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
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
Tripal v3 Web Services

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

```json
"@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
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

Search results

158525 results found

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...
Administrative Interface
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

Update Local Host

Local Elasticsearch Server Health

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

<table>
<thead>
<tr>
<th>EPOCH</th>
<th>TIMESTAMP</th>
<th>CLUSTER</th>
<th>STATUS</th>
<th>NODE TOTAL</th>
<th>NODE DATA</th>
<th>SHARDS</th>
<th>PRI</th>
<th>RELO</th>
<th>INIT</th>
<th>UNASSIGN</th>
<th>PENDING TASKS</th>
<th>MAX_TASK_WAIT_TIME</th>
<th>ACTIVE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1554297379</td>
<td>09:16:19</td>
<td>hardwoodgenomics</td>
<td>green</td>
<td>1</td>
<td>1</td>
<td>15</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>100.0%</td>
</tr>
</tbody>
</table>
## List of Available Indices

<table>
<thead>
<tr>
<th>INDEX NAME</th>
<th>INDEXED TABLE</th>
<th>EXPOSED</th>
<th>EDIT</th>
<th>DELETE</th>
<th>UPDATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>entities</td>
<td>Indexes Tripal Entities</td>
<td>public</td>
<td>Edit</td>
<td>Delete</td>
<td>Update</td>
</tr>
<tr>
<td>website</td>
<td>Indexes Drupal Nodes</td>
<td>public</td>
<td>Edit</td>
<td>Delete</td>
<td>Update not available</td>
</tr>
<tr>
<td>gene_search_index</td>
<td>chado.feature</td>
<td>public</td>
<td>Edit</td>
<td>Delete</td>
<td>Update</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>LABEL</th>
<th>MACHINE NAME</th>
<th>PRIORITY SETTING</th>
</tr>
</thead>
<tbody>
<tr>
<td>AED</td>
<td>null__aed</td>
<td>Low priority</td>
</tr>
<tr>
<td>EAED</td>
<td>null__eaed</td>
<td>Low priority</td>
</tr>
<tr>
<td>QI</td>
<td>null__qi</td>
<td>Low priority</td>
</tr>
<tr>
<td>AED</td>
<td>null__aed</td>
<td>Low priority</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>local__abbreviation</td>
<td>High priority</td>
</tr>
<tr>
<td>Abstract</td>
<td>tpub__abstract</td>
<td>Low priority</td>
</tr>
<tr>
<td>Accession</td>
<td>data_accession</td>
<td>Low priority</td>
</tr>
<tr>
<td>Age</td>
<td>tripal__age</td>
<td>Low priority</td>
</tr>
</tbody>
</table>
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

Search is performed asynchronously using ajax calls to each server.

### Available Databases

<table>
<thead>
<tr>
<th>Logo</th>
<th>Database</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="HWG" /></td>
<td>HWG</td>
</tr>
<tr>
<td><img src="image" alt="Citrus Genome Database" /></td>
<td>Citrus Genome Database</td>
</tr>
<tr>
<td><img src="image" alt="TreeGenes" /></td>
<td>TreeGenes</td>
</tr>
</tbody>
</table>
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
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
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.
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
- Pangeneome 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

AgBioData
PAG in person working group
- Jill Wegrzyn
- Emily Grau
- Andrew Farmer
- Marcela Karey Tello-Ruiz
- Justin Elser
- Sylva Donaldson
- Cyril Pommier