Data sharing and data federation in genetics, genomics and breeding databases - Current status, needs and future directions

Monica Poelchau, Stephen Ficklin, Rie Sadohara, Peter Selby, Taner Sen, Andrew Farmer, Jennifer Clarke

Award Abstract # 2126334
The Data Federation Working Group

- **Chair**: Jennifer Clarke - Professor of Statistics, Professor of Food Science and Technology, UNL
- **Co-Chair**: Andrew Farmer - VP Research, NCGR
- Olusola Afuwape - Post Graduate Student, University of Lagos
- Justin Elser - Research Associate, Oregon State
- Stephen Ficklin - Associate Professor, WSU
- Andrew Olson - Computational Science Manager, CSHL
- Maria Palombini, IEEE
- Monica Poelchau - Geneticist, USDA-ARS National Agricultural Library
- Rie Sadohara - Researcher, MSU & University of Minnesota
- Peter Selby - BrAPI Project Coordinator, Cornell University
- Taner Sen - GrainGenes Lead Scientist, USDA-ARS
Movement of Data into and among Databases

User-generated data

Structured data

Databases describe and structure data (Biocuration)

Data is shared among databases

Databases provide data access

Scientists generate and describe data

Data is shared among databases

Databases provide data access

Structured data

Structured data

Structured data

Databases provide data access
Data Sharing Spectrum

Slide credit: Peter Selby
Data Federation Working Group Questions

• What is the current status of AgBioData on the data sharing spectrum?
• What level of data sharing among databases does AgBioData 1) need and 2) want?
AgBioData data sharing assessment goals

1. **Current data sharing.** Assess the level of data sharing that AgBioData member databases *have*. What data and metadata are being shared, and how they are being shared?

2. **Desired data sharing.** Assess the level of data sharing that AgBioData members *want* - is there a discrepancy between existing and desired data sharing for databases and their stakeholders?

3. **Barriers towards data sharing.** Determine barriers towards advancing to the desired data-sharing level; and

4. **Technology awareness.** Gauge AgBioData members’ level of awareness of data sharing technologies, and need for or interest in training.
AgBioData data sharing assessment - Methodology

- 20 questions aligned to the above goals
- Sent out to AgBioData database members in July 2022 via Google form
- Responses were collected through August 2022
33 responses from individuals representing 38 databases or resources (out of 42 AgBioData member databases)
Results – current level of data sharing

Does your database share data with other databases, systems, or tools?

- No
- No other systems consume it
- Yes, shared only with specific tools
- Yes

32 responses

- No: 3 (9.4%)
- No other systems consume it: 21 (65.6%)
- Yes, shared only with specific tools: 7 (21.9%)
- Yes: 1 (3.1%)

Do you import, link, or share data programmatically from another database?

- No
- Yes

23 responses

- No: 3 (13.0%)
- Yes: 20 (87.0%)
Results – current level of data sharing

What data types do you share with other databases?

Number of responses: 29
Results – current level of data sharing

What mechanism(s) do you currently use for sharing?

Number of responses: 29
4 respondents don’t use machine-readable methods
AgBioData data sharing assessment goals

*Current data sharing.* A wide variety of data is being shared. Many but not all databases provide machine-readable access to at least some of their data.
Results – desired level of data sharing

How important is it to **you** to make your database more discoverable and available?

How important is it to **your user community** to make your database more discoverable and available?

How high is it in your **development priorities** to make your database more discoverable and available, given the financial and time cost associated with it?

Number of responses: 33

![Bar chart showing responses to the questions about the importance and development priorities regarding database discoverability and availability.](chart)
Results – desired level of data sharing

What types of data do you wish you could share more effectively from your database?

What types of data do you wish you could access from other databases?

Number of responses: 13
Desired data sharing. There is still the need for increased data sharing, in particular for phenotypic data. Seven out of 38 databases stated that phenotypic and phenomic data are still challenging to share. Increased stakeholder awareness may help.
What do you feel are the biggest blockers to successful data sharing in your community?

Number of responses: 32
AgBioData data sharing assessment goals

**Barriers towards data sharing.** Time and resources, funding and ROI, lack of data standards, and technical knowledge were the main identified barriers.
Results – Technology awareness

What data sharing technologies are you familiar with?

What data sharing technologies would you appreciate learning more about?
AgBioData data sharing assessment goals

*Technology awareness.* There are multiple areas of technology that the AgBioData community could learn more about.
Recommendations

• **Identify solutions to funding problems.** Funding (62.5%) was cited as a main barrier to data sharing. The AgBioData’s sustainability working group may help provide solutions to funding problems for databases.

• **Data sharing training for database personnel.** Technical knowledge (47%) was also a substantial barrier to data sharing, and specific areas were identified as training priorities. AgBioData has initiated a new working group focused on data federation training (contact AgBioData if you’re interested in joining!)

• **Stakeholder education on the benefits of data sharing.** Promoting an understanding of data sharing and discoverability in the user/stakeholder community may help divert resources towards improved data sharing. The new training working group may also approach this.
Recommendations

• **Focus on improvements to phenotypic data sharing.** Phenotypic data represents an extremely diverse class of data. AgBioData should prioritize improvements for specific phenotypic data types and formats in future working groups. We are planning on implementing a phenotypic data working group in year 3 of the RCN grant.

• **Continue work on standards improvement.** Lack of data standards (50%) was cited a substantial barrier. Identification, promotion, or development of data standards should be prioritized (see the previous AgBioData GFF3 working group).
Thank you for your attention!

Questions?

Award Abstract
# 2126334