FAIR Scientific Literature (FSL) Working Group Update

PAG 2025
1/10/2025
Leonore Reiser

Members


Leyla Cabugos
Librarian, Cal Poly


Jenna Daenzer
GSA


Leonore Reiser
TAIR curator


Sook Jung
Asst Research
Professor. GDR


David Molik
Computational
Biologist,
USDA ARS


Daniela Raciti
Exec Editor,
microPublication &
Wormbase Curator


Jacqueline Campbell
Geneticist, USDA ARS
SoyBase curator


Adam Wright
Software Engineer
Wormbase.Reactome

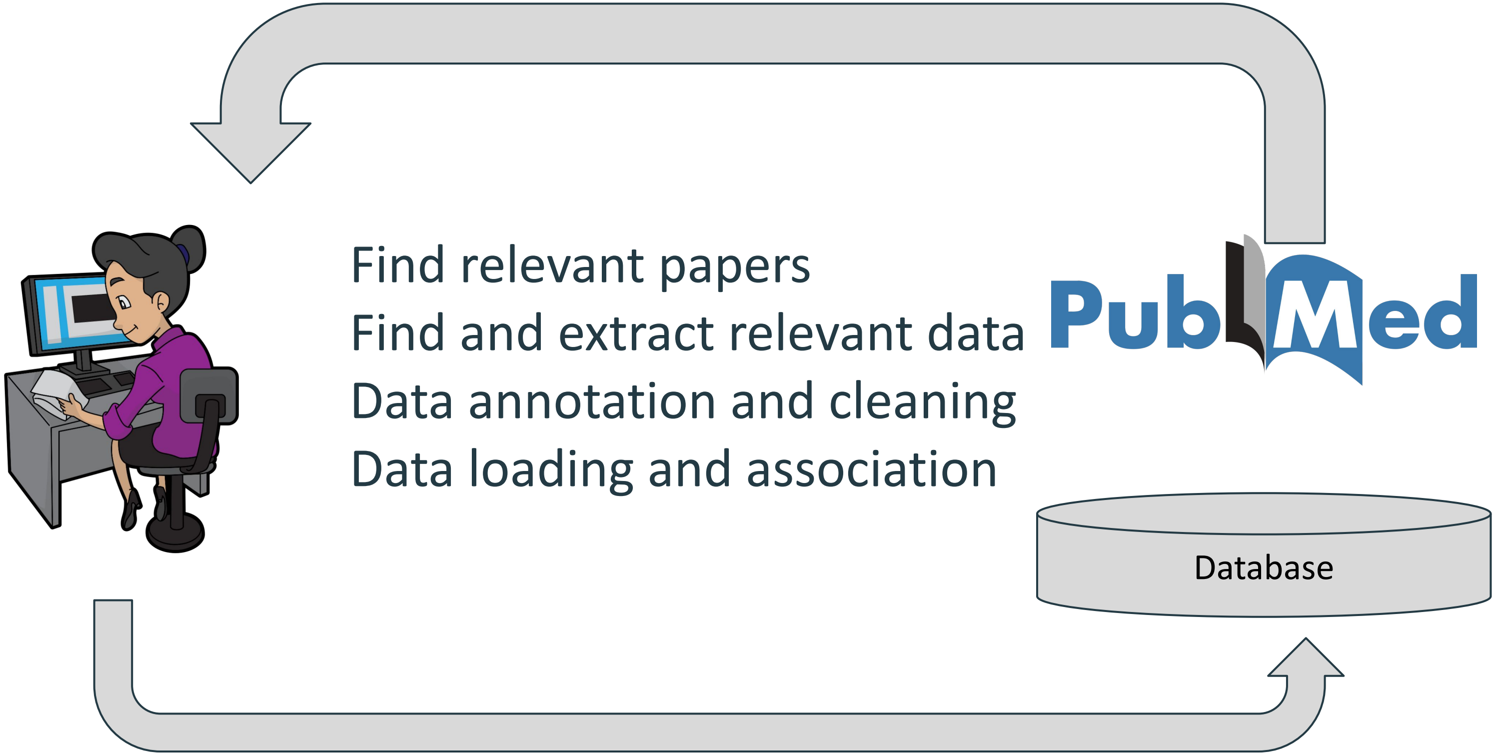

Karen Yook
Exec Editor,
microPublication &
Wormbase Curator

We gratefully acknowledge support from NSF, grant #2126334

FSL Working Group Goals

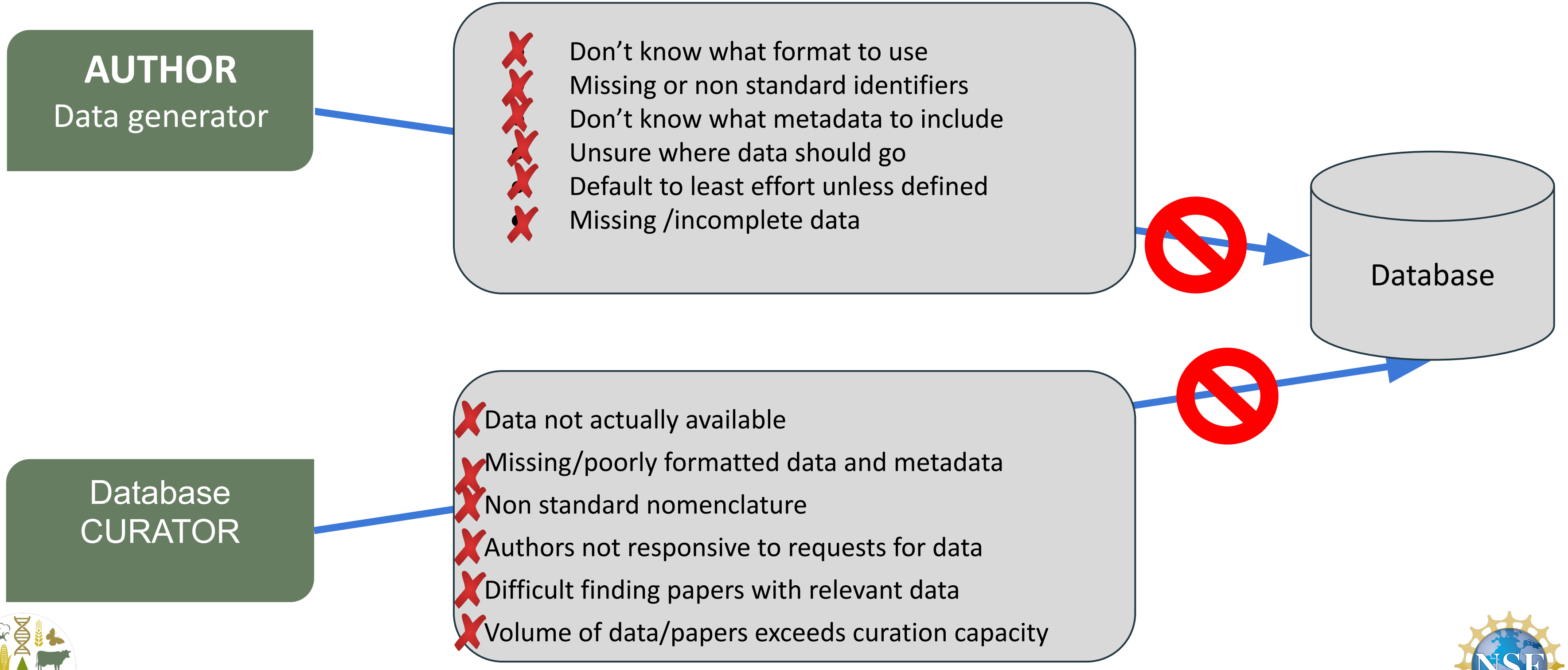
- Identify bottlenecks in the publication-curation pipeline.
- Identify sets of existing or desired tools or biocuration resources to increase literature curation throughput and accuracy.
- Publish recommendations and a roadmap for authors and publishers to increase the FAIRness of research.

What do we mean by publication-curation pipeline



Find relevant papers
Find and extract relevant data
Data annotation and cleaning
Data loading and association

Publication-curation workflow bottlenecks



AUTHOR
Data generator

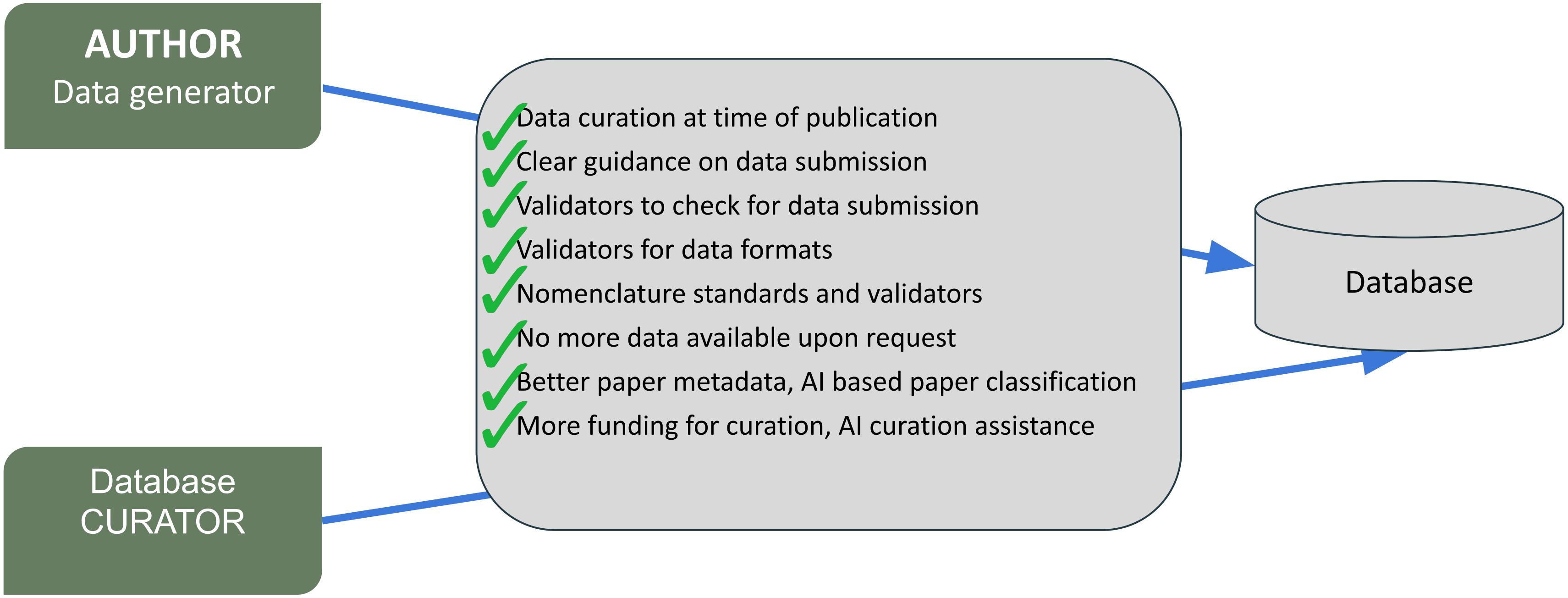
- ✗ Don’t know what format to use
- ✗ Missing or non standard identifiers
- ✗ Don’t know what metadata to include
- ✗ Unsure where data should go
- ✗ Default to least effort unless defined
- ✗ Missing /incomplete data

Database

Database
CURATOR

- ✗ Data not actually available
- ✗ Missing/poorly formatted data and metadata
- ✗ Non standard nomenclature
- ✗ Authors not responsive to requests for data
- ✗ Difficult finding papers with relevant data
- ✗ Volume of data/papers exceeds curation capacity

A better literature curation workflow



AUTHOR
Data generator

- ✓ Data curation at time of publication
- ✓ Clear guidance on data submission
- ✓ 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 assistance

Database
CURATOR

Database

Stakeholders: Challenges and barriers

Researcher

- **Not knowing where data should go**
- Time consuming submission process
- Not knowing how to format data / metadata

Publisher

- **Not knowing where data should go**
- Lack of easy verification of data availability
- Authors often do not want data available before publication

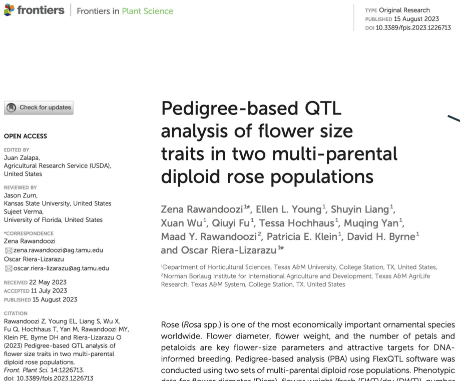
Research Librarian

- **Not knowing where data should go**
- Lack of engagement with researcher
- High volume of data and not enough time

Funder

- **Not knowing where data should go**
- Different programs have different repositories
- Proposal reviews don’t know how to evaluate DMPs

First Goal: Data to be FAIR from the start



frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 15 August 2023
DOI 10.3389/fpls.2023.1226713

Check for updates

OPEN ACCESS

EDITED BY
Juan Zalapa,
Agricultural Research Service (USDA),
United States

REVIEWED BY
Jason Zurn,
Kansas State University, United States
Sujeet Verma,
University of Florida, United States

*CORRESPONDENCE
Zena Rawandoozi
✉ zena.rawandoozi@ag.tamu.edu
Oscar Riera-Lizarazu
✉ oscar.riera-lizarazu@ag.tamu.edu

RECEIVED 22 May 2023
ACCEPTED 11 July 2023
PUBLISHED 15 August 2023

CITATION
Rawandoozi Z, Young EL, Liang S, Wu X, Fu Q, Hochhaus T, Yan M, Rawandoozi MY, Klein PE, Byrne DH and Riera-Lizarazu O (2023) Pedigree-based QTL analysis of flower size traits in two multi-parental diploid rose populations.
Front. Plant Sci. 14:1226713.
doi: 10.3389/fpls.2023.1226713

Pedigree-based QTL analysis of flower size traits in two multi-parental diploid rose populations

Zena Rawandoozi[1*], Ellen L. Young[1], Shuyin Liang[1], Xuan Wu[1], Qiuyi Fu[1], Tessa Hochhaus[1], Muqing Yan[1], Maad Y. Rawandoozi[2], Patricia E. Klein[1], David H. Byrne[1] and Oscar Riera-Lizarazu[1*]

[1]Department of Horticultural Sciences, Texas A&M University, College Station, TX, United States, [2]Norman Borlaug Institute for International Agriculture and Development, Texas A&M AgriLife Research, Texas A&M System, College Station, TX, United States

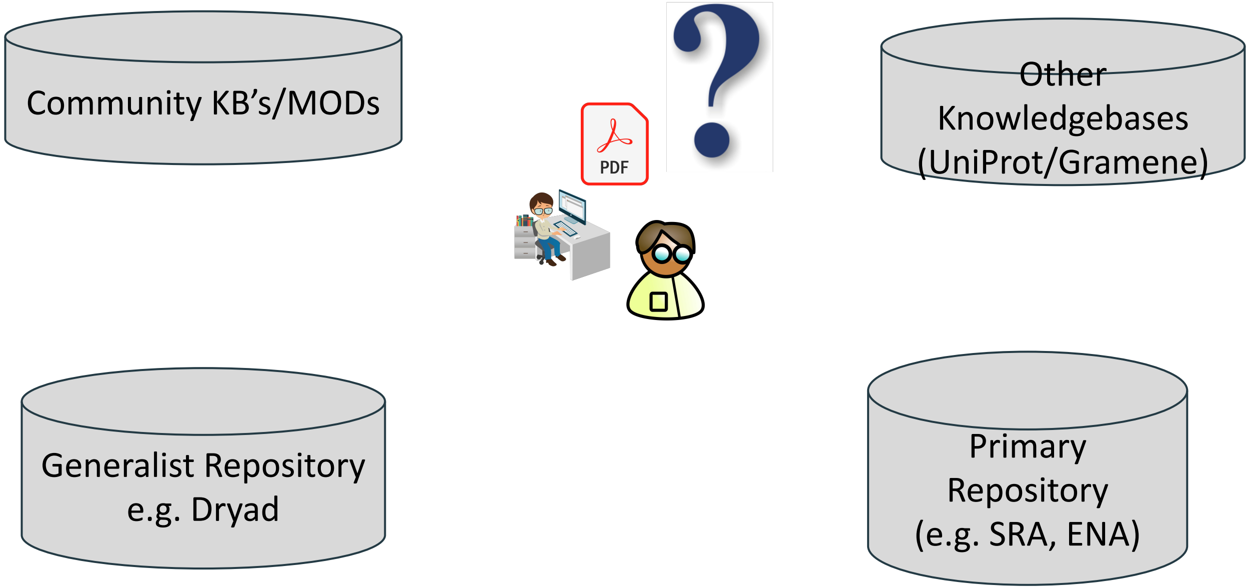
Rose (*Rosa* spp.) is one of the most economically important ornamental species worldwide. Flower diameter, flower weight, and the number of petals and petaloids are key flower-size parameters and attractive targets for DNA-informed breeding. Pedigree-based analysis (PBA) using FlexQTL software was conducted using two sets of multi-parental diploid rose populations. Phenotypic

DATA

METADATA

GDR

Many places Ag data can go and be found



Community KB's/MODs

Other Knowledgebases (UniProt/Gramene)

Generalist Repository e.g. Dryad

Primary Repository (e.g. SRA, ENA)

Generating a tool to help scientists get their data into the correct database

- **What are the AgBioData Databases and their crop/data focus?**
- **What types of data do each database maintain?**
    - map terms to Mesh or EDAM ontology terms
- **Does the database accept community submissions or do they only take data from other repositories or a mix?**
- **But wait, isn't there already a tool that already does the trick?**
    - FAIR sharing not granular enough
- ***What about data that should go into primary data repositories?***
    - *Ex. variations - should go to NCBI, ENA, EVA, can AgBioData and their databases act as brokers to get the data there?*
- ***What about data with no community database?***

Solution: Drupal Database Selector Module

Option 1: Filter by species

The Database Finder tool

Use this tool to find out which AgBio Member Database, or related repository, collects data you generate and want to submit, such as Genomic, Genetic, and Breeding Data of Agricultural crops or livestock. View list of data types and definitions.

This tool will also help you find member AgBioDatabase(s) or main genomic database(s) such as NCBI that house data you want to find.

Filter the list of databases by selecting Organism(s) and/or Datatype(s) you want to submit or find. As you select these options, the relevant Database(s) on the right will be highlighted. Click on one or more databases from this subset to view and compare additional details at the bottom of the page.

Database Finder to submit Agricultural Genomic, Genetic, Breeding Data

Evaluate database options to submit your data

Filter DBs By 1. Organism 2. Data to Submit 3. Data to Find

1. Organism

- ☐ All plants
- ☑ Arabidopsis
- ☐ Legumes
- ☐ Citrus
- ☐ Cotton
- ☐ Livestock
- ☐ Maize
- ☐ Poaceae
  - ☐ Maize
  - ☐ Poaceae

Select services you would like to compare.

- Animal QTLdb ○ The Animal Quantitative Trait Loci (QTL) Database
- CGD ○ Citrus Genome Database
- CottonGen ○ Cotton Database Resources
- GDR ○ Genome Database for Rosaceae
- GrainGenes ○ A Database for Triticeae and Avena
- Gramene ○ **Comparative plant genomics for crops and model organisms**
- LIS ○ the Legume Information System
- MaizeGDB ○ Maize Genetics and Genomics Database
- SorghumBase ○ Genomic resources for the sorghum research community
- SoyBase ○ Integrating Genetics and Genomics to Advance Soybean Research
- TAIR ○ **The Arabidopsis Information Resources**

Resources

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

Data type definitions (EDAM and other ontologies)

Data Type Definition

Submitted by amarrano on Fri, 12/13/2024 - 06:45

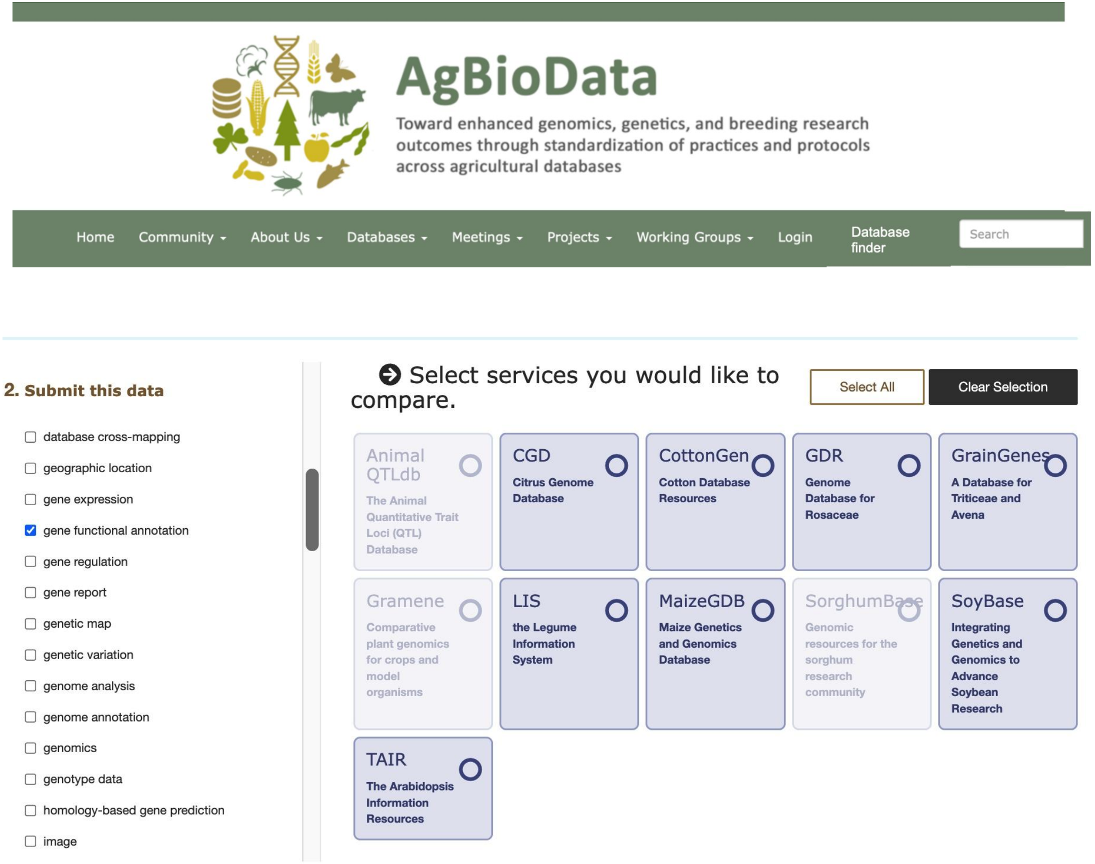
Data types and their definition listed in the Database Finder Tool.

More details on the data types can be found by clicking the data type term.

Term	Definition
Database cross-mapping	Accession numbers (or other identifiers) for entities/records found in other databases are captured and mapped to records in this database.
Gene expression	The analysis of levels and patterns of synthesis of gene products (proteins and functional RNA), including interpretation in functional terms of gene expression data.
Gene functional annotation	Ontologies (e.g., Gene Ontology Biological Process) or other controlled vocabularies are used to annotate functions of genes.
Gene regulation	The analysis of gene expression regulation.
Gene report	A summary report for a particular locus or gene that might include the gene name, description, and function of a gene, such as its encoded protein, or functional classification of the gene sequence according to the encoded protein(s).
Genetic map	A map showing the relative positions of genetic markers in a nucleic acid sequence, based on estimating non-physical distance such as recombination frequencies.
Genetic variation	Stable mutations in a nucleotide sequence, including alleles, naturally occurring mutations such as single base nucleotide substitutions, deletions and insertions, RFLPs, and other polymorphisms.
Genome analysis	Sequence feature, structure, variation, function, and evolution at a genomic scale.
Genome annotation	Use of ontologies or controlled vocabularies to annotate genome sequence.
Genomics	Whole genomes of one or more organisms, or genomes in general, such as meta-information on genomes, genome projects, gene names, etc.
Genotype data	A human-readable collection of information about the set of genes (or allelic forms) present in an individual, organism, or cell and associated with a specific physical characteristic or a report concerning an organism's traits and phenotypes.
Geographic location	A report about localization of the isolation of biological material, e.g., country or coordinates.
Homology-based gene prediction	Use of homology analysis to predict a gene or gene component(s).
Image	Data (typically biological or biomedical) that has been rendered into an image, typically for display on screen.
Image annotation	Use of ontologies or controlled vocabularies to annotate image data.
Metabolic pathway prediction	Predictions with influence on a metabolic pathway.
Metabolomics	The systematic identification and quantitation of all the metabolic products of a cell, tissue, organ, or organism under varying conditions as well as metabolite patterns in biological samples and or in correlation with xenobiotic challenge and disease states.
Metadata for other analysis	Structured data elements used to describe ancillary data.

Option 2

Filter by data to submit



2. Submit this data

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

Select services you would like to compare.

- Animal QTLdb ○ The Animal Quantitative Trait Loci (QTL) Database
- CGD ○ Citrus Genome Database
- CottonGen ○ Cotton Database Resources
- GDR ○ Genome Database for Rosaceae
- GrainGenes ○ A Database for Triticeae and Avena
- Gramene ○ Comparative plant genomics for crops and model organisms
- LIS ○ the Legume Information System
- MaizeGDB ○ Maize Genetics and Genomics Database
- SorghumBase ○ Genomic resources for the sorghum research community
- SoyBase ○ Integrating Genetics and Genomics to Advance Soybean Research
- TAIR ○ The Arabidopsis Information Resources

Option 3

Filter by data type to find

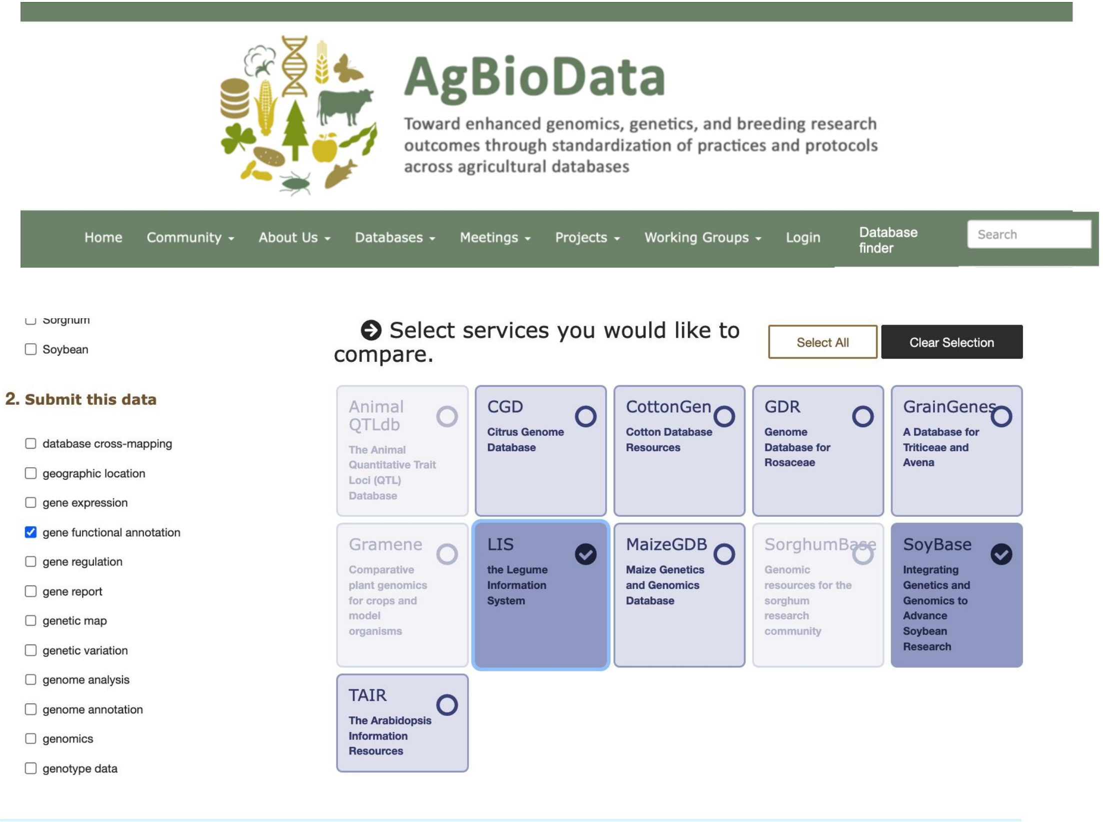
2. Submit this data

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

Select services you would like to compare.

- Animal QTLdb ○ — The Animal Quantitative Trait Loci (QTL) Database
- CGD ○ — Citrus Genome Database
- CottonGen ○ — Cotton Database Resources
- GDR ○ — Genome Database for Rosaceae
- GrainGenes ○ — A Database for Triticeae and Avena
- Gramene ○ — Comparative plant genomics for crops and model organisms
- LIS ○ — the Legume Information System
- MaizeGDB ○ — Maize Genetics and Genomics Database
- SorghumBase ○ — Genomic resources for the sorghum research community
- SoyBase ○ — Integrating Genetics and Genomics to Advance Soybean Research
- TAIR ○ — The Arabidopsis Information Resources

After applying filters select resource(s) to view



☐ Sorghum

☐ Soybean

2. Submit this data

☐ database cross-mapping

☐ geographic location

☐ gene expression

☑ gene functional annotation

☐ gene regulation

☐ gene report

☐ genetic map

☐ genetic variation

☐ genome analysis

☐ genome annotation

☐ genomics

☐ genotype data

Select services you would like to compare.

○ **Animal QTLdb** The Animal Quantitative Trait Loci (QTL) Database

○ **CGD** Citrus Genome Database

○ **CottonGen** Cotton Database Resources

○ **GDR** Genome Database for Rosaceae

○ **GrainGenes** A Database for Triticeae and Avena

○ **Gramene** Comparative plant genomics for crops and model organisms

✓ **LIS** the Legume Information System

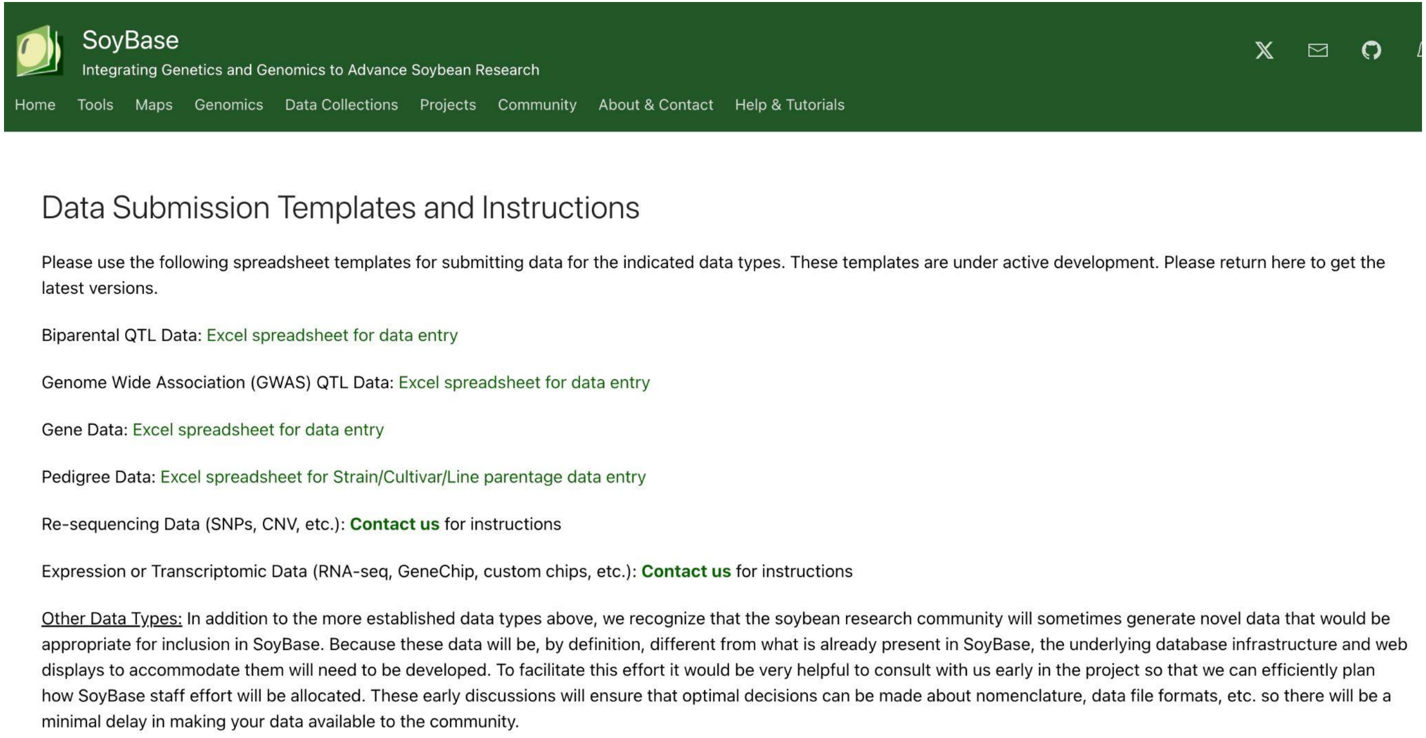
○ **MaizeGDB** Maize Genetics and Genomics Database

○ **SorghumBase** Genomic resources for the sorghum research community

✓ **SoyBase** Integrating Genetics and Genomics to Advance Soybean Research

○ **TAIR** The Arabidopsis Information Resources

Database submission instructions



Data Submission Templates and Instructions

Please use the following spreadsheet templates for submitting data for the indicated data types. These templates are under active development. Please return here to get the latest versions.

Biparental QTL Data: Excel spreadsheet for data entry

Genome Wide Association (GWAS) QTL Data: Excel spreadsheet for data entry

Gene Data: Excel spreadsheet for data entry

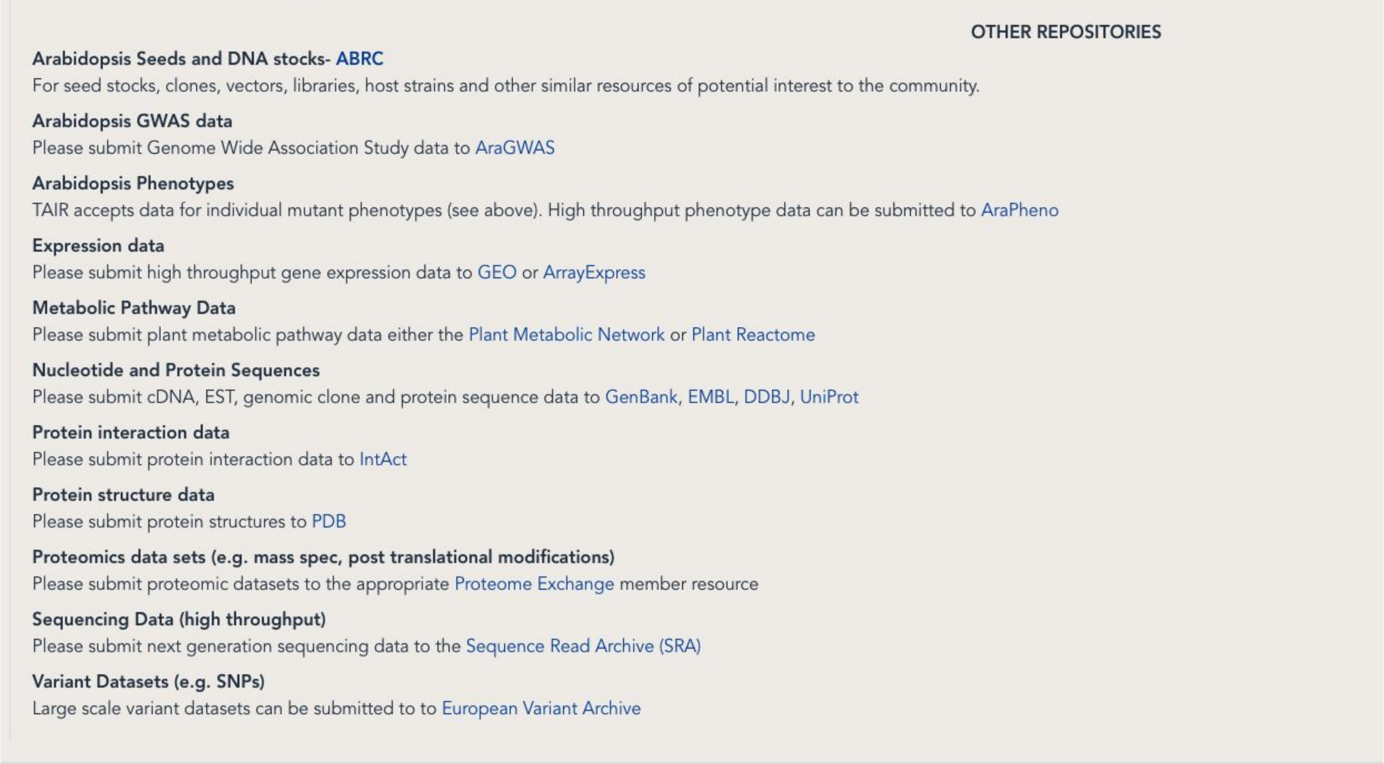
Pedigree Data: Excel spreadsheet for Strain/Cultivar/Line parentage data entry

Re-sequencing Data (SNPs, CNV, etc.): **Contact us** for instructions

Expression or Transcriptomic Data (RNA-seq, GeneChip, custom chips, etc.): **Contact us** for instructions

Other Data Types: In addition to the more established data types above, we recognize that the soybean research community will sometimes generate novel data that would be appropriate for inclusion in SoyBase. Because these data will be, by definition, different from what is already present in SoyBase, the underlying database infrastructure and web displays to accommodate them will need to be developed. To facilitate this effort it would be very helpful to consult with us early in the project so that we can efficiently plan how SoyBase staff effort will be allocated. These early discussions will ensure that optimal decisions can be made about nomenclature, data file formats, etc. so there will be a minimal delay in making your data available to the community.

What about hosted data?



OTHER REPOSITORIES

Arabidopsis Seeds and DNA stocks- ABRC
For seed stocks, clones, vectors, libraries, host strains and other similar resources of potential interest to the community.

Arabidopsis GWAS data
Please submit Genome Wide Association Study data to AraGWAS

Arabidopsis Phenotypes
TAIR accepts data for individual mutant phenotypes (see above). High throughput phenotype data can be submitted to AraPheno

Expression data
Please submit high throughput gene expression data to GEO or ArrayExpress

Metabolic Pathway Data
Please submit plant metabolic pathway data either the Plant Metabolic Network or Plant Reactome

Nucleotide and Protein Sequences
Please submit cDNA, EST, genomic clone and protein sequence data to GenBank, EMBL, DDBJ, UniProt

Protein interaction data
Please submit protein interaction data to IntAct

Protein structure data
Please submit protein structures to PDB

Proteomics data sets (e.g. mass spec, post translational modifications)
Please submit proteomic datasets to the appropriate Proteome Exchange member resource

Sequencing Data (high throughput)
Please submit next generation sequencing data to the Sequence Read Archive (SRA)

Variant Datasets (e.g. SNPs)
Large scale variant datasets can be submitted to to European Variant Archive

https://www.arabidopsis.org/submit/overview

Database Finder Tool

Creativity Extension !!

- Temporary deployment on NRSP10 site
- Finalize data collection from member DBs
- Manually enter values
- Update AgBioData Site Drupal version to install module
- Potentially making software modifications for better searching and display

Stakeholders: Challenges and barriers

Researcher

- ~~Not knowing where data should go~~
- Time consuming submission process
- Not knowing how to format data / metadata

Publisher

- ~~Not knowing where data should go~~
- Lack of easy verification of data availability
- Authors often do not want data available before publication

Research Librarian

- ~~Not knowing where data should go~~
- Lack of engagement with researcher
- High volume of data and not enough time

Funder

- ~~Not knowing where data should go~~
- Different programs have different repositories
- Proposal reviews don’t know how to evaluate DMPs

A tool that can guide people to the appropriate database, now what?

- Publish and disseminate
- Work with each database to help guide datatype submission in a consistent manner
    - Create a similar choice tool for each database that focuses on the datatypes they accept
    - Add recommendations for data that should go into other repos (NCBI, Genome warehouses, ENA, EVA) that AgBioData resources draw upon
    - Help databases build training modules for submitting data >> add to AgBioData curriculum
- FAIRness guidelines for community databases
    - -> FAIR Data Practices (maizegdb.org) https://www.maizegdb.org/FAIRpractices
- Add information on metadata/data formats

(How) can software/AI help?

- Automate analysis of papers
    - Assist editors and authors in determining what data there is to deposit
    - Guide on file formats and metadata
    - Guide to repositories
- Data and metadata formatting and validation
- Validation of data submission and remove embargo flags when paper is published
- AI alone is not replacing a subject matter expert – curator in the loop!

Your feedback is most welcome!

Stakeholders: Challenges and barriers

Researcher

- Not knowing where data should go
- Time consuming submission process
- Not knowing how to format data / metadata

Publisher

- Not knowing where data should go
- Lack of easy verification of data availability
- Authors often do not want data available before publication

Research Librarian

- Not knowing where data should go
- Lack of engagement with researcher
- High volume of data and not enough time

Funder

- Not knowing where data should go
- Different programs have different repositories
- Proposal reviews don’t know how to evaluate DMPs

Stakeholders: Challenges and barriers

Researcher

- Not knowing where data should go
- Time consuming submission process
- Not knowing how to format data / metadata

Publisher

- Not knowing where data should go
- Lack of easy verification of data availability
- Authors often do not want data available before publication

Research Librarian

- Not knowing where data should go
- Lack of engagement with researcher
- High volume of data and not enough time

Funder

- Not knowing where data should go
- Different programs have different repositories
- Proposal reviews don’t know how to evaluate DMPs

Authors should put their data in Community Databases, but….

✓ **here may be more than one Database for a community (or none at all)**

- ○ Wheat Data can be found in Triticeae toolbox; GrainGenes; Gramene

✓ **ll Databases do not host the same data types**

- ○ Triticeae toolbox -> SNP, phenotypic, pedigree; GrainGenes -> genomes/tracks, images, maps, curated data

✓ **ll Databases do not allow data submission from individuals**

- ○ Gramene only takes data from other databases, Triticeae toolbox only takes from the Wheat Coordinated Agricultural Project (Wheat CAP), GrainGenes allows community submissions

Data should go in Community Databases, but….

- **There may be more than one Database for a community**
    - Wheat Data can be found in Triticeae toolbox, GrainGenes, and Gramene
- **All Databases do not host the same data types**
    - Triticeae toolbox → SNP, phenotypic, pedigree
    - GrainGenes → genomes/tracks, images, maps, curated data
- **Only SOME Databases allow data submission from individuals**
    - Gramene ← other databases
    - Triticeae toolbox ← Wheat Coordinated Agricultural Project (Wheat CAP)
    - GrainGenes ← Community submissions
- **There may not be a Database for the community**
    - Ex. Vegetable Crops (ex. broccoli)