FAIR Scientific Literature (FSL) Working Group Update

PAG 2025
1/10/2025
Leonore Reiser

Current 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

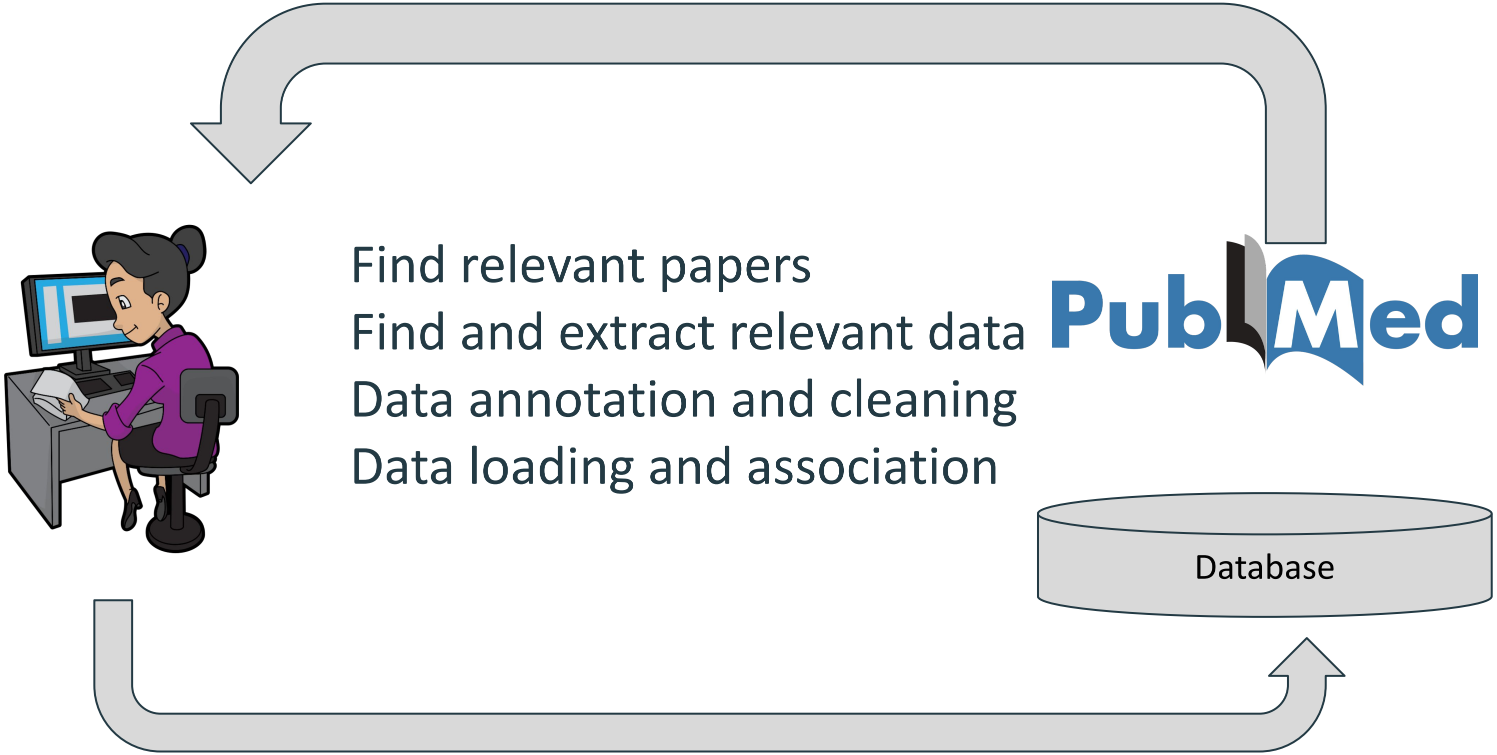
Karen Yook
Exec Editor,
microPublication &
Wormbase Curator

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



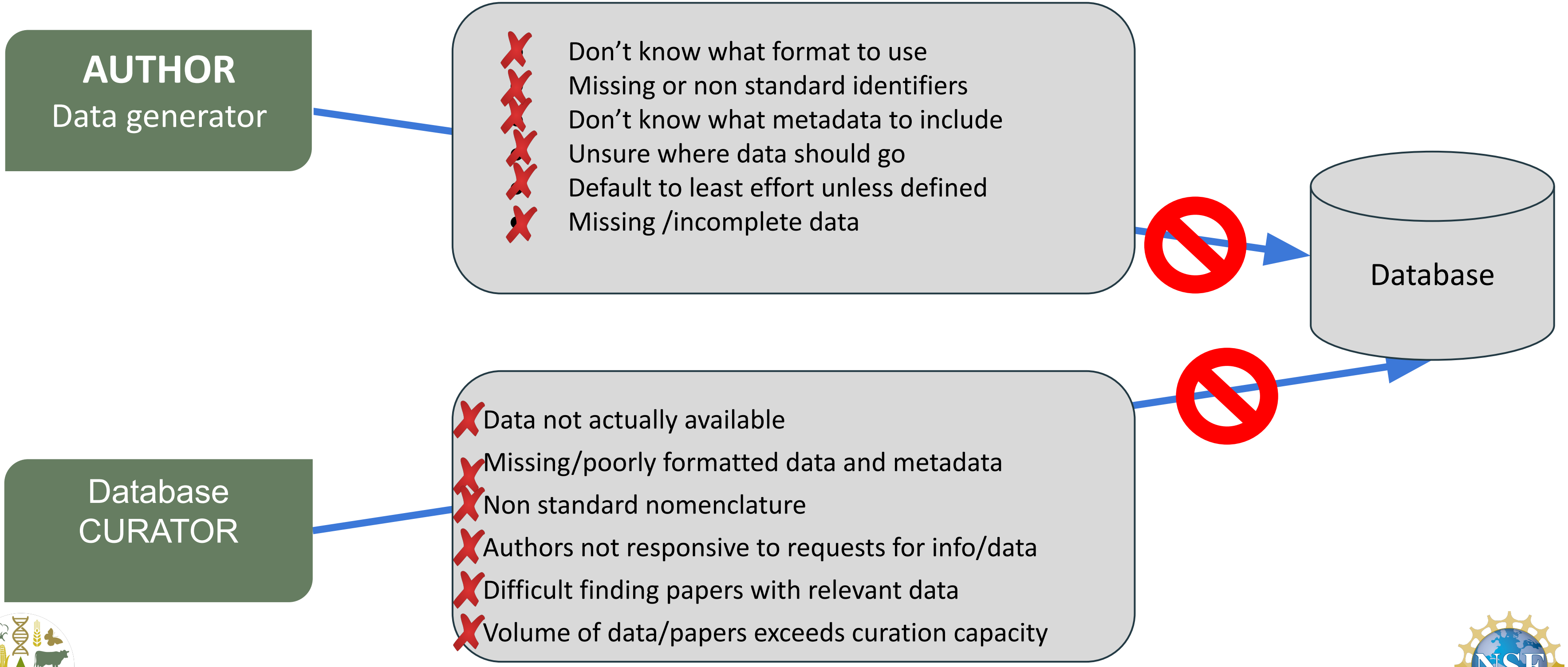
Find relevant papers

Find and extract relevant data

Data annotation and cleaning

Data loading and association

Publication-curation workflow bottlenecks



AUTHOR
Data generator

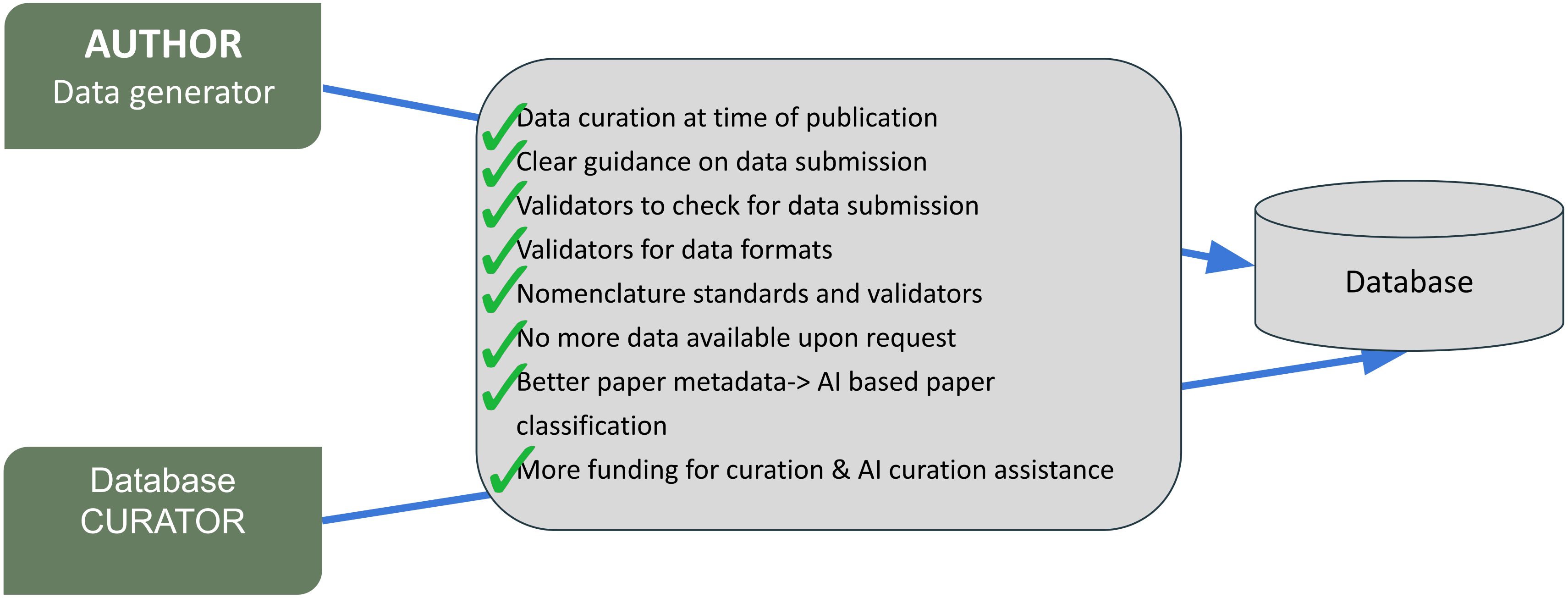
- ✗ Don’t know what format to use
- ✗ Missing or non standard identifiers
- ✗ Don’t know what metadata to include
- ✗ Unsure where data should go
- ✗ Default to least effort unless defined
- ✗ Missing /incomplete data

Database

Database
CURATOR

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

A better literature curation workflow



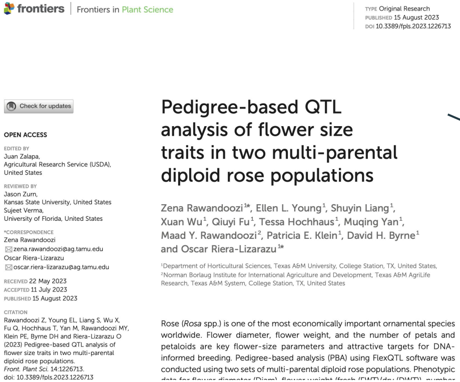
AUTHOR
Data generator

- ✓ 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

Database
CURATOR

First Goal: Ag 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 for flower diameter (Diam), flower weight (fresh (FWT)/dry (DWT)), number

DATA

TEXT Document

METADATA

TEXT Document

GDR

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

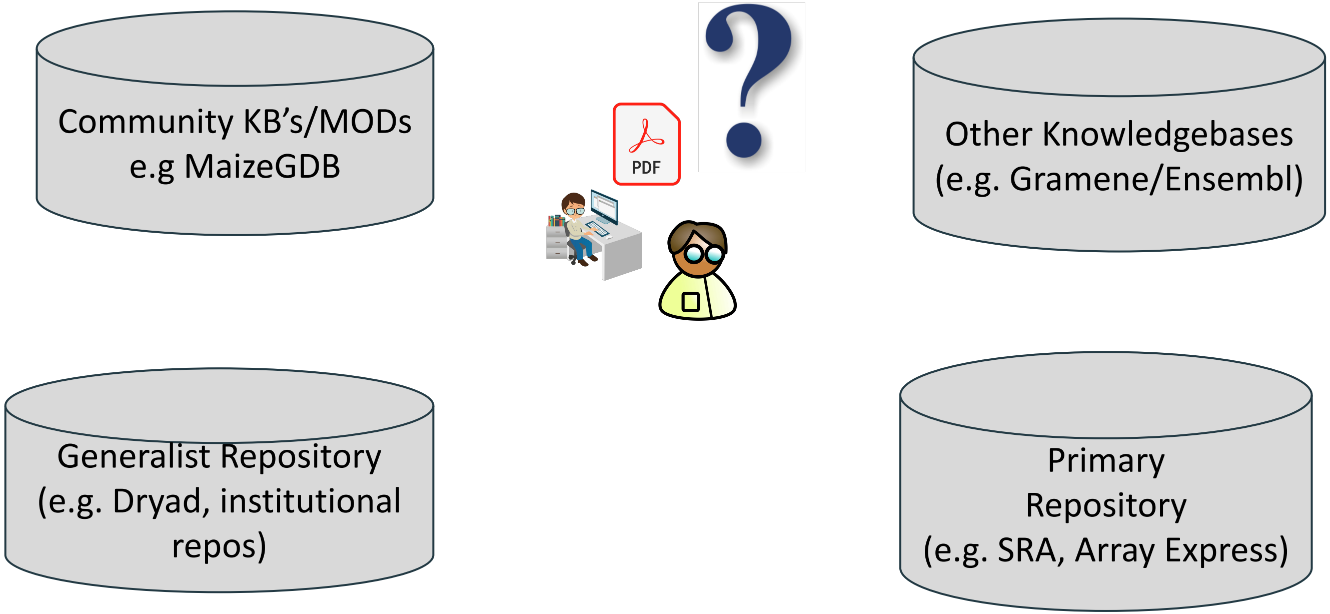
Research Librarian

- **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 reviewers don’t know how to evaluate

Many places Ag data can be found



Community KB's/MODs
e.g MaizeGDB

Other Knowledgebases
(e.g. Gramene/Ensembl)

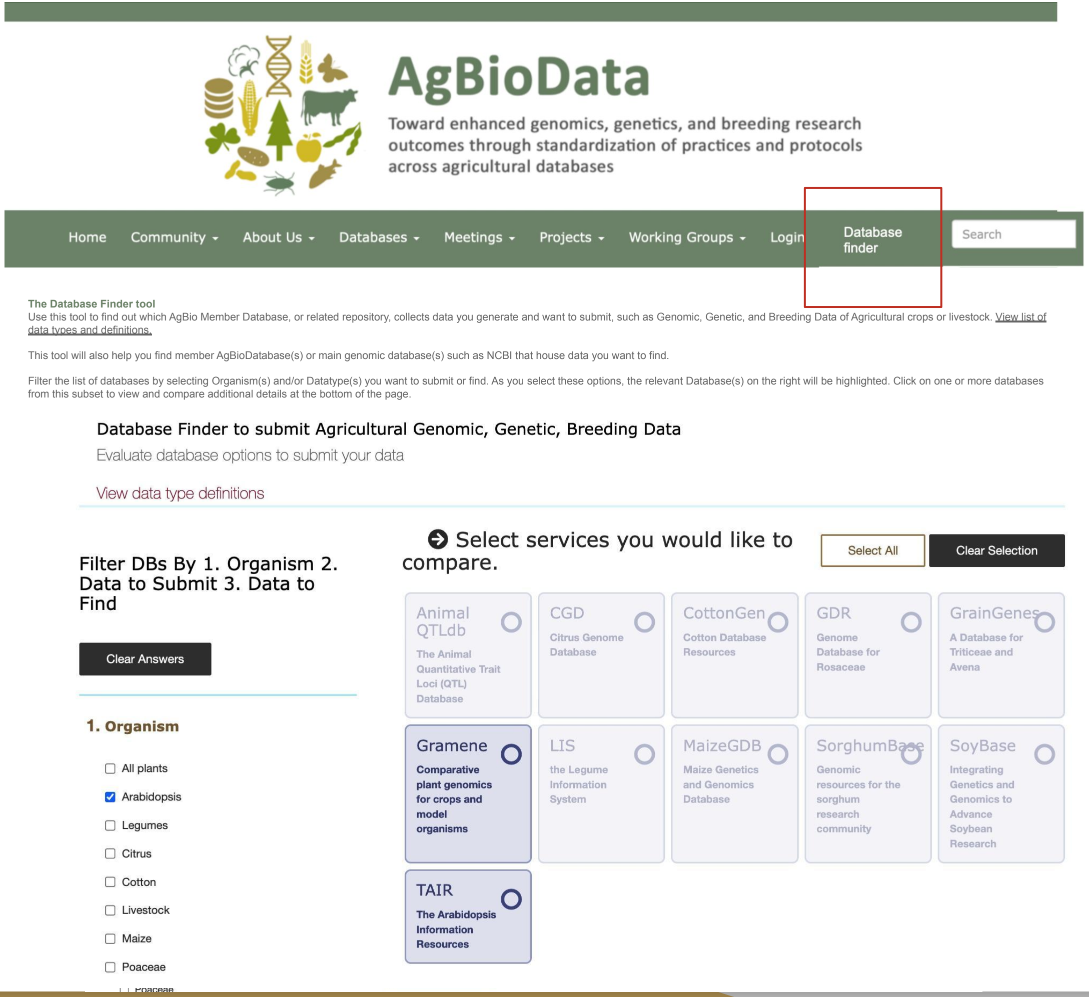
Generalist Repository
(e.g. Dryad, institutional repos)

Primary Repository
(e.g. SRA, Array Express)

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?**
  - start with G2P table (https://doi.org/10.1093/database/baad088)
- **Does the database accept community submissions, do they only curate take data from other sources, 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



AgBioData
Toward enhanced genomics, genetics, and breeding research outcomes through standardization of practices and protocols across agricultural databases
Home
Community
About Us
Databases
Meetings
Projects
Working Groups
Login
Database finder
Search

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

View data type definitions

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

Clear Answers

1. Organism

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

Poaceae

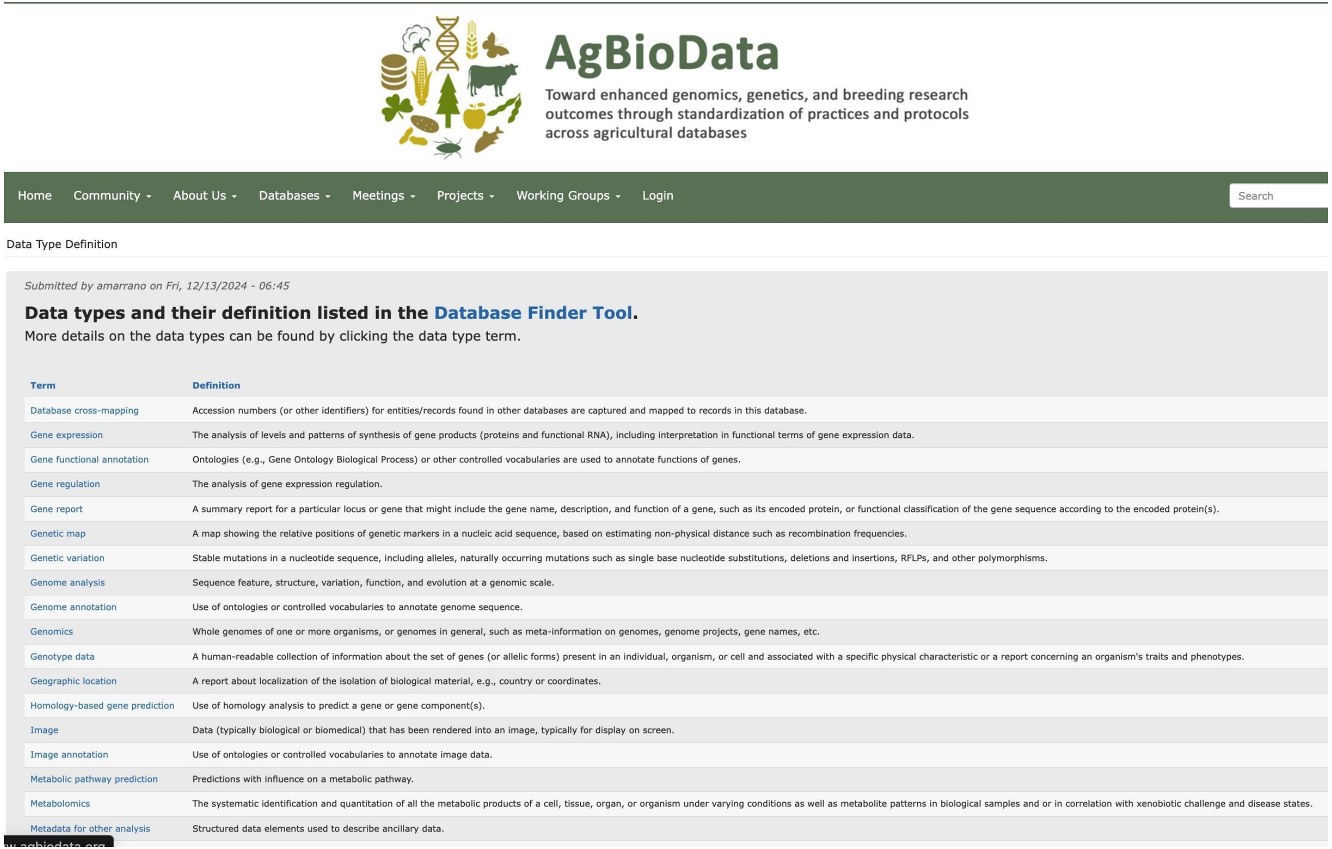
Select services you would like to compare.

Select All | Clear Selection

- ○ 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

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

Data type definitions (MESH, EDAM and other ontologies)


AgBioData
Toward enhanced genomics, genetics, and breeding research outcomes through standardization of practices and protocols across agricultural databases
Home
Community
About Us
Databases
Meetings
Projects
Working Groups
Login
Search

Data Type Definition

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

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

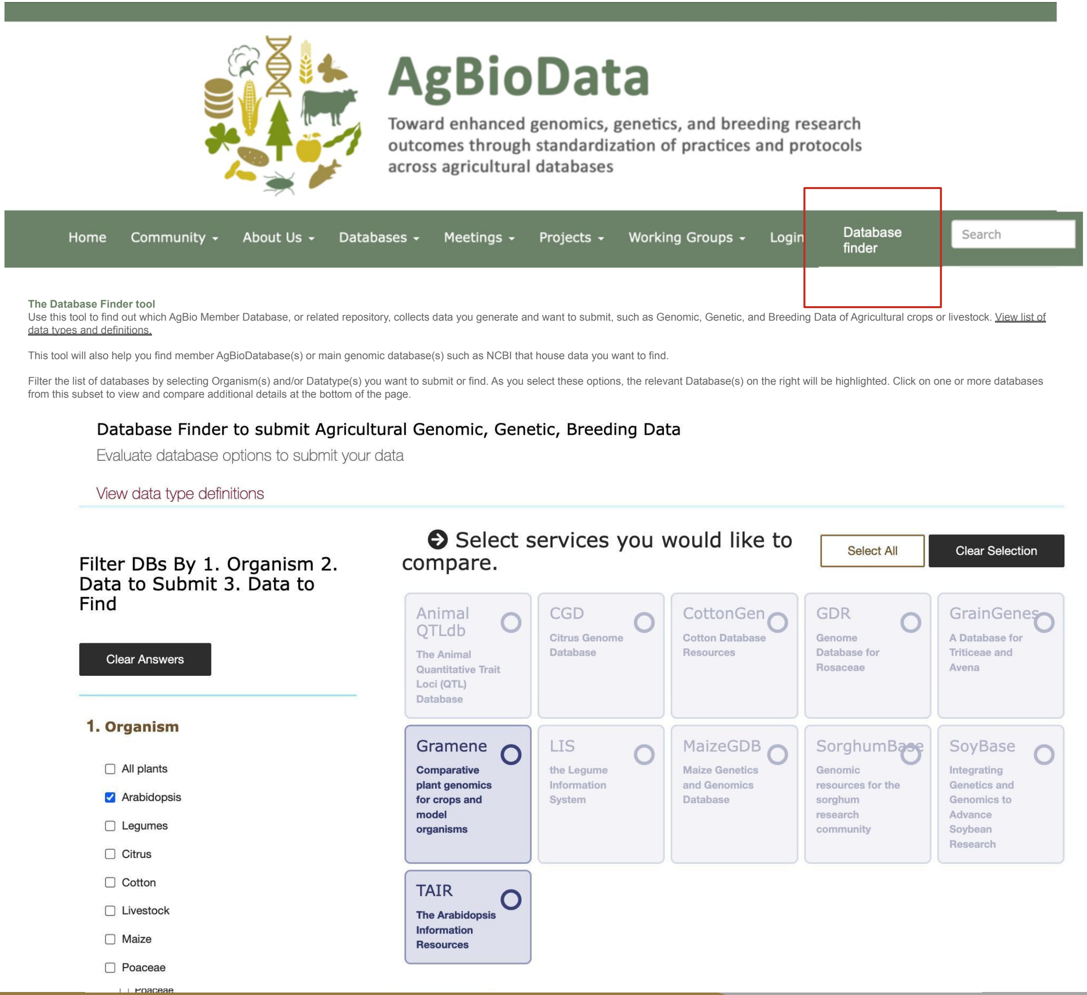
More details on the data types can be found by clicking the data type term.

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.

w.agbiodata.org

Option 1:

Filter by species


AgBioData
Toward enhanced genomics, genetics, and breeding research outcomes through standardization of practices and protocols across agricultural databases
Home
Community
About Us
Databases
Meetings
Projects
Working Groups
Login
Database finder
Search

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

View data type definitions

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

1. Organism

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

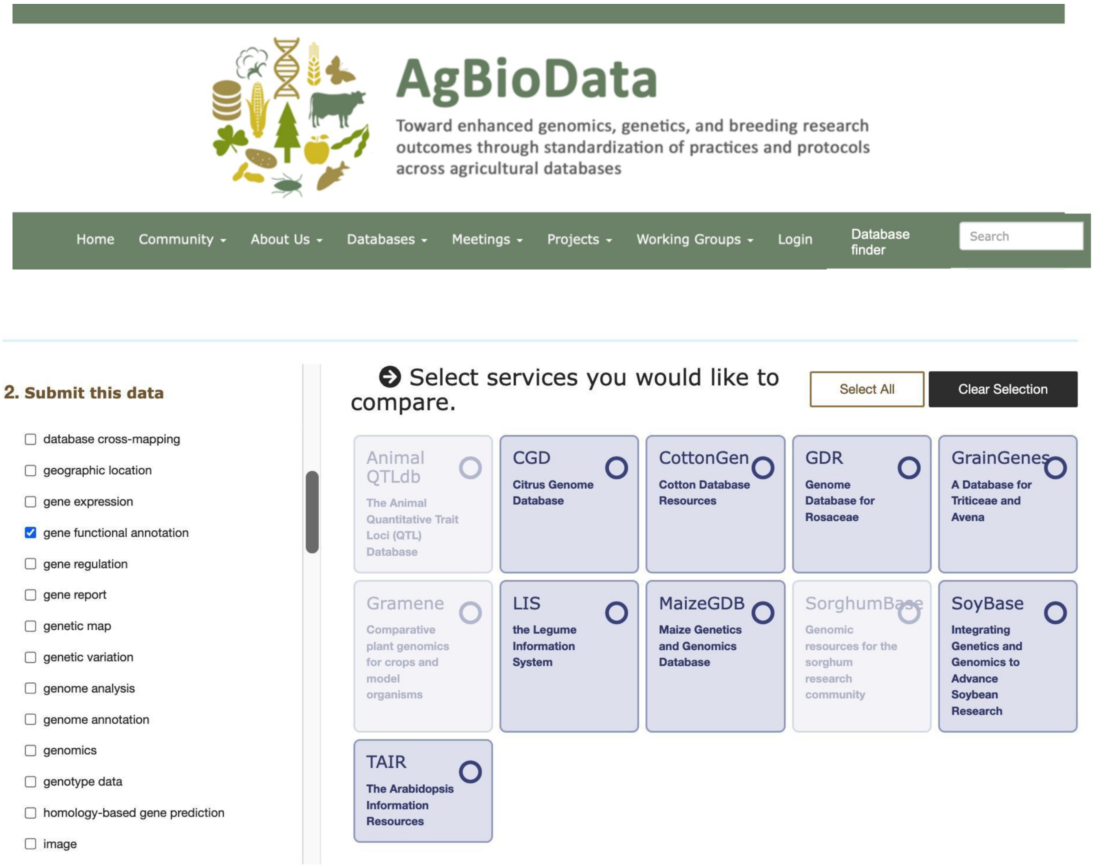
SoyBase ○ Integrating Genetics and Genomics to Advance Soybean Research

TAIR ○ The Arabidopsis Information Resources

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

Option 2

Filter by data to submit



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

2. Submit this data

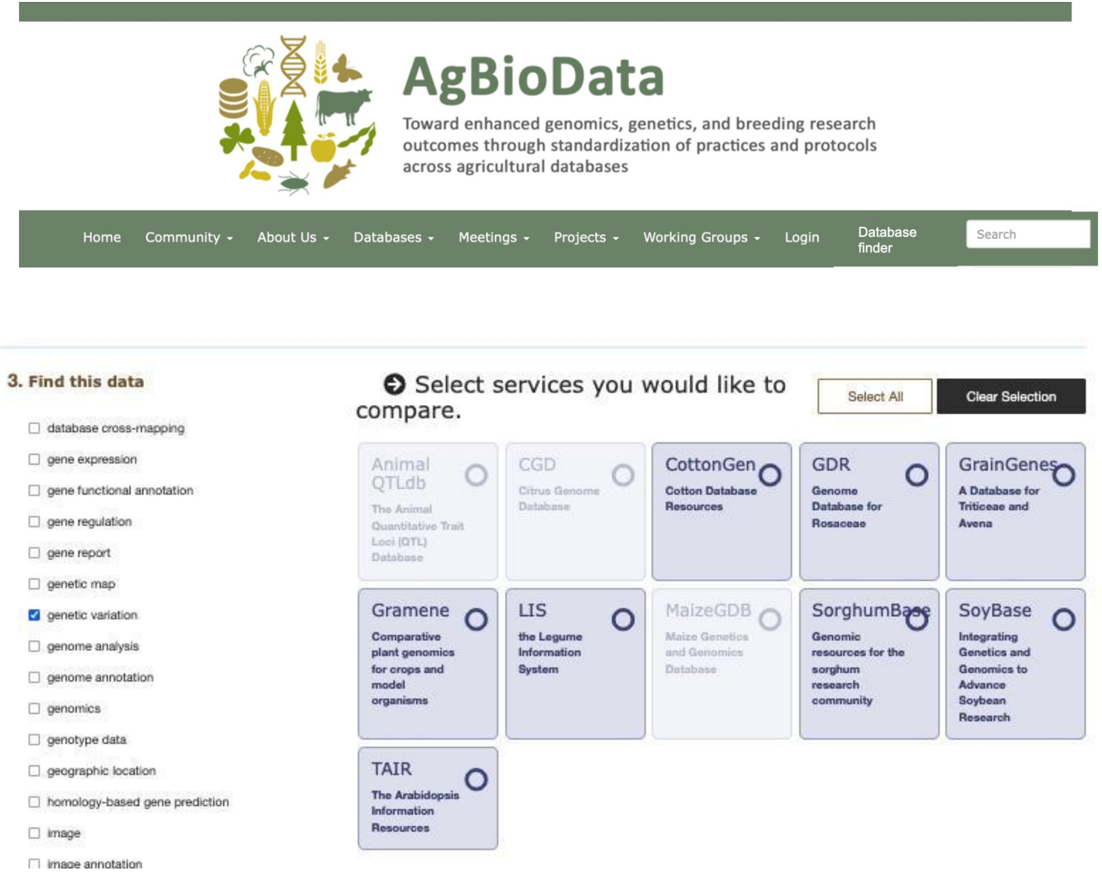
- ☐ database cross-mapping
- ☐ geographic location
- ☐ gene expression
- ☑ gene functional annotation
- ☐ gene regulation
- ☐ gene report
- ☐ genetic map
- ☐ genetic variation
- ☐ genome analysis
- ☐ genome annotation
- ☐ genomics
- ☐ genotype data
- ☐ homology-based gene prediction
- ☐ image

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


AgBioData
Toward enhanced genomics, genetics, and breeding research outcomes through standardization of practices and protocols across agricultural databases
Home
Community ▾
About Us ▾
Databases ▾
Meetings ▾
Projects ▾
Working Groups ▾
Login
Database finder
Search
Select All
Clear Selection

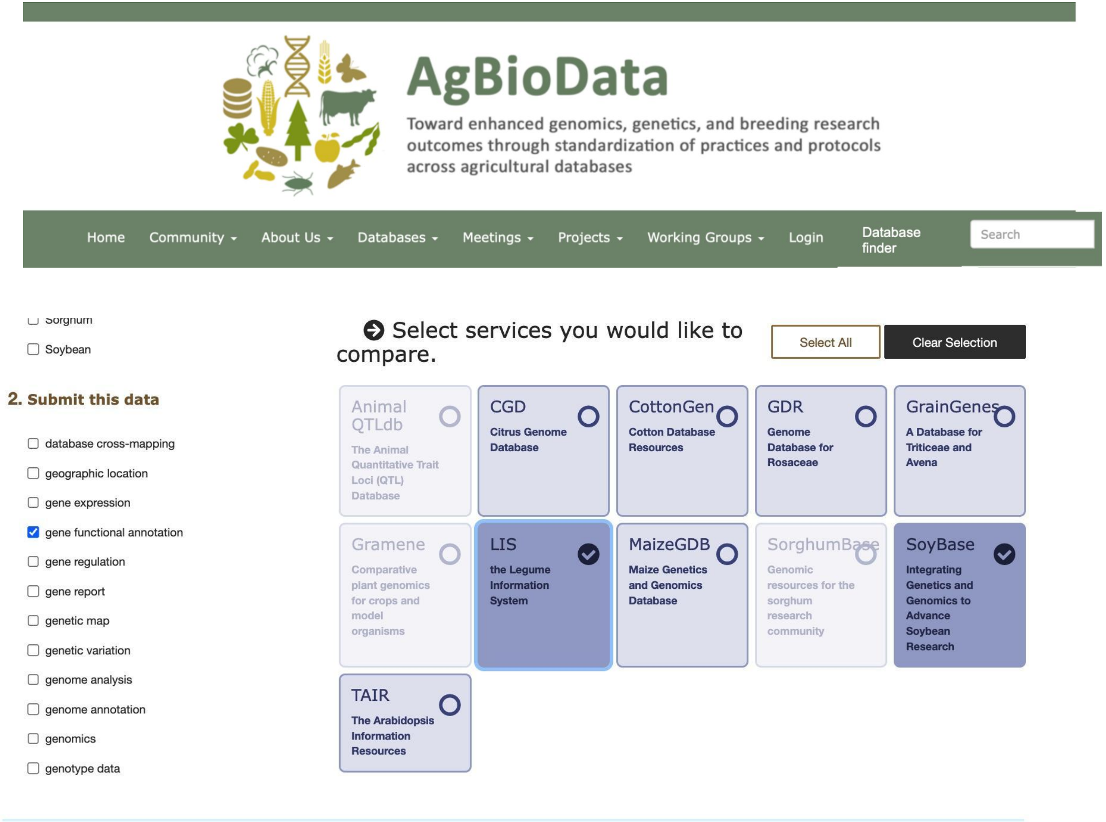
3. Find this data

- □ 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
- □ image annotation

➔ 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



AgBioData

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

□ 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

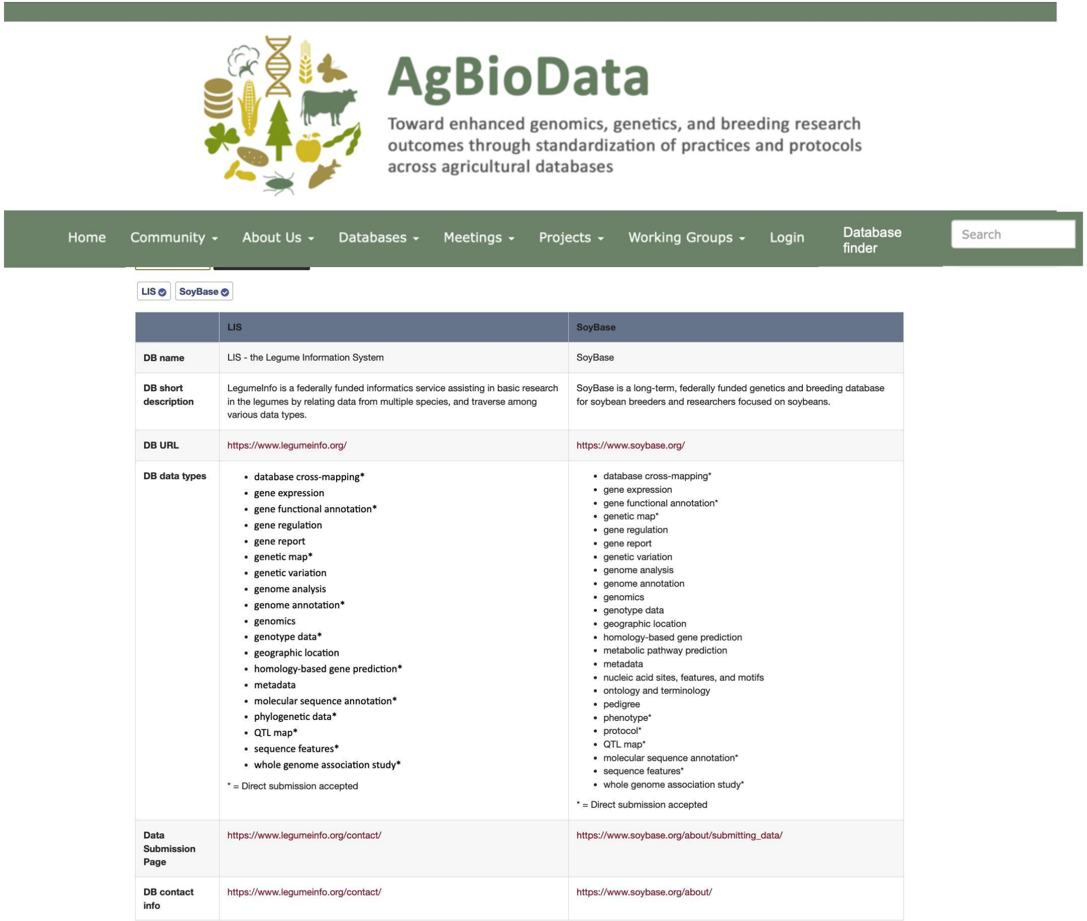
○ MaizeGDB – Maize Genetics and Genomics Database

○ SorghumBase – Genomic resources for the sorghum research community

✓ SoyBase – Integrating Genetics and Genomics to Advance Soybean Research

○ TAIR – The Arabidopsis Information Resources

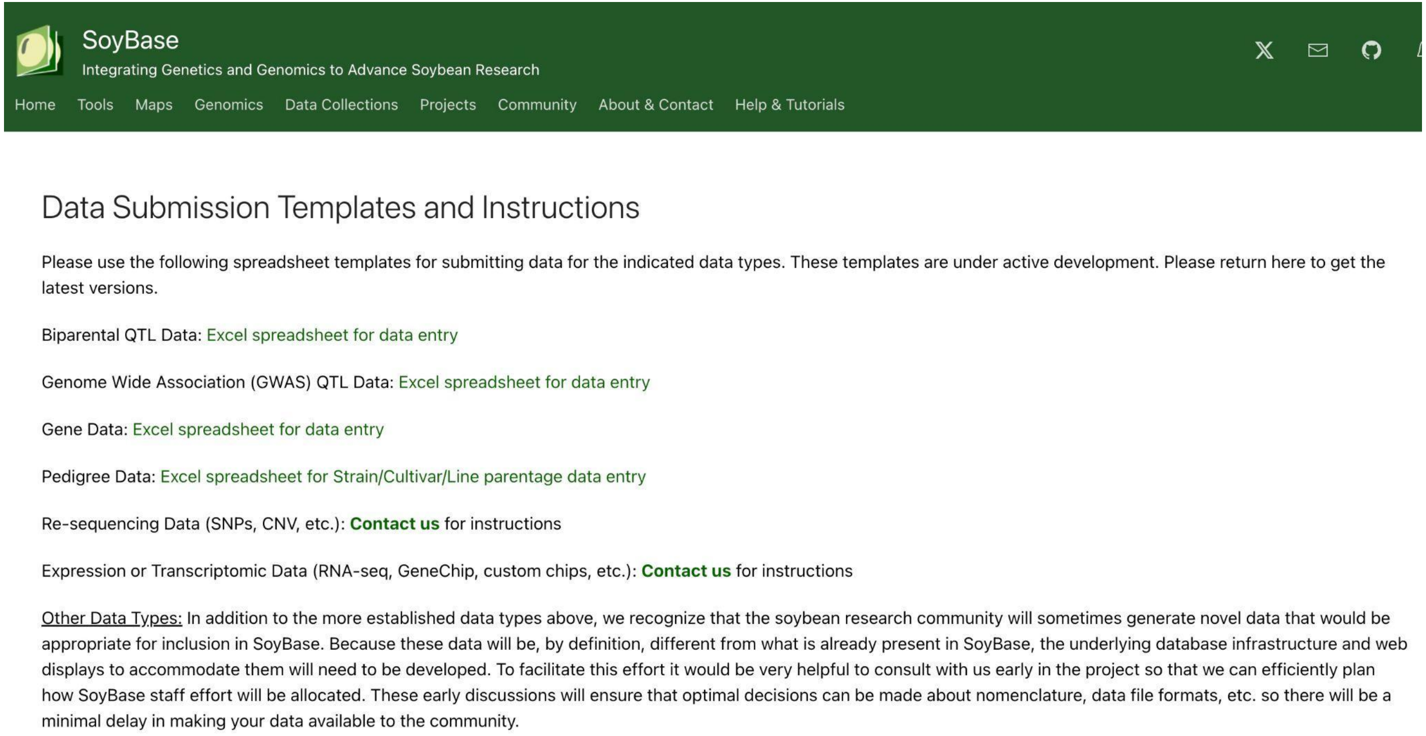
Table display



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

	LIS	SoyBase
DB name	LIS - the Legume Information System	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.	SoyBase is a long-term, federally funded genetics and breeding database for soybean breeders and researchers focused on soybeans.
DB URL	https://www.legumeinfo.org/	https://www.soybase.org/
DB data types	• 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* • metadata • molecular sequence annotation* • phylogenetic data* • QTL map* • sequence features* • whole genome association study* * = Direct submission accepted	• database cross-mapping* • gene expression • gene functional annotation* • genetic map* • gene regulation • gene report • genetic variation • genome analysis • genome annotation • genomics • genotype data • geographic location • homology-based gene prediction • metabolic pathway prediction • metadata • nucleic acid sites, features, and motifs • ontology and terminology • pedigree • phenotype* • protocol* • QTL map* • molecular sequence annotation* • sequence features* • whole genome association study* * = Direct submission accepted
Data Submission Page	https://www.legumeinfo.org/contact/	https://www.soybase.org/about/submitting_data/
DB contact info	https://www.legumeinfo.org/contact/	https://www.soybase.org/about/

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

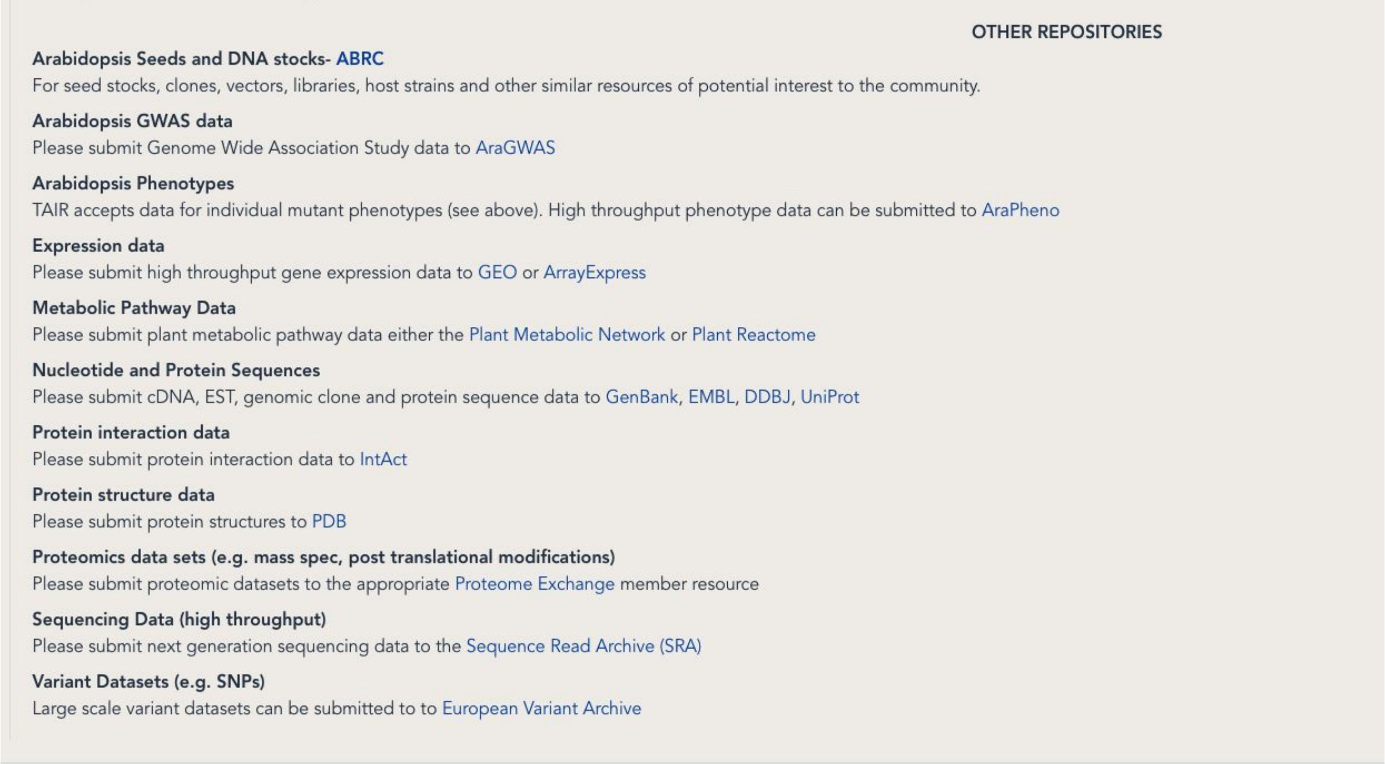
Pedigree Data: Excel spreadsheet for Strain/Cultivar/Line parentage data entry

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-woo hoo !!

- 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

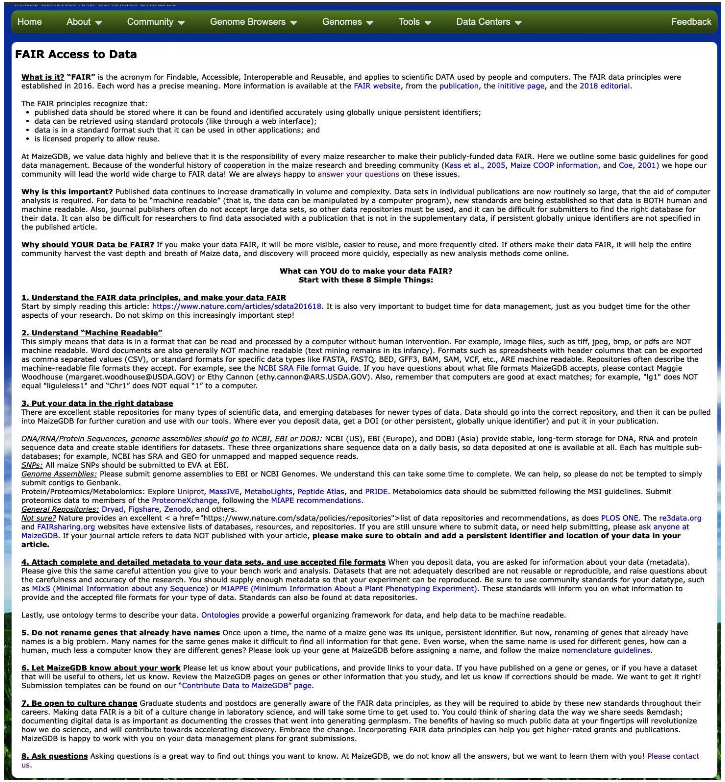
A tool that can guide people to the appropriate database, now what?

- Publish and disseminate
- Add information about data and metadata standards and formats
- Add recommendations for data that should go into other repos that AgBioData DBs draw from

- Work with each database to help guide datatype submission in a consistent manner
    - Help databases build training modules for submitting data >> add to AgBioData curriculum
    - Maybe create a similar choice tool for each database that focuses on the datatypes they accept
    - FAIRness guidelines for community databases

    - FAIR Data Practices (maizegdb.org) https://www.maizegdb.org/FAIRpractices

Example of FAIRness guides



FAIR Access to Data

What is it? "FAIR" is the acronym for Findable, Accessible, Interoperable and Reusable, and applies to scientific DATA used by people and computers. The FAIR data principles were established in 2016. Each word has a precise meaning. More information is available at the FAIR website, from the publication, the initive page, and the 2018 editorial.

The FAIR principles recognize that:

- published data should be stored where it can be found and identified accurately using globally unique persistent identifiers;
- data can be retrieved using standard protocols (like through a web interface);
- data is in a standard format such that it can be used in other applications; and
- is licensed properly to allow reuse.

At MaizeGDB, we value data highly and believe that it is the responsibility of every maize researcher to make their publicly-funded data FAIR. Here we outline some basic guidelines for good data management. Because of the wonderful history of cooperation in the maize research and breeding community (Kass et al., 2005, Maize COOP information, and Coe, 2001) we hope our community will lead the world wide charge to FAIR data! We are always happy to answer your questions on these issues.

Why is this important? Published data continues to increase dramatically in volume and complexity. Data sets in individual publications are now routinely so large, that the aid of computer analysis is required. For data to be "machine readable" (that is, the data can be manipulated by a computer program), new standards are being established so that data is BOTH human and machine readable. Also, journal publishers often do not accept large data sets, so other data repositories must be used, and it can be difficult for submitters to find the right database for their data. It can also be difficult for researchers to find data associated with a publication that is not in the supplementary data, if persistent globally unique identifiers are not specified in the published article.

Why should YOUR Data be FAIR? If you make your data FAIR, it will be more visible, easier to reuse, and more frequently cited. If others make their data FAIR, it will help the entire community harvest the vast depth and breath of Maize data, and discovery will proceed more quickly, especially as new analysis methods come online.

What can YOU do to make your data FAIR?
Start with these 8 Simple Things:

1. Understand the FAIR data principles, and make your data FAIR
Start by simply reading this article: https://www.nature.com/articles/sdata201618. It is also very important to budget time for data management, just as you budget time for the other aspects of your research. Do not skimp on this increasingly important step!

2. Understand "Machine Readable"
This simply means that data is in a format that can be read and processed by a computer without human intervention. For example, image files, such as tiff, jpeg, bmp, or pdfs are NOT machine readable. Word documents are also generally NOT machine readable (text mining remains in its infancy). Formats such as spreadsheets with header columns that can be exported as comma separated values (CSV), or standard formats for specific data types like FASTA, FASTQ, BED, GFF3, BAM, SAM, VCF, etc., ARE machine readable. Repositories often describe the machine-readable file formats they accept. For example, see the NCBI SRA File format Guide. If you have questions about what file formats MaizeGDB accepts, please contact Maggie Woodhouse (margaret.woodhouse@USDA.GOV) or Ethy Cannon (ethy.cannon@ARS.USDA.GOV). Also, remember that computers are good at exact matches; for example, "lg1" does NOT equal "liguleless1" and "Chr1" does NOT equal "1" to a computer.

3. Put your data in the right database
There are excellent stable repositories for many types of scientific data, and emerging databases for newer types of data. Data should go into the correct repository, and then it can be pulled into MaizeGDB for further curation and use with our tools. Where ever you deposit data, get a DOI (or other persistent, globally unique identifier) and put it in your publication.

DNA/RNA/Protein Sequences, genome assemblies should go to NCBI, EBI or DDBJ: NCBI (US), EBI (Europe), and DDBJ (Asia) provide stable, long-term storage for DNA, RNA and protein sequence data and create stable identifiers for datasets. These three organizations share sequence data on a daily basis, so data deposited at one is available at all. Each has multiple sub-databases; for example, NCBI has SRA and GEO for unmapped and mapped sequence reads.
SNPs: All maize SNPs should be submitted to EVA at EBI.
Genome Assemblies: Please submit genome assemblies to EBI or NCBI Genomes. We understand this can take some time to complete. We can help, so please do not be tempted to simply submit contigs to Genbank.
Protein/Proteomics/Metabolomics: Explore Uniprot, MassIVE, MetaboLights, Peptide Atlas, and PRIDE. Metabolomics data should be submitted following the MSI guidelines. Submit proteomics data to members of the ProteomeXchange, following the MIAPE recommendations.
General Repositories: Dryad, Figshare, Zenodo, and others.
Not sure? Nature provides an excellent < a href="https://www.nature.com/sdata/policies/repositories">list of data repositories and recommendations, as does PLOS ONE. The re3data.org and FAIRsharing.org websites have extensive lists of databases, resources, and repositories. If you are still unsure where to submit data, or need help submitting, please ask anyone at MaizeGDB. If your journal article refers to data NOT published with your article, **please make sure to obtain and add a persistent identifier and location of your data in your article.**

4. Attach complete and detailed metadata to your data sets, and use accepted file formats When you deposit data, you are asked for information about your data (metadata). Please give this the same careful attention you give to your bench work and analysis. Datasets that are not adequately described are not reusable or reproducible, and raise questions about the carefulness and accuracy of the research. You should supply enough metadata so that your experiment can be reproduced. Be sure to use community standards for your datatype, such as MIxS (Minimal Information about any Sequence) or MIAPPE (Minimum Information About a Plant Phenotyping Experiment). These standards will inform you on what information to provide and the accepted file formats for your type of data. Standards can also be found at data repositories.

Lastly, use ontology terms to describe your data. Ontologies provide a powerful organizing framework for data, and help data to be machine readable.

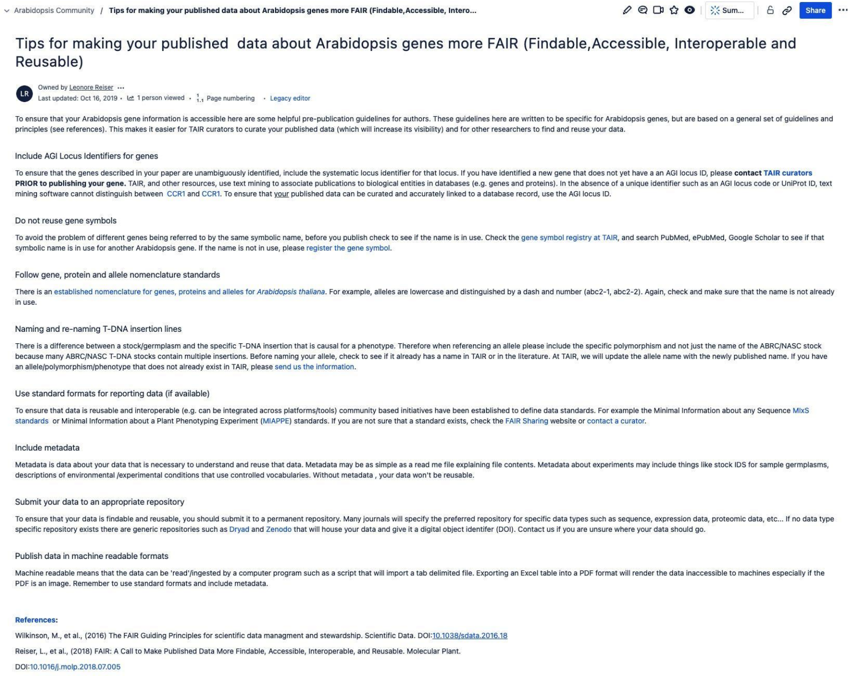
5. Do not rename genes that already have names Once upon a time, the name of a maize gene was its unique, persistent identifier. But now, renaming of genes that already have names is a big problem. Many names for the same genes make it difficult to find all information for that gene. Even worse, when the same name is used for different genes, how can a human, much less a computer know they are different genes? Please look up your gene at MaizeGDB before assigning a name, and follow the maize nomenclature guidelines.

6. Let MaizeGDB know about your work Please let us know about your publications, and provide links to your data. If you have published on a gene or genes, or if you have a dataset that will be useful to others, let us know. Review the MaizeGDB pages on genes or other information that you study, and let us know if corrections should be made. We want to get it right! Submission templates can be found on our "Contribute Data to MaizeGDB" page.

7. Be open to culture change Graduate students and postdocs are generally aware of the FAIR data principles, as they will be required to abide by these new standards throughout their careers. Making data FAIR is a bit of a culture change in laboratory science, and will take some time to get used to. You could think of sharing data the way we share seeds &emdash; documenting digital data is as important as documenting the crosses that went into generating germplasm. The benefits of having so much public data at your fingertips will revolutionize how we do science, and will contribute towards accelerating discovery. Embrace the change. Incorporating FAIR data principles can help you get higher-rated grants and publications. MaizeGDB is happy to work with you on your data management plans for grant submissions.

8. Ask questions Asking questions is a great way to find out things you want to know. At MaizeGDB, we do not know all the answers, but we want to learn them with you! Please contact us.

https://www.maizegdb.org/FAIRpractices


Arabidopsis Community / Tips for making your published data about Arabidopsis genes more FAIR (Findable,Accessible, Intero...
Sum...
Share
LR

Tips for making your published data about Arabidopsis genes more FAIR (Findable,Accessible, Interoperable and Reusable)

Owned by Leonore Reiser
Last updated: Oct 16, 2019 · 1 person viewed · Page numbering · Legacy editor

To ensure that your Arabidopsis gene information is accessible here are some helpful pre-publication guidelines for authors. These guidelines here are written to be specific for Arabidopsis genes, but are based on a general set of guidelines and principles (see references). This makes it easier for TAIR curators to curate your published data (which will increase its visibility) and for other researchers to find and reuse your data.

Include AGI Locus Identifiers for genes

To ensure that the genes described in your paper are unambiguously identified, include the systematic locus identifier for that locus. If you have identified a new gene that does not yet have a an AGI locus ID, please **contact TAIR curators PRIOR to publishing your gene.** TAIR, and other resources, use text mining to associate publications to biological entities in databases (e.g. genes and proteins). In the absence of a unique identifier such as an AGI locus code or UniProt ID, text mining software cannot distinguish between CCR1 and CCR1. To ensure that your published data can be curated and accurately linked to a database record, use the AGI locus ID.

Do not reuse gene symbols

To avoid the problem of different genes being referred to by the same symbolic name, before you publish check to see if the name is in use. Check the gene symbol registry at TAIR, and search PubMed, ePubMed, Google Scholar to see if that symbolic name is in use for another Arabidopsis gene. If the name is not in use, please register the gene symbol.

Follow gene, protein and allele nomenclature standards

There is an established nomenclature for genes, proteins and alleles for *Arabidopsis thaliana*. For example, alleles are lowercase and distinguished by a dash and number (abc2-1, abc2-2). Again, check and make sure that the name is not already in use.

Naming and re-naming T-DNA insertion lines

There is a difference between a stock/germplasm and the specific T-DNA insertion that is causal for a phenotype. Therefore when referencing an allele please include the specific polymorphism and not just the name of the ABRC/NASC stock because many ABRC/NASC T-DNA stocks contain multiple insertions. Before naming your allele, check to see if it already has a name in TAIR or in the literature. At TAIR, we will update the allele name with the newly published name. If you have an allele/polymorphism/phenotype that does not already exist in TAIR, please send us the information.

Use standard formats for reporting data (if available)

To ensure that data is reusable and interoperable (e.g. can be integrated across platforms/tools) community based initiatives have been established to define data standards. For example the Minimal Information about any Sequence MIxS standards or Minimal Information about a Plant Phenotyping Experiment (MIAPPE) standards. If you are not sure that a standard exists, check the FAIR Sharing website or contact a curator.

Include metadata

Metadata is data about your data that is necessary to understand and reuse that data. Metadata may be as simple as a read me file explaining file contents. Metadata about experiments may include things like stock IDS for sample germplasms, descriptions of environmental /experimental conditions that use controlled vocabularies. Without metadata , your data won't be reusable.

Submit your data to an appropriate repository

To ensure that your data is findable and reusable, you should submit it to a permanent repository. Many journals will specify the preferred repository for specific data types such as sequence, expression data, proteomic data, etc... If no data type specific repository exists there are generic repositories such as Dryad and Zenodo that will house your data and give it a digital object identifer (DOI). Contact us if you are unsure where your data should go.

Publish data in machine readable formats

Machine readable means that the data can be 'read'/ingested by a computer program such as a script that will import a tab delimited file. Exporting an Excel table into a PDF format will render the data inaccessible to machines especially if the PDF is an image. Remember to use standard formats and include metadata.

References:

Wilkinson, M., et al., (2016) The FAIR Guiding Principles for scientific data managment and stewardship. Scientific Data. DOI:10.1038/sdata.2016.18

Reiser, L., et al., (2018) FAIR: A Call to Make Published Data More Findable, Accessible, Interoperable, and Reusable. Molecular Plant.

DOI:10.1016/j.molp.2018.07.005

https://bit.ly/3BT0FPY

(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!

Feedback is welcome

Let us know if you want to join our working group

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)