FAIR Scientific Literature (FSL) Working Group Update

AgBioData Community Workshop
4/30/2024

Members (meeting since 1/23/23, generally biweekly)

Ruth Isaacson
GSA

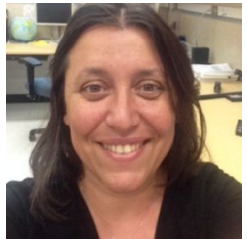
Leyla Cabugos
Librarian, Cal Poly

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

Daniel Morris , Professor (advisory)

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.

Stakeholders: Challenges and barriers

Researcher

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

Publisher

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

Research Librarian

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

Funder

- Not knowing where data should wind up
- 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

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

Research Librarian

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

Funder

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

Data should go in Community Databases, but....

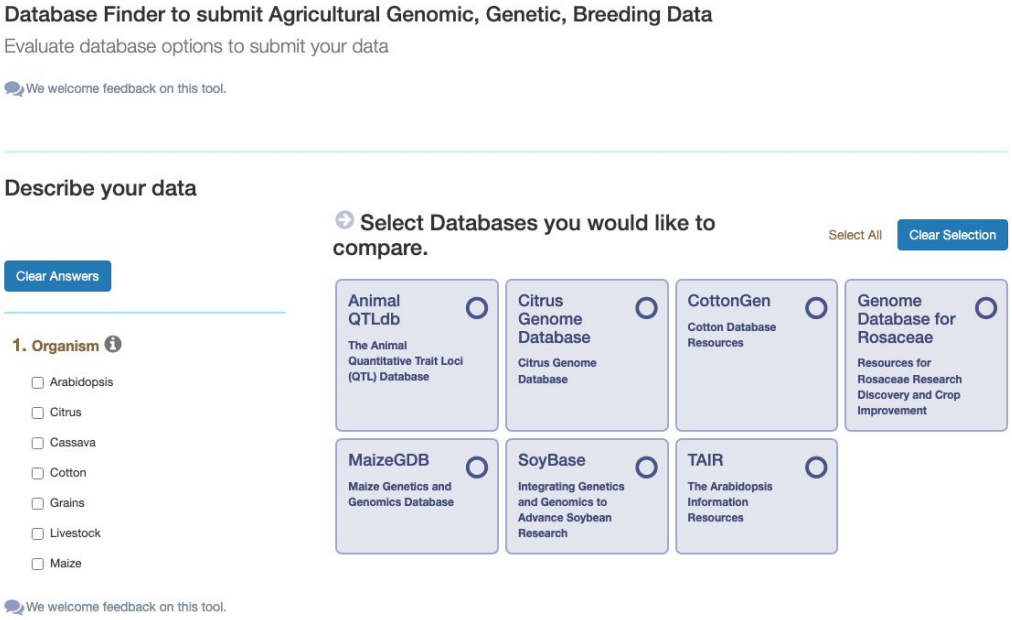
- **There may be more than one Database for a community**
  - Wheat Data can be found in Triticeae toolbox, GrainGenes, and Gramene
- **All Databases do not host the same data types**
  - Triticeae toolbox → SNP, phenotypic, pedigree
  - GrainGenes → genomes/tracks, images, maps, curated data
- **Only 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)

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 allow community submission or do they only take data from other repositories?**
- **What about data that should go into NonCommunity 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?**
- **Is there already a tool that already does the trick?**

Database Finder Tool (PROTOTYPE): where should authors put (or find) their data

https://dev.fruitandnutlist.org/finder



Database Finder Tool (PROTOTYPE): where should authors put (or find) their data

- Should live on the AgBioData site and contain information for all member databases

- Needs developer support
  - Drupal model - needs to be updated to AgBioData Site Drupal version
  - Not easy to enter values and thus new databases
  - Some features are not needed as well as there are a couple features that would be nice

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

- ✓ **There may be more than one Database for a community (or none at all)**
  - Wheat Data can be found in Triticeae toolbox; GrainGenes; Gramene
- ✓ **All Databases do not host the same data types**
  - Triticeae toolbox -> SNP, phenotypic, pedigree; GrainGenes -> genomes/tracks, images, maps, curated data
- ✓ **All 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

Stakeholders: Challenges and barriers

Researcher

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

Publisher

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

Research Librarian

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

Funder

- ~~Not knowing where data should wind up~~
- 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?

- 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
  - Help with recommendations for data to go into other repos (NCBI, Genome warehouses, ENA, EVA)
  - Help databases build modules for submitting data (Education Working Group?)
- Assess AgBioData Consortium Databases for FAIRness
  - Develop Best Practices -> FAIR Data Practices (maizegdb.org) https://www.maizegdb.org/FAIRpractices

Breakout

Tool Feedback

Log in and try the tool (5 min)

- How easy is it to use?
- Should we add AgBioData member databases that do not accept data directly from authors? We could still display what data types are available in the comparison table.
- Should we add non-member databases such as NCBI to direct users to submit raw sequences, etc?
- What else needs to be included in the tool.
- How well does the Database Finder Tool address the needs of the community?

Other Questions

1. How can this be supported for development?
2. What should we prioritize next?
   a. Work with each database to help guide datatype submission in a consistent manner
      i. Create a similar choice tool for each database that focuses on the datatypes they accept
      ii. Help with recommendations for data to go into other repos (NCBI, Genome warehouses, ENA, EVA)
      iii. Help databases build modules for submitting data (Education Working Group?)
   b. Assess AgBioData Consortium Databases for FAIRness
      i. Develop Best Practices -> FAIR Data Practices (maizegdb.org) https://www.maizegdb.org/FAIRpractices
   c. Data formats for harmonization

Tool Name (Slido, or poll through zoom)

- ❏ DB Finder (DataBase Finder)
- ❏ FAIR Traffic Control