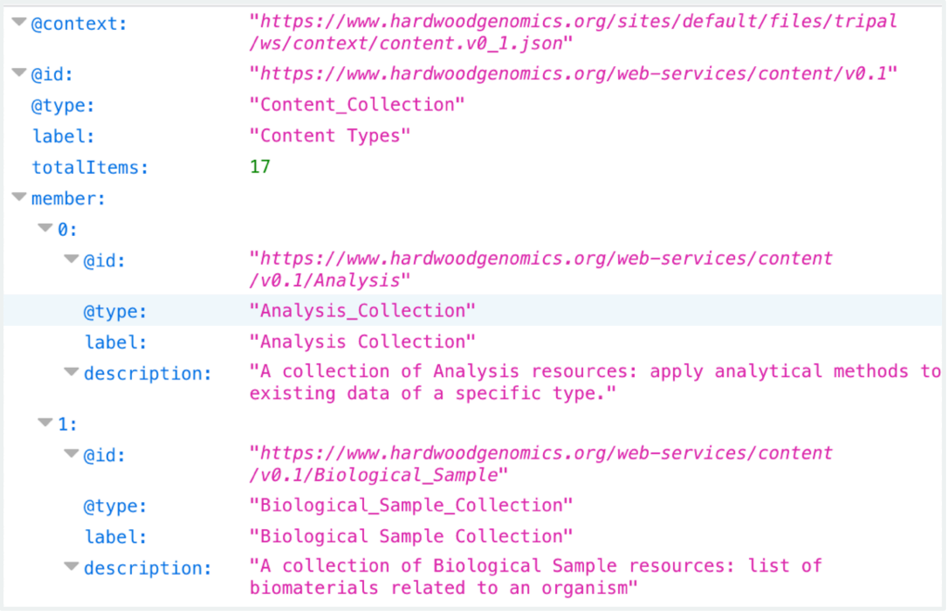
Banana Genome Hub

A Next-Generation Information System for Musa genomics

Medicago truncatula Genome Database

Tripal v3 Web Services

- RESTful
- Discoverable
- Searchable
- Use controlled vocabularies to ensure maximal interoperability.



```
▼@context:          "https://www.hardwoodgenomics.org/sites/default/files/tripal
                    /ws/context/content.v0_1.json"
▼@id:               "https://www.hardwoodgenomics.org/web-services/content/v0.1"
 @type:             "Content_Collection"
 label:             "Content Types"
 totalItems:        17
▼member:
  ▼0:
    ▼@id:           "https://www.hardwoodgenomics.org/web-services/content
                    /v0.1/Analysis"
     @type:         "Analysis_Collection"
     label:         "Analysis Collection"
    ▼description:   "A collection of Analysis resources: apply analytical methods to
                    existing data of a specific type."
  ▼1:
    ▼@id:           "https://www.hardwoodgenomics.org/web-services/content
                    /v0.1/Biological_Sample"
     @type:         "Biological_Sample_Collection"
     label:         "Biological Sample Collection"
    ▼description:   "A collection of Biological Sample resources: list of
                    biomaterials related to an organism"
```

What Web Services Is and Is Not

Difficult to implement for non-Tripal databases- different architectures and underlying storage = lots of coding!

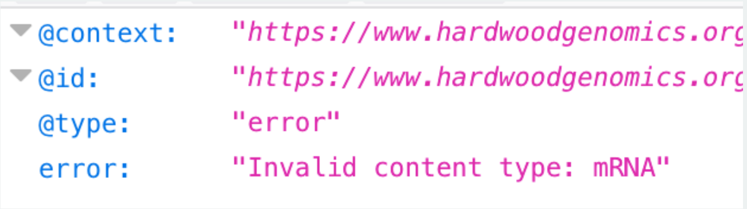
Slow searching

Great for computers and developers, but less useful for users directly (must know structure!)

To exchange data among sites, site developers must be able to predict what users want to find and integrate

Follow the manual: Filter all mRNA to include only those from the genus Acer construct the following URL:

`https://www.hardwoodgenomics.org/web-services/content/v0.1/mRNA?organism,genus=Acer`



```
▼@context:    "https://www.hardwoodgenomics.org
▼@id:         "https://www.hardwoodgenomics.org
 @type:       "error"
 error:       "Invalid content type: mRNA"
```

Search

Elasticsearch is an open source search engine

- Fast searching and handling of large volumes of data
- Largely scalable
- Sorts by relevance to search terms
- Extensively documented and tested

Learn more at www.elastic.co

elasticsearch

Lucene

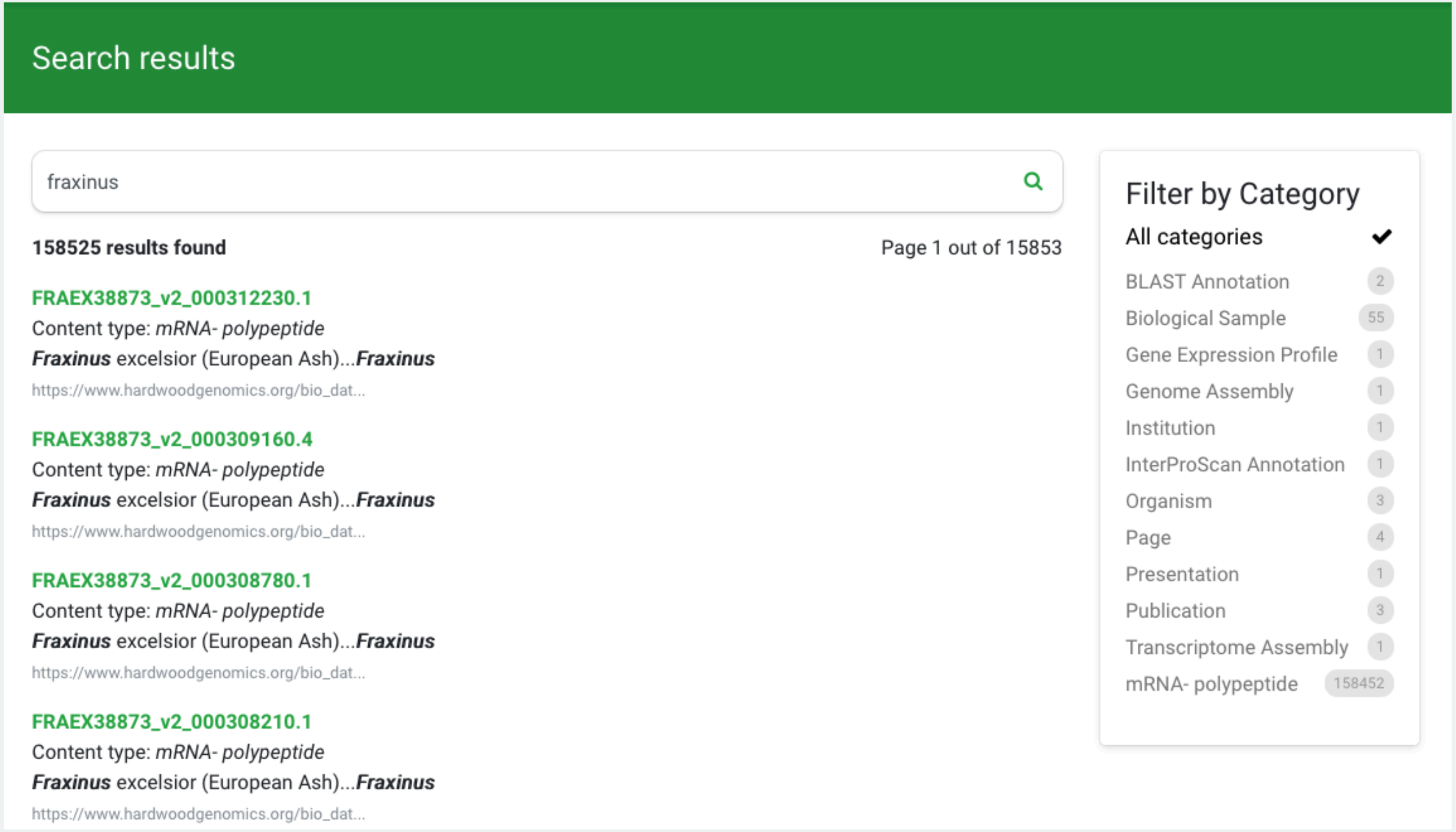
JSON

Tripal Elasticsearch

- A Tripal extension that provides a user-friendly interface to index large genomic data
- Provides default indices that work “out of the box”
- Highly customizable
    - Allows administrators to create custom indices and search forms

Basic Local Search

Website Search of all content



Search results

fraxinus

158525 results found

Page 1 out of 15853

FRAEX38873_v2_000312230.1
Content type: *mRNA- polypeptide*
Fraxinus excelsior (European Ash)...***Fraxinus***
https://www.hardwoodgenomics.org/bio_dat...

FRAEX38873_v2_000309160.4
Content type: *mRNA- polypeptide*
Fraxinus excelsior (European Ash)...***Fraxinus***
https://www.hardwoodgenomics.org/bio_dat...

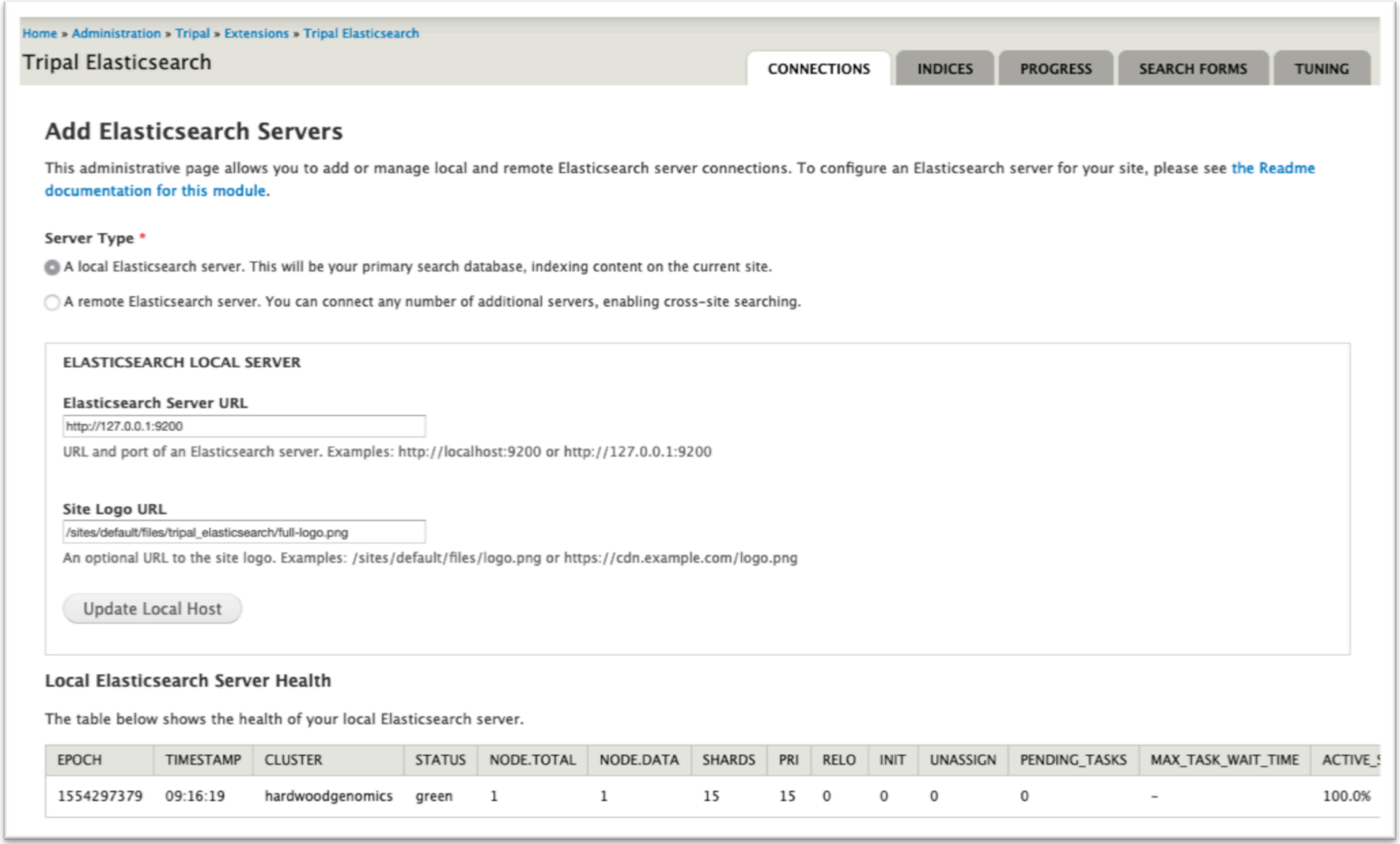
FRAEX38873_v2_000308780.1
Content type: *mRNA- polypeptide*
Fraxinus excelsior (European Ash)...***Fraxinus***
https://www.hardwoodgenomics.org/bio_dat...

FRAEX38873_v2_000308210.1
Content type: *mRNA- polypeptide*
Fraxinus excelsior (European Ash)...***Fraxinus***
https://www.hardwoodgenomics.org/bio_dat...

Filter by Category

All categories	✓
BLAST Annotation	2
Biological Sample	55
Gene Expression Profile	1
Genome Assembly	1
Institution	1
InterProScan Annotation	1
Organism	3
Page	4
Presentation	1
Publication	3
Transcriptome Assembly	1
mRNA- polypeptide	158452

Administrative Interface


Home » Administration » Tripal » Extensions » Tripal Elasticsearch
Tripal Elasticsearch
CONNECTIONS
INDICES
PROGRESS
SEARCH FORMS
TUNING

Add Elasticsearch Servers

This administrative page allows you to add or manage local and remote Elasticsearch server connections. To configure an Elasticsearch server for your site, please see the Readme documentation for this module.

Server Type *

● A local Elasticsearch server. This will be your primary search database, indexing content on the current site.

○ A remote Elasticsearch server. You can connect any number of additional servers, enabling cross-site searching.

ELASTICSEARCH LOCAL SERVER

Elasticsearch Server URL

http://127.0.0.1:9200

URL and port of an Elasticsearch server. Examples: http://localhost:9200 or http://127.0.0.1:9200

Site Logo URL

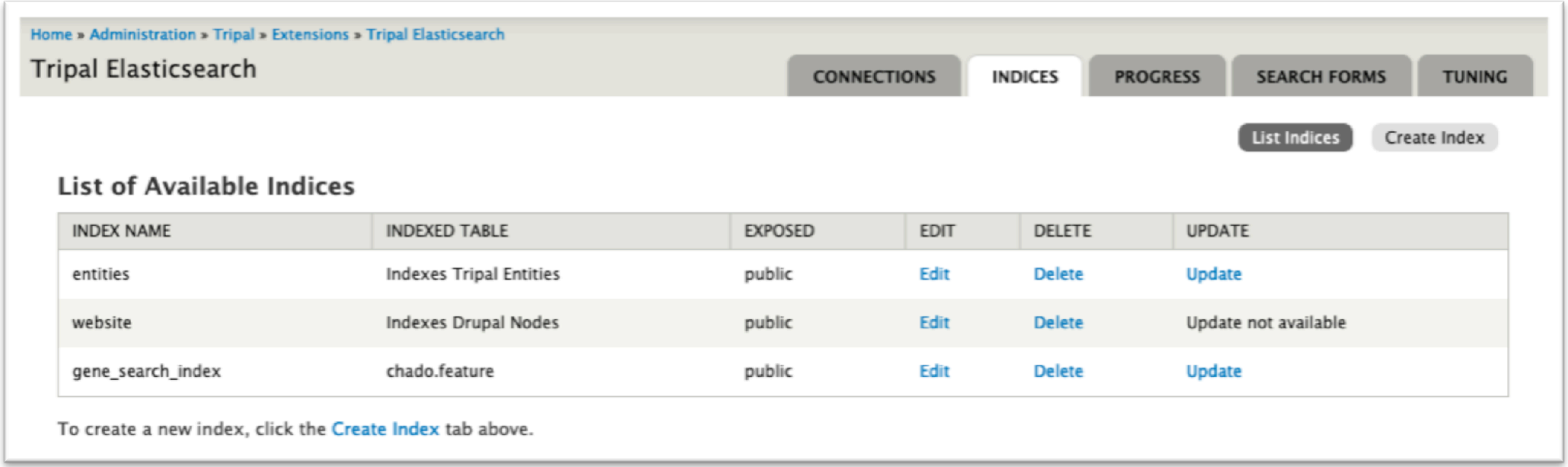
/sites/default/files/tripal_elasticsearch/full-logo.png

An optional URL to the site logo. Examples: /sites/default/files/logo.png or https://cdn.example.com/logo.png

Local Elasticsearch Server Health

The table below shows the health of your local Elasticsearch server.

EPOCH	TIMESTAMP	CLUSTER	STATUS	NODE.TOTAL	NODE.DATA	SHARDS	PRI	RELO	INIT	UNASSIGN	PENDING_TASKS	MAX_TASK_WAIT_TIME	ACTIVE_S
1554297379	09:16:19	hardwoodgenomics	green	1	1	15	15	0	0	0	0	-	100.0%

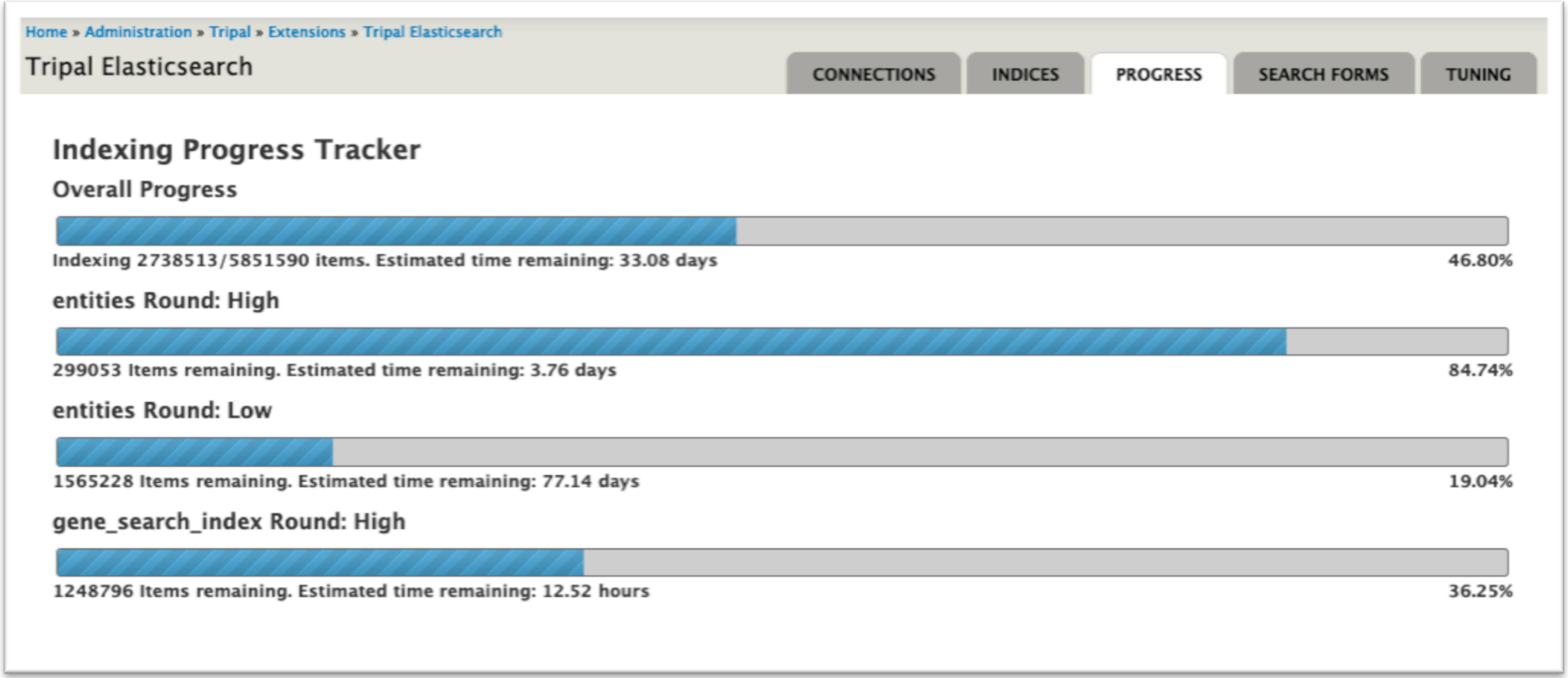


Tripal Elasticsearch

List of Available Indices

INDEX NAME	INDEXED TABLE	EXPOSED	EDIT	DELETE	UPDATE
entities	Indexes Tripal Entities	public	Edit	Delete	Update
website	Indexes Drupal Nodes	public	Edit	Delete	Update not available
gene_search_index	chado.feature	public	Edit	Delete	Update

To create a new index, click the Create Index tab above.


Home » Administration » Tripal » Extensions » Tripal Elasticsearch
Tripal Elasticsearch
CONNECTIONS
INDICES
PROGRESS
SEARCH FORMS
TUNING

Indexing Progress Tracker

Overall Progress

Indexing 2738513/5851590 items. Estimated time remaining: 33.08 days 46.80%

entities Round: High

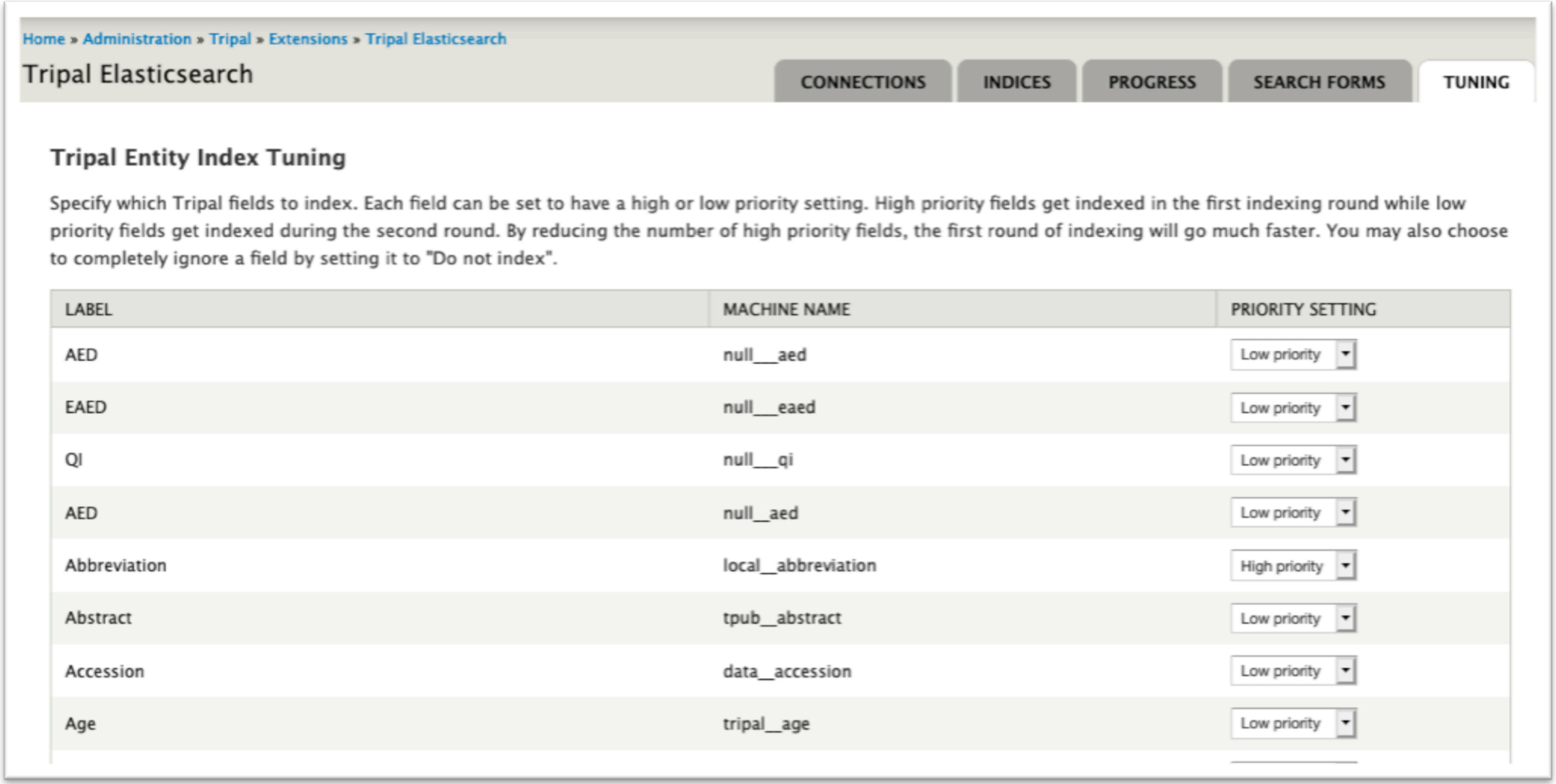
299053 Items remaining. Estimated time remaining: 3.76 days 84.74%

entities Round: Low

1565228 Items remaining. Estimated time remaining: 77.14 days 19.04%

gene_search_index Round: High

1248796 Items remaining. Estimated time remaining: 12.52 hours 36.25%



Home » Administration » Tripal » Extensions » Tripal Elasticsearch
CONNECTIONS
INDICES
PROGRESS
SEARCH FORMS
TUNING

Tripal Elasticsearch

Tripal Entity Index Tuning

Specify which Tripal fields to index. Each field can be set to have a high or low priority setting. High priority fields get indexed in the first indexing round while low priority fields get indexed during the second round. By reducing the number of high priority fields, the first round of indexing will go much faster. You may also choose to completely ignore a field by setting it to "Do not index".

LABEL	MACHINE NAME	PRIORITY SETTING
AED	null___aed	Low priority
EAED	null___eaed	Low priority
QI	null___qi	Low priority
AED	null__aed	Low priority
Abbreviation	local__abbreviation	High priority
Abstract	tpub__abstract	Low priority
Accession	data__accession	Low priority
Age	tripal__age	Low priority

Search as a Service

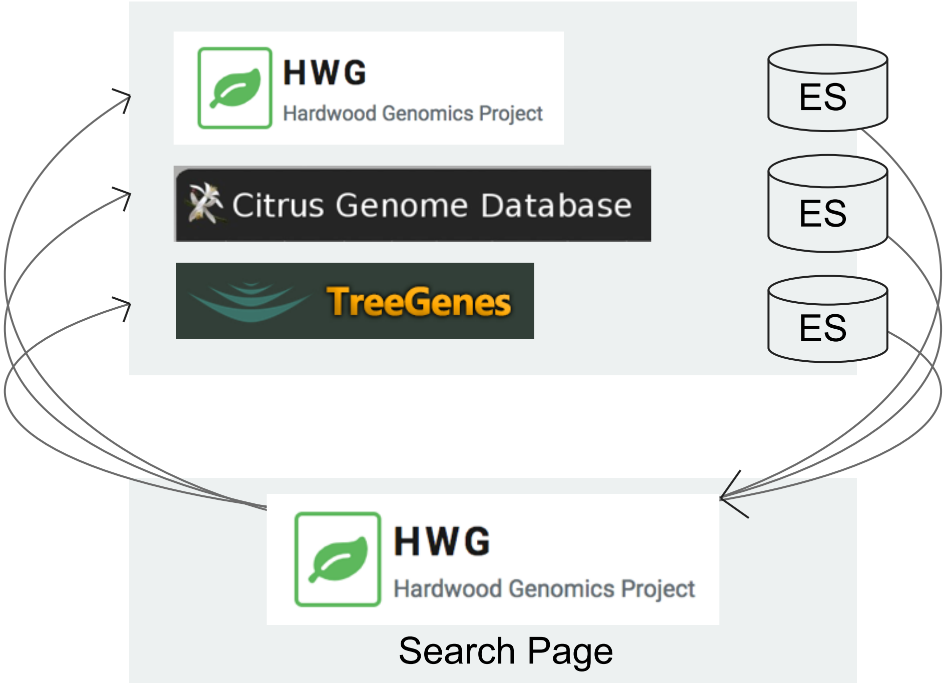
ElasticSearch can expose a searchable index online

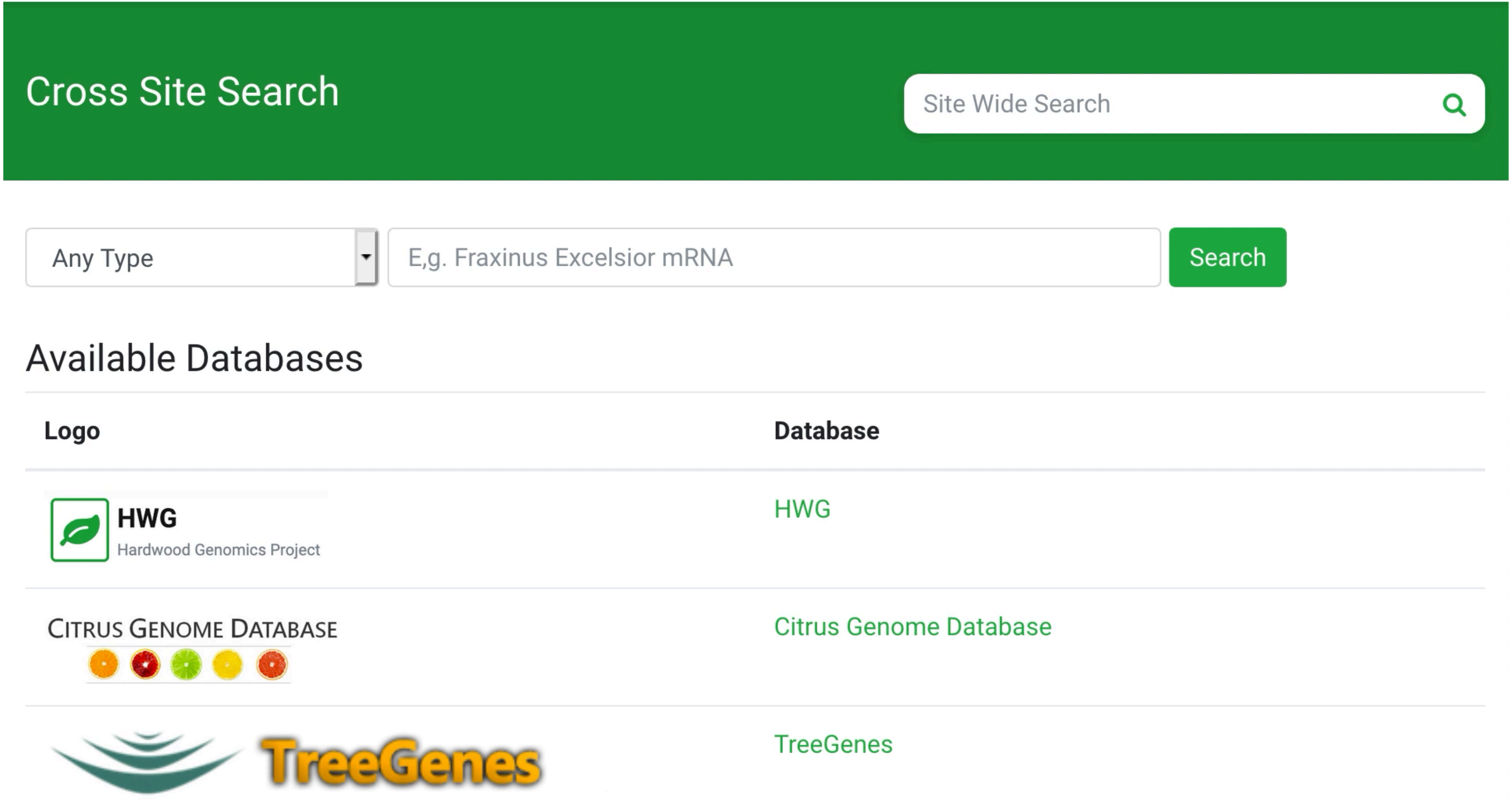
The ElasticSearch engine can use these public indices to find and aggregate data across sites

Search as a service

And search as a form of data federation!

“Cross site search”





Cross Site Search

Site Wide Search

Any Type

E,g. Fraxinus Excelsior mRNA

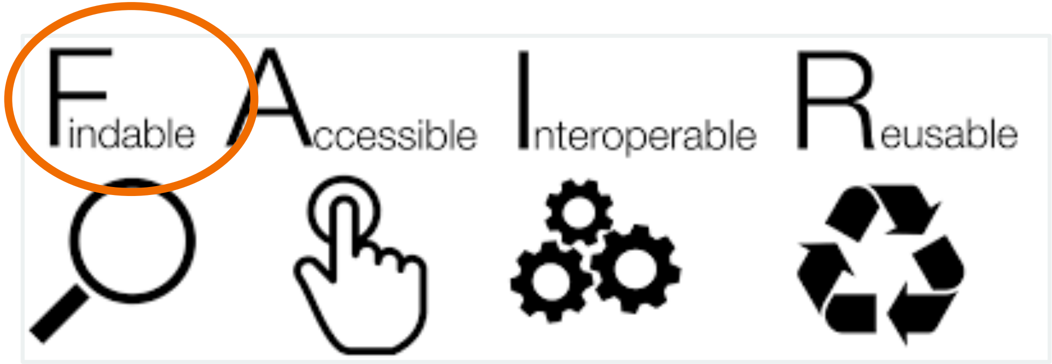
Search

Available Databases

Logo	Database
HWG Hardwood Genomics Project	HWG
Citrus Genome Database	Citrus Genome Database
TreeGenes	TreeGenes

Search is a complementary tool for data federation and exchange

- Directly benefits users
- Not just for Tripal!
- Relatively quick to implement across any online website or storage backend
  - Not limited to relational databases!



Structuring Data

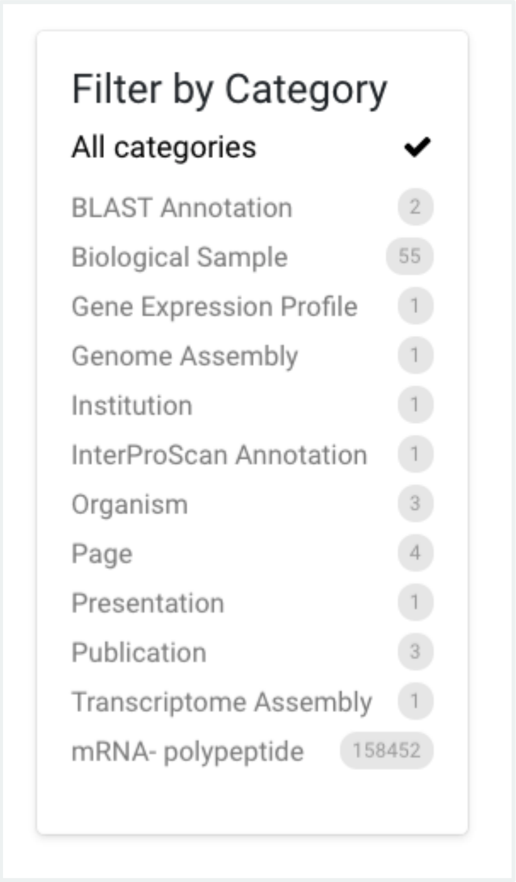
Structure makes data better!

Tripal Elasticsearch stores tokenized information free of HTML clutter

This enables faceted searching and filtering of search results

Currently only available for internal search

Working on implementing for cross site search



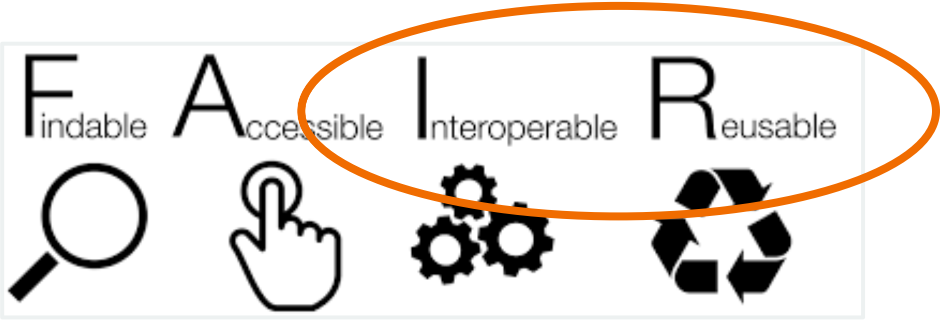
Filter by Category

All categories ✓

BLAST Annotation	2
Biological Sample	55
Gene Expression Profile	1
Genome Assembly	1
Institution	1
InterProScan Annotation	1
Organism	3
Page	4
Presentation	1
Publication	3
Transcriptome Assembly	1
mRNA- polypeptide	158452

More work still to be done

- How to add structure across other types of data storage?
  - Web services?
  - JSON/Schema.org?
- Offer access to structured and unstructured data

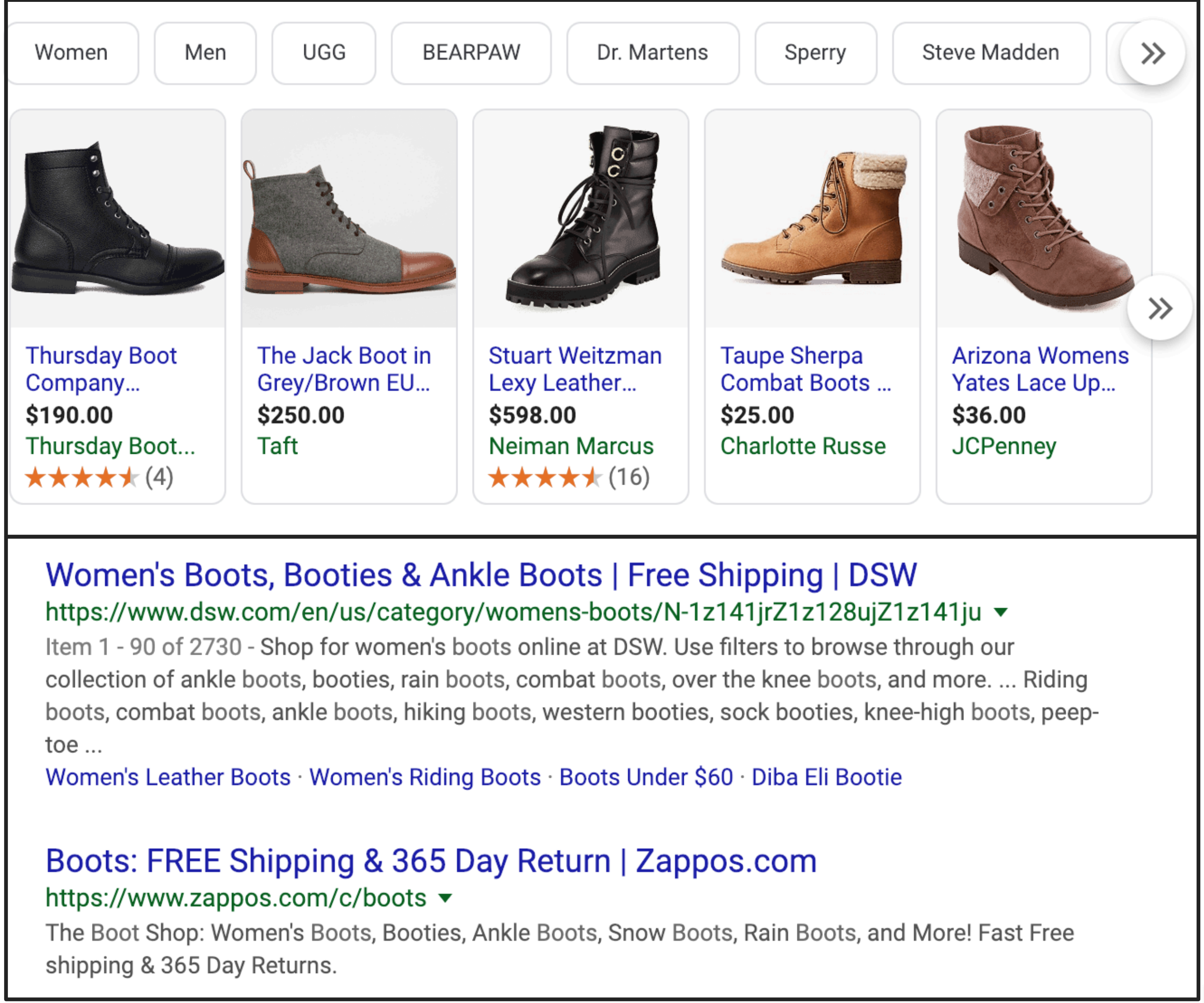


Google

+

schema.org

Example



Thursday Boot Company...
$190.00
Thursday Boot...
★★★★★ (4)

The Jack Boot in Grey/Brown EU...
$250.00
Taft

Stuart Weitzman Lexy Leather...
$598.00
Neiman Marcus
★★★★★ (16)

Taupe Sherpa Combat Boots ...
$25.00
Charlotte Russe

Arizona Womens Yates Lace Up...
$36.00
JCPenney

Women's Boots, Booties & Ankle Boots | Free Shipping | DSW
https://www.dsw.com/en/us/category/womens-boots/N-1z141jrZ1z128ujZ1z141ju ▾
Item 1 - 90 of 2730 - Shop for women's boots online at DSW. Use filters to browse through our collection of ankle boots, booties, rain boots, combat boots, over the knee boots, and more. ... Riding boots, combat boots, ankle boots, hiking boots, western booties, sock booties, knee-high boots, peep-toe ...
Women's Leather Boots · Women's Riding Boots · Boots Under $60 · Diba Eli Bootie

Boots: FREE Shipping & 365 Day Return | Zappos.com
https://www.zappos.com/c/boots ▾
The Boot Shop: Women's Boots, Booties, Ankle Boots, Snow Boots, Rain Boots, and More! Fast Free shipping & 365 Day Returns.

Structured results

- Can be filtered

- Can be sent to other services

Unstructured results

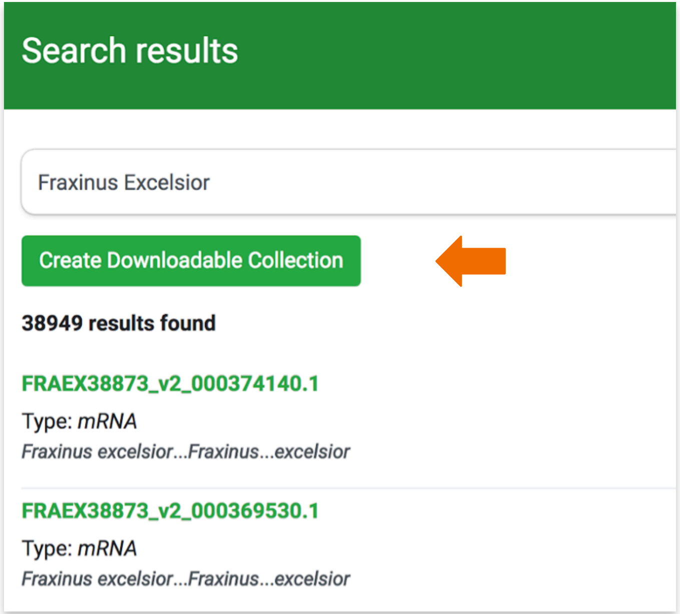
- Can still be found and explored by a user

Tripal Features for even more Interoperability and Reusability

Data from searches can be placed into collections by users

Collections can be downloaded as proper format types (fasta for sequences, vcf for variants)

Collections can be sent to a Galaxy workflow for analysis



Search results

Fraxinus Excelsior

38949 results found

FRAEX38873_v2_000374140.1

Type: *mRNA*

Fraxinus excelsior...Fraxinus...excelsior

FRAEX38873_v2_000369530.1

Type: *mRNA*

Fraxinus excelsior...Fraxinus...excelsior

AgBioData

Data Sharing using Web Services Working Group

- Identify the current methods of data exchange within and across AgBioData databases
- Explore community opinions on data sharing needs and priorities
- Identify a set of partners with interest and throughput to actually implement some concrete examples
- Develop a set of recommended best practices for data exchange
- Promote best practices for data exchange

PAG in person meeting

We have lots of methods of sharing data but few are commonly used across many resources

- BrAPI
- Search engines – Solr, ElasticSearch
- FTP
- Bioschema (needs additional structure!)
- Custom built APIs

PAG in person meeting

We have lots of needs and priorities!

- Increase discoverability/findability of services
- Connecting among different data types
- People structure and store the same types of data in different ways (lack of standards and/or many standards)
- Standards are difficult to validate - gff, chado, vcf - groups use them differently
- Phenotypes – lack of structure
- Pangenome support - moving between assemblies, gene ids, locations, etc
- Enrich Europe/US/Other collaboration and crosstalk
- JSON-LD may be a convergence point

This list was produced by 8 people.

We need a survey!

PAG in person meeting

Proposed Action Plan

- Survey!
- Develop a set of recommended best practices for data exchange
- Try to incorporate as many people in the conversation as possible
- Encourage use of the recommended best practices by developing demonstrations and proof of concept data sharing examples
- Identify a set of partners with interest and throughput to actually implement some concrete examples (concrete work in addition to discussions)

Summary

Join the Data Sharing group…. We communicate well!

We need partners to help figure out data exchange standards and implementations.

Its ok to be in more than one group!

https://www.agbiodata.org/

Acknowledgements

- National Science Foundation
- University of Tennessee
  - Margaret Staton, co-PI
  - Abdullah Almsaeed
  - Bradford Condon
  - Ming Chen
- Washington State University
  - Stephen Ficklin, PI
  - Dorrie Main, co-PI
- University of Connecticut
  - Jill Wegrzyn, co-PI

CITRUS GENOME DATABASE

Resources for citrus genomics, genetics, breeding and disease research

AgBioData
PAG in person working group

- Jill Wegrzyn
- Emily Grau
- Andrew Farmer
- Marcela Karey Tello-Ruiz
- Justin Elser
- Sylva Donaldson
- Cyril Pommier