eResearch Australasia 2024

Taking stock of biological data resources

Keeva Connolly, Matt Andrews, Peter Brenton, Simon Checksfield, Christopher Mangion, Winnie Mok, Caitlin Ramsay, Sarah Richmond, Goran Sterjov, Nigel Ward, and Kathryn Hall

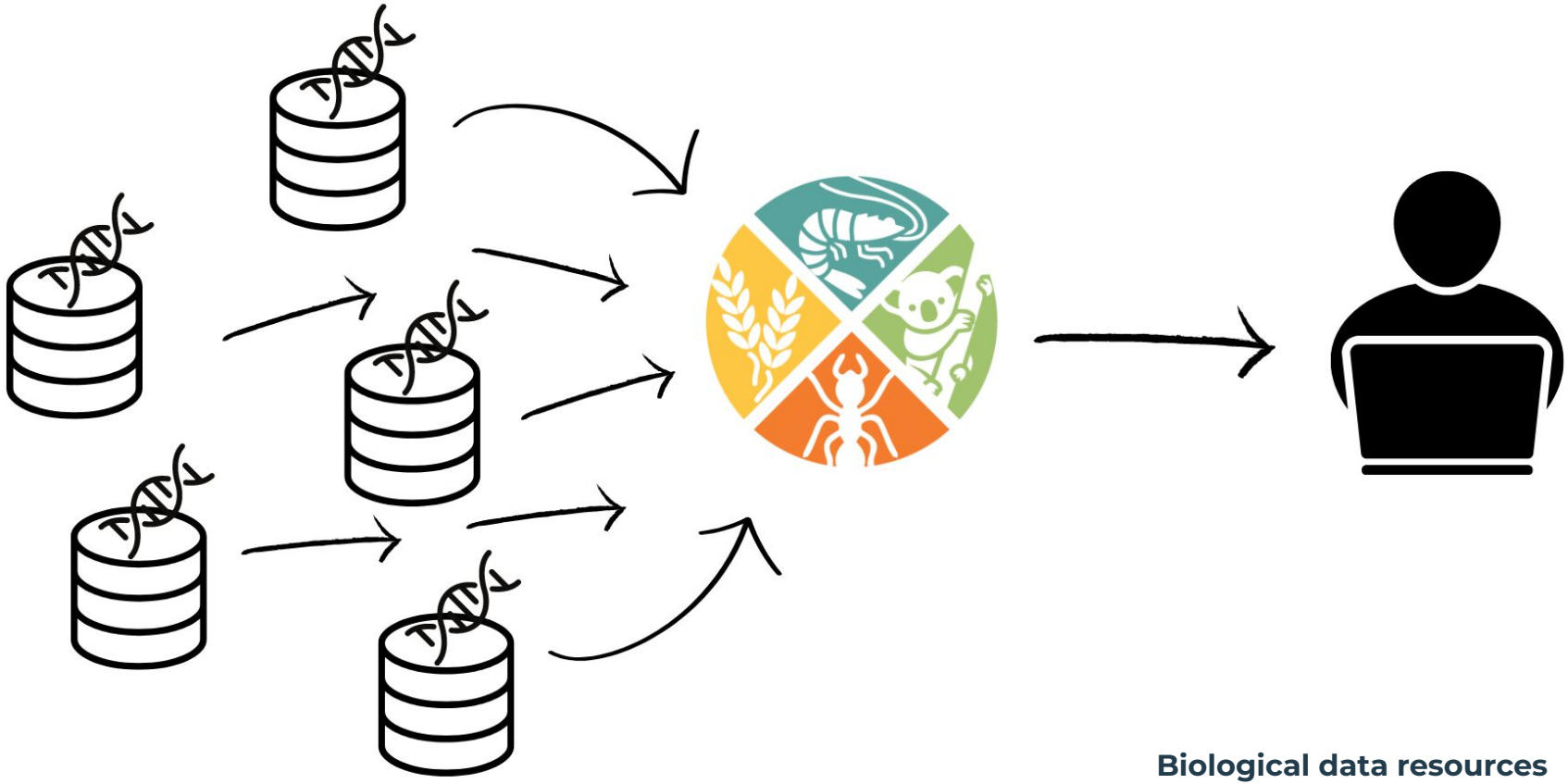
Atlas of Living Australia
Australian BioCommons
Bioplatforms Australia

Biological data (biodata) resources

Definition: any biological, life science, or biomedical database that archives research data generated by scientists, or functions as a knowledgebase by adding value to scientific data by aggregation, processing, and expert curation.

Source: Global Biodata Coalition

The Australian Reference Genome Atlas



ARGA is a genomic data indexing platform

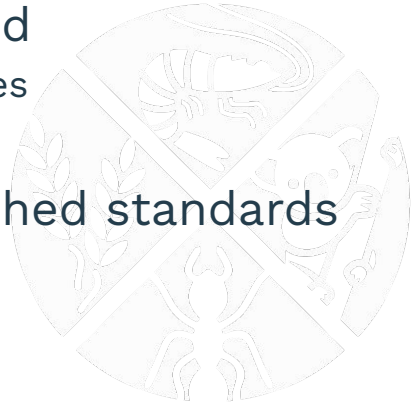
Challenges for ARGA

1. Identify biodata resources

2. Determine whether their scope overlaps with ARGA
   - Data types, location, taxa

3. Determine how their data are sourced or generated
   - Primary research data, curated datasets, secondary analyses

4. Identify whether data are described using established standards
   - Including data schemas and ontologies

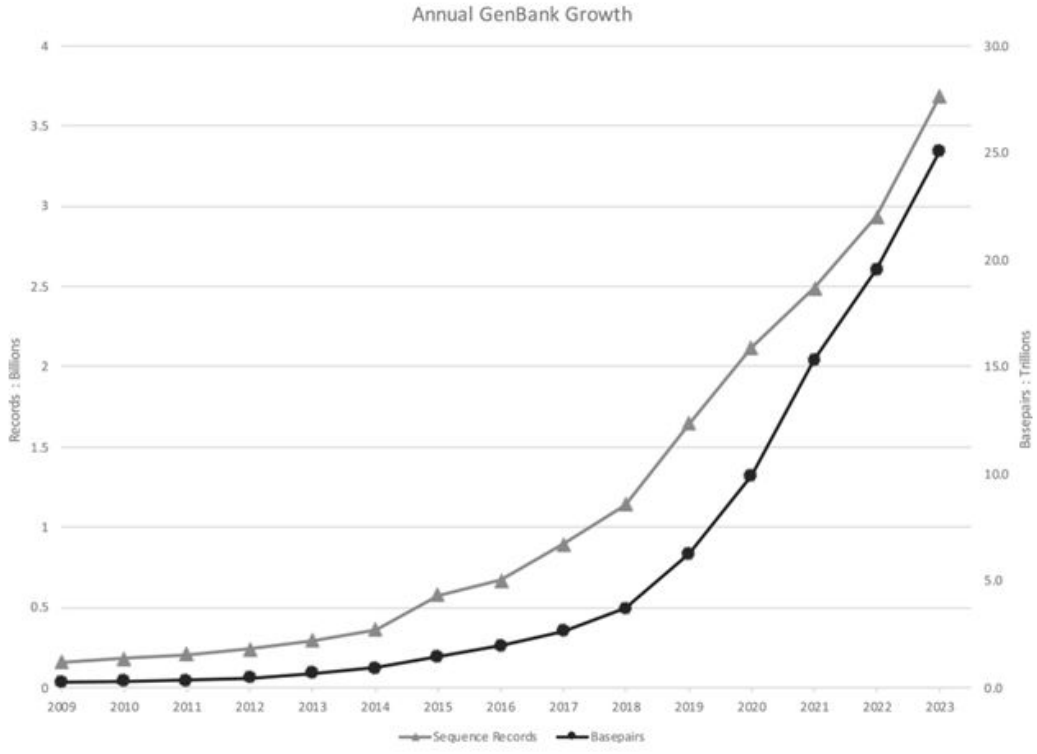


There are many biodata resources

Major global resources:

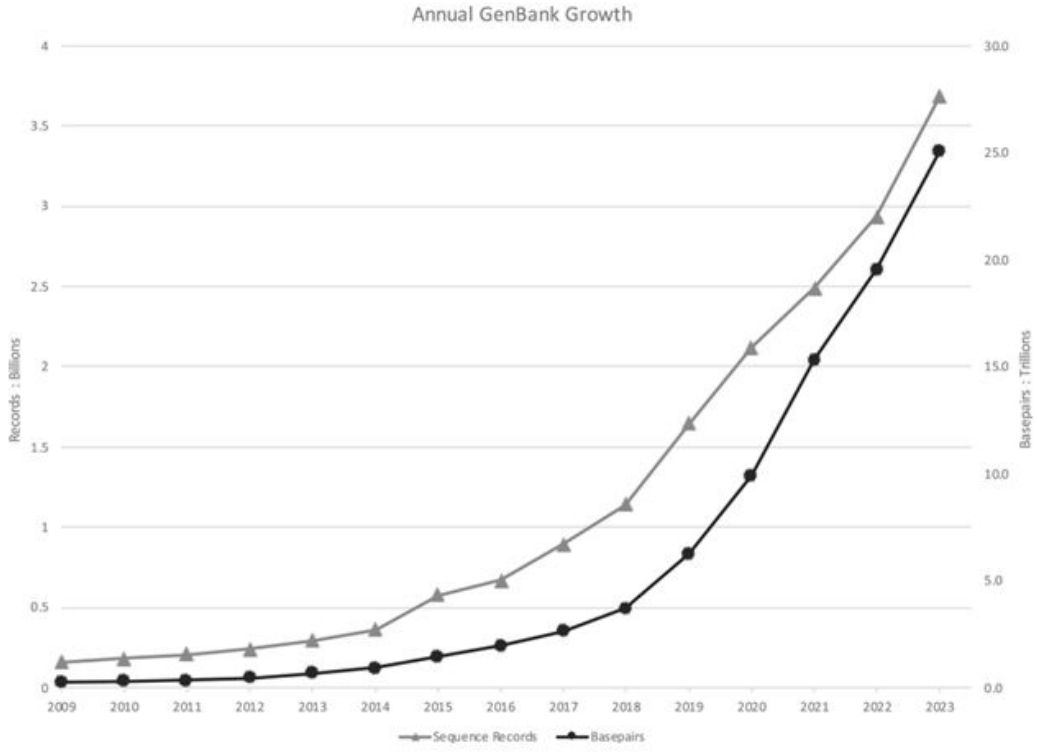
- Genomic sequence data - International Nucleotide Sequence Database Collaboration (INSDC) [NCBI, EMBL, DDBJ]
- Species occurrences - Global Biodiversity Information Facility (GBIF)
- Protein structural data - Protein Data Bank (PDB)
- Taxonomic data - Catalogue of Life (CoL)
- Genome annotation data - Ensembl

Data generation is accelerating



Eric W Sayers, Mark Cavanaugh, Karen Clark, Kim D Pruitt, Stephen T Sherry, Linda Yankie, Ilene Karsch-Mizrachi, GenBank 2024 Update, *Nucleic Acids Research*, Volume 52, Issue D1, 5 January 2024, Pages D134–D137, https://doi.org/10.1093/nar/gkad903

Data generation is accelerating

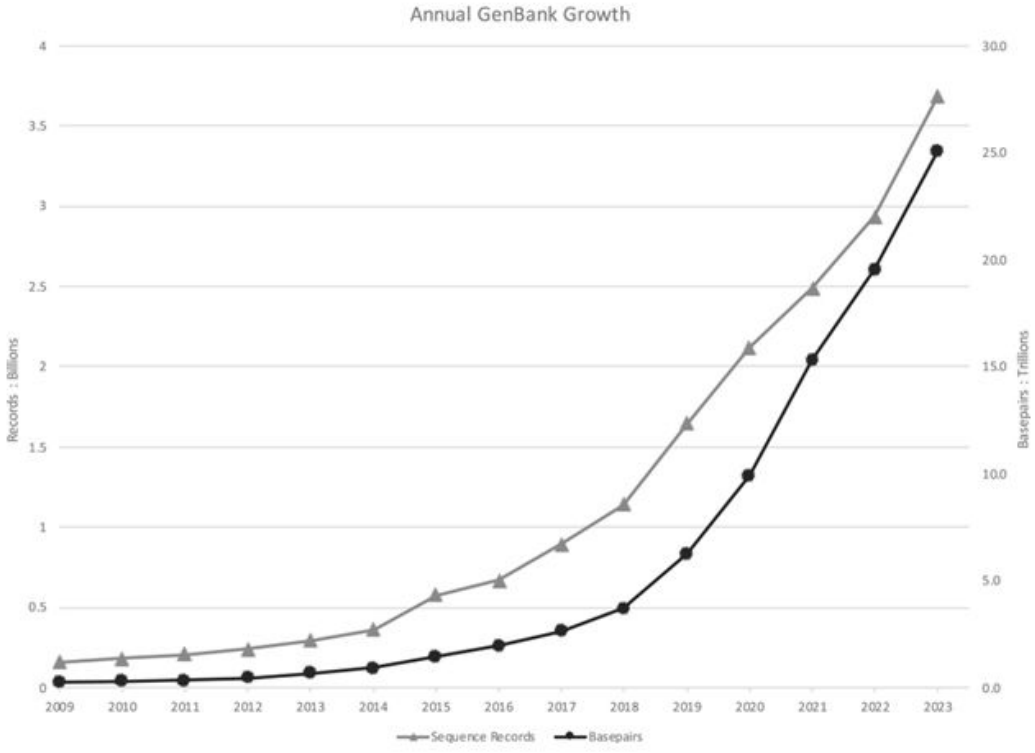


"Difficult to find data"

Eric W Sayers, Mark Cavanaugh, Karen Clark, Kim D Pruitt, Stephen T Sherry, Linda Yankie, Ilene Karsch-Mizrachi, GenBank 2024 Update, *Nucleic Acids Research*, Volume 52, Issue D1, 5 January 2024, Pages D134–D137, https://doi.org/10.1093/nar/gkad903

Data generation is accelerating

"Unsuited to certain data types"

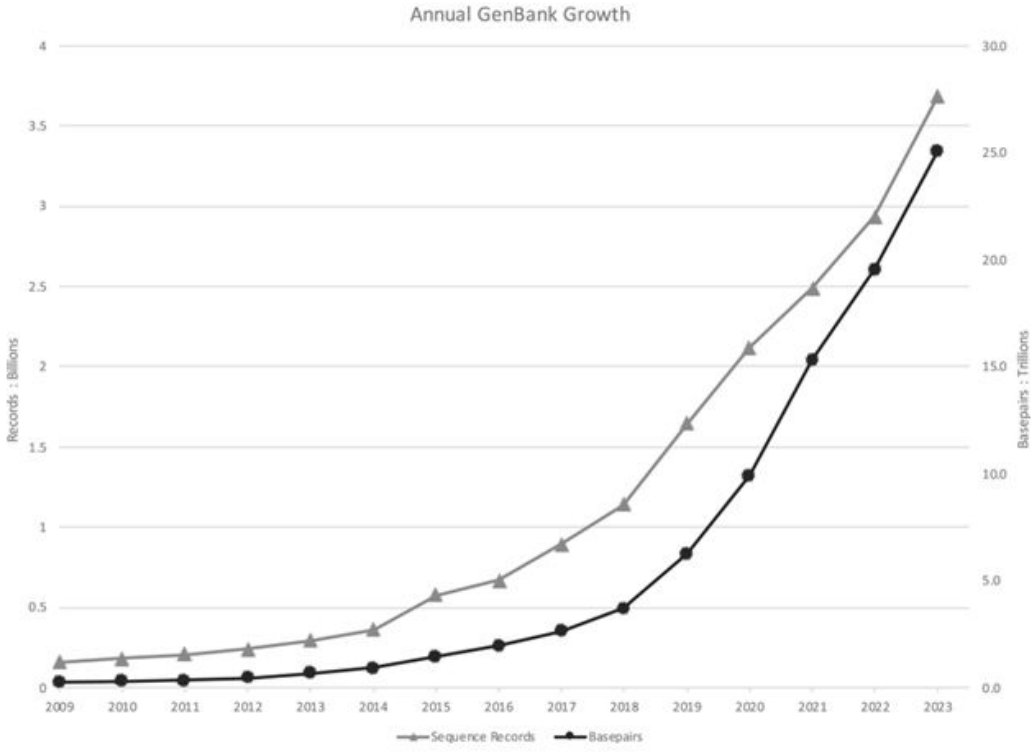


"Difficult to find data"

Eric W Sayers, Mark Cavanaugh, Karen Clark, Kim D Pruitt, Stephen T Sherry, Linda Yankie, Ilene Karsch-Mizrachi, GenBank 2024 Update, *Nucleic Acids Research*, Volume 52, Issue D1, 5 January 2024, Pages D134–D137, https://doi.org/10.1093/nar/gkad903

Data generation is accelerating

“Unsuited to certain data types”



“Difficult to find data”

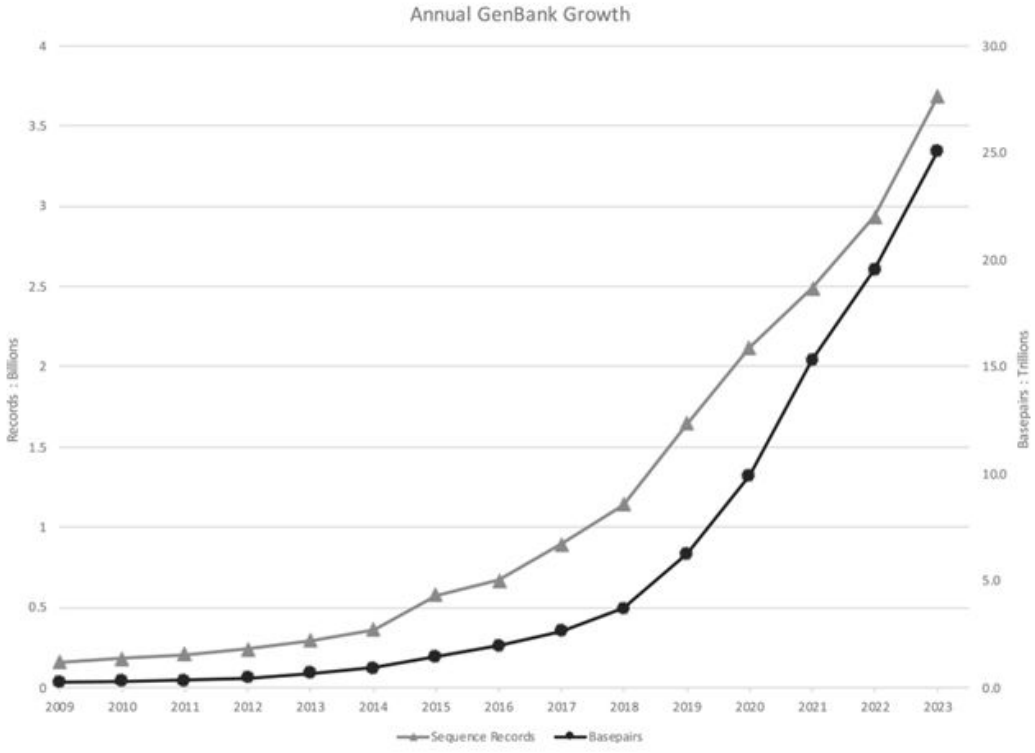
“Difficult to submit to”

Eric W Sayers, Mark Cavanaugh, Karen Clark, Kim D Pruitt, Stephen T Sherry, Linda Yankie, Ilene Karsch-Mizrachi, GenBank 2024 Update, *Nucleic Acids Research*, Volume 52, Issue D1, 5 January 2024, Pages D134–D137, https://doi.org/10.1093/nar/gkad903

Data generation is accelerating

"Unsuited to certain data types"

"Missing or unstandardised metadata"



"Difficult to find data"

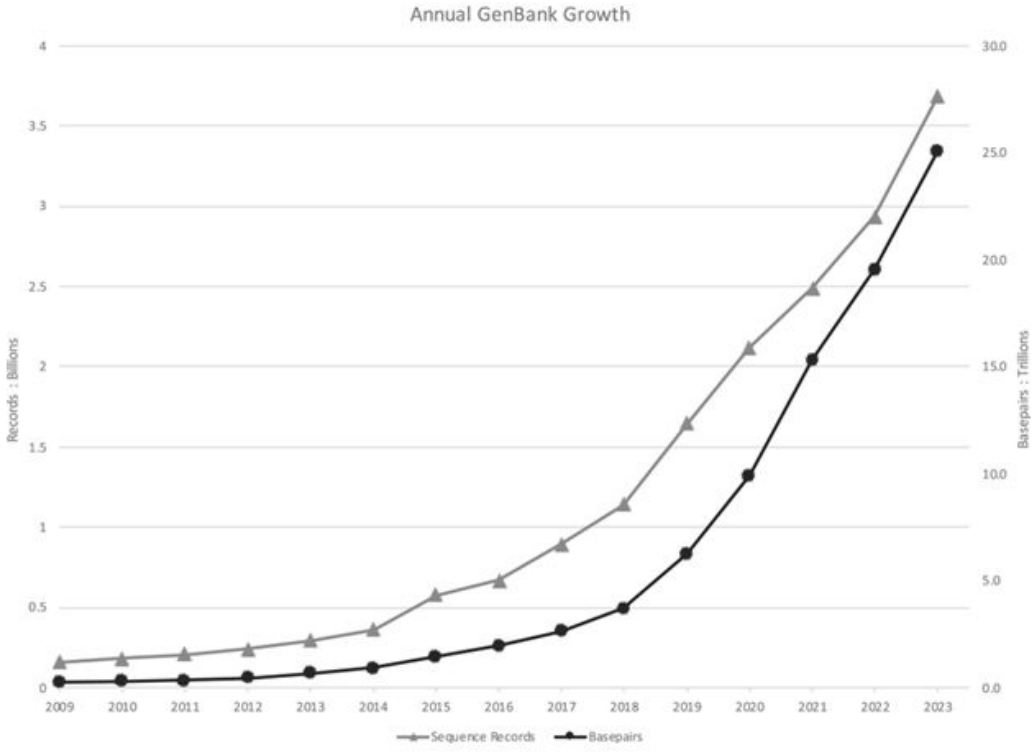
"Difficult to submit to"

Eric W Sayers, Mark Cavanaugh, Karen Clark, Kim D Pruitt, Stephen T Sherry, Linda Yankie, Ilene Karsch-Mizrachi, GenBank 2024 Update, *Nucleic Acids Research*, Volume 52, Issue D1, 5 January 2024, Pages D134–D137, https://doi.org/10.1093/nar/gkad903

Data generation is accelerating

"Unsuited to certain data types"

"Missing or unstandardised metadata"



"Difficult to find data"

"Difficult to submit to"

"Erroneous data"

Eric W Sayers, Mark Cavanaugh, Karen Clark, Kim D Pruitt, Stephen T Sherry, Linda Yankie, Ilene Karsch-Mizrachi, GenBank 2024 Update, *Nucleic Acids Research*, Volume 52, Issue D1, 5 January 2024, Pages D134–D137, https://doi.org/10.1093/nar/gkad903

Alternatives to major biodata resources

- Institutional and generalist repositories - for example:
  - University data repositories
  - Dryad, Figshare, Zenodo
- Specialised biodata resources - often restricted by:
  - Data type
    - BOLD repository for barcodes and genetic markers
  - Taxon
    - ReefGenomics for data from corals, sponges and their symbionts
  - Location
    - Aotearoa Genomic Data Repository for data from Aotearoa/New Zealand
  - Specific research project or large-scale analysis
    - GenomeArk S3 repository for data from the Vertebrate Genome Project

There are many biodata resources

Resources cataloguing distributed biodata resources:

- FAIRsharing.org - 2,253 databases
- re3data - 1,868 ‘life sciences’ databases
- Database Commons - 6,919 biological databases
- *NAR* Molecular Biology Database Collection - 1,959 databases

Global Biodata Coalition resource inventory

PLOS ONE

RESEARCH ARTICLE

A machine learning-enabled open biodata resource inventory from the scientific literature

Heidi J. Imker[1,2]*, **Kenneth E. Schackart, III**[1,3], **Ana-Maria Istrate**[4], **Charles E. Cook**[1]

1 Global Biodata Coalition, Strasbourg, France, **2** University Library, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **3** Department of Biosystems Engineering, The University of Arizona, Tucson, Arizona, United States of America, **4** Chan Zuckerberg Initiative, Redwood City, California, United States of America

Global Biodata Coalition resource inventory

- Analysed a literature corpus from Europe PMC (published 2011-2021)
- Used a natural language processing model to identify publications expected to describe biodata resources from titles and abstracts
- Extracted resource names and URLs using named entity recognition
- The final inventory contained 3,112 resources

Analysing the GBC inventory

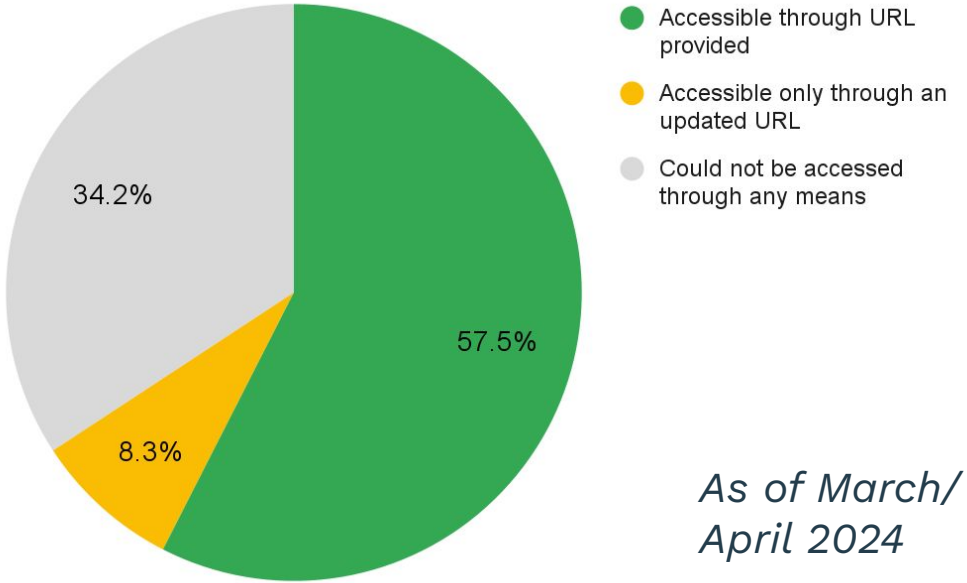
1. Confirm the resource is still available

2. Categorise:
   a. Relevant taxa
   b. Content types
   c. Research area

3. For a subset of relevant resources:
   a. Identify data source/s
   b. Identify data standards and schemas

Resource availability

1. Confirm the resource is still available



As of March/ April 2024

Biodata resource sustainability

Resource sustainability requires:

- Ongoing data curation, processing and/or management
- Compute and/or storage capacity
- Database and website maintenance

This depends on prolonged funding.

The Global Biodata Coalition is currently targeting mechanisms to support a more sustainable research infrastructure.

Taxonomic diversity

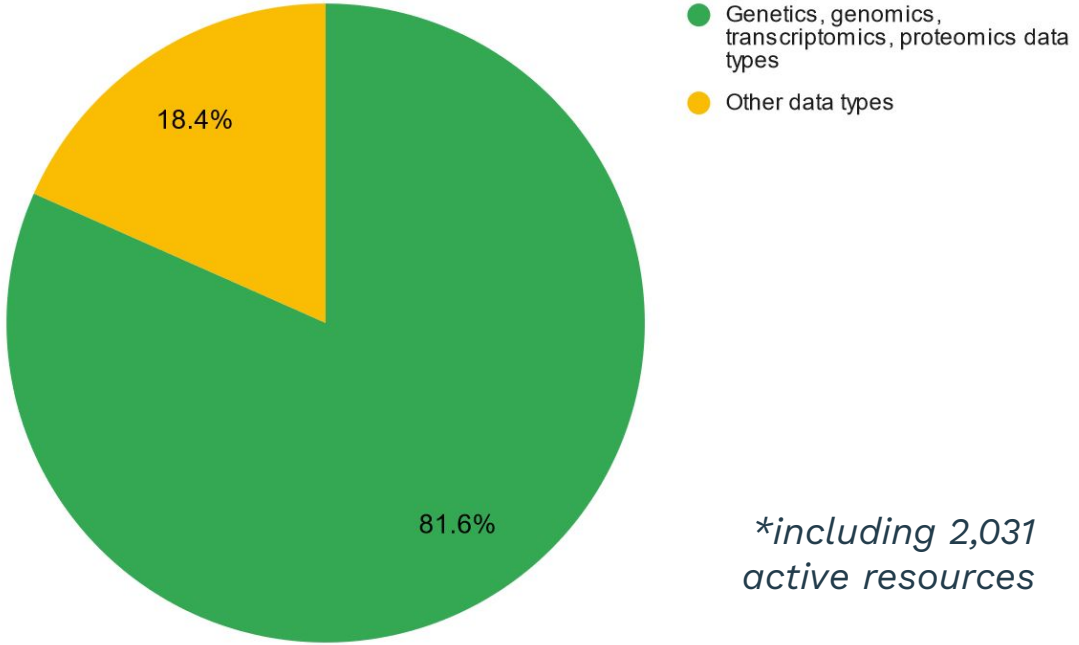
2. Categorise:
   a. Relevant taxa
   b. Content types
   c. Research area

Relevant taxa	#	%
Humans	922	55.3%
Model organisms	444	26.6%
Plants/algae	366	22.0%
Animals	288	17.3%
Bacteria	236	14.2%
Fungi	145	8.7%
Viruses	108	6.5%
Archaea	86	5.2%
Protists	67	4.0%
Chromists	7	0.4%

**including 1,667 resources for which taxa could be identified*

Content type diversity

2. Categorise:
   a. Relevant taxa
   b. Content types
   c. Research area



**including 2,031 active resources*

Research area diversity

2. Categorise:
   a. Relevant taxa
   b. Content types
   c. Research area

Research domain	#	%
Clinical (all domains)	786	38.7%
Gene annotation/expression	629	31.0%
Molecular interactions	331	16.3%
Clinical (host genetics)	288	14.2%
Clinical (other)	250	12.3%
Clinical (drugs/therapeutics)	221	10.9%
Physical structure	202	9.9%
Feature annotation	181	8.9%
Biodiversity	169	8.3%

**including all 2,031 active resources*

Research area diversity

2. Categorise:
   a. Relevant taxa
   b. Content types
   c. Research area

Research domain	#	%
Gene annotation and expression	510	41.0%
Molecular interactions	253	20.3%
Physical structure	166	13.3%
Protein and RNA annotation	156	12.5%
Biodiversity	134	10.8%
Agriculture	132	10.6%
Phylogenetics and evolution	130	10.4%
Genetic regulation	105	8.4%
Comparative genomics	66	5.3%

**including 1,245 non-clinical resources*

Data sources

3. For a subset of relevant resources:
   a. Identify data source/s
   b. Identify data standards and schemas

Data source	#	%
Database/s (analysed/transformed)	233	48.9%
Literature (extracted)	129	27.1%
Database/s (extracted)	99	20.8%
Primary data	69	14.5%
User submitted	53	11.1%
Literature (analysed/transformed)	23	4.8%

**including 476 ARGA-relevant resources*

Primary and secondary data resources

Primary data resources: databases or repositories archiving primary research data.

Secondary data resources: databases or knowledgebases containing externally sourced data, for the purposes of either:

- aggregation, curation, and/or standardisation; or
- downstream analysis to generate novel data.

Data standards

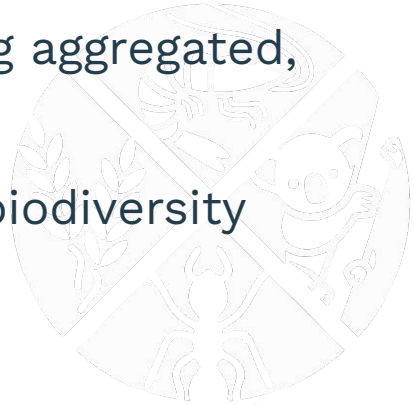
3. For a subset of relevant resources:
    a. Identify data source/s
    b. Identify data standards and schemas

Schema	#	%
Genomics Standards Consortium (MIxS)	7	1.5%
Biodiversity standards (Darwin Core, ABCD, GGBN)	5	1.1%
Combined genomics and biodiversity standards	3	0.6%
Other minimum information standards	13	2.7%

**including 476 ARGA-relevant resources*

Conclusions

- Resources face major sustainability challenges
- Trends
  - Taxa: human and model organism groups
  - Content types: genes and gene product-related data
  - Research domains: clinical, followed by less application-focussed areas
- A majority of secondary data resources (containing aggregated, standardised, curated, or derived data)
- Very few resources adhere to major genomics or biodiversity standards



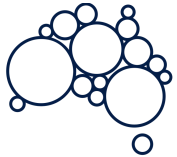
ARGA Partnerships

The Australian Reference Genome Atlas (ARGA) is an NCRIS-enabled platform powered by the Atlas of Living Australia (ALA), in collaboration with Bioplatforms Australia and the Australian BioCommons, with investment from the Australian Research Data Commons (ARDC) (https://doi.org/10.47486/DC011). ARGA integrates data sourced from a number of international repositories, including NCBI GenBank, EMBL-ENA and Bioplatforms Australia.

ARGA
Australian Reference Genome Atlas

Atlas of Living Australia
ala.org.au

Australian BioCommons



BIOPLATFORMS AUSTRALIA

Australian Research Data Commons

NCRIS
National Research Infrastructure for Australia
An Australian Government Initiative

ARGA Development Team

Caitlin Ramsay	Atlas of Living Australia	Software Engineer
Christopher Mangion	Atlas of Living Australia	Data Engineer
Goran Sterjov	Atlas of Living Australia	Software Engineer
Jack Brinkman	Atlas of Living Australia	Software Engineer
Matt Andrews	Atlas of Living Australia	Systems Support
Mok	Australian BioCommons	UX/UI Designer
Keeva Connolly	Australian BioCommons	Scientific Business Analyst
Kathryn Hall	Atlas of Living Australia	Project Manager