

FAIR Scientific Literature Is Not What You Think: How to Know Where Data Goes

Plant and Animal Genomes Conference (PAG 31)
01-12-24



FAIR Scientific Literature WG Goals

- Identify bottlenecks in the publication-curation pipeline.
- Identify sets of existing or desired tools or biocuration resources to increase literature curation throughput and accuracy.
- Publish recommendations and a roadmap for authors and publishers to increase the FAIRness of published research.



Members



Katheryn Buble,
Washington State
University



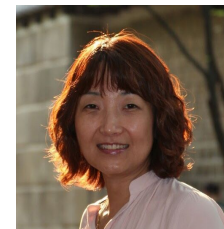
Leyla Cabugos
Librarian, Cal Poly



Jenna Daenzer
GSA



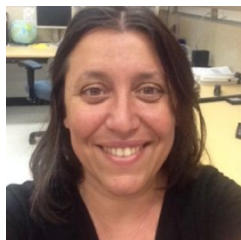
Leonore Reiser
TAIR curator



Sook Jung
Asst Research
Professor. GDR



David Molik
Computational
Biologist,
USDAARS



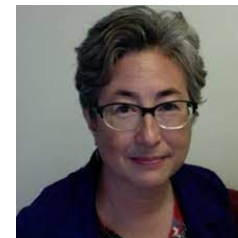
Daniela Raciti
Exec Editor,
microPublication &
Wormbase Curator



Jacqueline Campbell
Geneticist, USDA ARS
SoyBase curator



Adam Wright
Software Engineer
Wormbase.Reactome

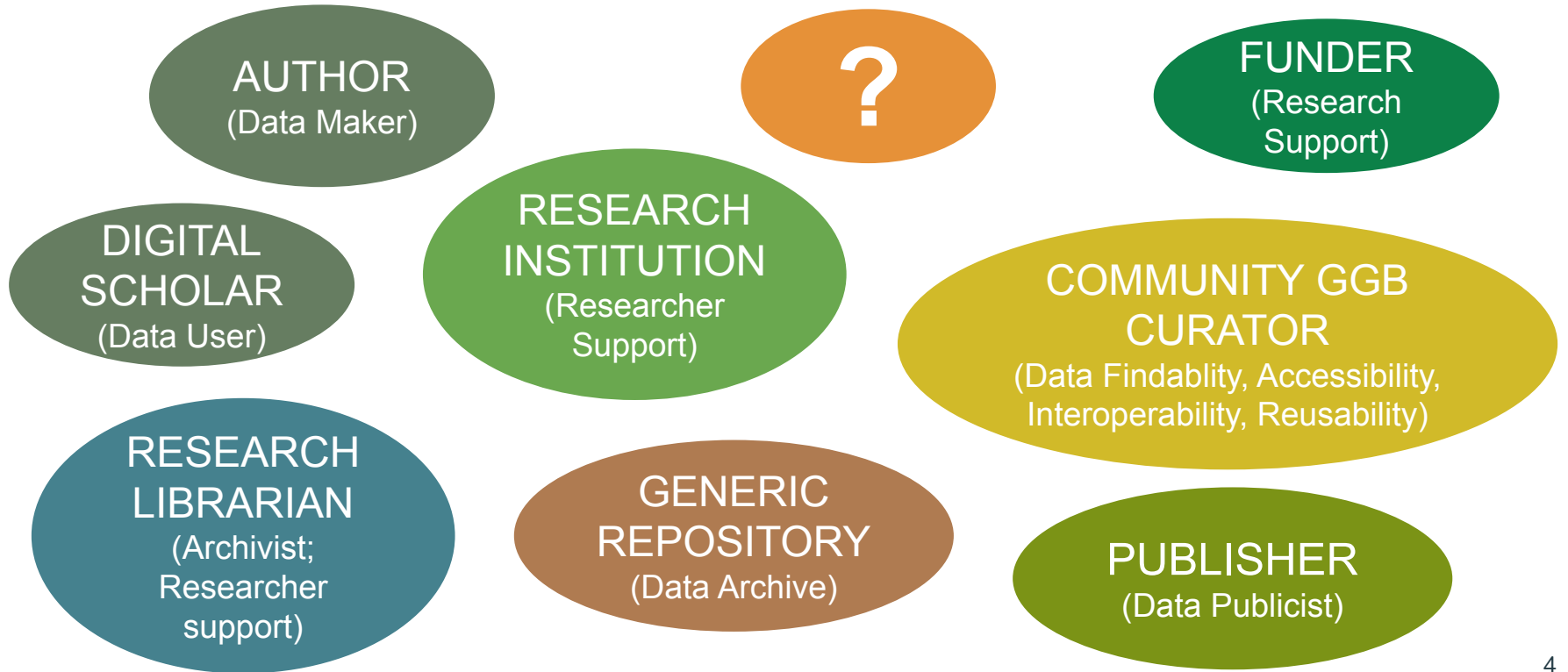


Karen Yook
Exec Editor,
microPublication &
Wormbase Curator

Daniel Morris , Professor (advisory)



Stakeholders involved w/ data management



Persona	Motivated By	Challenged By	Possible Incentives	Possible tools	Notes
GGB Database Curator	Desire to present comprehensive, integrated data to user community. Need to maintain a desired resource that community values	Always having to process data post-publication. High volume of data to process and not enough time or curation power (\$ money). Difficult to find relevant datasets, poorly or incorrectly formatted data and metadata. Lack of verification from Publishers that data is submitted. Data availability statements that allow 'data on request'. Lack of responsiveness from authors.	Increased community use, increased value by community. \$ for curation services.	Publicly accessible resources that authors and journals can use to determine what data is included and where it should go. DataSeer is an example of something that could work but needs more information. Easy to use software that facilitates data submission /co-curation and professional curator review as part of publication pipeline. Better metadata about papers (species, entities, datatypes) to make published data more findable.	
Researcher	Professional rewards, recognition (citations, invitations to speak, more funding), moving science forward.	Not knowing where data should go, not knowing how to properly format data and metadata. Time consuming submission processes. Long delays sometimes between data collection and write up may mean some information is lost (e.g. people leave lab), unclear on benefits of data sharing or consequences of not.	Public rewards/recognition for data sharing. Increased citations for research, better position for future funding.	Easy to use software that facilitates data submission /co-curation and professional curator review as part of publication pipeline. Better tools for tracking data reuse and reporting on how often shared data is accessed and remixed. Publicly accessible resources that authors and journals can use to determine what data	
Funder	Advancing science through thoughtful allocation of funds. Increasing US national competitiveness, food security.	Not knowing all the places data should go, different program areas have different repos and specialist knowledge. Rely on reviewers who don't necessarily have the specific knowledge to evaluate DMPs well.	Ability to track memo compis quantifiable m value of funds		
Publisher	[R] Best practices for data sharing; ability to promote efforts toward FAIR; community needs and moving science forward (note that as a society journal we likely have different views than some publishers, but I tried to think "general")	Not knowing all the places data should go, different areas have different repos and specialist knowledge. Rely on reviewers who don't necessarily have the specific knowledge to evaluate data availability statements. Lack of easy verification of data availability. Additional cost of adding more curation to publishing pipeline. Authors often don't want data available before publication - need for reviewer links/reviewer tokens.	Increased imp data is reused citations. (Soc also be motiva community an		

Persona	Motivated By	Challenged By	Possible Incentives	Possible tools	Notes
Generalist Repository Curator	Increased user base in terms of both submissions and downloads. Being part of critical infrastructure. Repository of record.	Lack of familiarity with data types and specialist repositories, not sure what the data is or where it should go, lack of specialized knowledge. High volume of data to process and not enough time or curation power limits the depth of curation.		Easy to use software that facilitates data submission /co-curation and professional curator review as part of publication pipeline. Better tools for tracking data reuse and reporting on how often shared data is accessed and remixed. Publicly accessible resources that authors and journals can use to determine what data is included and where it should go (if not in generalist repo).	
Research Librarian	Need to assist researchers in developing and complying with DMP, desire to properly archive data, desire to support best practices in data discovery and research conduct.	Lack of familiarity with data types and specialist repositories, not sure what the data is or where it should go, lack of specialized knowledge. High volume of data to process and not enough time or curation power limits the depth of curation. Lack of engagement with Research faculty when they need it	Current information and access to other stakeholders	Better tool's for tracking data reuse and reporting on how often shared data is accessed and remixed. Publicly accessible resources that authors and journals can use to determine what data is included and where it should go.	
Meta-analyst/Digital Scholar	This is a subtype of researcher, they are pulling together various journal articles in a review. They are either summarizing a body of work, or trying to find a previously unrecognized trend in the corpus. They are motivated by professional rewards, and more so then a general researcher moving a field forward. They are interested in standardized data. The meta-analyst would think about Journal articles AS data.	Lack of machine readable data, lack of standardized data, lack of reported data, lack of association of data in an article and its repository, and access to APIs		Standardized data reporting in journals, standardized journal article formatting, AI summary systems.	
Research Institutions	I see this as the umbrella over Research Librarian, Scholar, and Researcher. The Research Institution is motivated by supporting their researchers in complying with funder mandates.	Lack of budget; many different players (library, research office, and or IT department) with competing needs; lack of clarity about what is required?	Increased ability to track DMP players (library, research office, and grant award/tracking compliance. Demonstrating research impacts. Increasing research capability.	Most Institutions probably have DMPs, but they don't enforce/track/etc, therefore need machine actionable DMPs. They likely have the tools, but need more focus on their collaborations btwn departments and funding.	

Current publication-curation challenges

AUTHOR
Data Maker

Databases



Current publication-curation challenges

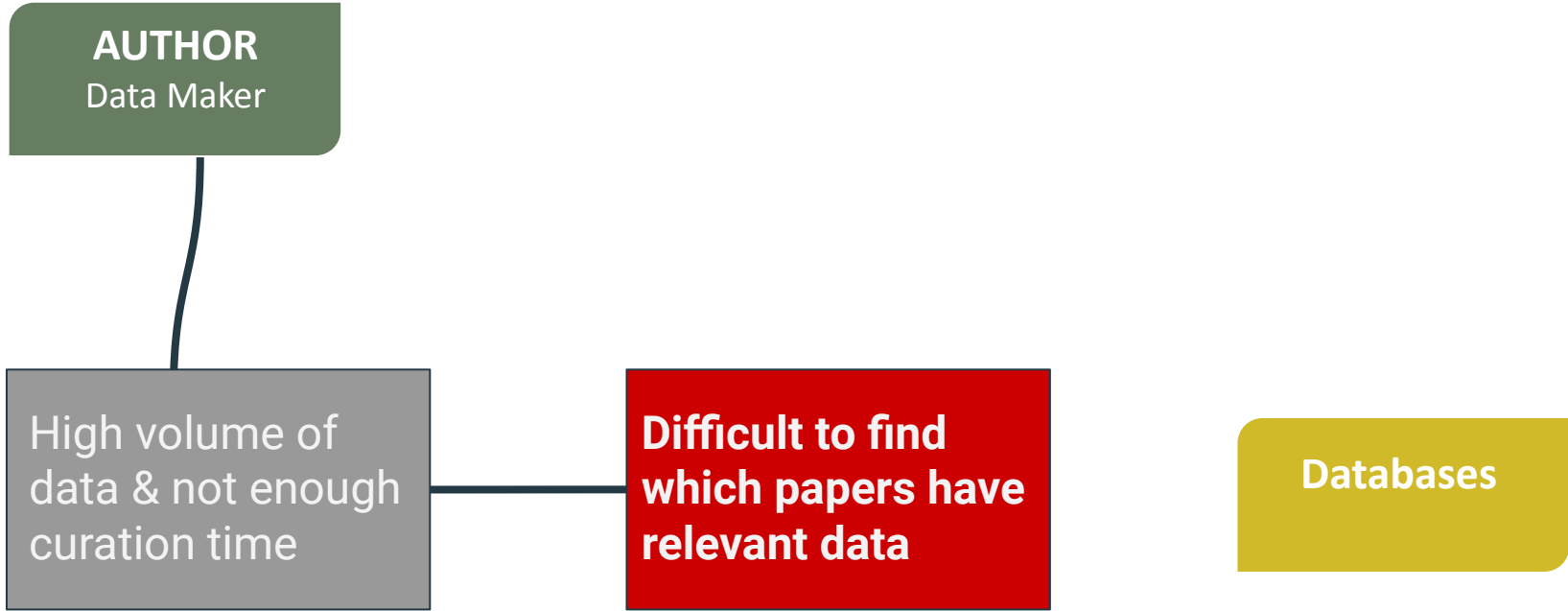
AUTHOR
Data Maker

**High volume of
data & not enough
curation time**

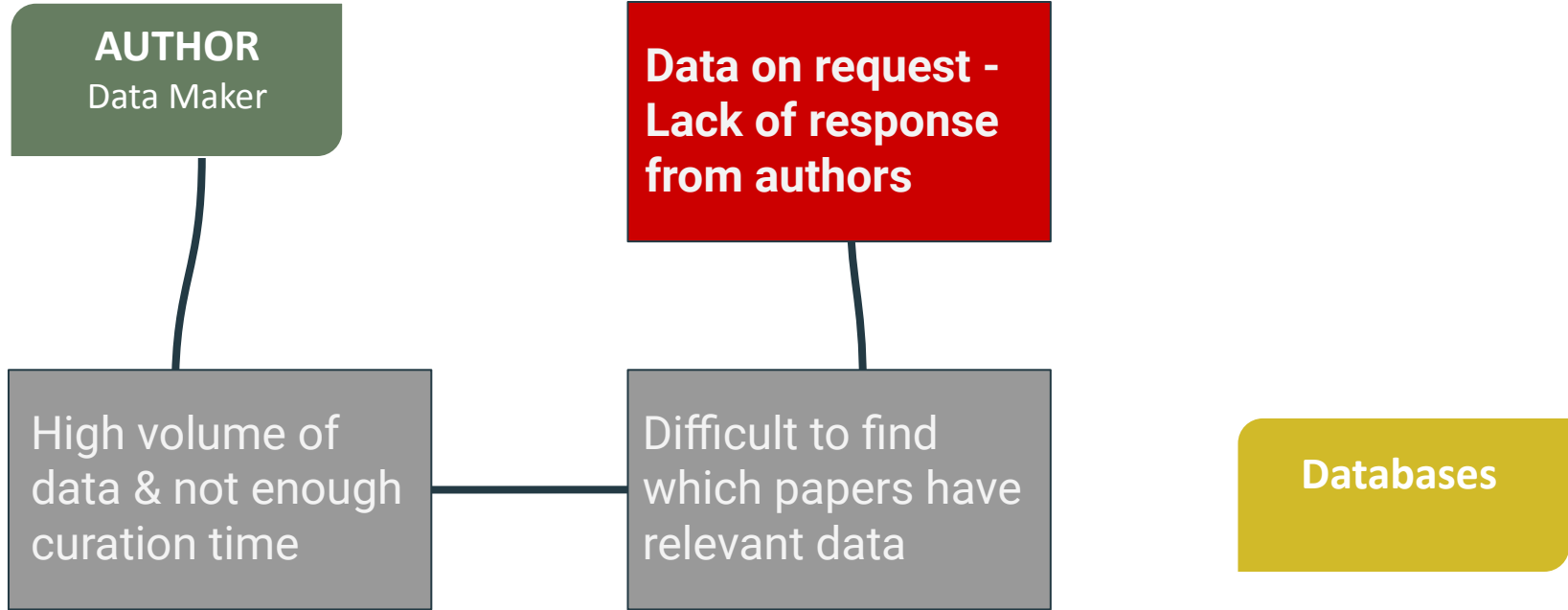
Databases



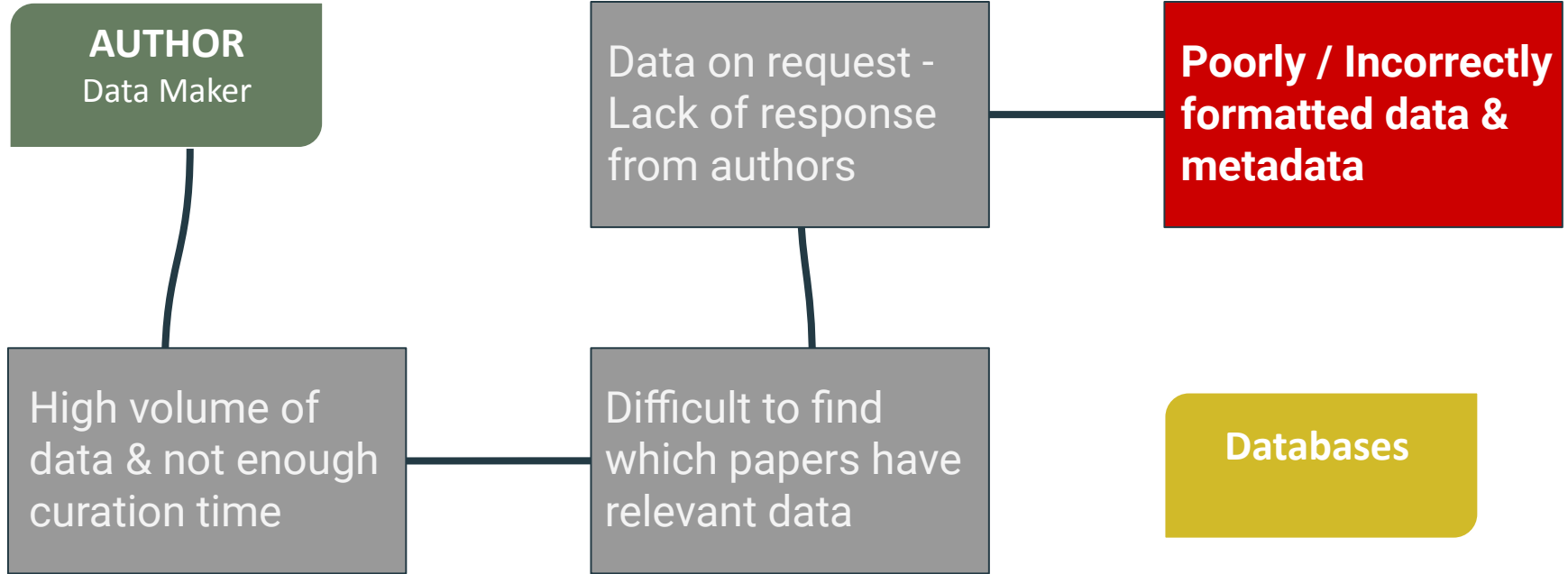
Current publication-curation challenges



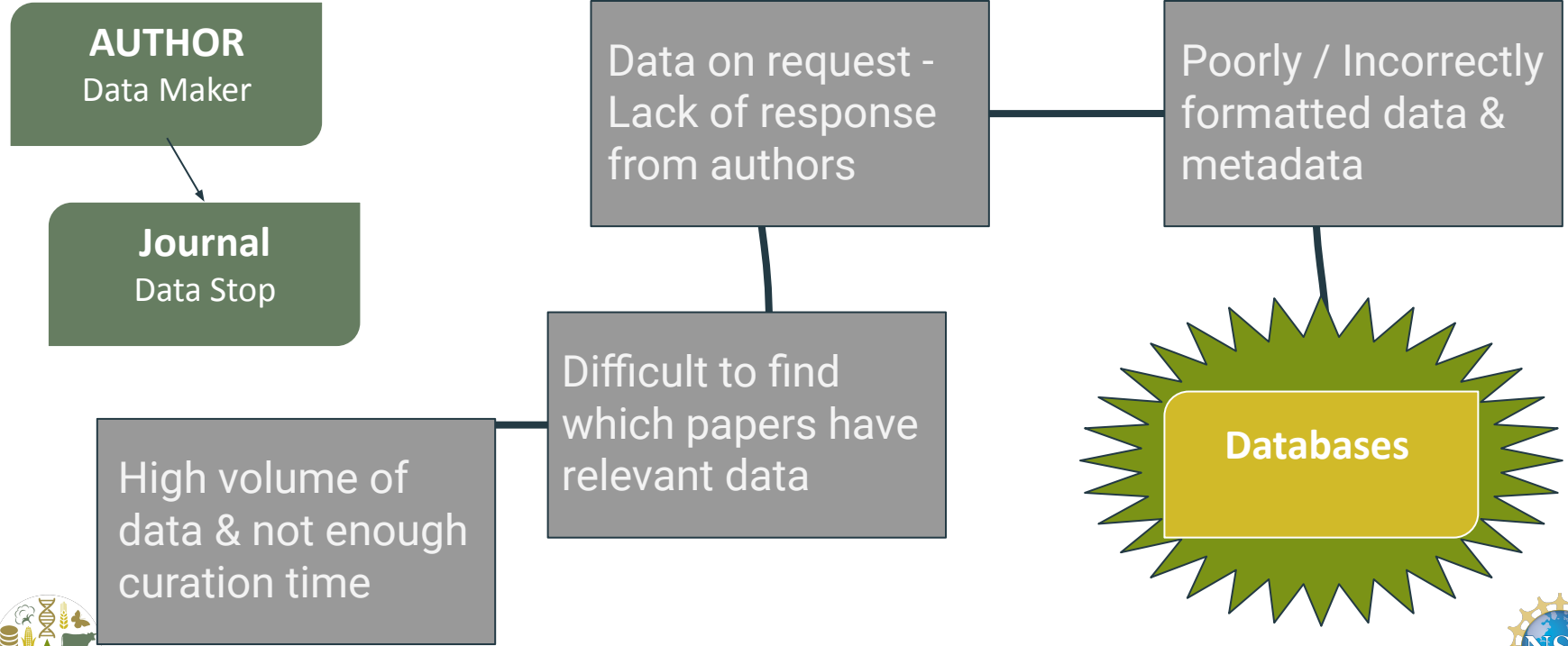
Current publication-curation challenges



Current publication-curation challenges



Current data publishing workflow



Aim for a BETTER workflow

AUTHOR
Data Maker

Data on request -
Lack of response
from authors

Poorly / Incorrectly
formatted data &
metadata

High volume of
data & not enough
curation time

Difficult to find
which papers have
relevant data

Databases



Four Stakeholder challenges & barriers

Researcher

- **Not knowing where data should go**
- Time consuming submission process
- Not knowing how to format data / metadata

Publisher

- **Not knowing where data should go**
- Lack of easy verification of data availability
- Authors often do not want data available before publication

Librarian

- **Not knowing where data should go**
- Lack of familiarity with data
- Lack of engagement with researcher

Funder

- **Not knowing where data should go**
- Different programs have different repositories
- Proposal reviews don't know how to evaluate DMPs



Four Stakeholders Resources & incentives

Researcher

- **Resources to determine where data should go**
- Public rewards / recognition for data sharing
- Increased citations for future funding and job security

Publisher

- Increased impact factors when data is reused
- Better tools for tracking data reuse & sharing
- **Resources to determine where data should go**

Librarian

- Increased engagement with other stakeholders
- Better tools for tracking data reuse & sharing
- **Resources to determine where data should go**

Funder

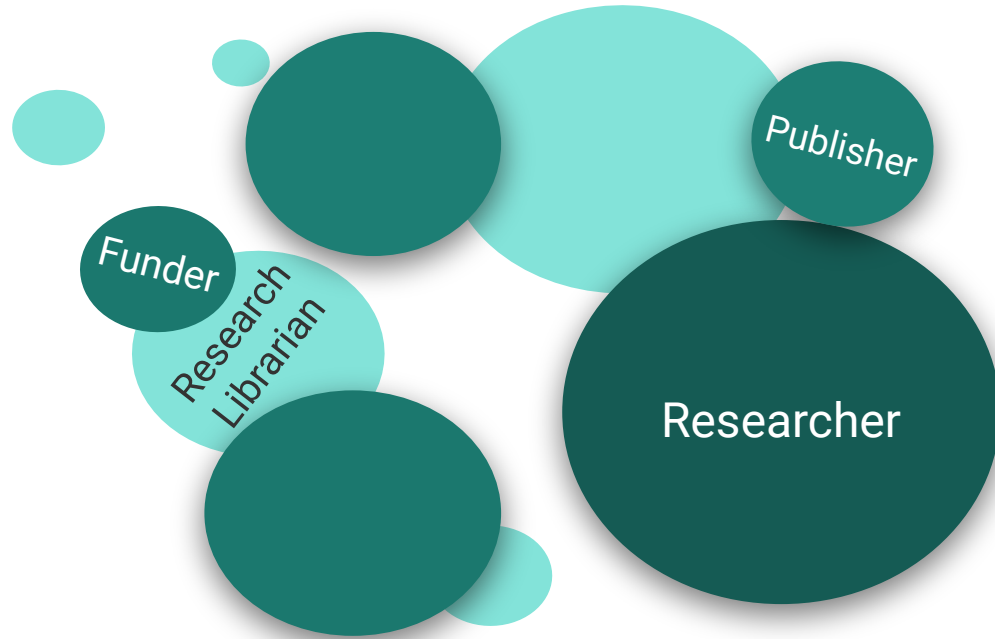
- FAIR education for reviewers & awardees
- Better tools for tracking data reuse & sharing
- Ability to track FAIR data & having quantifiable metrics



AgBioData can bridge the workflow



Focus on 'before or during' publication not afterwards (FAIR from the start)



The Data Guide

Creating tools for stakeholders to know where data go

A tool from which the most appropriate location/database can be derived for a dataset

Data Storage Finder | Research | +

https://finder.research.cornell.edu/storage

Data Storage Finder

Evaluate options for data storage at Cornell

- All services presented on this finder tool are vetted and supported by Cornell University.
- To explore data options available to Weill Cornell Medicine Cornellians please visit the [WCMC storage wizard](#).
- We welcome [feedback](#) on this tool.

Describe your data

Answer these questions to help identify data storage services that are suitable for your needs. Checking these boxes will change the list of available services. If you are uncertain how to answer, leave the question blank to maximize your resulting options.


[Clear Answers](#)

Select data storage services you would like to compare. [Select All](#) [Clear Selections](#)

- What is the classification of your data?**
 - Public / Low Risk
 - Sensitive / Moderate Risk
 - Confidential or Restricted / High Risk
 - HIPAA-Regulated
- Do you need backups, snapshots or replication of your data?**
 - I need one or more backup/snapshot copies of the data, and need to be able to restore data from previous points in time (high durability).
 - I need to have replicate copies of the data to minimize downtime (high availability).
- How much data do you have and how fast will it grow?**
 - Unlikely to exceed 1TB in 2 years
 - Greater than 1TB or likely to exceed in 2 years
- Do you have special performance needs?**
 - I am likely to have more than 1,000 files in a single directory within two years.
 - My data interactions demand high transaction or transfer rates.

Amazon Web Services Elastic Block Store Storage for use with Amazon EC2	Amazon Web Services Elastic File System Storage for use with multiple Amazon EC2 instances	Amazon Web Services Glacier Cloud based archival storage	Amazon Web Services S3 Flexible, scalable object storage
BioHPC Cloud Storage for BioHPC lab computing services	CAC Archival Storage Single copy, non-mountable storage	CAC Red Cloud Storage Storage for Red Cloud compute instances	CCSS Research Servers Storage for CCSS computing environment
CISER Data & Reproduction Archive Publicly shared and restricted data and code packages repository	CUGIR Publicly shared geospatial data storage	CUL eCommons Publicly shared data repository	Cornell Box Online file sharing and collaboration
Cornell Restricted Access Data Center Storage for CRADC (confidential) computing environment	EZ-Backup Static Storage Archival storage and backup storage	Google Drive Cornell Google Workspace file storage and sharing	Kaltura Video Platform Service Flexible, scalable video and multi-media storage (customizable)
Kaltura Video on Demand Video and multi-media storage	LabArchives Online electronic lab notebook	Microsoft OneDrive Cloud storage for individual use or sharing with specific individuals	Microsoft SharePoint Cloud storage for sharing with a broad audience
Microsoft Shared Libraries Cloud storage for group and project collaboration	Microsoft Teams Cloud storage for group and project collaboration with online meetings and chat	Open Science Framework Online project management repository	Shared File Services File sharing between users and computers
Shared File Services - Confidential File sharing between users and computers for	WCM Block Storage High performance storage attached to centrally hosted servers (WCM only)	WCM File Sharing Secure network storage (NFS/CIFS) for research computing (WCM only)	WCM Red Cloud Secure Storage Secure storage for Red Cloud compute instances

Finder | Fruit and Nut Cultivars x +
https://dev.fruitandnutlist.org/finder



Fruit and Nut Cultivars Database Home Browse Crops About How to Contribute Contact Us Log In

Finder

Database Finder to submit Agricultural Genomic, Genetic, Breeding Data
Evaluate database options to submit your data

We welcome feedback on this tool.

Describe your data

We welcome feedback on this tool.

Select Databases you would like to compare. Select All Clear Selection

Clear Answers

1. Crop

- Arabidopsis
- Citrus
- Rosaceae Crop
- Cassava

2. Data Type

Citrus Genome Database

Citrus Genome Database

Citrus Greening

Citrus Greening Database

Genome Database for Rosaceae


Resources for Rosaceae Research Discovery and Crop Improvement

TAIR

The Arabidopsis Information Resources

Supported by a partnership of ARS, USDA NIFA, ASHS, NIBSP10, Industry and US Land Grant Universities

Finder | Fruit and Nut Cultivars x +
https://dev.fruitandnutlist.org/finder



Fruit and Nut Cultivars Database Home Browse Crops About - How to Contribute Contact Us Log In

Finder

Database Finder to submit Agricultural Genomic, Genetic, Breeding Data
Evaluate database options to submit your data

We welcome feedback on this tool.

Describe your data

Select Databases you would like to compare. Select All Clear Selection

Clear Answers

1. Crop

- Arabidopsis
- Citrus
- Rosaceae Crop
- Cassava

2. Data Type

We welcome feedback on this tool.

Citrus Genome Database

Citrus Genome Database

Citrus Greening

Citrus Greening Database

Genome Database for Rosaceae

Resources for Rosaceae Research Discovery and Crop Improvement

TAIR

The Arabidopsis Information Resources

Supported by a partnership of APS, USDA-NIFA, ASHS, NPS10, Industry and U.S. Land Grant Universities

	A	B	C	D	E	F
1	Resource Type	Species/Crop	Resource	Database URL		
2	Community DB	Arabidopsis	TAIR	https://www.arabidopsis.org/		
3	Community DB	Cassava	CassavaBase	https://www.cassavabase.org/		
4	Community DB	Citrus	Citrus Genome Da	https://www.citrusgenomedb.org/		
5	project Database	Citrus Species and pathogens.Citru	Citrus Greening	https://www.citrusgreening.org/		
6	Community DB	Cotton	CottonGen	https://www.cottongen.org/		
7	Community DB	Cucurbit	Cucurbit Genomic	http://cucurbitgenomics.org/		
8	Community DB	Forest trees	TreeGenes	https://treegenesdb.org		
9	deprecated	(this has been closed and data fold	Hardwood Genom	http://www.hardwoodgenomics.org/		
10	Community DB	Grains	GrainGenes	https://wheat.pw.usda.gov		
11	general plant	Any plant	Gramene	https://www.gramene.org/		
12	Community DB	Sorghum	SorghumBase	https://www.sorghumbase.org/		
13	Community DB	Wheat	Triticeae toolbox,	https://wheat.triticeaetoolbox.org/		
14	Project Database	Wheat	WheatIS	http://www.wheatis.org/		
15	Project Database	Rice	KitBase	http://kitbase.ucdavis.edu/		
16	Community DB	Legumes	KnowPulse	https://knowpulse.usask.ca/		
17	Community DB	Legumes	Legume Informati	https://www.legumeinfo.org/		
18	Community DB	Peanut	PeanutBase	https://peanutbase.org		
19	Community DB	Pulses	Pulse Crop Databa	https://www.pulsedb.org/		
20	Community DB	Soybean	Soybase	https://www.soybase.org/		
21	Community DB	Maize	MaizeGDB	https://maizegdb.org/		
22	Community DB	Musa	MusaBase	https://www.musabase.org/		
23	Community DB	Rosaceae	Genome Database	https://www.rosaceae.org/		
24	Community DB	Solanaceae	Sol Genomics	https://solgenomics.net/		
25	Community DB	Sweet Potato	SweetPotatoBase	https://www.sweetpotatobase.org/		
26	Community DB	Vaccinium	Genome Database	https://www.vaccinium.org/		
27	Community DB	Yam	YamBase	https://www.yambase.org/		
28	general	Plants and Animals	AgBase	https://agbase.arizona.edu/		
29	general plant	Any Plant	Bio-Analytic Reso	https://bar.utoronto.ca/		
30		Livestock	Animal QTL	http://www.animalgenome.org/cgi-bin/QTLdb/		
31	Community DB	Insects	i5Kworkspace	https://i5k.nal.usda.gov/		
32						

	A	B	C	D	G	H	I	U
1	Species/Crop (NCBI taxon ID	Database	Data Category...	Ontology terms	Ontology terms _Karen Sorted	Data Type (Data types that can be fo Data	Metadata Requirements
2			CassavaBase		delete	?	analysis results	analysis model, algorithm, input data, results
3	Citrus / Diaphorina citri / Ca. Liberibacter as		Citrus Greening			?	analysis results	analysis model, algorithm, input data, results
4	Maize	NCBI:txid381138	MaizeGDB		delete - category too narrow	?	Data Sets	
5			CassavaBase	germplasm	pedigree	metadata for other analysis	pedigree data	female parent, male parent, type
6	Citrus / Diaphori	NCBI:txid121845	Citrus Greening		pedigree	metadata for other analysis	pedigree data	female parent, male parent, type
7	Cotton		CottonGen		accession	metadata for other analysis	stocks	
8	Soybean	NCBI:txid3850	Soybase		pedigree	metadata for other analysis	Pedigree Data (Strain/Cultivar/Line parentage)	
9			TAIR			metadata for other analysis	Stocks	
10	Forest trees		TreeGenes		accession	agricultural science branch request	Plant PopGen (GeoReferenced Plants)	
11	Grains		Triticeae toolbox,		delete	metadata for other analysis	Trials	
12			TAIR			Database cross-mapping	External links	
13			Citrus Genome Da	transcriptome/...	Gene expression	Gene expression	gene expression	
14	Rosaceae	NCBI:txid3745	Genome Database	transcriptome/...	Gene expression	Gene expression	gene expression	NCBI BioProject and BioSample IDs
15	Vaccinium	NCBI:txid13750	Genome Database	transcriptome/...	Gene expression	Gene expression	gene expression	NCBI BioProject and BioSample IDs
16			i5k Workspace@	transcriptome...		Gene expression	gene expression/mapped RNA-Seq	NCBI requirements
17	Maize	NCBI:txid381132	MaizeGDB	transcriptome/...	delete - indirect submission	Gene expression	Gene Expression	reference, alignments, conditions
18	Soybean	NCBI:txid3852	Soybase	transcriptome/...	Gene expression	Gene expression	Expression of Transcriptomic Data (RNA-seq, GeneChip, custom Chips, etc)	
19			TAIR	transcriptome/...	Gene expression	Gene expression	Expression Data	
20	Grains		Triticeae toolbox,	transcriptome/...		Gene expression	Gene Expression from EBI Expression Atlas	
21	Cotton		CottonGen	gene function	gene functional annotation	gene functional annotation	gene function annotation	author (ORCID), GO, evidence code
22	Rosaceae	NCBI:txid3745	Genome Database	gene function	gene functional annotation	gene functional annotation	gene function annotation	author (ORCID), GO, evidence code
23			i5k Workspace@	gene function	gene functional annotation	gene functional annotation	gene function annotation	Internally generated metadata. Chado analysis fields: N
24	Maize	NCBI:txid381125	MaizeGDB	gene function	gene functional annotation	gene functional annotation	gene function annotation	reference, ontology, description, evidence code
25	Arabidopsis	NCBI:txid3702	TAIR	genomics	gene functional annotation	agricultural science branch request	gene function annotation	author (ORCID), PMID, GO, evidence code
26	Maize	NCBI:txid381135	MaizeGDB	gene function	delete - indirect submission	Gene regulation	Epigenomic, DNA-binding, and gene regulati	GFF, BED, reference
27			i5k Workspace@	gene function	gene report	gene report	gene structure and metadata	https://i5k.nal.usda.gov/i5k-workspace-gene-and-prote
28			TAIR	gene function	gene report	gene report	Gene Structure Updates	
29			Citrus Genome Da	GWAS/QTL/map	genetic map	genetic map	genetic maps	experiments, germplasm, marker
30	Cotton		CottonGen	GWAS/QTL/map	genetic map	genetic map	genetic maps	experiments, germplasm, marker
31	Rosaceae	NCBI:txid3745	Genome Database	GWAS/QTL/map	genetic map	genetic map	genetic maps	experiments, germplasm, marker
32	Vaccinium	NCBI:txid13753	Genome Database	GWAS/QTL/map	genetic map	genetic map	genetic maps	experiments, germplasm, marker
33	Grains		GrainGenes	GWAS/QTL/map	genetic map	genetic map	genetic maps	experiments, germplasm, marker
34	Legumes	NCBI:txid3805	Legume Informati	GWAS/QTL/map	genetic map	genetic map	genetic maps	experiments, germplasm, marker
35				GWAS/QTL/map	genetic map	genetic map	genetic maps	experiments, germplasm, marker

A	B	C	D	E	F
Category	Ontology Term(s)	Ontology ID	ID		
	sequence record	http://edamontology.org/data_0849			
	protein interaction data	http://edamontology.org/data_0906			
	Citation	http://edamontology.org/data_0970			
	genetic map	http://edamontology.org/data_1278			
	Protein structure	http://edamontology.org/data_1460			
	QTL map	http://edamontology.org/data_1860			
	external links	http://edamontology.org/data_1883			
	accession	http://edamontology.org/data_2091			
	person identifier	http://edamontology.org/data_2118			
	Phylogenetic Data	http://edamontology.org/data_2523			
	protocol	http://edamontology.org/data_2531			
	pathway or network	http://edamontology.org/data_2600			
	Image	http://edamontology.org/data_2968			
	genome annotation	http://edamontology.org/operation_0362			
	Genome assembly	http://edamontology.org/operation_0525			
	gene functional annotation	http://edamontology.org/operation_3672			
	Ontology and terminology	http://edamontology.org/topic_0089			
	proteomics	http://edamontology.org/topic_0121			
	Protein targeting and localisation	http://edamontology.org/topic_0140			
	Gene expression	http://edamontology.org/topic_0203			
	gene regulation	http://edamontology.org/topic_0204			
	protein modification	http://edamontology.org/topic_0601			
	Molecular interactions, pathways and networks	http://edamontology.org/topic_0602			
	metabolomics	http://edamontology.org/topic_3172			
	epigenomics	http://edamontology.org/topic_3173			
	structural variation	http://edamontology.org/topic_3175			
	phenomics	http://edamontology.org/topic_3298			
	raw data	http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C142863			
	genotype data	http://purl.bioontology.org/ontology/MESH/D005838			
	pedigree	http://purl.bioontology.org/ontology/MESH/D010375			
	phenotype data	http://purl.bioontology.org/ontology/MESH/D010641			
	whole genome association study	http://purl.obolibrary.org/obo/NCIT_C93020 (ALSO http://edamontology.org/topic_3517)			
	genetic interaction	http://purl.obolibrary.org/obo/VariO_0237	VariO:0237		

Next Steps - formatting metadata

Hmm...

White Paper SOP for guiding data management

Provide checklist, Best ways for Authors, Reviewers, Editors, Staff (production)