

# FAIR from the Start: Empowering Data Sharing in Agricultural Research

FAIR Scientific Literature  
Working Group Update

AgBioData Workshop

PAG33

1/11/2026

Leonore Reiser



# Members



Leyla Cabugos  
Librarian, Cal Poly



Jacqueline Campbell  
Geneticist, USDAARS  
SoyBase curator



Jenna Daenzer  
Peer review Coordinator  
Genetics Society of  
America



Sook Jung  
Asst Research  
Professor. GDR



David Molik  
Center for Scholarly  
Publishing, Kansas  
State University  
Department head



Daniela Raciti  
Exec Editor,  
microPublication &  
Wormbase Curator



Leonore Reiser  
TAIR curator



Adam Wright  
Software Engineer  
WormBase, Reactome



Karen Yook  
Exec Editor,  
microPublication &  
Wormbase Curator



We gratefully acknowledge support from NSF, grant #2126334



# FSL Working Group Goals



Identify bottlenecks in the publication-curation pipeline.

- Identify sets of existing or desired tools or biocuration resources to increase literature curation throughput and accuracy.
- Publish recommendations and a roadmap for authors and publishers to increase the FAIRness of research.



# A better literature curation workflow

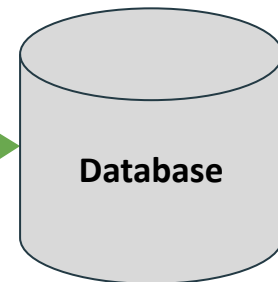
## AUTHOR

Data generator



## Database CURATOR

- ✓ Data curation at time of publication
- ✓ Clear guidance on data submission & databases
- ✓ Validators to check for data submission
- ✓ Validators for data formats
- ✓ Nomenclature standards and validators
- ✓ No more data available upon request
- ✓ Better paper metadata, AI based paper classification
- ✓ More funding for curation, AI curation assistants



# FSL Working Group Goals

- Identify bottlenecks in the publication-curation pipeline.




Identify sets of existing or desired tools or biocuration resources to increase literature curation throughput and accuracy.

- Publish recommendations and a roadmap for authors and publishers to increase the FAIRness of research.



# Tool: AgBio Database Finder Drupal Module



**AgBioData**  
Toward enhanced genomics, genetics, and breeding research outcomes through standardization of practices and protocols across agricultural databases

Home About Us Engage **Database Finder** Databases Meetings Projects Working Groups Login

## AgBioDatabase Finder Tool

Find the right AgBioData Member database for your agricultural genomic, genetic, or breeding data. Additional repository collections can be searched and found at **FAIRsharing**.

**How to Use This Tool**

**Select Your Organism:** Filter databases by choosing the crop or livestock species you're working with from the dropdown menu.

**Choose Your Purpose:**

**To Submit Data:** Select Option 2 and choose the type of data you want to deposit (e.g., genome sequences, QTL data, breeding records)

**To Find Data:** Select Option 3 and choose the type of data you're looking for

## Option 1: Filter by species

<https://www.agbiodata.org/databasefinder>  
<https://doi.org/10.17912/micropub.biology.001896>

Answer these questions to help identify databases that are suitable for your needs

Click one or more tiles to view/compare database details below.

Select All Clear Selections

Animal QTLdb  
Animal Quantitative Trait Loci Database

AquaMine  
Aquatic eumetazoan species genome data mine resource

BAR  
Bio-Analytic Resource for Plant Biology

BGD  
Bovine Genome Database

CGD  
Database for Citrus Genomics, Genetics and Breeding Research

CorrDB  
Animal Trait Correlation Database

CottonGen  
Cotton Database Resources

FAANGMine  
Integrates genome assemblies and currently available annotation data for FAANG species

GDR  
Database for Rosaceae Genomics, Genetics and Breeding Research

GDV  
Database for Vaccinium Genomics, Genetics and Breeding Research

GrainGenes  
A Database for Triticeae and Avena

Gramene  
Comparative plant genomics for crops and model organisms

HGD  
Hymenoptera Genome Database

LIS  
the Legume Information System

MaizeGDB  
Maize Genetics and Genomics Database

MaizeMine  
Integrated queryable database for Zea mays genetics and genomics

PCD  
Database for Pulse Genomics, Genetics and Breeding Research

Plant Reactome  
Gramine pathways

Planteome  
Ontologies and annotations of plant phenotypes and traits

SGN  
Solgenomics

SorghumBase  
Genomic resources for the sorghum research community

SoyBase  
Integrating Genetics and Genomics to Advance Soybean Research

TAIR  
The Arabidopsis Information Resource

TreeGenes  
Forest tree genome database

**1. organism**

- ☐ Arabidopsis
- ☐ Bacteria
- ☐ Bovinae
- ☐ Citrus
- ☐ Cotton
- ☐ Fish
- ☐ Forest trees
- ☐ Grasses
- ☐ Insects
- ☒ Legumes (chickpea, soy and lentil)
- ☐ Generalist Animal DB
- ☐ Generalist Plant DB
- ☐ Maize
- ☐ Poaceae (e.g., wheat, rice, barley, etc.)
- ☐ Rosaceae (e.g., apple, pear, almond, etc.)
- ☐ Solanaceae (e.g., tomato, eggplant, potato, etc.)
- ☐ Sorghum
- ☐ Soybean
- ☐ Vaccinium (e.g., blueberry, cranberry, etc.)

<https://github.com/CU-CommunityApps/CD-finder>

# Option 2: Filter by ...

Data to submit

OR

Data to find

### 2. Data to submit

- ☐ database cross-mapping
- ☐ geographic location
- ☐ gene expression
- ☐ gene functional annotation
- ☐ gene regulation
- ☐ gene report
- ☒ genetic map
- ☐ genetic variation
- ☐ genome analysis

### 3. Data to find

- ☐ database cross-mapping
- ☐ gene expression
- ☐ gene functional annotation
- ☐ gene regulation
- ☐ gene report
- ☐ genetic map
- ☐ genetic variation
- ☐ genome analysis
- ☐ genome annotation
- ☐ genomics
- ☐ genotype data
- ☐ geographic location
- ☐ homology-based gene prediction
- ☐ image

Click one or more tiles to view/compare database details below.

Select All Clear Selections

<b>Animal QTLdb</b> Animal Quantitative Trait Loci Database	<b>AquaMine</b> Aquatic eumetazoan species genome data mine resource	<b>BAR</b> Bio-Analytic Resource for Plant Biology	<b>BGD</b> Bovine Genome Database	<b>CGD</b> Database for Citrus Genomics, Genetics and Breeding Research
<b>CorrDB</b> Animal Trait Correlation Database	<b>CottonGen</b> Cotton Database Resources	<b>FAANGMine</b> Integrates genome assemblies and currently available annotation data for FAANG species	<b>GDR</b> Database for Rosaceae Genomics, Genetics and Breeding Research	<b>GDV</b> Database for Vaccinium Genomics, Genetics and Breeding Research
<b>GrainGenes</b> A Database for Triticeae and Avena	<b>Gramene</b> Comparative plant genomics for crops and model organisms	<b>HGD</b> Hymenoptera Genome Database	<b>LIS</b> the Legume Information System	<b>MaizeGDB</b> Maize Genetics and Genomics Database
<b>MaizeMine</b> Integrated queryable database for Zea mays genetics and genomics	<b>PCD</b> Database for Pulse Genomics, Genetics and Breeding Research	<b>Plant Reactome</b> Gramene pathways	<b>Planteome</b> Ontologies and annotations of plant phenotypes and traits	<b>SGN</b> Solgenomics
<b>SorghumBase</b> Genomic resources for the sorghum research community	<b>SoyBase</b> Integrating Genetics and Genomics to Advance Soybean Research	<b>TAIR</b> The Arabidopsis Information Resource	<b>TreeGenes</b> Forest tree genome database	<b>i5k</b> i5k Workspace@NAL

# Apply filters and compare resources

## 2. Data to submit

- ☐ database cross-mapping
- ☐ geographic location
- ☐ gene expression
- ☐ gene functional annotation
- ☐ gene regulation
- ☐ gene report
- ☒ genetic map
- ☐ genetic variation

**CottonGen** ○

Cotton Database Resources

**FAANGMine** ○

Integrates genome assemblies and currently available annotation data for FAANG species

**GDR** ○

Database for Rosaceae Genomics, Genetics and Breeding Research

**Gramene** ○

Comparative plant genomics for crops and model organisms

**HGD** ○

Hymenoptera Genome Database

**LIS** ✔

the Legume Information System

**PCD** ✔

Database for Pulse Genomics, Genetics and Breeding Research

**Plant Reactome** ○

Gramene pathways

**Planteome** ○

Ontologies and annotations of plant phenotypes and traits

**SoyBase** ✔

Integrating Genetics and Genomics to Advance Soybean Research

**TAIR** ○

The Arabidopsis Information Resource

**TreeGenes** ○

Forest tree genome database

	LIS <span>○</span>	PCD <span>○</span>	SoyBase <span>○</span>
DB name	LIS - the Legume Information System	Pulse Crop Database (PCD)	SoyBase
DB short description	Legumeinfo is a federally funded informatics service assisting in basic research in the legumes by relating data from multiple species, and traverse among various data types.	The Pulse Crop Database (PCD), formerly the Cool Season Food Legume Database (CSFL), has been developed by the Main Bioinformatics Laboratory at Washington State University in collaboration with the USDA-ARS Grain Legume Genetics and Physiology Research Unit, the USDA-ARS Plant Germplasm Introduction and Testing Unit, the USA Dry Pea and Lentil Council, Northern Pulse Growers and allied scientists in the US and across the world, to serve as a resource for Genomics-Assisted Breeding (GAB). GAB offers tools to identify genes related to traits of interest, among other methods to optimize plant breeding efficiency and research, by providing relevant genomic, genetic, and breeding information and analysis. Therefore, tools such as JBrowse and MapViewer can be found in this database, as well as key resources to provide access to the annotation of available transcriptome data, helping pulse breeders and researchers to succeed in their programs.	SoyBase is a long-term, federally funded genetics and breeding database for soybean breeders and researchers focused on soybeans.
DB URL	<a href="https://www.legumeinfo.org/">https://www.legumeinfo.org/</a>	<a href="https://www.pulsedb.org/">https://www.pulsedb.org/</a>	<a href="https://www.soybase.org/">https://www.soybase.org/</a>
DB data types	<ul style="list-style-type: none"> <li>• database cross-mapping*</li> <li>• gene expression</li> <li>• gene functional annotation*</li> <li>• gene regulation</li> <li>• gene report</li> <li>• genetic map*</li> <li>• genetic variation</li> <li>• genome analysis</li> <li>• genome annotation*</li> <li>• genomics</li> <li>• genotype data*</li> <li>• geographic location</li> <li>• homology-based gene prediction*</li> <li>• metadata</li> <li>• molecular sequence annotation*</li> <li>• phylogenetic data*</li> <li>• QTL map*</li> <li>• sequence features*</li> <li>• whole genome association study*</li> </ul>	<ul style="list-style-type: none"> <li>• database cross-mapping</li> <li>• gene expression*</li> <li>• gene functional annotation*</li> <li>• gene regulation*</li> <li>• gene report*</li> <li>• genetic map*</li> <li>• genetic variation*</li> <li>• genome analysis</li> <li>• genome annotation*</li> <li>• genotype data*</li> <li>• homology-based gene prediction*</li> <li>• image*</li> <li>• metabolic pathway prediction</li> <li>• ontology and terminology</li> <li>• phenotype*</li> <li>• QTL*</li> <li>• whole genome association study*</li> </ul> <p>* = Direct submission accepted</p>	<ul style="list-style-type: none"> <li>• database cross-mapping*</li> <li>• gene expression</li> <li>• gene functional annotation*</li> <li>• genetic map*</li> <li>• gene regulation</li> <li>• gene report</li> <li>• genetic variation</li> <li>• genome analysis</li> <li>• genome annotation</li> <li>• genomics</li> <li>• genotype data</li> <li>• geographic location</li> <li>• homology-based gene prediction</li> <li>• metabolic pathway prediction</li> <li>• metadata</li> <li>• nucleic acid sites, features, and motifs</li> <li>• ontology and terminology</li> <li>• pedigree*</li> <li>• phenotype*</li> </ul>



# What about source/data type repositories?

- Add more tiles/databases to the tool?
- Create a collection in FAIRsharing?
- Maintain a separate list ordered by data type?
- *AI/Chatbot?*



# (How) can AI (tools) help make published data more FAIR?

- What is the future of publishing and co-ordinated data curation in the age of AI?
- Can we make the data and metadata in papers more machine readable/FAIR?
- A freely accessible, universal AI tool that:
  - Scans papers and identifies types of data
  - Extracts experiment metadata
  - Validates data and metadata formats
  - Checks against nomenclature and other standards
  - Readies data for submission to appropriate repository



# How to engage with publishers and journals?

How to  
introduce  
data curation  
into the  
publishing  
pipeline

How to connect  
journals to community  
databases



Community Workshop April 2025

## Incentives



Apicuron Credits for authors



FAIRness badges for papers/journals

## Workflow changes



Pre-submission or pre-print curatorial review



Advanced vetting to speed peer review

Start with society journals

Journals need assurance that recommended  
repos are FAIR



# Increasing FAIRness of AgBioData Member Databases

- FAIR sharing record completion
- Assess FAIRness of databases
  - What minimum expectation and standards for member databases
  - How can AgBioData support meeting those expectation?



# FSL Working Group Goals

- Identify bottlenecks in the publication-curation pipeline.
- Identify sets of existing or desired tools or biocuration resources to increase literature curation throughput and accuracy.
- Publish recommendations and a roadmap for authors and publishers to increase the FAIRness of research.



# Thoughts for Discussion

- What if the roadmap was an *interactive tool* not a static white paper?
- How could an integrated curation- publication data pipeline be made sustainable?

