Data sharing and data federation in genetics, genomics and breeding databases

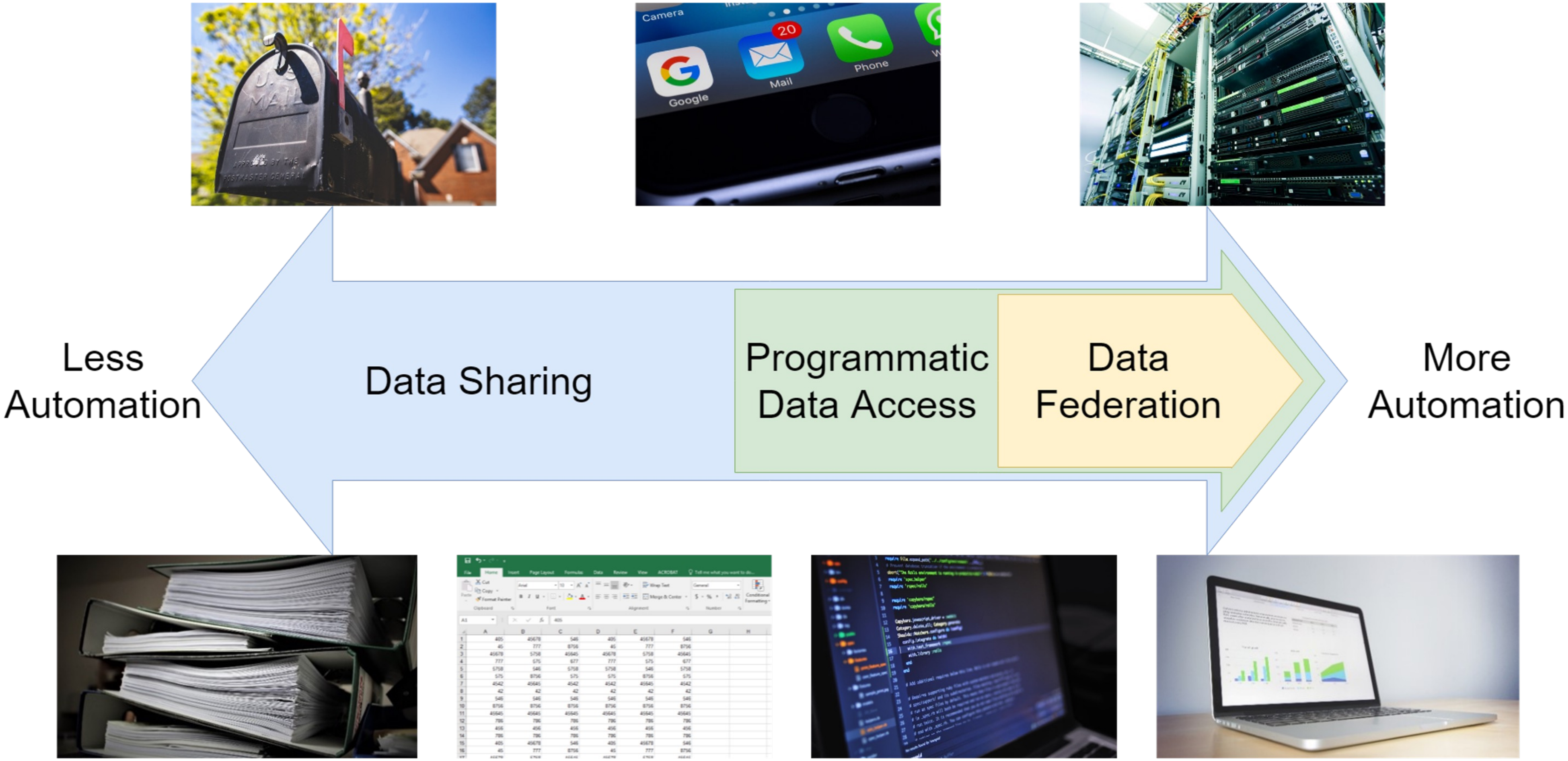
Monica Poelchau, Stephen Ficklin, Rie Sadohara, Peter Selby, Taner Sen, Andrew Farmer, Jennifer Clarke

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

Data Sharing Spectrum



Less Automation

Data Sharing

Programmatic Data Access

Data Federation

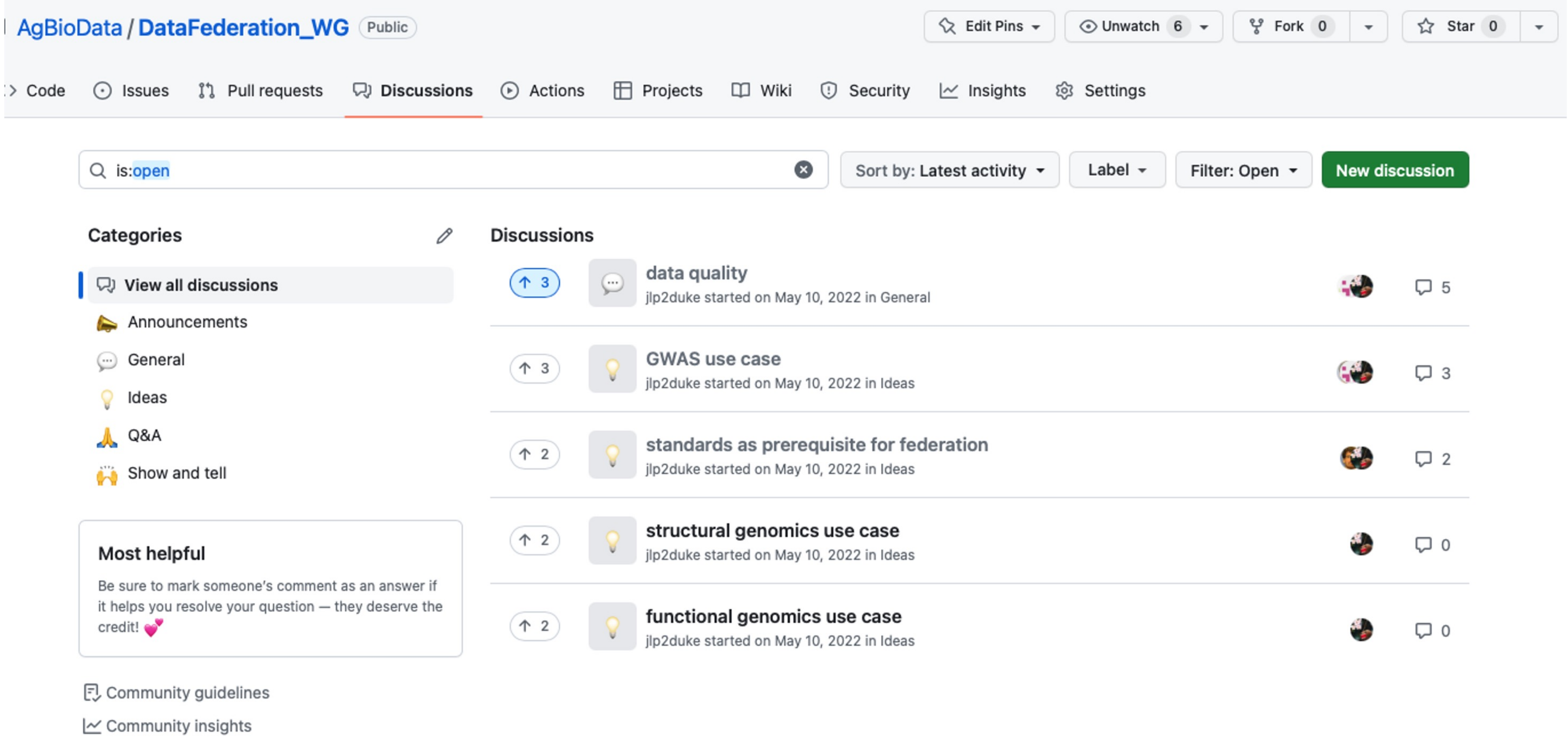
More Automation

Slide credit: Peter Selby

Data Federation Working Group Questions

- What use cases for data sharing among databases does the AgBioData community have?
- 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?

Data sharing use cases



Categories

- View all discussions
- 📣 Announcements
- 💬 General
- 💡 Ideas
- 🙏 Q&A
- 🙌 Show and tell

Most helpful

Be sure to mark someone's comment as an answer if it helps you resolve your question — they deserve the credit! 💞

Community guidelines

Community insights

Discussions

Votes	Discussion	Comments
3	data quality jlp2duke started on May 10, 2022 in General	5
3	GWAS use case jlp2duke started on May 10, 2022 in Ideas	3
2	standards as prerequisite for federation jlp2duke started on May 10, 2022 in Ideas	2
2	structural genomics use case jlp2duke started on May 10, 2022 in Ideas	0
2	functional genomics use case jlp2duke started on May 10, 2022 in Ideas	0

https://github.com/AgBioData/DataFederation_WG/discussions

AgBioData data sharing assessment goals

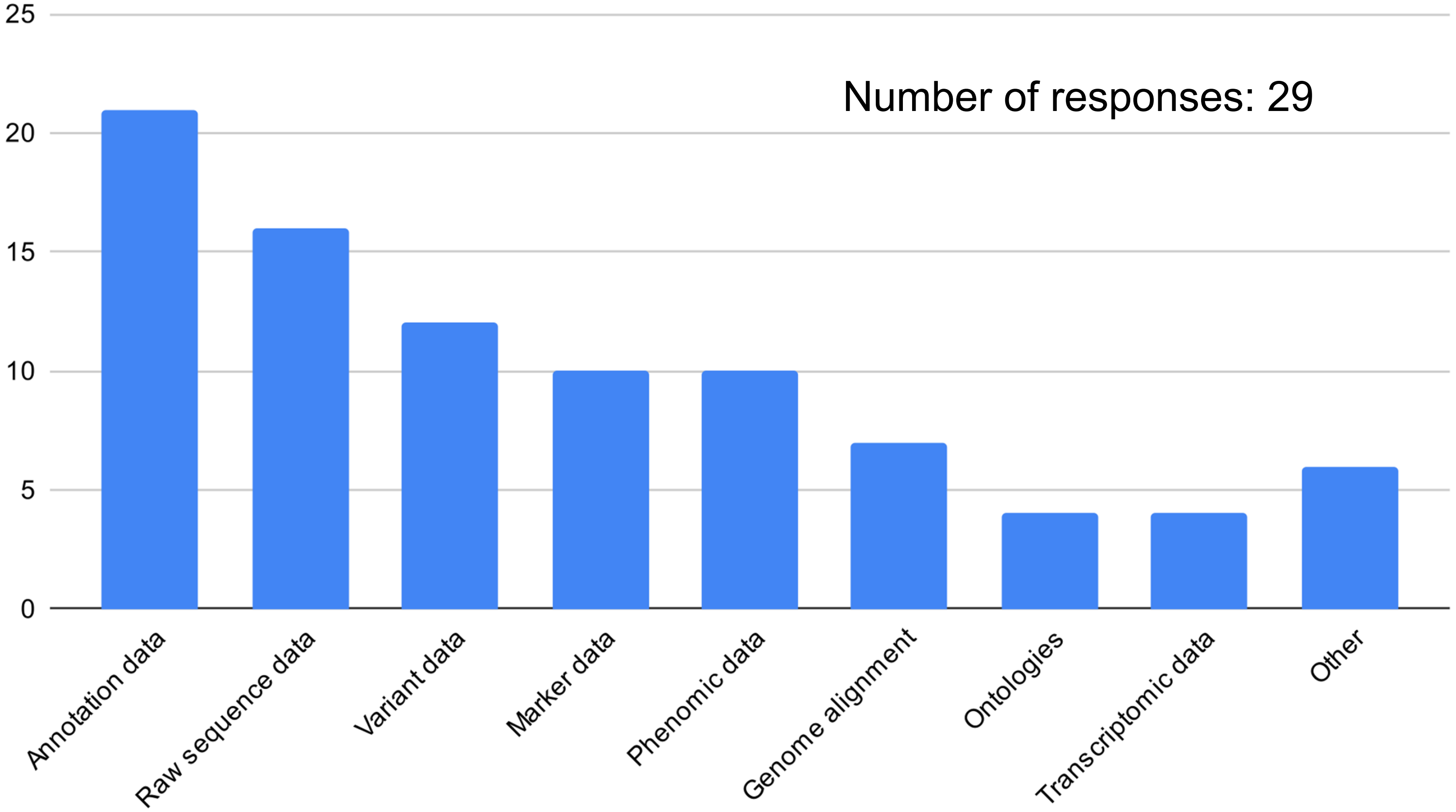
1. ***Current data sharing.*** Assess the level of data sharing that AgBioData member databases ***have***. What data and metadata are being shared, and how they are being shared?
2. ***Desired data sharing.*** Assess the level of data sharing that AgBioData members ***want*** - is there a discrepancy 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
- Collaborated with Ontologies working group

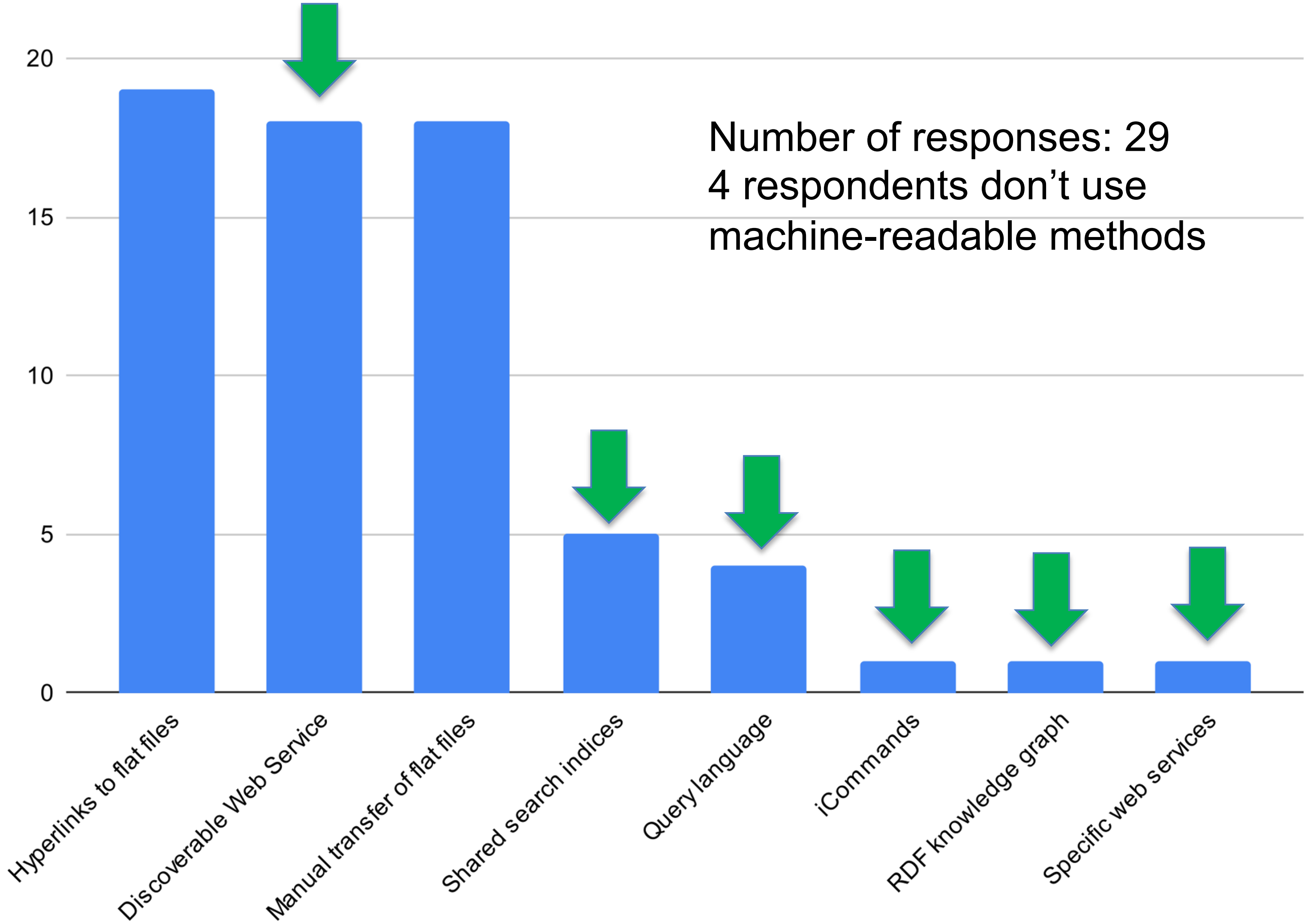
Results – current level of data sharing

What data types do you share with other databases?



Results – current level of data sharing

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



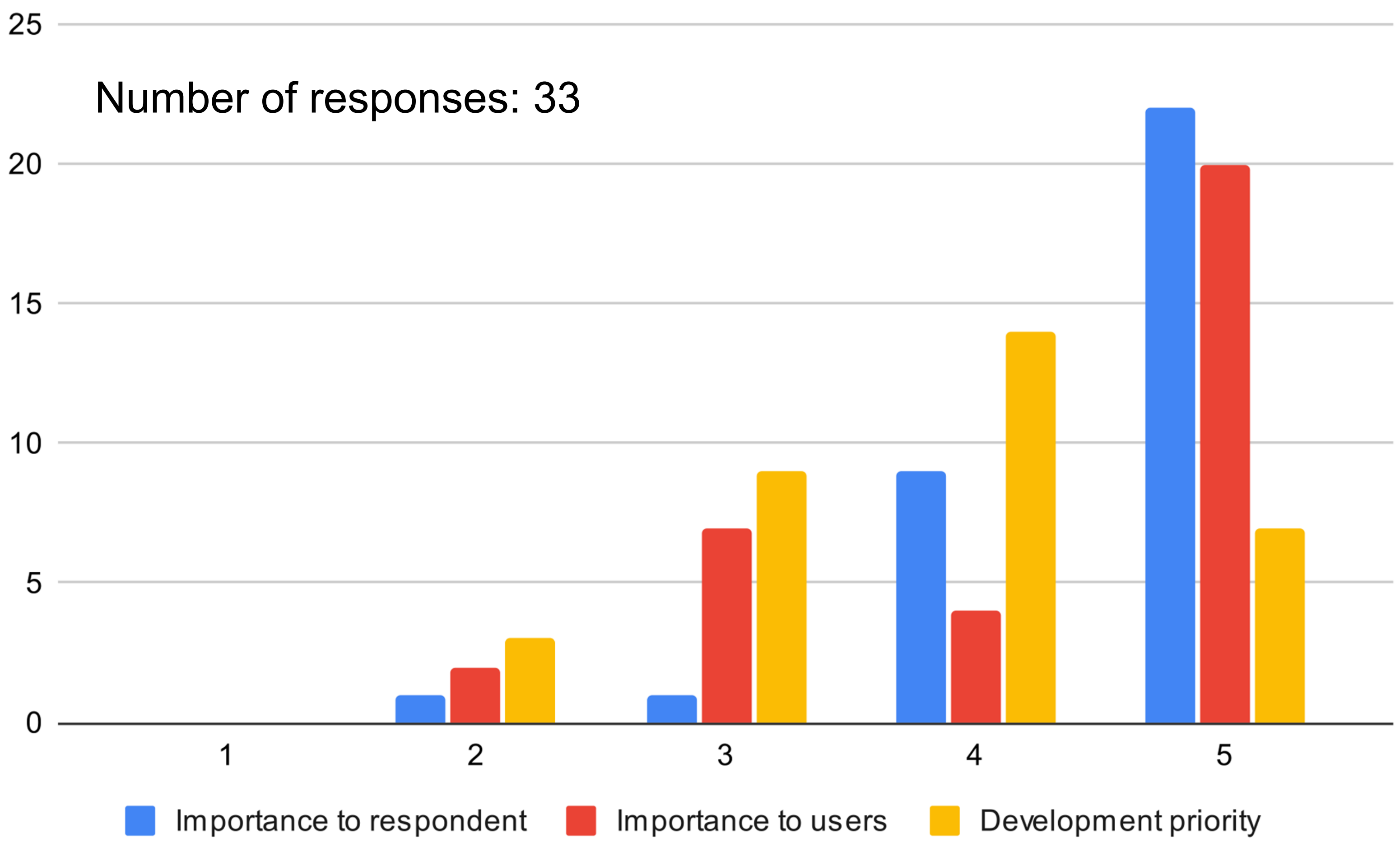
Number of responses: 29
4 respondents don’t use machine-readable methods

Results – desired level of data sharing

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

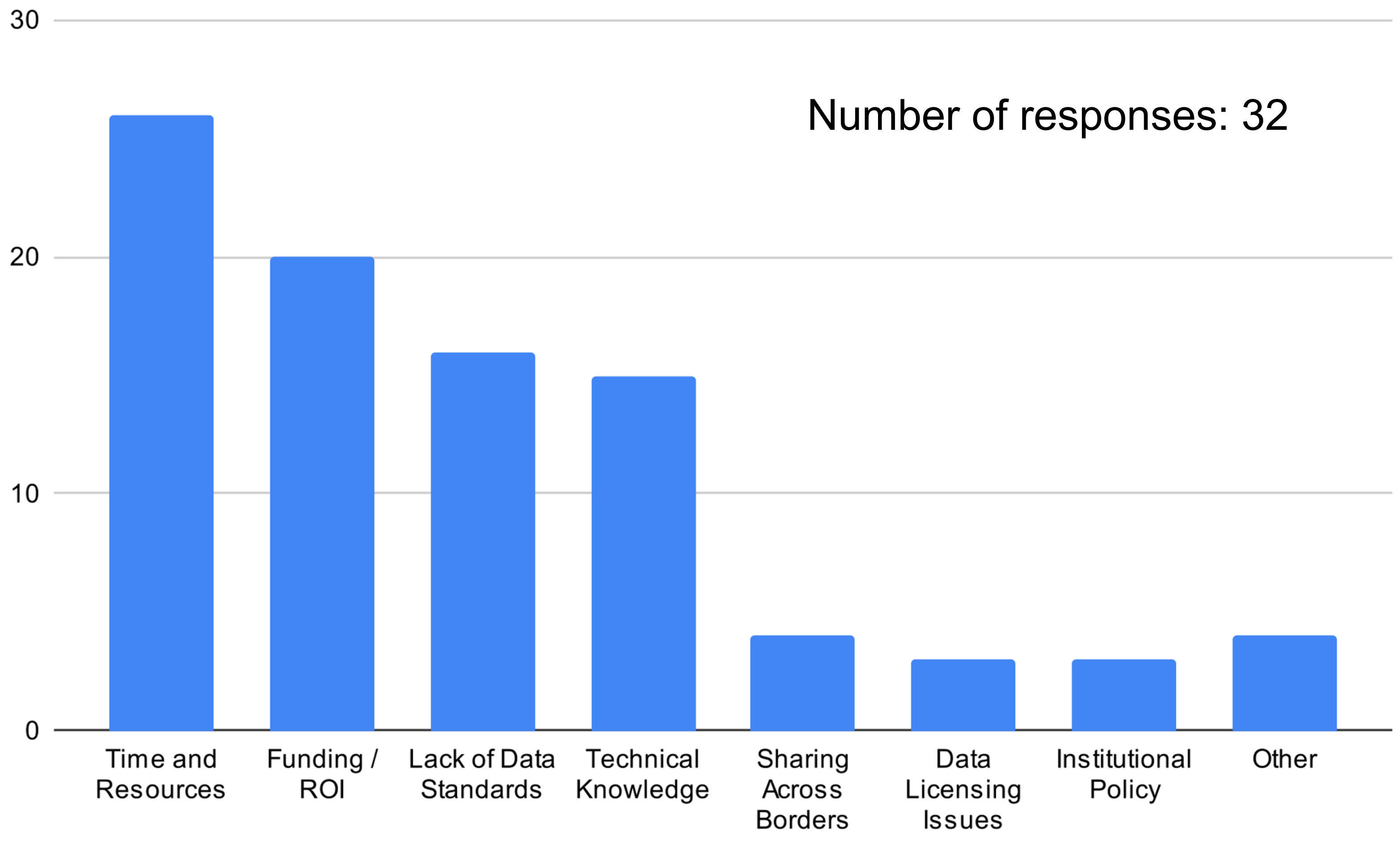
How important is it to your user community to make your database more discoverable and available?

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

Results – Barriers to success

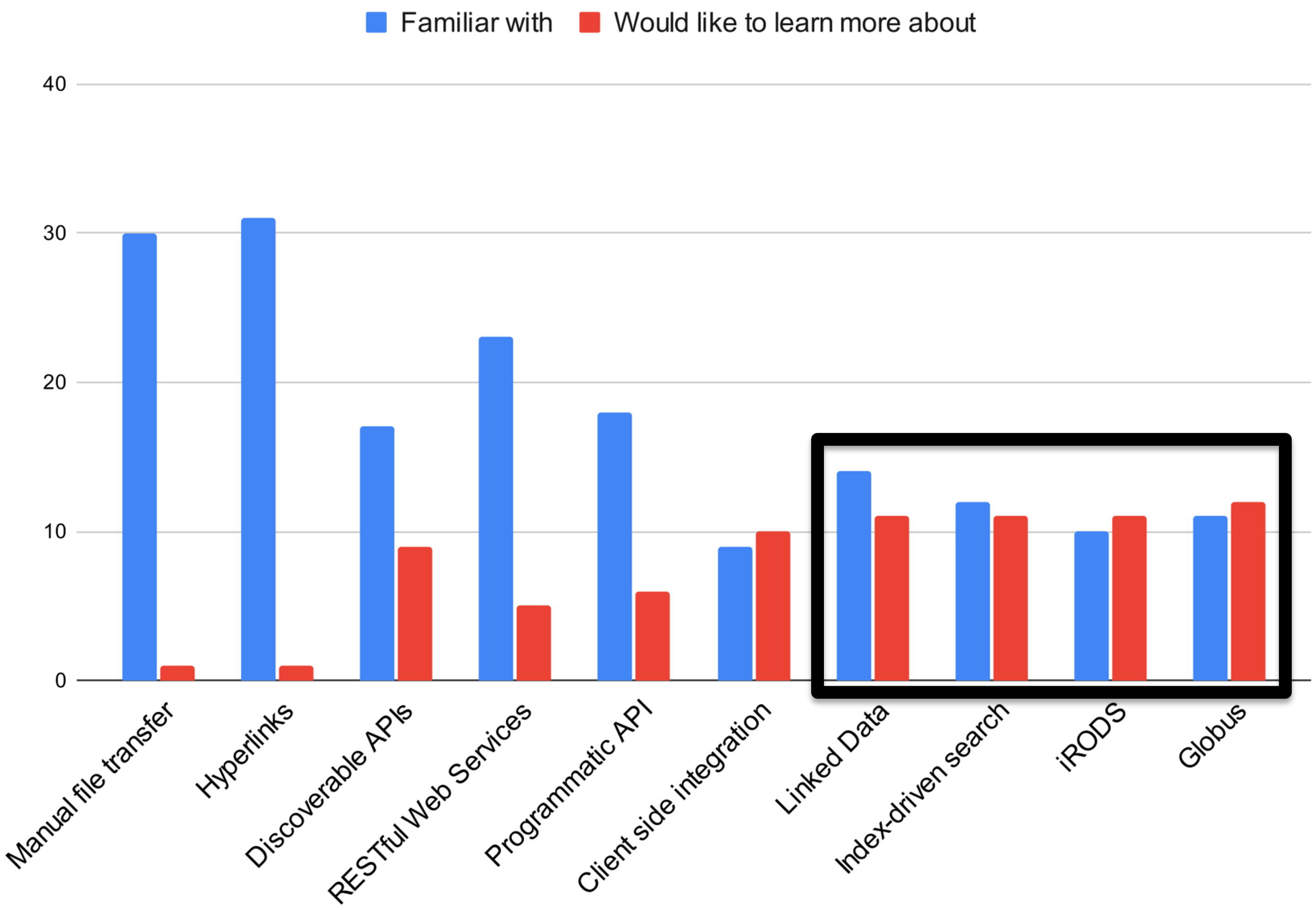
What do you feel are the biggest blockers to successful data sharing in your community?



Results – Technology awareness

What data sharing technologies ***are you familiar with***?

What data sharing technologies ***would you appreciate learning 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.*** 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. AgBioData should prioritize improvements for specific phenotypic data types and formats in future working groups.
- ***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).

Working Group Challenges

- Time zones/internet connectivity
- Deciding on working group focus
- Getting use cases
- Developing a shared vocabulary, e.g., federation vs. sharing

What went well

- Cool and friendly group of colleagues that support open dialogue
- Helpful feedback from participants in AgBioData Community Workshop
- Coordination of survey with ontology group (manuscript)
- Survey led to new WG on data federation training

Specific questions for the AgBioData community

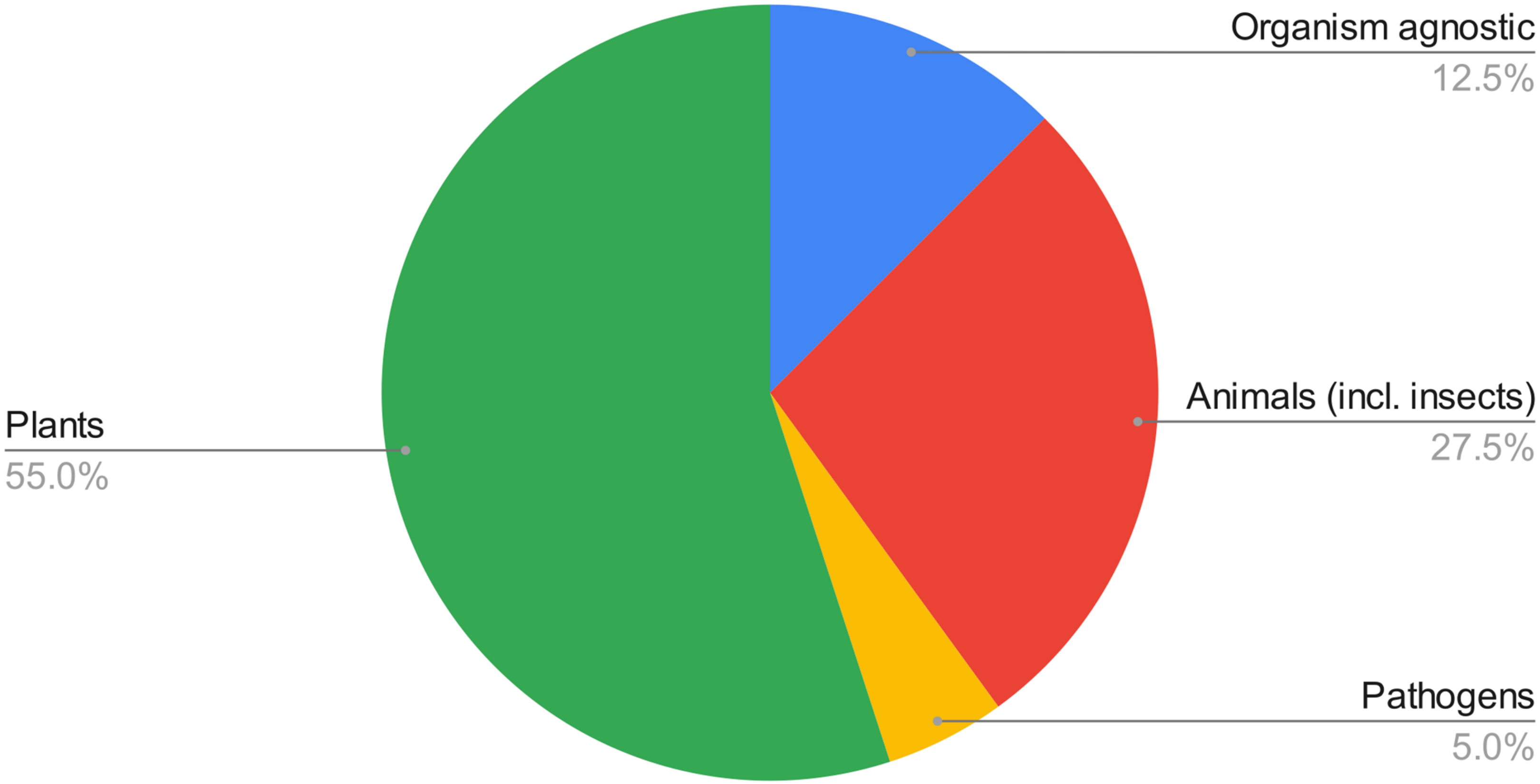
- Are there other communities/groups outside of AgBioData working on data sharing technologies that we should be aware of?
- For which types of data does our community have the most urgent need for standards?
- Which use cases would provide the most compelling demonstration of the benefits of data federation?

Thank you for your attention!

Award Abstract
# 2126334

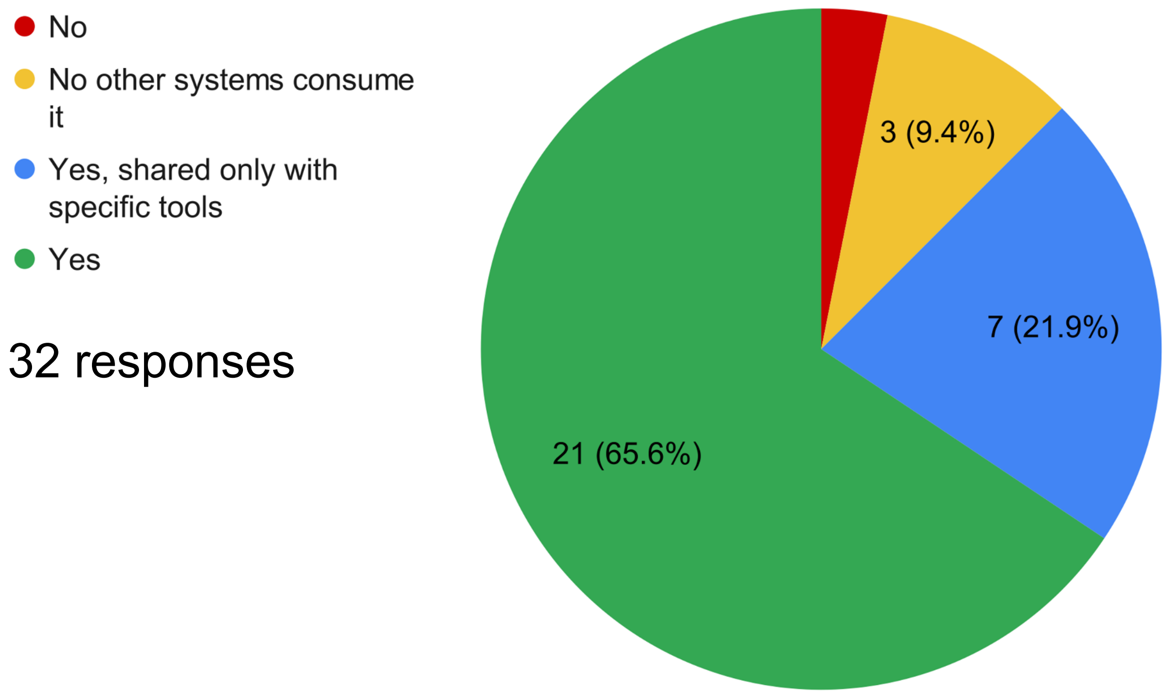
Respondents

33 responses from individuals representing 38 databases or resources (out of 42 AgBioData member databases)

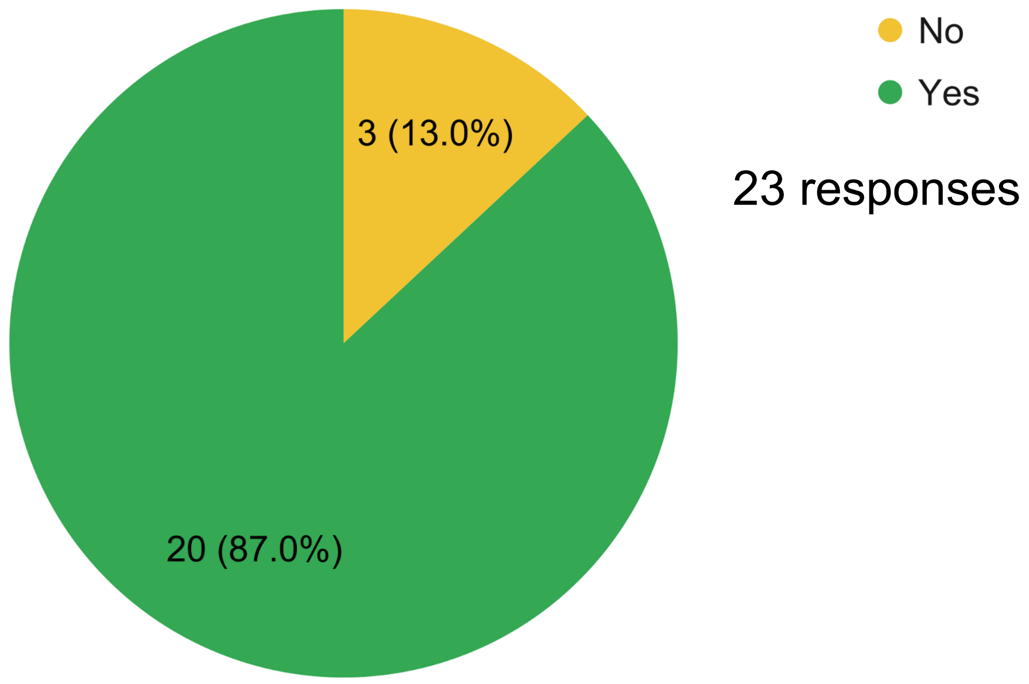


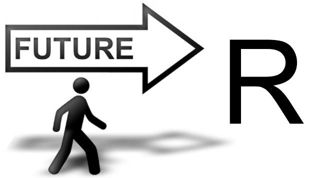
Results – current level of data sharing

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



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

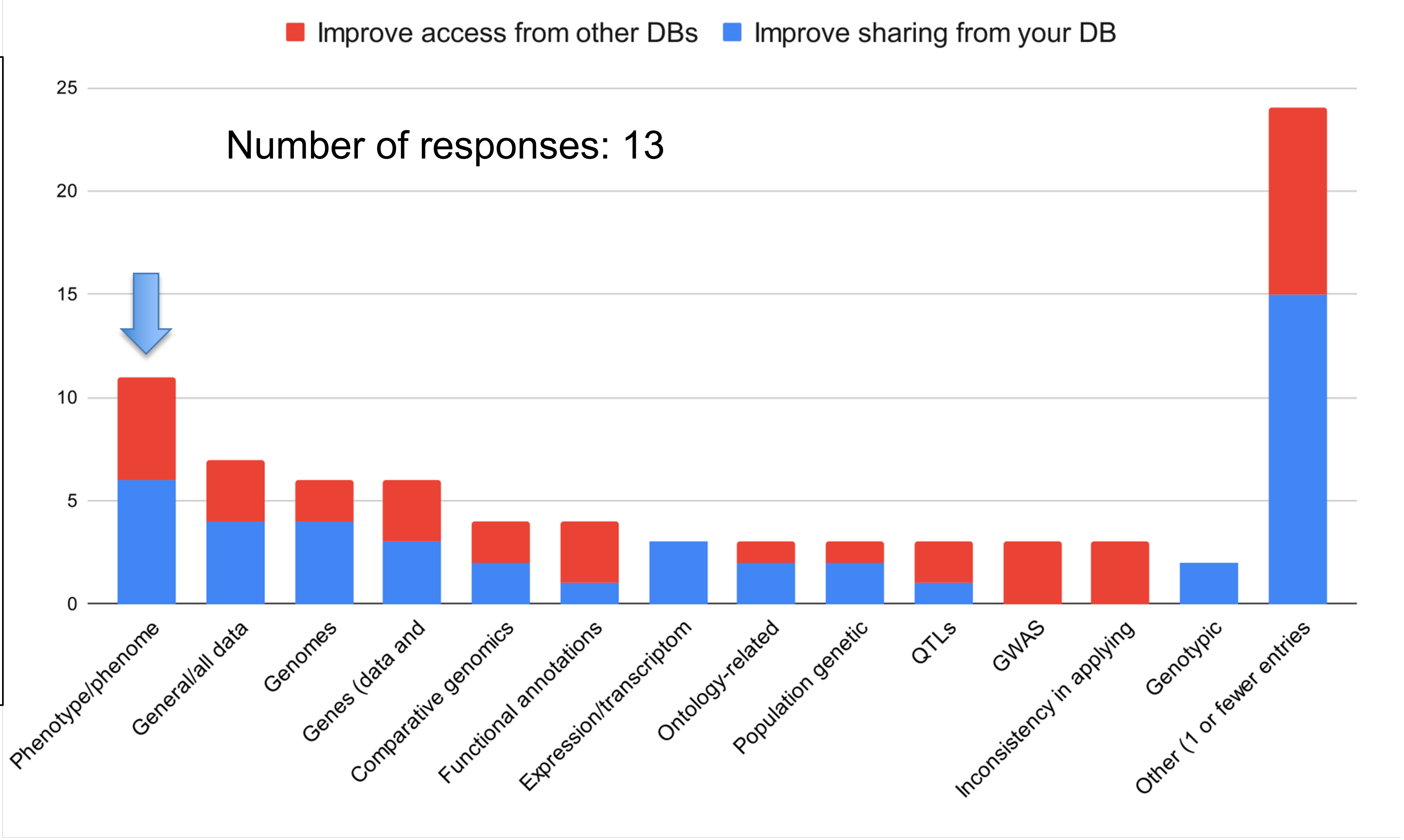




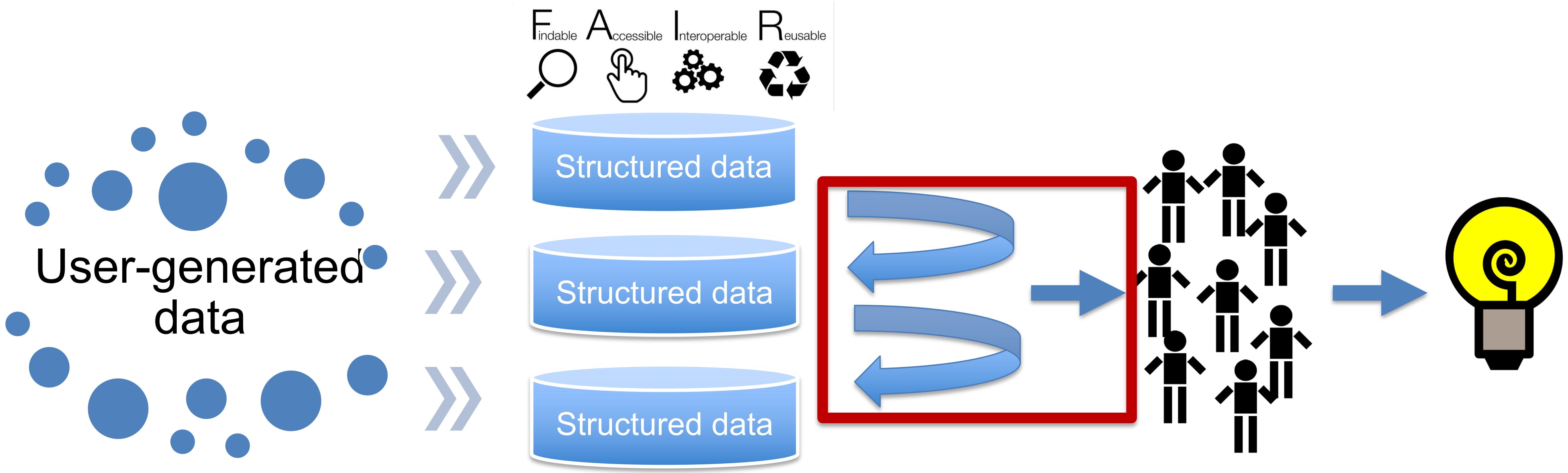
Results – desired level of data sharing

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?



Movement of Data into and among Databases



User-generated data

Findable Accessible Interoperable Reusable

Structured data

Structured data

Structured data

Scientists generate and describe data

Databases describe and structure data (Biocuration)

Data is shared among databases

Databases provide data access