Perspectives on how to improve sequence data reuse from AgBioData’s data reuse working group

Dr. James Koltes

Chair, Data Reuse Working Group

Iowa State University

DATA REUSE

USDA NIFA

United States Department of Agriculture
National Institute of Food and Agriculture

AG2PI
Agricultural Genome to Phenome Initiative

Seed Grant Award #s:
2020-70412-32615
2021-70412-35233

Image attribution: Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, et al. (2014) Troubleshooting Public Data Archiving: Suggestions to Increase Participation. PLoS Biol 12(1): e1001779. doi:10.1371/journal.pbio.1001779

Motivation for data reuse WG

[redacted]

No dataset is perfect

Data are often underutilized due to insufficient metadata and other challenges related to the FAIR data standards

Objective for WG: to identify bottlenecks in data reuse and critical needs to propose solutions for agricultural genomics community

Personal interest: reusing public sequence data to link genotype to phenotype (*USDA-AG2PI* project to build genome annotation through data reuse)

Data reuse topics addressed by the WG (addressing barriers)

1. Data quality concerns
2. Addressing Meta-data challenges
3. Making public data truly public
4. Interoperability
5. Data ownership
6. Considerations for data reuse for users with different skill levels
7. DEI in data reuse

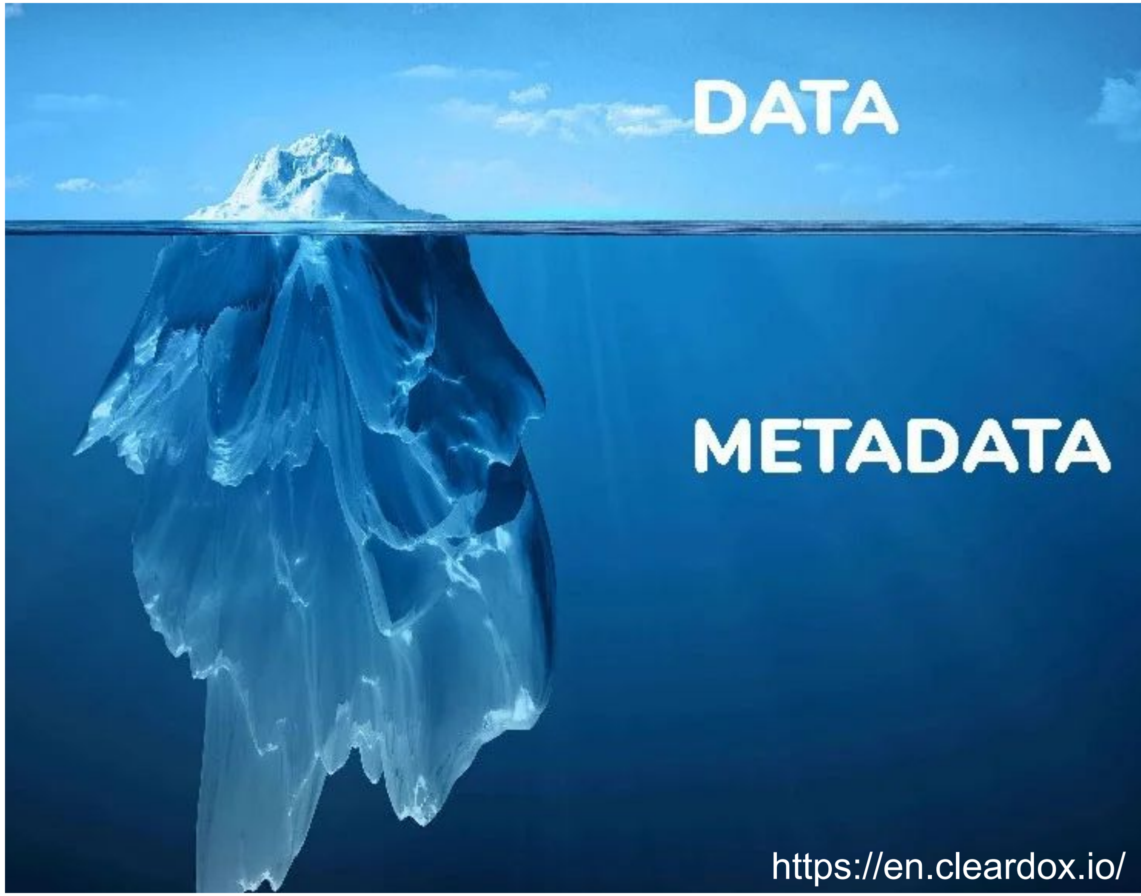
Barriers to data reuse & recommendations to overcome them

Data quality: standards as a solution

- Data made publicly available regardless of quality □ a (subjective) decision on suitability for reuse
- Factors to assess:
  - coverage,
  - depth,
  - technical and biological replication,
  - tissue type,
  - sample collection method,
  - extraction method and library preparation,
  - sequencing technology,
  - other dataset properties
- Limited scope of existing standards (even for common data types and organisms)
- □ difficult to obtain experimental and computational protocols for informed reuse & meta-analyses
- **More protocols, pipelines, and statistical standards needed in the agricultural genomics field**

On the road to complete metadata: incentives

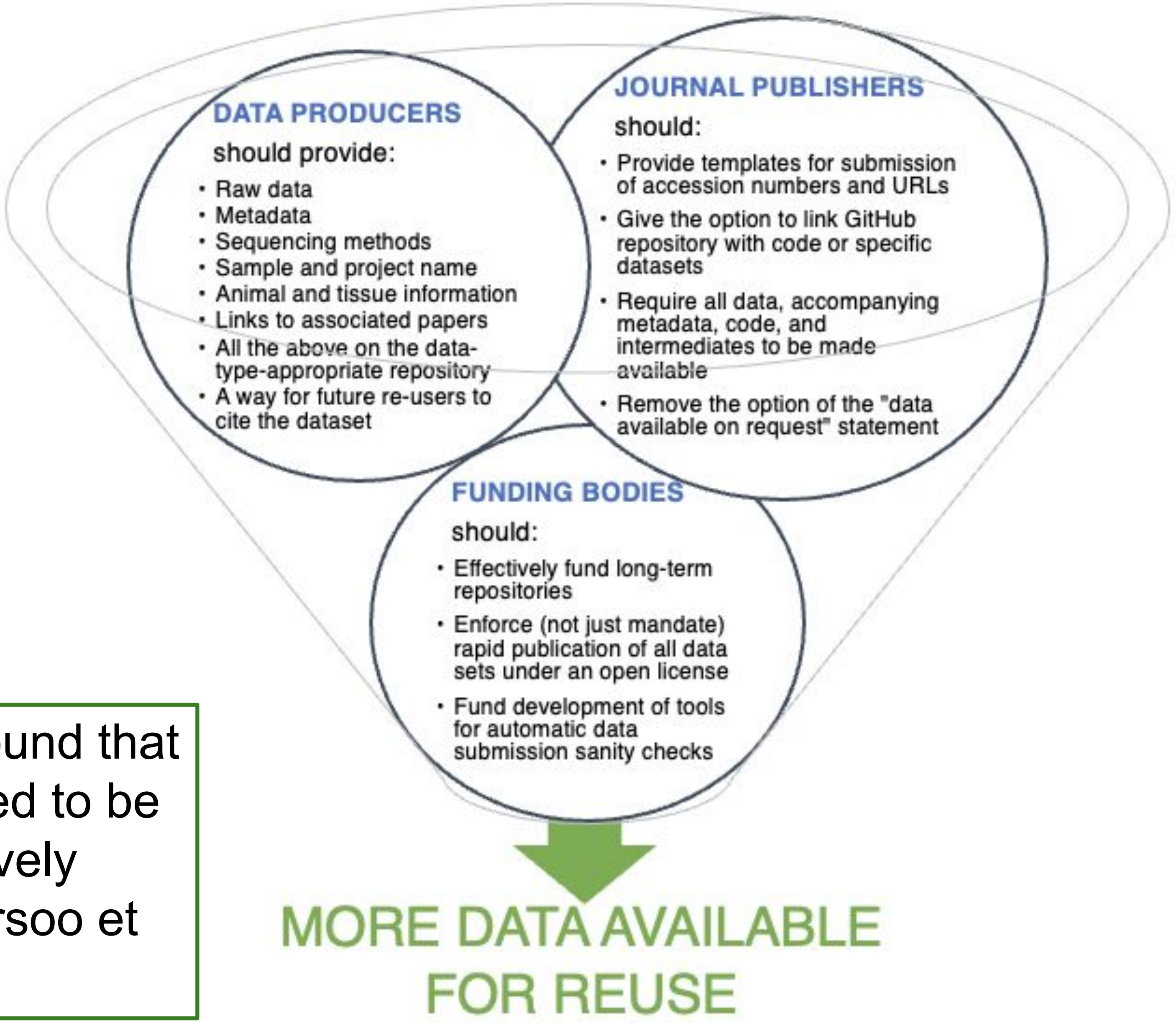
- Limited/incomplete/missing metadata submission templates
- Submission requires work □ Trade-off between collecting all some metadata via a lenient submission system and mandating comprehensive metadata
- **Incentives are needed! E.g., data citations…**



https://en.cleardox.io/

Bridging the data availability gap:
a role for all stakeholders

Survey of *Science* and *Nature* in 2021 found that an alarming **less than 50% of data** stated to be "available upon request" could be effectively obtained from the original authors (Tedersoo et al., 2021)



DATA PRODUCERS should provide:

- Raw data
- Metadata
- Sequencing methods
- Sample and project name
- Animal and tissue information
- Links to associated papers
- All the above on the data-type-appropriate repository
- A way for future re-users to cite the dataset

JOURNAL PUBLISHERS should:

- Provide templates for submission of accession numbers and URLs
- Give the option to link GitHub repository with code or specific datasets
- Require all data, accompanying metadata, code, and intermediates to be made available
- Remove the option of the "data available on request" statement

FUNDING BODIES should:

- Effectively fund long-term repositories
- Enforce (not just mandate) rapid publication of all data sets under an open license
- Fund development of tools for automatic data submission sanity checks

MORE DATA AVAILABLE FOR REUSE

Towards interoperability: data formatting

- Our community has converged on (meta)data standards for data file types:

  FASTQ, SAM/BAM, VCF, GTF, GFF3, BED, …

- Issue: ~~lack of standards~~ □ consistency of use
- Reference genome mapping can be an issue down the line
- **“Backwards compatibility”: outdated lab and sequencing methodology can be combated through extensive metadata** (https://www.protocols.io)
- **Rapid/efficient methods are needed to compare many annotation sets/ (pan)genomes**

Data ownership & sharing requirements

- Challenge: Having access to relevant, affordable study populations from breeding companies that can also be shared publicly as sequence or genotype data

- **Already many sharing requirements + 2026 mandate to make research funded by the USA government publicly available**

- **Need to continue to develop computational approaches to protect industry data while allowing it to be used for research**

User skill level & resource availability

- A recent study (LaFlamme et al., 2022) shows that skill or perceived ability was identified by many participants as a major factor influencing reuse behavior.
- 2017-2018 global survey: most scientists exhibited “high and mediocre risk data practices” (Tenopir et al., 2020).
- US-based institutions: computational resources likely not the limiting factor □ it’s skill level
- **Education programs for awareness-raising and good practice training needed**
- **Incentives (!):**DataWorks! Prize (https://www.herox.com/dataworks)

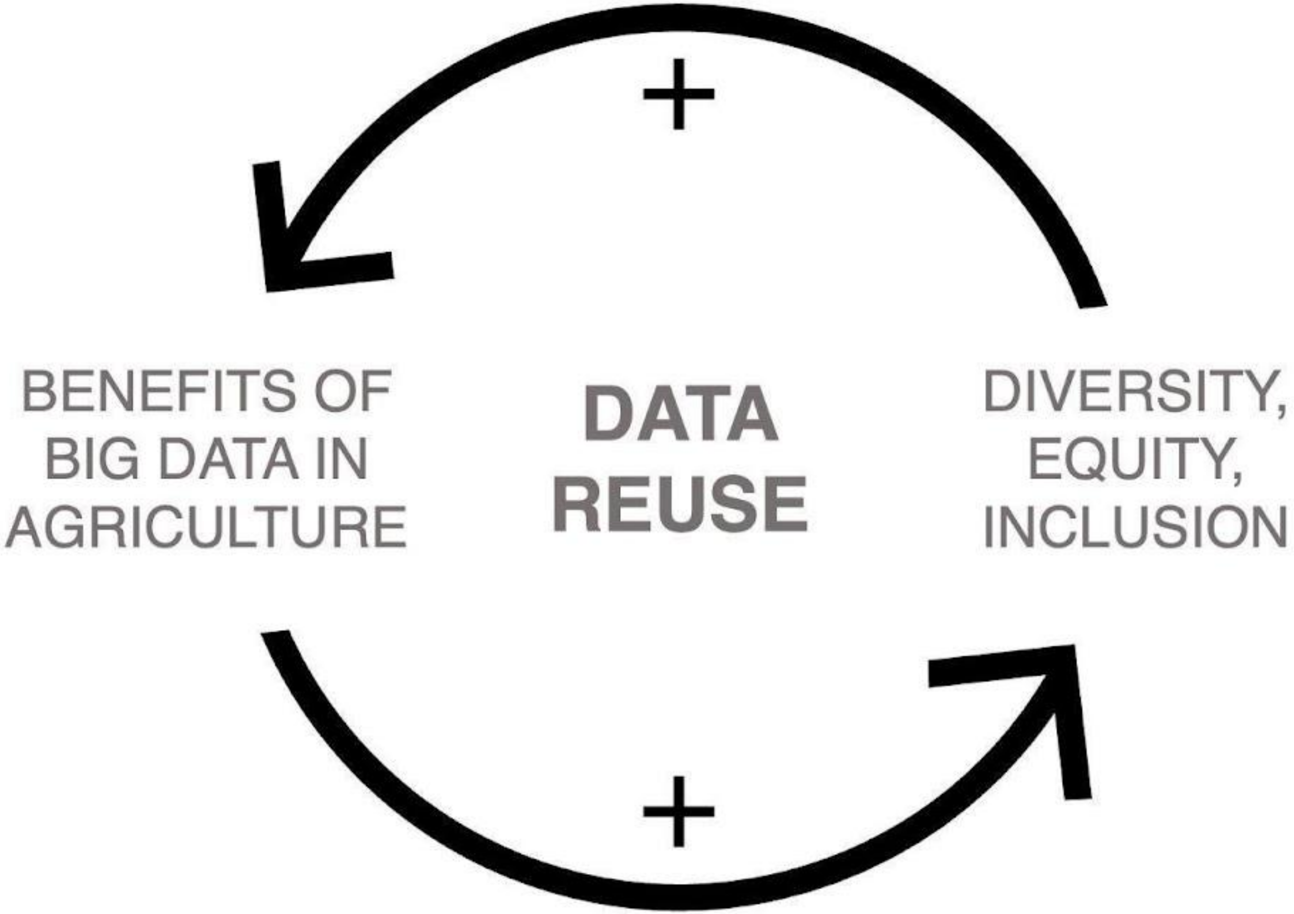
THE FUTURE OF DATA REUSE

The importance and benefits of equity and inclusion

Take-aways and looking ahead

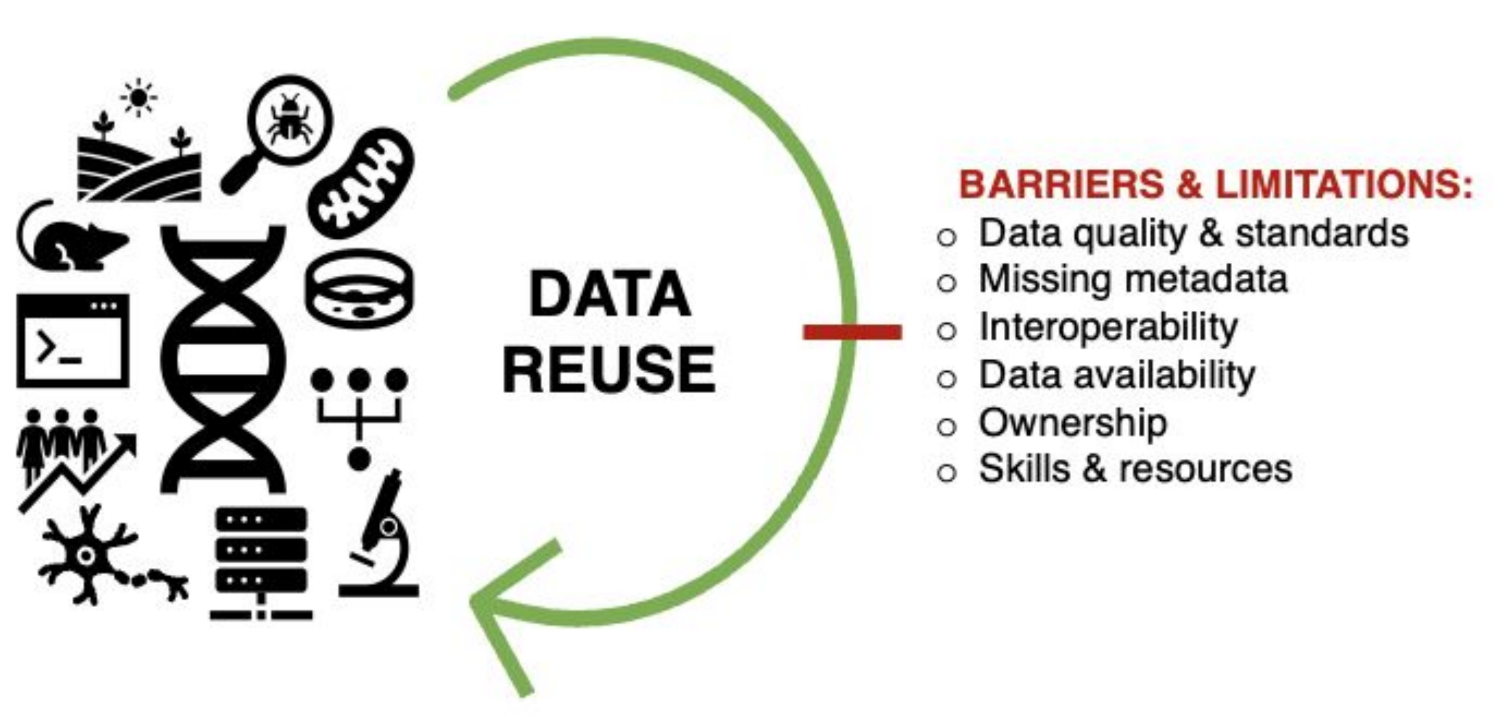
The importance and benefits of equity and inclusion

- ***Diversity breeds innovation***
- Reuse requires computational capacity, internet access, digital literacy, and proficiency in dominant languages
- Data sovereignty: https://localcontexts.org


+
BENEFITS OF BIG DATA IN AGRICULTURE
DATA REUSE
DIVERSITY, EQUITY, INCLUSION
+

Take-aways and looking ahead

Take-aways



BARRIERS & LIMITATIONS:

- Data quality & standards
- Missing metadata
- Interoperability
- Data availability
- Ownership
- Skills & resources

Also, there are many opportunities!

Looking ahead

- ***The future of data reuse is bright and exciting!***
- Integration of datasets
- AI and ML
- Emerging data types:

  Gaps & Opportunities:
  - Phenomes (including sensing data), metabolomes, proteomes, interactomes, enviromes, microbiomes, lipidomes, and glycomes

  Who should tackle & How?
  - New WG? Sensing & microbiomes first?

Outcomes/ Deliverables: WG’s white paper

https://doi.org/10.20944/preprints202401.0780.v1

Alenka Hafner, Penn State University (WG co-chair)

Victoria DeLeo - Bowery Farming

Cecilia Deng- The New Zealand Institute for Plant and Food Research Limited

Christine G. Elsik - University of Missouri

Damarius Fleming, USDA-ARS

Peter W. Harrison, European Bioinformatics Institute

Ted Kalbfleisch, University of Kentucky

Bruna Petry, Iowa State University

Boas Pucker, TU Braunschweig

Elsa H Quezada-Rodríguez, Universidad Autónoma Metropolitana-Xochimilco; Universidad Nacional Autónoma de México

Christopher K. Tuggle, Iowa State University

James Koltes, Iowa State University (WG chair)

Thank you for your Attention!

https://boardagenda.com/2021/08/05/are-you-asking-the-right-questions-of-your-data-team/

Acknowledgments

AgBioData

United States Department of Agriculture
National Institute of Food and Agriculture

AG2PI
Agricultural Genome to Phenome Initiative

Seed Grant Award #s:
2020-70412-32615
2021-70412-35233