# Facilitating data re-use for genome-enabled communities

Monica Poelchau and Chris Childers

USDA-ARS National Agricultural Library

https://i5k.nal.usda.gov/

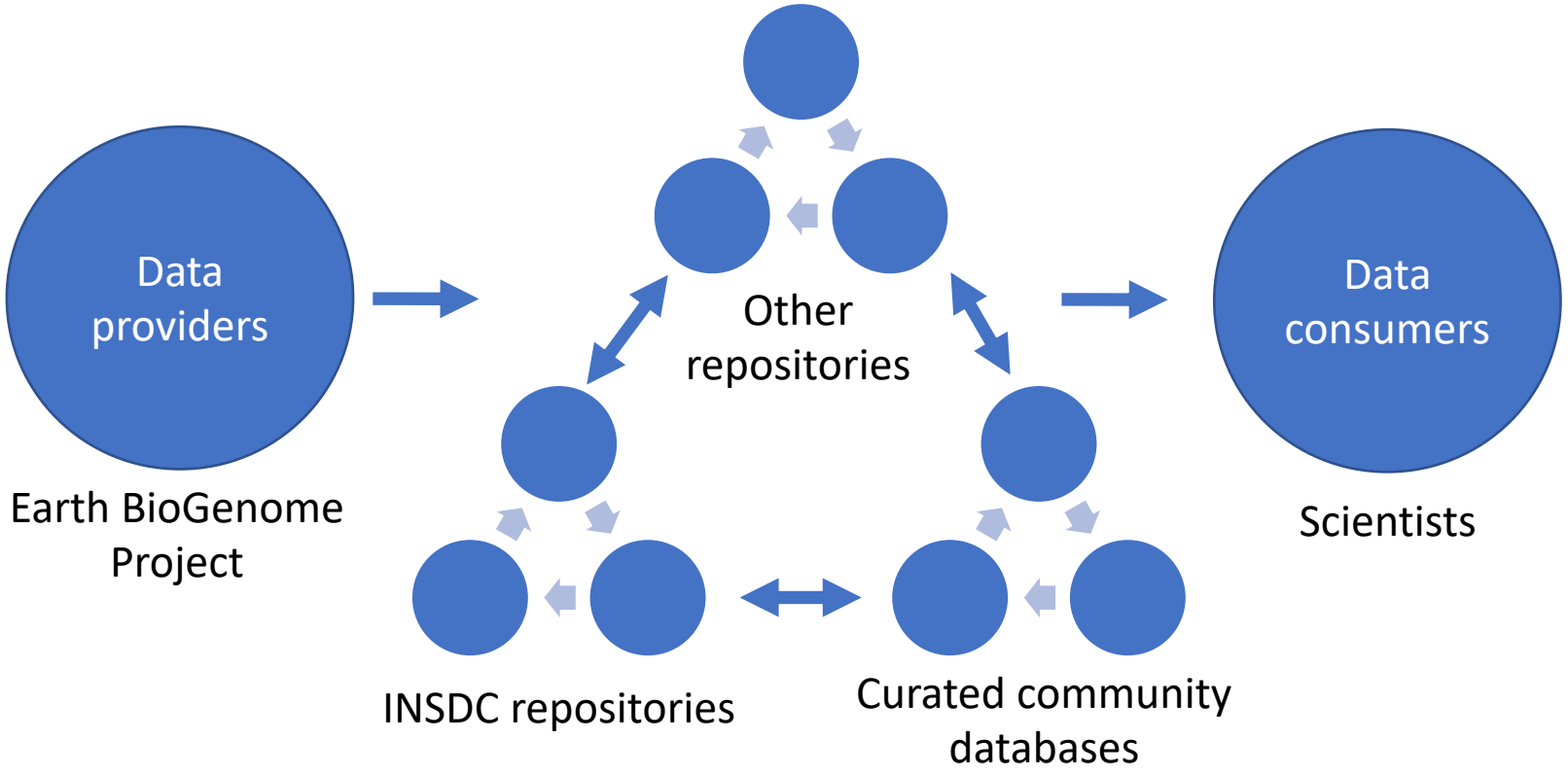
USDA

# Acknowledgements


- The AgBioData consortium and steering committee (https://www.agbiodata.org/);
- USDA-NAL's Knowledge Services Division (https://www.nal.usda.gov/main/data);
- The Earth Biogenome Project IT and informatics subcommittee (https://www.earthbiogenome.org/subcommittee-it-informatics);
- The AgBioData GFF3 working group;
- Alliance for Genome Resources (https://www.alliancegenome.org/);
- And many others!


USDA

# Genome-enabled communities need efficient data re-use




USDA

# 1. Follow (or create) community standards and best practices

- File formats (e.g. GFF3 for genome annotations, VCF for variant data)
- Metadata standards (e.g. MIGS v5.0 for genome metadata)
- Nomenclature (e.g. the vertebrate gene nomenclature committee for gene names)

# 1. Follow (or create) community standards and best practices

Recommendations for databases:

- Join or follow a group leading standards development or implementation (e.g. AgBioData; RDA; EBP IT subcommittee; Elixir; Force11; FAIRsharing; Genomic Standards consortium; etc.)

Recommendations for data producers:

- Generate genome project data with submission to the appropriate database in mind, and find out early what the submission standards are

# 2. Rich, machine-readable metadata is critical for data re-use

- Metadata is information about data
  - Example: the assembly program(s) (=metadata) for your genome assembly (=data)
- Metadata is crucial for scientists other than the data generator to understand and re-use the dataset
- Description in a publication isn't sufficient – often behind a paywall, not machine-readable

# 2. Rich, machine-readable metadata is critical for data re-use

Recommendations for databases:

- Provide incentives to get metadata from data submitters (e.g. more metadata = more services)
- Use ontologies/CVs to the extent possible (EBI's ontology lookup service is helpful)

Recommendations for data generators:

- Embrace metadata!
- Record metadata as early as possible, and provide as much as you can. Useful resources for this are: CyVerse; COPO; Open Science Framework; etc.

# 3. Plan and budget for software development, operations and maintenance

- Sharing data requires software that stores and serves it.
- Biologists (myself included) often underestimate the cost, effort, and expertise it takes to develop, operate, and maintain software (in particular over longer periods of time)

# 3. Plan and budget for software development, operations and maintenance

Recommendations for databases:

- Use and contribute to existing open-source software if possible (e.g. the GMOD community)
- Follow best software development practices
- Budget… Good developers are expensive, and can be difficult to recruit and retain.
- Build for web services – these are essential for programmatic data sharing (see standards and metadata)
- Dialing in on user needs is essential to reduce software bloat and costs

Recommendations for data generators:

- Keep this in mind in grant proposal reviews

# Resources and References

- The AgBioData database recommendations: https://academic.oup.com/database/article/doi/10.1093/database/bay088/5096675
- The FAIR data principles: https://www.go-fair.org/fair-principles/
- The Research Data Alliance: https://www.rd-alliance.org/
- Elixir: https://elixir-europe.org/
- Force11: https://www.force11.org/
- FAIRsharing standards: https://fairsharing.org/standards/
- EBI's Ontology Lookup Service: https://www.ebi.ac.uk/ols/index
- Genomic Standards consortium: https://gensc.org/
- CyVerse: https://cyverse.org/
- COPO: http://copo-project.org/
- The Open Science Framework: https://www.cos.io/products/osf
- Generic Model Organism Database Project (GMOD): http://gmod.org/wiki/Main_Page
- The I5k Workspace@NAL: http://i5k.nal.usda.gov/