Sustainability Survey of AgBioData member databases

Prepared by Sabarinath Subramaniam, Director of Business Development, Phoenix Bioinformatics and Josh Young, Executive Director, Phoenix Bioinformatics

1. Introduction

1.1 About this report

This report was prepared by Phoenix Bioinformatics as partial fulfillment of the deliverables within NSF Award # 2126334, "RCN: Reimagining a Sustainable Data Network to Accelerate Agricultural Research and Discovery". Information for this report was gathered by Phoenix Bioinformatics between January 2022 and November 2022, with assistance of principal investigators of various AgBioData member databases. The final version of this report was submitted on Dec 19, 2022.

1.2 Disclaimer

The findings described in this report are highly dependent on the accuracy of the information provided to Phoenix Bioinformatics by AgBioData member databases and any errors in the underlying data will require correction before taking any recommended action.

1.3 About Phoenix Bioinformatics

Phoenix Bioinformatics was founded in 2013 as a nonprofit 501(c)3 organization. Our mission is to assist scientific data repositories and other research cyberinfrastructure components in developing innovative and sustainable funding support mechanisms to ensure their long-term sustainability. We pioneered our novel approach with TAIR, the Arabidopsis Information Resource, a widely used plant genome database, and were successful in replacing grant funding with a $1.1M/yr revenue stream from users without significantly impacting usage of the resource. Since that early success we have been working in partnership with a range of scientific resources to assist them in finding sustainable revenue streams. Current partners include BioCyc, a set of over 13,000 microbial genome and metabolic pathway databases; AgBase, a database of functional information for several agriculturally important plants and animals; and Repbase, a database of genomic DNA repeats and transposable elements.

2. Background

2.1 Genomic, Genetic and Breeding Databases GGB Databases

Genomic, Genetic and Breeding (GGB) databases serve and respond to research and breeding stakeholder communities to provide value-added curated data and tools that meet stakeholder needs. To ensure that researchers continue to have access to reliable, high quality, curated, and FAIR data in the future, GGB databases need to plan and develop infrastructure, strategies and tools to ensure long term sustainability of GGB data and GGB Databases. The AgBioData consortium (https://www.agbiodata.org) has agricultural biological databases with the mission of consolidating standards and best practices for acquiring, displaying, and reusing genomic, genetic, and breeding (GGB) data. Formed in 2015, the consortium involves 40 GGB databases and over 200 members, including database curators, researchers, librarians, and anybody that works with agricultural data.

2.2 AgBioData member Databases

Table 1 (Appendix 7.1) shows the list of AgBase member databases that was submitted with the RCN grant proposal. We excluded VectorBase because the resource is no longer part of the AgBioData consortium; TAIR and CyVerse, because they have implemented a sustainability plan based on subscriptions; and Araport because the database has become defunct with the tools and data now being hosted by other databases. Of the 40 AgBioData member databases, the survey was sent to 25 PIs representing 36 databases.

2.3 Rationale for sustainability efforts

Sustainability of GGB Databases and Resources has emerged as an important issue. Most GGB Databases rely on short-term funding for a majority of their operating costs, and are vulnerable to loss of personnel and knowledge if funding lapses, even while demand for their services by researchers continues to increase. Financial uncertainty and gaps in funding not only hampers efficient operation, but also limits long-term planning, potentially resulting in higher costs for data access. Loss of database funding can also lead to permanent loss of valuable data and software gathered and built at taxpayer or industry expense, and slows the progress of research . Also, while there is a need to ensure that the data for new genomes is made accessible in a timely and cost-effective manner, it is simply not feasible nor desirable to create a GGB Database for every species. Consolidated and standardized database resources are needed.
To plan for the future of GGB Databases and data gather feedback from member databases and stakeholder communities via surveys and assess different sustainability options ranging from support for individual database projects to federated /cost sharing models that can be applied across many databases.

3. RCN Aim 4.1: Self-assessment of the long term financial stability of member databases

This study focuses on achieving the goals set within Aim 4.1 of the RCN grant. This involves gathering data through written surveys of staff at all 36 AgBioData member databases. Our surveys are intended to capture cost of operations, staff level, sources of funding, usage level, data types, species and strains, stakeholders served and anticipated future needs. We will collect information on each GGB Database’s view of its sustainability and approaches to improve that sustainability. This data will provide us with both a picture of the current funding situation and the anticipated future needs.

4. Stakeholder surveys and interviews: Principal Investigators

4.1 Methods

We used the Survey Monkey platform (surveymonkey.com) to create an online survey whose link we emailed to principal investigators of a subset of all AgBiodata databases (Table 2, Appendix 6.2). Since the survey and interview questions were presented to AgBioData resource Principal Investigators, an IRB review was deemed not necessary. Principal investigators who did not respond to the surveys were contacted for phone interviews.

The survey was designed to make it easy to provide us with information to help us understand the following:

1. General information on each database
2. Database Content
3. Database funding and expenses
4. Database staffing
5. Database users
6. Past sustainability strategies and planning
7. Shared data, tools and resources

5. Survey results

5.1 Database content

This section asked for general information about each of the databases. Please note that the information presented here has been collected from surveys and phone interviews. So the data reflects only the 19 respondents, not the entirety of AgBioData member databases.

5.1.1 Number of years for which each database has been active

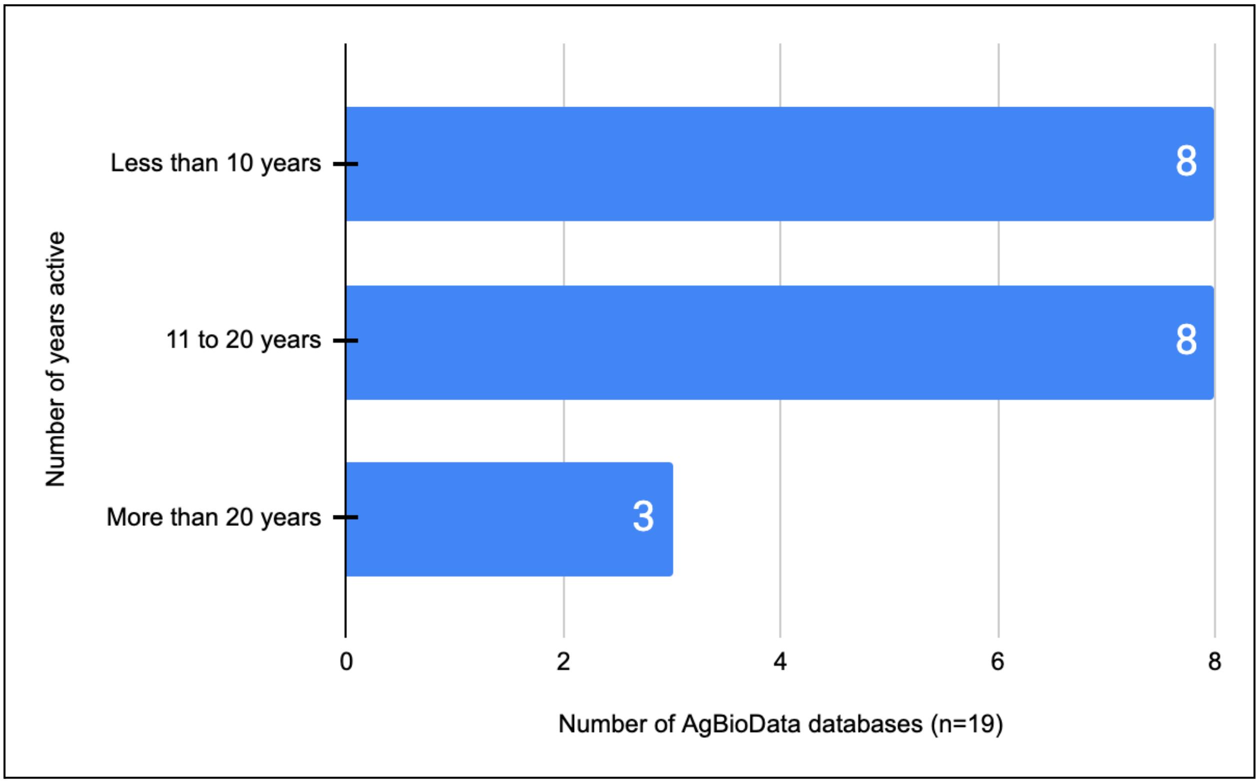

Figure 1. Distribution of the number of years each database has been active. n refers to the total number of responding databases.

5.1.2 Frequency of data update

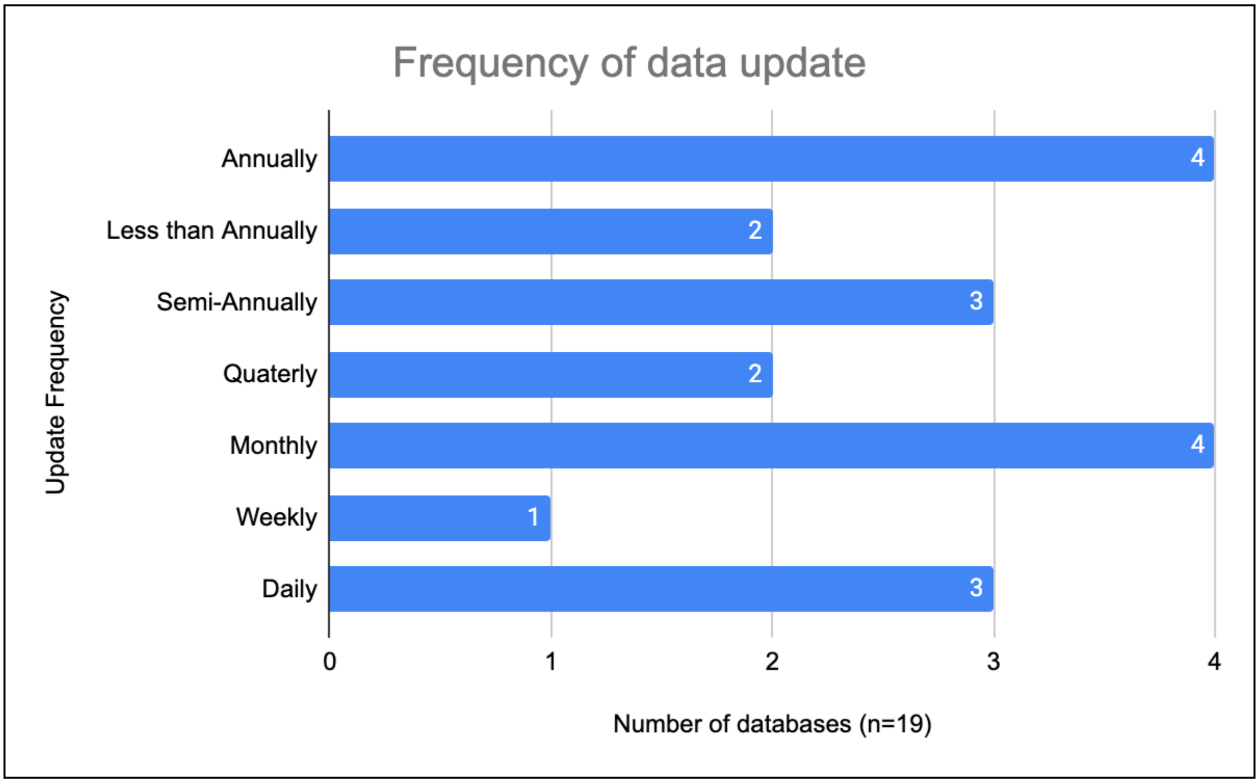

Figure 2. Data shown above collected from the survey question: How often do you update your data? n refers to the total number of responding databases.

5.2 Database Funding, Expenses and stakeholders

5.2.1: Funding security for next 3 years

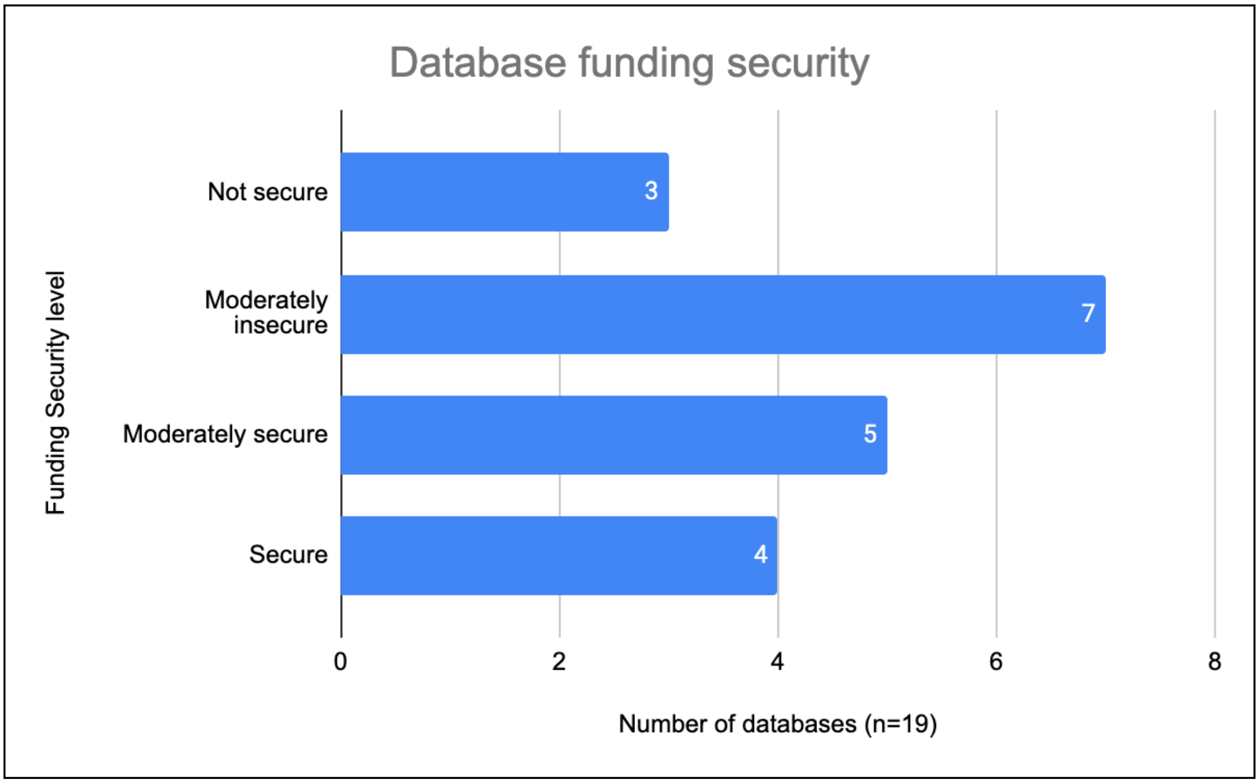

Figure 3. Data shown collected from the survey question: How would you rate the current level of funding security for your database(s) for the next 3 years? n refers to the total number of responding databases.

5.2.2 Funding sources for each database for past three years

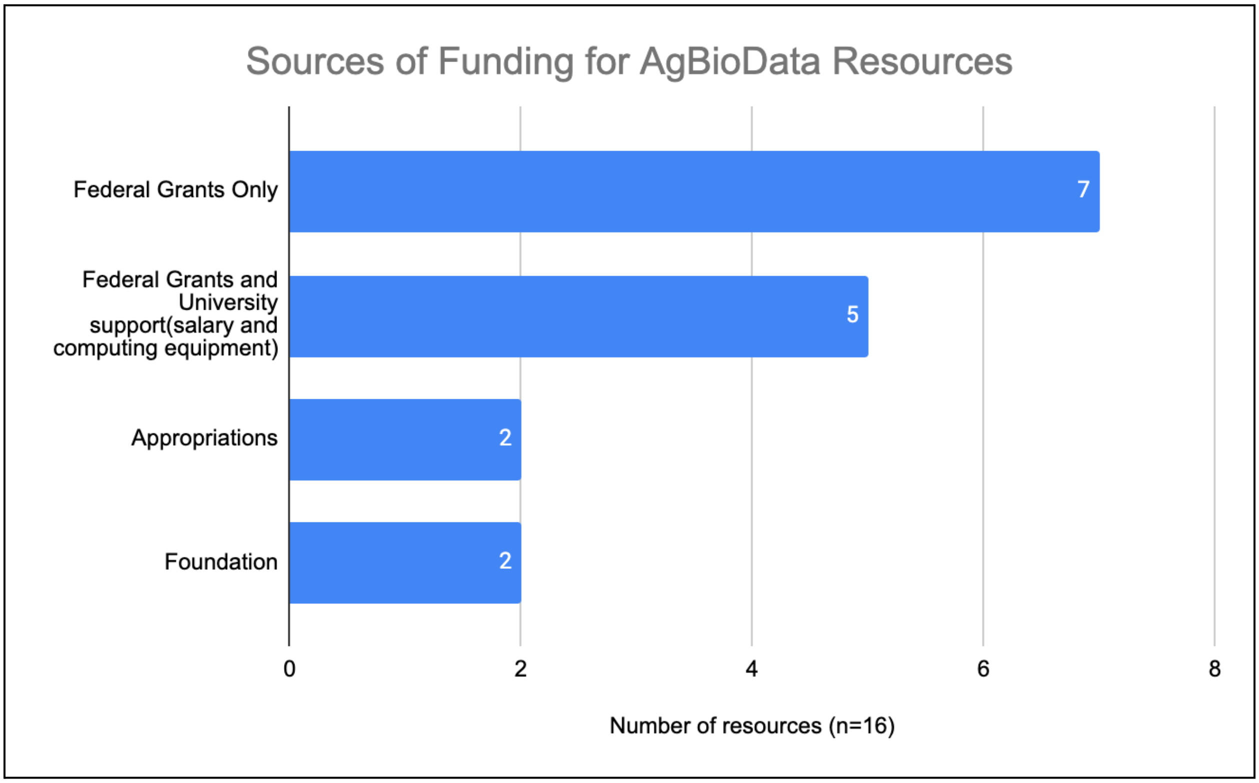

Figure 4. Our survey asked each database to list sources of funding for the next three years. 16 AgBioData databases provided an answer to this question.

5.2.3 Anticipated change in expenses over the next 3 years.

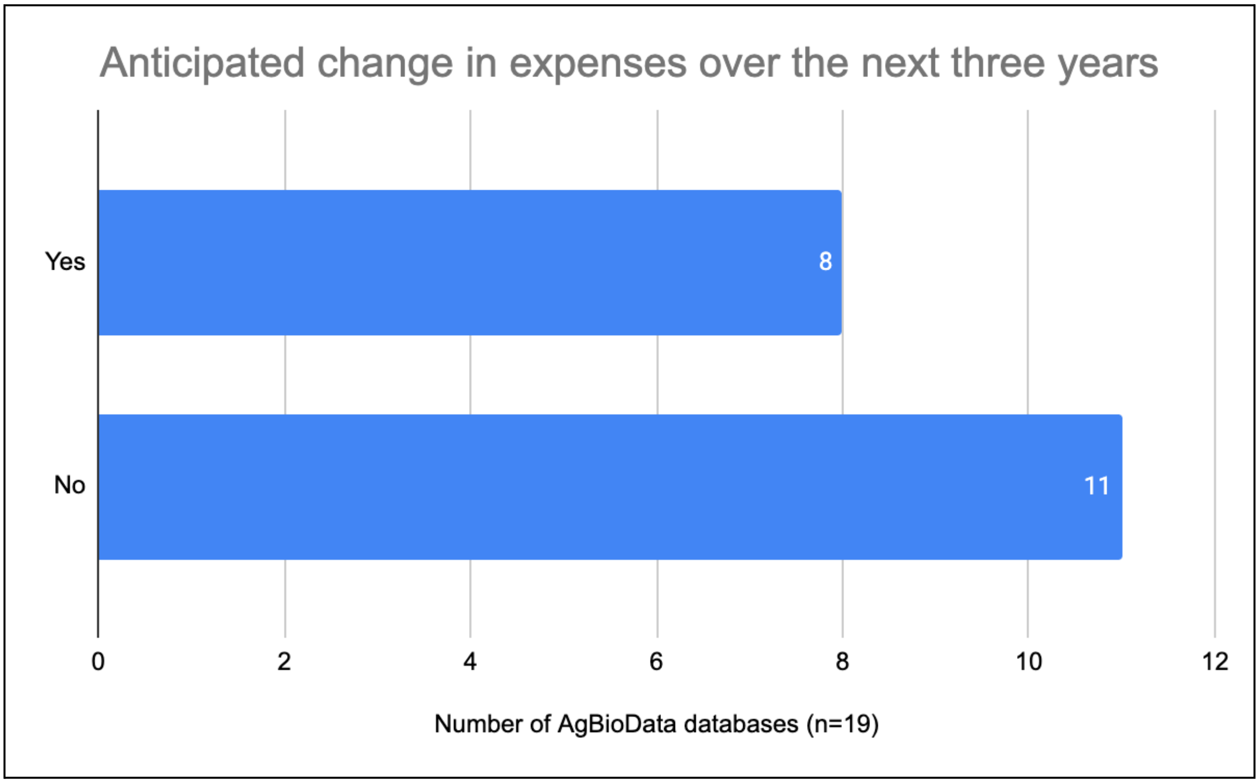

Figure 5. Responses to the question "Do you anticipate your annual expenses changing significantly over the next 3 years?" No directionality was requested (i.e. increase or decrease). n refers to the total number of responding databases.

5.3 Stakeholders of AgBioData databases

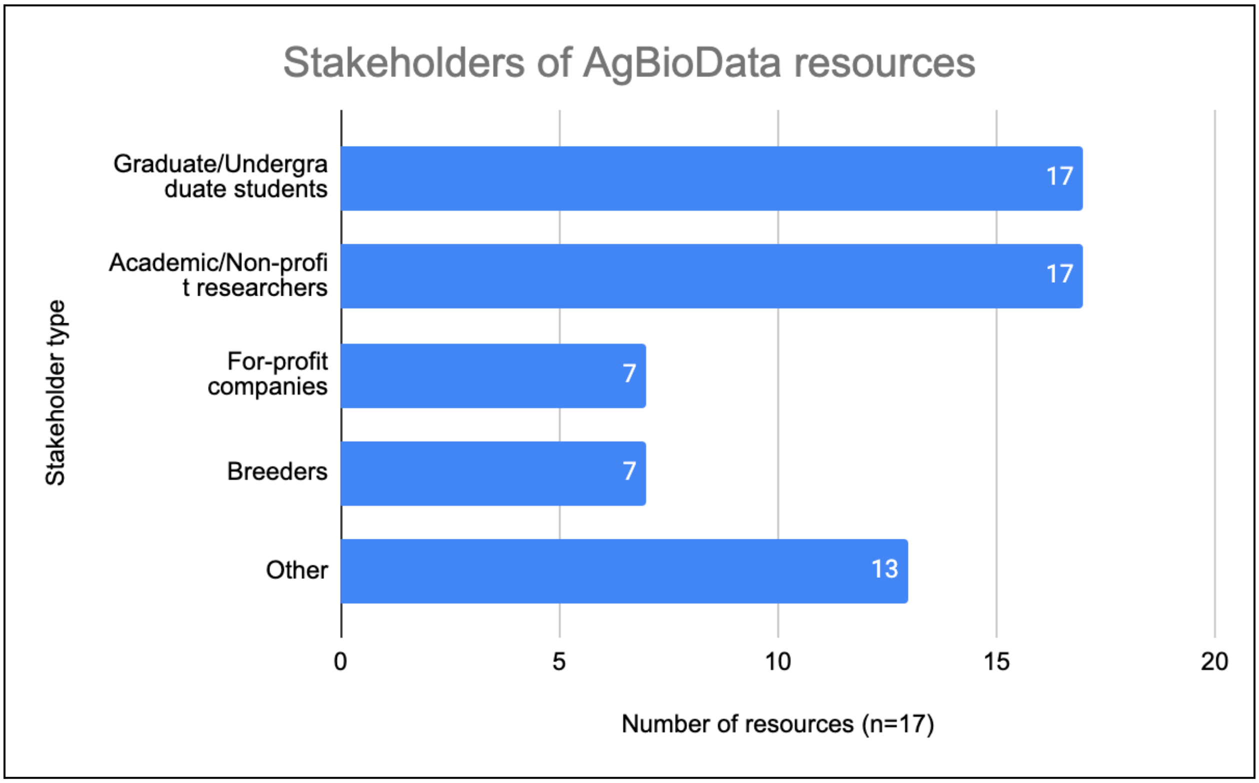

Figure 6. The “Other” category of stakeholders include K-12 students/educators (3 databases), Crop boards (3 databases), Regulatory officials (2 databases), Seed testing officials (2 databases), Intellectual property application examiners (2 databases), Federal government researchers (1 databases) and USDA-ARS administrators (1 databases). n refers to the total number of responding databases.

5.4 Sustainability strategies and placing

5.4.1 Usage data capture mechanism and User surveys

Having a mechanism to capture usage statistics is a very useful tool to identify sustainability options. Our survey asked whether the AgBioData resource had a mechanism for usage capture, such as Google Analytics. Out of the 19 databases that took the survey, 17 databases provided a response to this question. Out of 17 respondents, 16 databases indicated that they already have a mechanism for capturing data usage, while 1 resource said they do not.

Understanding the value that the data/tools within the resource represents to the users of the resource is very important to judge the willingness of the users to support a sustainability model for the resource. Educating the users about the value of the resource as well as bringing their attention to the potential consequences if the resource were to disappear are some of the key aspects of a user survey. We asked AgBioData databases if they have conducted user surveys in the past 2 years. Out of the 19 databases that returned the survey, 17 databases provided a response to this question. 11 databases indicated that they have conducted user surveys, while 6 databases said they have not.

5.4.2 Acceptable sustainability model(s) for each AgBioData database

We surveyed AgBioData databases about the types of sustainable funding models the users of the resource would be willing to support. Our survey asked respondents to select all models that would be acceptable to the users of the resource, from the following:

- Shared infrastructure (Chado DB, Tripal, BrAPI)
- Database federation (e.g Alliance for Genome Resources model)
- Subscriptions
- Voluntary contributions

Users were also presented with another option “Other” and a text box to describe the choice. Of the 19 AgBioData databases that returned the survey, 14 answered this question. Of those, acceptable models included: Seven indicated either shared infrastructure, Database federation and/or Subscriptions as models acceptable for their users. Four indicated Shared infrastructure, Database federation and voluntary contributions. Two databases selected Shared infrastructure as the only option. These two databases also wrote the following under "Other": USDA and ARS are currently consolidating cloud operations to share infrastructure and reduce operational costs. One chose voluntary contributions as the only option.

5.4.4 Steps taken by resource to reduce annual expenses

We surveyed AgBioData databases to understand any steps taken by the resource to reduce annual expenses (Table1).

Cost reduction step	Number of databases (n=15)
Open-source software, shared teams, shared computing databases across our projects	9
Everything they use is open source	2
Reduced biocuration and number of database releases, machine learning, shared database build and backend database and tool development and centralized data management	2
Running server in-house (no cloud fees but fixed one- time cost) and open source software	1
Adoption of open source software, shared teams with special skills, and other kinds of databases shared with other projects	1

Table 1. Cost reduction steps implemented by AgBioData databases.

5.5 Data redundancy

5.5.1 Sharing data with other databases

Data sharing is one way to reduce costs. We surveyed AgBioData databases to identify If the database shares data with other AgBioData databases and asked respondents to list the names of the other databases. 14 of the 19 respondents responded to this question listing at least one database they share data with. We grouped these 14 answers into two categories: Databases that contain data pertaining to a specific organism (e.g. TAIR) or databases that are part of a larger data repository (e.g. Ensembl).

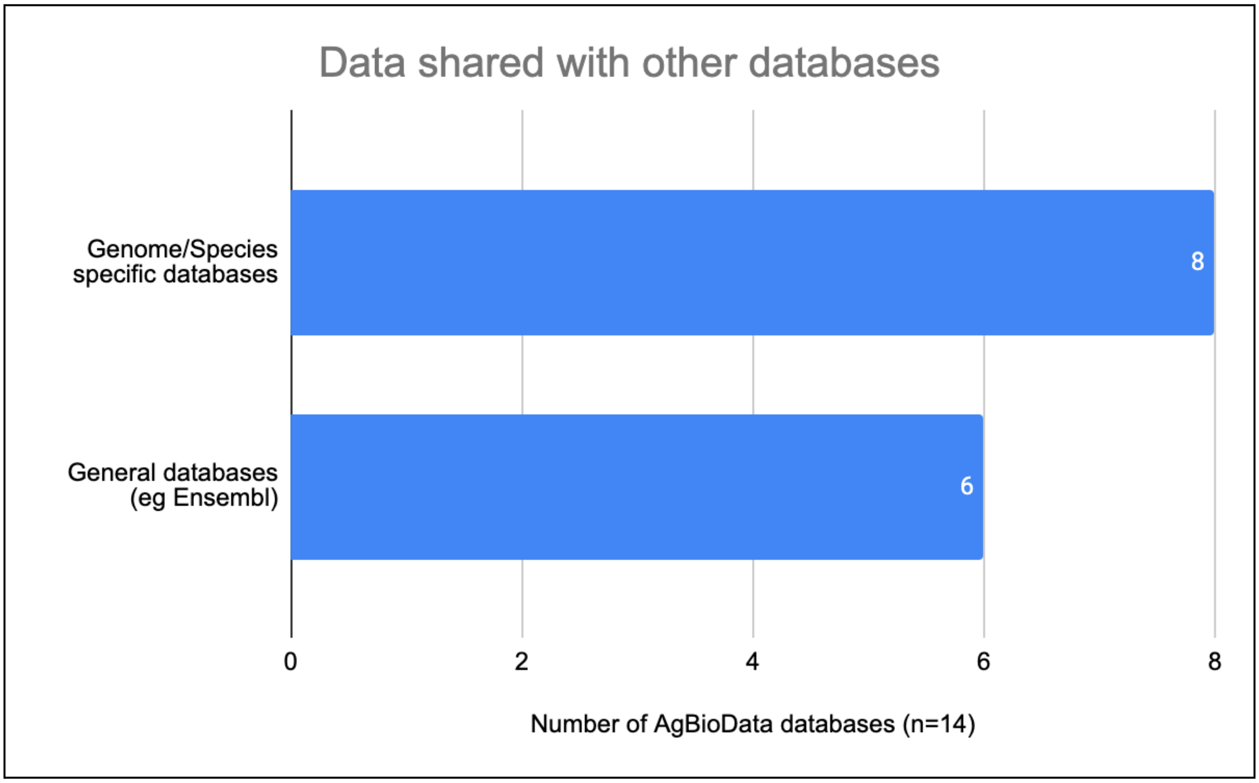

Figure 7. 8 of the respondents share their data with at least one other organism-specific database. 6 of the 14 respondents share their data with general data repositories. n refers to the total number of responding databases.

5.5.2 Data duplicated or archived in other databases

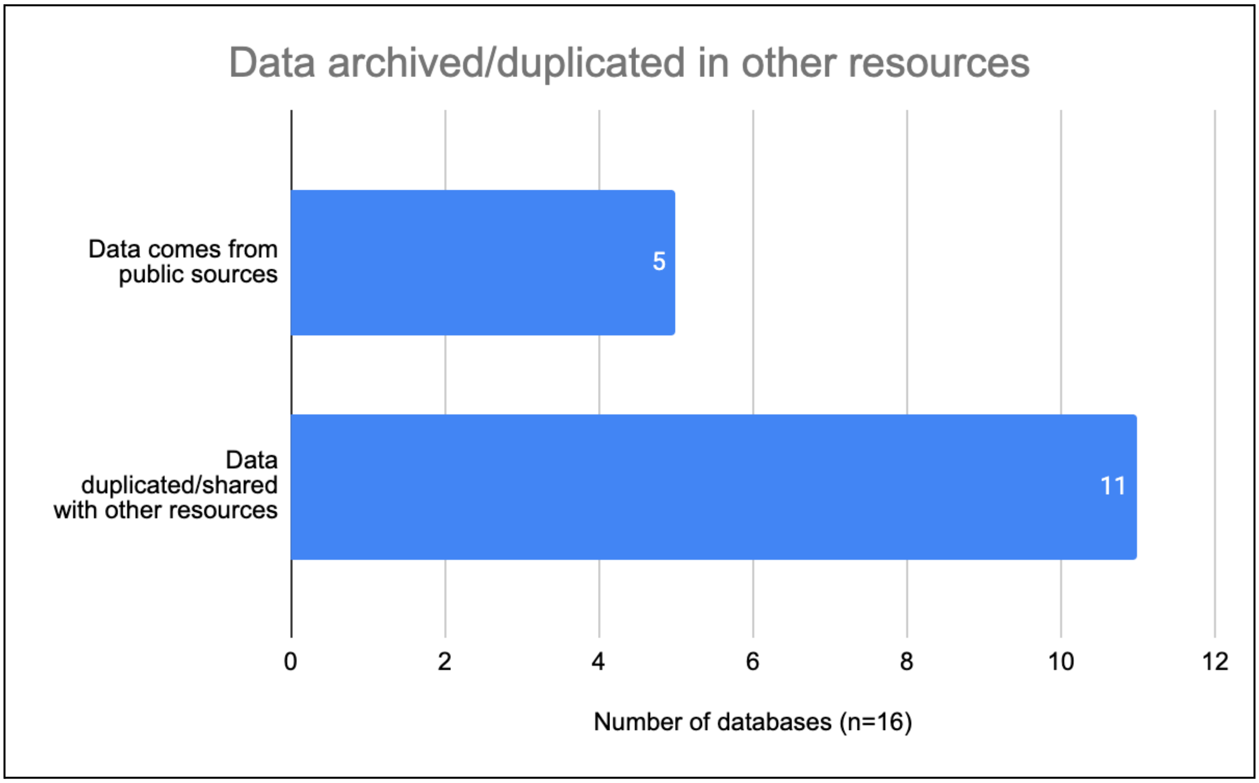

Figure 8. Our survey asked AgBioData databases to indicate if their data is archived or duplicated elsewhere. 16 of 19 respondents answered this question. 11 of the respondents have their curated data duplicated within another resource. The remaining 5 databases answered that they receive their data from external databases. n refers to the total number of responding databases.

6. Summary

Based on the results of our survey we would like to make the following recommendations to better understand the value of AgBioData databases among its users and to identify at least one viable sustainability strategy for each resource or the consortium as a whole.

Stakeholder surveys and interviews

Our survey identified multiple stakeholders for the respondent AgBIoData databases. Surveying or interviewing these stakeholders will give a clearer picture on stakeholder buy-in for any sustainability strategy and will help understand the value of each resource among its stakeholders.

Usage statistics and User surveys

15 databases have a mechanism to capture usage statistics. Identifying a viable sustainability strategy requires understanding user behavior and studying the usage statistics from the 15 databases would be a good starting point.

User surveys are a valuable tool in determining a viable sustainability strategy. Also 11 databases indicated they have conducted user surveys. These surveys would be a good start to understand user behavior. 14 out of 19 respondents identified at least one of our suggested sustainability models as appropriate for their user community. It would be valuable to do a user survey to gauge which of these strategies are acceptable for the users of these databases.

Data duplication/sharing

For databases that are sharing their data through public databases like Ensembl, it would be good to understand how their users access the data. eg: How many access the data from the primary resource versus the public repository?

5 databases said they get their data from public sources. Further discussion with these databases will help us understand what fraction of their data comes from public sources and is collected, curated and integrated. 11 respondents indicated their data is archived or duplicated in other databases.

Other

Our survey asked “Do you anticipate your annual expenses changing significantly over the next 3 years?” In hindsight we should have included elaboration on which direction (increase or decrease) the change was expected.

7. Appendices

7.1 Table 1. AgBioData consortium member focus and metrics. (Source: NSF RCN grant 2126334)

Database (citation) [age]	**Focus**	**# Users 2020**	**# Page Views 2020**	**# Citations 2015-2020**
AgBase (*16*) [19]	Animals, plants, microbes, parasites	3,500	15,514	428
Agroportal (*17*) [5]	Agronomy vocabularies and ontologies	5,000	44,775	122
Alfalfa Toolboxb [6]*	Alfalfa	-	-	24
Animal QTLdb (*13*) [32]	7 animal species	7,873	83,163	1,130
BGD (*18*) [18]	Cattle	2,793	21,520	108
CassavaBase (*19*) [11]	Cassava	5,939	54,479	84
CGD [12]	Citrus crops and pathogens	9,004	100,889	126
Citrusgreening (*20*) [5]	HLB, Asian citrus psyllid, citrus crops	4,018	11,999	46
CottonGen (*21*) [11]	Cotton species	21,215	321,846	477
CuGenDB (*22*)	Cucurbit crops	19,662	783,854	362
GDR (*23*) [19]	Rosaceae (apple, peach, etc.)	31,150	1,138,573	1,210
GDV [11]	Blueberry, cranberry	5,465	66,694	42
GrainGenes (*24*) [27]	Wheat, barley, rye, oats	39,093	475,268	1,560
Gramene (*25*) [22]	57 plant species	56,254	3,253,557	2,650
GRIN [46]	Crops and wild relatives	245,924	3,491,489	5,100
HWG(*26*) [14]	Forest trees and woody plants	6,009	37,868	126
HGD (*27*) [18]	hymenopteran insect species	1,462	24,103	139
i5K NAL (*28*) [8]	Arthropods	12,031	66,509	195
KitBase*	Rice cultivar KitaakeX	-	-	4
LIS (*29*) [17]	Legume species	18,000	294,000	218
MaizeGDB (*12*) [31]	Maize	84,291	1,837,055	1,530
MusaBase (*30*) [8]	Banana	3,002	23,715	12
PeanutBase (*31*) [22]	Peanut species	16,000	293,292	300
Planteome (*32*) [23]	Crops and other plants	10,683	38,061	84
PulseDB	Pulse crops	8,845	45,650	5
SGN (*19*) [20]	Solanaceae (tomato, etc.)	79,333	1,200,000	2,390
SoyBase (*33*) [23]	Soybean	30,609	746,200	1,270
SweetPotatoBase [8]	Sweet potato	28,803	1,810	10
T3 (*34*) [3]	Barley, oat, wheat	7,655	183,326	257
TAIR (*35*) [23]	Arabidopsis	487,973	13,514,068	8,450
TreeGenes (*36*) [27]	Forest trees	5,589	32,249	77
VectorBase (*37*)	Arthropod disease vectors	48,101	625,080	1,030
WheatIS*	wheat	-	-	50
YamBase (*38*) [8]	Yam	1,810	28,803	14
	Totals	1,307,086	28,855,409	29,630