AgBioData Consortium

2022 Survey of Genomic, Genetic, and Breeding (GGB) Database Team Members

Summary of Baseline Data

v1.2 (7/22/2022)

Survey Sample, Participant Characteristics, and Familiarity and Experience with GGB Databases

A total of 25 usable survey responses were received during Spring 2022, with complete or mostly complete responses to the substantive questions that were asked after the initial self-descriptors at the beginning of the survey. Sample sizes for specific questions may vary slightly since a few people may have left particular questions unanswered. In some of the following tables, reported percentages may not sum to 100 due to rounding.

The responding group of database team members was not a random sample of all team members, since the invitation to participate was circulated widely and those who responded were self-selected, i.e. they chose to participate. We can't know the extent to which this group may be representative of all team members, so we can't confidently generalize the responses of this group to the entire population of GGB database team members. We can, however, be confident that we know with reasonable clarity what this group of 25 GGB database team members reported about their experiences, opinions, and recommendations with regard to these databases.

Over half of respondents (60%) reported working in a university that offers related PhD degrees; 8% reported working in a Land Grant University State Agricultural Experiment Station (see Table 1). About a sixth of respondents (16%) reported working for USDA and 24% reported working for a non-profit organization. Others reported working in a primarily undergraduate institution (4%), a university offering related Masters but not PhD degrees (4%), or for other types of institutions. As displayed in Table 2, most reported their primary professional role as being a biocurator (32%), software developer (20%), research scientist (16%), bioinformatics professional (16%), or technical staff member (8%).

Most respondents (76%) reported a professional focus on plants, including major plant crops (20%), horticultural specialty crops (16%), and plants grown for other purposes besides human consumption (12%); see Table 3. Animals were a primary focus for 12% of respondents, including major livestock animals (4% of respondents). Pests, diseases, physiological stressors, and other threats were a focus for 20% of respondents,; 24% reported working on understanding wild organisms not directly used in agriculture. Respondents could indicate more than one area of focus.

As displayed in Table 4, more than 75 percent of participants "strongly" or "moderately" agreed that they are very familiar with the concept of FAIR data and could explain this to others (84%), and similarly familiar with FAIR data management practices (76%), and how to make data accessible to researchers (76%). More than 50 percent gave similar ratings to their familiarity with how to make data reusable (72%), how to make data easily findable (60%), and how to make data interoperable (60%).

Table 1. Organizational Affiliations of Survey Respondents

Which terms best describe your organization? *(mark all that apply)*	
16%	US Department of Agriculture (USDA)
8%	Land Grant University State Agricultural Experiment Station (SAES)
60%	University offering related PhD degrees
4%	University offering related Masters (but not PhD) degrees
4%	Primarily undergraduate institution (PUI)
0%	Minority-serving institution (MSI)
0%	Historically Black College or University (HBCU)
0%	Private company or industry organization
24%	Nonprofit organization
0%	Other organization

Table 2. Primary Professional Role of Survey Respondents

Please indicate your primary role:	
32%	Biocurator
20%	Software developer
16%	Research scientist
16%	Bioinformatics professional
8%	Technical staff member
0%	Post-doc
0%	Graduate student
0%	Undergraduate student
12%	Other: *"Principal Investigator" "Project manager and biocurator"*

Table 3. Professional Focus of Survey Respondents

Please indicate your primary area(s) of professional focus:	
76%	Plants
12%	Animals
20%	Major plant crops grown as food for people on a large percentage of farm land (e.g. maize, soybeans, rice etc.)
16%	Horticultural, specialty, or other plant crops grown as food for people on a smaller percentage of farm land (e.g. vegetables, fruits, nuts, etc.)
12%	Plants grown for animal feed, fiber, lumber, industrial use, or ecosystem services/cover crops
4%	Major livestock animals grown as food for people on a large percentage of ranch/farm land (e.g. beef or dairy cattle, pork, chicken, turkey)
0%	Minor livestock animals such as sheep or goats; fish, shellfish or other aquatic animals raised as food for people; honey bees, etc.
24%	Wild organisms not directly used in agriculture (as potential germplasm sources for agriculture or for understanding biology and ecosystems independent of potential agricultural applications)
20%	Pests, diseases, physiological stressors, other threats to agriculture or ecosystems
12%	Other: *"Forestry" "Information technology" "Ontologies"*

Table 4. Team Member Familiarity with Implementation of FAIR Data Practices

Based on your knowledge of and experience with FAIR data and GGB databases, please rate how much you agree with these statements:	Strongly Disagree	Moderately Disagree	Slightly Disagree	Slightly Agree	Moderately Agree	Strongly Agree
I am very familiar with the concept of FAIR data (Findable, Accessible, Interoperable and Reusable) and could explain this to others.	4 %	-	12	-	40	44
I am very familiar with FAIR data management (what metadata to collect, how to use standardized names and methods, etc.) and could supervise implementation of these practices.	4	4	8	8	52	24
I am very familiar with how to make data easily findable for others.	4	-	12	24	24	36
I am very familiar with how to make data accessible to researchers.	4	-	4	16	48	28
I am very familiar with how to make data reusable.	4	8	4	12	36	36
I am very familiar with how to make data interoperable.	4	-	16	20	44	16
Please add comments, extensions, clarifications or recommendations related to any of the questions above: ▪ *I am familiar with current available options, which are not always fully FAIR themselves.* ▪ *There are still many data formats that are limited in interoperability.*						

Note. N = 25. Each row of numbers contains the proportion (percentage) of participants who gave each response to the question. Row percentages may not add to exactly 100% due to rounding.

Baseline Ratings and Recommendations On Implementing FAIR Data Practices in GGB Databases

GGB database team members were asked to rate their level of agreement with a series of statements about their experiences, observations and opinions regarding the current status of FAIR data practices in GGB databases. They were also asked to rate their priorities for future development of FAIR data practices in these databases, and to provide related comments and written recommendations. The distributions of their responses are displayed in Tables 5 – 8; some highlights are noted below.

In some cases, ratings on multiple survey questions will eventually be combined into composite index scores, which will provide more reliable, psychometrically useful baseline measures to compare with follow up data that will be collected near the end of the current project. This step has not yet been done for this report; such indices will not be very useful until follow up data are available for formal statistical comparisons, but some notes about this are included below.

The first three questions displayed in Table 5 ask about baseline perceptions of the extent to which the GGB databases provide good guidelines for users on some key FAIR data practices. The highest ratings were given to guidelines related to file formats for preparing and submitting data, while the lowest ratings were given to guidelines on how to provide information about how the data have been generated and cleaned, how missing data has been handled, and other process transparency issues.

Response patterns to the other questions in Table 5 may provide insight into other relative strengths and weaknesses of GGB databases as perceived by team members. Relatively high ratings, for example were given to the use of cost-sharing efficiencies such as open source software, the availability of learning resources for users, the quality of germplasm collections data curation, the consistent use of standard nomenclatures, and the use of common file formats. Meanwhile, relatively low ratings were given for documentation of how the data have been generated and cleaned, how missing data has been handled, and other process transparency issues; interconnectedness and interoperability between related databases; and ease of finding and combining related but not previously integrated data on organisms of interest.

Although the distributions of responses to most of these questions are somewhat skewed, with a majority of respondents already reporting some level of agreement, there is room for improvement (greater agreement), so hopefully these measures will be sensitive enough to detect any differences at the follow up. (Though the follow up survey will include a different group of respondents, so any interpretation of differences from baseline to follow up should be carefully framed in that context.) It appears that the 2022 data has reasonable psychometric properties and will serve as a useful baseline measure for the project, though the small sample size is concerning and may limit the confidence with which any inferences can be made about change over time as the project unfolds.

There are various ways to scan these ratings to identify which areas appear to be most ripe for improvement. We could add the two (or three) right-most columns to produce a quick sum that would show those items with the highest ratings (GGB database strengths), or conversely, add the two (or three) left-most columns to produce a quick sum that would show those items with the lowest ratings (weaknesses). For now, a quick color-coding scheme has been applied to give some indication of how this might be interpreted; darker blue indicates relative strengths, darker red indicates relative weaknesses that could be important targets for improvement. Participants were offered a chance to make open-ended comments on the issues, but none were received.

Table 5. Baseline Appraisal of the Status of FAIR Data Implementation in GGB Databases

	Based on your knowledge of and experience with FAIR data and your GGB database, please rate how much you agree with these statements:	Strongly Disagree	Moderately Disagree	Slightly Disagree	Slightly Agree	Moderately Agree	Strongly Agree
C1	The GGB databases I work for provide good guidelines on what metadata to provide when preparing/submitting data.	-	4 %	8	33	21	33
C2	The GGB databases I work for provide good guidelines on what file formats to use when preparing/submitting data.	-	4	4	17	21	54
C3	The GGB databases I work for provide good guidelines on how to provide information about how the data have been generated and cleaned, how missing data has been handled, and other process transparency issues.	-	17	8	38	33	4
C4	The GGB databases I work for employ cost-sharing efficiencies such as reusable open-source software.	4	-	4	8	38	46
C5	The GGB databases I work for provide useful resources for learning how to use them: tutorials, FAQs, how-to documents and videos, etc.	-	-	12	12	56	20
C6	GGB databases I work for are presently well curated.	-	8	8	8	48	28
C7	The GGB databases I work for highlight the importance of FAIR data principles for researchers and provides educational resources to help researchers understand and follow FAIR practices.	4	8	12	32	24	20
C8	GGB databases I work for currently include thorough, up to date collections and documentation of genetic and genomic information of target organisms.	-	4	4	16	52	24
C9	GGB databases I work for currently include thorough documentation of how the data have been generated and cleaned, how missing data has been handled, and other process transparency issues.	-	4	28	20	40	8
C10	Standard nomenclatures are used consistently by the databases I work for, making it easier for researchers to find relevant data.	-	4	16	4	36	40
C11	Metadata is used consistently by the databases I work for.	-	-	13	29	42	17
C12	Common file formats are used consistently by the databases I work for to make it easier to share and integrate data from different sources.	-	4	8	12	28	48
C13	Tools and processes now used for contributing data to my databases are easy to use and facilitate accurate, efficient, thorough uploading of needed data and metadata according to FAIR principles.	4	13	-	25	42	17
C14	Tools and processes for finding and retrieving data from my databases are currently easy to use and facilitate accurate, efficient transfer of all needed data and metadata according to FAIR principles.	4	4	8	29	38	17
C15	At this point in time, reliable interconnectedness and interoperability between related databases make it easy to combine and integrate data in new ways to address new questions.	12	8	28	32	4	16
C16	Currently, attempts to find and combine related but not previously integrated data on organisms of interest are rarely blocked by technical difficulties (incompatible differences in curation, data structure etc.)	12	16	28	12	12	20

Note. N = 22 to 25. Each row of numbers contains the proportion (percentage) of participants who gave each response to the question. Row percentages may not add to exactly 100% due to rounding.

Baseline Priorities for Further Development of FAIR Data Practices in GGB Databases

Survey participants were asked to rate the importance of five potential priorities for improved data curation; their responses are summarized in Table 6. All five were rated as being "very important" or "highest priority" by more than half of respondents.

The highest ratings were given to application of community standards by publishers and enforcement of data submission requirements," which was rated as "highest priority" by 50 percent of respondents and as "a very important priority" by another 46 percent of respondents.

Most of the other priorities were rated as being "very important" or "highest priority" by more than two thirds of respondents. Training and support on FAIR data for database users received the lowest priority ratings from these team members, with 58 percent rating this topic as being "very important" or "highest priority."

Additional comments on these priorities are listed verbatim in Table 7.

Table 6. Baseline Priorities for Further Development of FAIR Data Practices in GGB Databases

Please indicate what you believe to be the most pressing priorities for improved data curation:		Not a Priority for Funding or Development	A Minor Priority	A Somewhat Important Priority	A Very Important Priority	Highest Priority for Funding or Development
D1	Timely and up-to-date availability of curated data	4 %		17	50	29
D2	Application of community standards by publishers and enforcement of data submission requirements	4			46	50
D3	Professional incentives from funders (e.g. enforcement of compliance with DMPs, credit for compliance with DMPs, inclusion of metrics for data sharing in evaluation)	4		29	25	42
D4	Training material for FAIR data (for database professionals)	8	4	17	42	29
D5	Training and support on FAIR data for database users	8	4	29	25	33

Table 7. Comments on Priorities for Further Development of FAIR Data Practices in GGB Databases

Please add any comments about your answers above or your recommendations for priorities. What are the most important avenues for improving the GGB databases you are aware of?
▪ *Fostering more opportunities for dialog between database users and maintainers; greater encouragement for interoperability between separately managed resources.* ▪ *Incentives and support for scientists contributing data, funding for curation on the database side.*

Recommendations for Topics and Formats of Training Opportunities for Users of GGB Databases

Survey participants were asked an open-ended question about "what sort of training opportunities/formats or content/topics for users of these databases would be most helpful?" Their verbatim answers are listed in Table 8. Color coded themes that were developed based on the comments from stakeholders (in their separate survey) are listed below; these have been applied in Table 8, since several of these themes are represented in the team member comments and no new themes emerged from their comments.

- Synchronous (live) educational events online such as webinars or asynchronous online video presentations, demonstrations or tutorials (these are coded together because after being recorded, events like webinars can be cut into segments and made into brief asynchronous, "static" video recordings. It is often helpful to design and organize webinars or similar online presentations with this segmentation and re-use in mind.) (6 comments)
- Static online educational materials such as tutorials or manuals (5 comments)
- Online courses – again, these can be synchronous and/or asynchronous or static, and an initial synchronous event or event series can later be posted online as an asynchronous or static learning resource. (No comments from team members, theme retained from stakeholder survey comments.) (Note that curriculum materials developed for an online or face to face course can then be repurposed and provided to instructors for use in their own courses; curriculum and assistance for course instructors was specifically mentioned in some comments.)
- Face to face workshops, stand alone or in conjunction with conferences or meetings (1 comment.)
- Synchronous "office hours" or asynchronous "discussion forums" in which people can ask questions and get timely advice and assistance (No comments from team members, theme retained from stakeholder survey comments.)
- The importance of regularly updating any static online tutorials, manuals, or similar materials so that they reflect the current status of the databases they reference, even as those databases are updated. (1 comment, theme retained from stakeholder survey comments)

As noted in the stakeholder survey summary, in some cases it may be difficult to classify specific comments into adjacent themes, for example, a comment about "workshops" could mean online or face-to-face workshops. It may be best to combine the first two themes since the boundary between an online event and a later offering of the recorded version of that event is fuzzy and perhaps not important. The counts presented above might have some error but the themes themselves may still be useful, especially those that were also present in the stakeholder survey comments. In addition, there are some individual comments that were not echoed by others and thus did not become part of a theme, but that may be worth considering, so it may be a good idea for readers to review all of the verbatim comments to find worthwhile suggestions or revise their personal interpretation of the themes above.

The comments and the coded themes above were dominated by format issues, e.g. how and where to deliver the topics and content. However, the comments in Table 8 also contain a few recommendations for specific topics and content that would be useful. These are repeated here for ease of viewing:

- *How to submit data.*
- *Why FAIR data helps individual researchers and science overall, and how community databases are essential to FAIR data.*
- *User-driven topics of interest.*

Table 8. Recommendations for Training of GGB Database Users

What sort of training **opportunities/formats** or **content/topics** for users of these databases would be most helpful? (e.g. more static online tutorials or manuals, videos, live webinars, online courses over a period of weeks, face to face workshops, training on specific topics, etc.)
▪ *How to submit data, via online tutorials and in person short workshops at conferences. Undergraduate and graduate curriculum - why FAIR data helps individual researchers and science overall, and how community databases are essential to FAIR data.*
▪ *Live webinars.*
▪ *More static online tutorials or manuals.*
▪ *Live webinars featuring user-driven topics of interest.*
▪ *Videos, online hands-on workshops to encourage user learning and participation.*
▪ *Detailed and searchable manuals.*
▪ *Online tutorial and manuals.*
▪ *Live webinars, videos that are current.*
▪ *Clearly-written online tutorials and manuals, live webinars/online training to supplement.*
▪ *Online tutorials are great because they are available when needed.*

At the end of the survey, participants were asked to offer any last comments or recommendations. No further comments from database team members were received.