Sustainability RoadMap for AgBioData databases

Sabarinath Subramaniam
Phoenix Bioinformatics
1. **Self-assessment of the long term financial stability of member databases**

2. **Conduct a detailed analysis and modeling of sustainability solutions for representative AgBioData member databases.**

3. **Develop a roadmap for Genomic, Genetic and Breeding (GGB) Database sustainability to ensure data persistence and resource longevity (Years 2, 3, 4)**
Self-assessment of the long term financial stability of member databases

Goals:
● Gather data through written surveys and interviews with staff at all member databases.
● Capture cost of operations, staff level, sources of funding, usage level, data types, species and strains, stakeholders served and anticipated future needs.
● Collect information on each GGB Database’s view of its sustainability and approaches to improve that sustainability.
● Understand the current funding situation and the anticipated future needs.
PI surveys

AgBioData databases

AgBioData database PIs
Goal 1: Identify stakeholders
Stakeholder surveys

Identifying all stakeholder groups.

Recognizing the value that the database represents to stakeholders.

Understanding stakeholder attitudes toward sustainability models.
Stakeholders of AgBioData resources

PI survey results

Stakeholders of AgBioData resources

- Graduate/Undergraduate students: 17
- Academic/Non-profit researchers: 17
- For-profit companies: 7
- Breeders: 7
- Other: 13

Number of resources (n=17)
Goal 2: Urgency of action
Current level of funding security

PI survey results

Urgency of action
Anticipated change in expenses over next 3 years

PI survey results
What happens to your data if your database disappears?

- External repository: 40.0%
- External source: 25.0%
- Local/Cloud Backup: 15.0%
- Dissappear entirely: 5.0%
- Data/tools lost: 15.0%

Urgency of action

Data persistence
Goal 3: Database-specific facts/stats
What more do we need?

- **Database size**
  - How much resources are needed to hold the data?
  - How can the data be organized/distributed more efficiently?
  - What opportunities are there to consolidate the data between databases?

- **Rate of data growth**
  - How much resource allocation would be needed to sustain the current rate of data growth?

- **Data content**
  - What are the types of datasets within each database?
  - What, if any, are opportunities for shared data content between databases?
  - What are new datasets/types that would add value to the database?
  - How much of your data is public?
  - How much of it is curated in-house?
  - How much of the data is imported from other resources?
Database-specific facts/stats

Number of years active

PI survey results

Number of AgBioData databases (n=19)

- Less than 10 years: 8
- 11 to 20 years: 8
- More than 20 years: 3
PI survey results

Database-specific facts/stats

Frequency of data updates

<table>
<thead>
<tr>
<th>Update Frequency</th>
<th>Number of databases (n=19)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annually</td>
<td>4</td>
</tr>
<tr>
<td>Less than Annually</td>
<td>2</td>
</tr>
<tr>
<td>Semi-Annually</td>
<td>3</td>
</tr>
<tr>
<td>Quarterly</td>
<td>2</td>
</tr>
<tr>
<td>Monthly</td>
<td>4</td>
</tr>
<tr>
<td>Weekly</td>
<td>1</td>
</tr>
<tr>
<td>Daily</td>
<td>3</td>
</tr>
</tbody>
</table>
Goal 4: Cost-sharing strategies
PI survey results

Steps to reduce expenses

Cost-sharing strategies

Cost-cutting measures undertaken

Open-source software: 10
Shared computing resources: 9
Shared data processing, pipelines and scripts: 8
Shared teams: 6

Number of AgBioData databases
Data shared with other databases

Cost-sharing strategies

PI survey results

Data shared with other databases

<table>
<thead>
<tr>
<th>Genome/Species specific databases</th>
<th>General databases (e.g., Ensembl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>6</td>
</tr>
</tbody>
</table>

Number of AgBioData databases (n=14)
Goal 5: User experience
Acceptable sustainability models

User experience

PI survey results

Sustainability models acceptable to users

- Shared infrastructure: 12
- Database federation: 8
- Subscriptions: 7
- Voluntary contributions: 4

Number of AgBioData databases
User experience

Usage capture mechanism

PI survey results

Databases with Usage capture mechanism

- Yes: 15
- No: 1

Number of AgBioData databases
User experience

Private data workspace

PI survey results

Do you hold private data for users (e.g. in a dedicated workspace?)

- **Yes**: 8
- **No**: 8
- **NA**: 2

Number of databases
What does the survey tell us?

- Never too early to plan for a sustainable future. 50:50 split in level of funding security. So it’s never too early to look into sustainable options.
- For a majority of databases the curated part of their data will no longer be available if the database disappeared completely. So sustainability solutions need to be considered with curated datasets in mind. Curated datasets are a great add in value to your database. So the more curated datasets you can add to your databases the better.
Recommendations

- **Mechanism to capture usage statistics**: If your database does not have a usage capture mechanism, implement one.
- **Have an authentication system**: This allows users to store private data, workflows, and analysis. Shared authentication system
- **Database participation is key.**
  - We need databases to be more responsive to our surveys.
  - The more data we have the more we can develop a solution that works for all AgBIoData databases as a group. Model depends on data.
  - If you prefer one on one conversations like an interview style, we will work with you on that.
● **User surveys**
  ○ Users: academic and commercial
  ○ Do you have any user surveys completed in the last 2 years?
    ■ Important to understand if the user mindset has changed.

● **Use Cases**
  ○ How do stakeholders utilize the data/tools within the database?
  ○ Where are opportunities for shared use cases?

● **Surveys of other stakeholders including:**
  ○ Employees
  ○ Crop commissions
  ○ Funding agencies

● Where are untapped opportunities for the data/tools within the databases?
1. Funding models:
   a. Voluntary membership
   b. Data deposit fees
   c. Subscriptions
   d. Freemium models
   e. Donations

2. Cost sharing:
   a. Shared curation or software teams
   b. Community Curation
   c. Shared infrastructure
Goals of the Sustainability Working Group

1. **New funding sources.**
   a. Assess the viability of various sustainability models
   b. Recommend sustainability models for AgBioData databases, either as a consortium or for individual databases.

2. **Cost-saving measures.**
   a. Determine effective strategies for cost sharing including sharing data storage, FTEs, Tools etc.
   b. Evaluate the idea of a centralized approach to resource management benefitting all member databases.

3. **Policy or procedure changes related to grant funding.**
   a. Increase the ability of databases to obtain a share of research funding that supports data generation.
   b. Examine the role that a central coordinating entity could play in assisting with sustainability.
Acknowledgements

AgBioData Steering committee
Annarita Marrano
Leonore Reiser
Monica Poelchau
Meg Staton
Jaqueline Campbell
Peter Harrison
Sook Jung
Sunita Kumari
John McNamara
Sushma Naithani

RCN NSF Award 2126334
Phoenix Bioinformatics Family

Tech Team
- Swapnil
- Trilok
- Xingguo
- Karthik

Sales
- Amin

Finance
- Connie

Science Team
- Tanya
- Leonore

Business Development
- Zach
- Shabari

Director
- Josh

Annarita
- Brooke
Contact information and PAG30 schedule

- YouTube: @phoenixbioinformatics8986
- Facebook: phoenixbioinformatics
- Twitter: @phoenix_bio
- Email: shabari@phxbio.org

Join us for our workshop on sustainability:

Data Resource Sustainability and Funding
Monday, Jan 16 4:00 PM @ Palm 7
EXPERTISE

Expert  Consulting  Knowledge  Teamwork  Advice  Trust  Research

CONSULTING

GOALS  KNOWLEDGE  EXPERT

SUCCESS  POTENTIAL  SUPPORT

PARTNERSHIP

TRUST  PERFORMANCE  TEAMWORK

SUCCESS  BUSINESS  PLAN

WIN-WIN  COLLABORATION
Partners, Sustainability consulting and Supported resources

- tair
-UtfC)
- GBE Base
- MOPH OBANK
- CIPRES
- CYVERSE
- USDA
- NRSP10
- AgBioData
- Persephone
- OpenMRI
- CoGe
- LPI MAPS
- ALZFORUM
- The MEME Suite
- Sol Genomics Network

Timeline:
- 2013
- 2016
- 2018
- 2019
- 2020
- 2021
- 2022
- 2023

Today
Principal Investigator survey goals

- Identify stakeholders
- Determine effective cost-sharing strategies
- Database-specific facts/stats
- Urgency of action
- User experience
Volume of data from scientific research is rapidly increasing

Data Volume + Increased requirements for access = Increased resources for data storage and analysis

Digital data resources (DDRs) allow for collaborative discovery

Biological data resource sustainability
Challenges to Digital Data Resource Sustainability

**FUNDING**
- DDRs are growing rapidly while funding flatlines.
- Grants focus on funding new research at the expense of saving prior data.
- Heavy reliance on a single source of funding, which is not sustainable.

**Data Management**
- Lack of data standards, Increasing computing costs, Rapidly increasing volume of data
- Making current data available.

**DDR community/ user culture**
- Cultural resistance to data sharing within their user community.
- Misconceptions that computing is free.
- Data sharing and management is typically not valued for career advancement.

**Unpredictable future**
- Understanding and predicting future sustainability challenges is daunting
- Future data sharing policies and funding need buy in from all stakeholders.
Working groups

- Sustainability
- Data Reuse
- Public Genetic Resources
- DEI_Recruiting
- FAIR Scientific Literature
- Data Federation Training
Urgency of action

Sources of funding for AgBioData databases

PI survey results

Sources of Funding for AgBioData Resources

- Federal Grants: 12
- Appropriations: 2
- Foundation: 1

Number of resources (n=15)
Data archived elsewhere

PI survey results

Data archived/duplicated in other resources

- Data comes from public sources: 5
- Data duplicated/shared with other resources: 11

Number of databases (n=16)