Harmonising metadata to improve data discoverability and interoperability

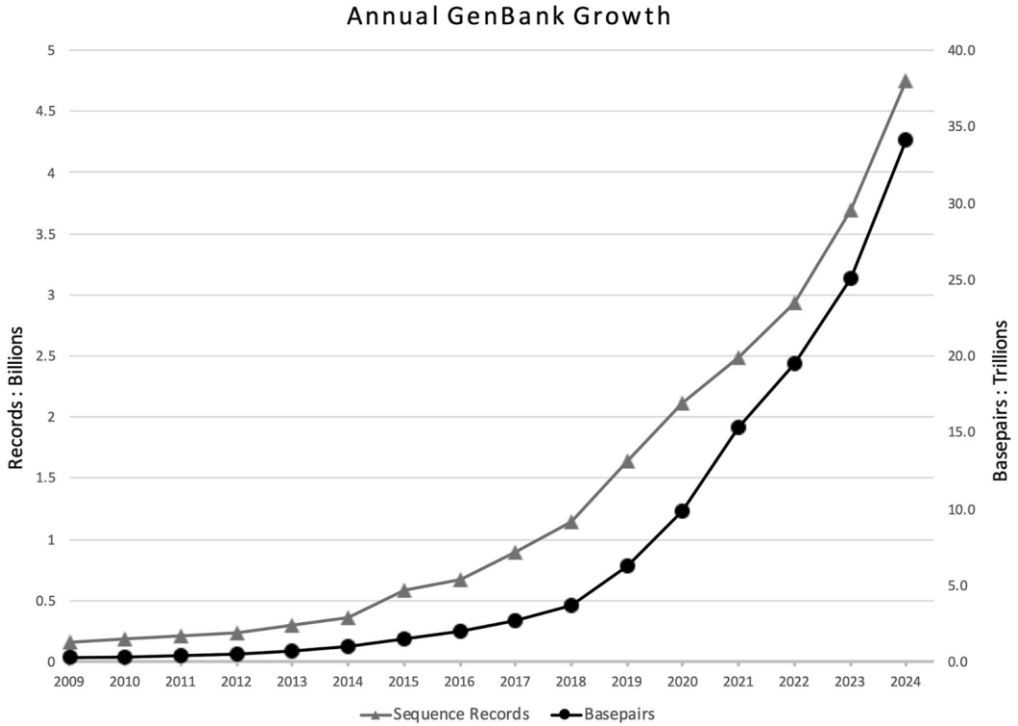
eResearch Australasia
22 October 2025

Keeva Connolly[1,2], Matt Andrews[3], Peter Brenton[3], Jack Brinkman[3], Christopher Mangion[3], Emily Marshall[1], Winnie Mok[1], Lisa Phippard[1], Sarah Richmond[4], Goran Sterjov[3], Nigel Ward[1], Tom Harrop[1], Kathryn Hall[3]

[1]Australian BioCommons, [2]QCIF, [3]Atlas of Living Australia, [4]Bioplatforms Australia

Australian BioCommons

Genomic data growth is accelerating



Sayers, E. W., Cavanaugh, M., Frisse, L., Pruitt, K. D., Schneider, V. A., Underwood, B. A., Yankie, L., & Karsch-Mizrachi, I. (2025). GenBank 2025 update. *Nucleic acids research*, *53*(D1), D56–D61. https://doi.org/10.1093/nar/gkae1114

Australian BioCommons

New biological databases are being released

- Each year, more data resources are published online.
- As of October 2025, there are:
  - **1,974** life sciences databases registered in RE3data.org
  - **2,039** life sciences databases registered in FAIRsharing.org
  - **2,236** databases registered in the NAR Molecular Biology Database Collection
  - **7,354** biological databases registered in Database Commons

More data → more opportunity for reuse

- Genome assemblies can be used as reference genomes to **investigate population structure** and diversity
- Genetic marker sequences can be used in **species identification**
- Genome annotations can be used in comparative genomics to **identify divergent genes** between taxa

More data → more opportunity for reuse

- Combining and enriching genomic data can provide **more contexts for reuse** - e.g. sequence data records with occurrence information can be used to:
  - devise breeding programs,
  - trace species invasion routes,
  - investigate clues about selective pressures.
- More comprehensive metadata can also help researchers **trust** data

There is no single common data standard

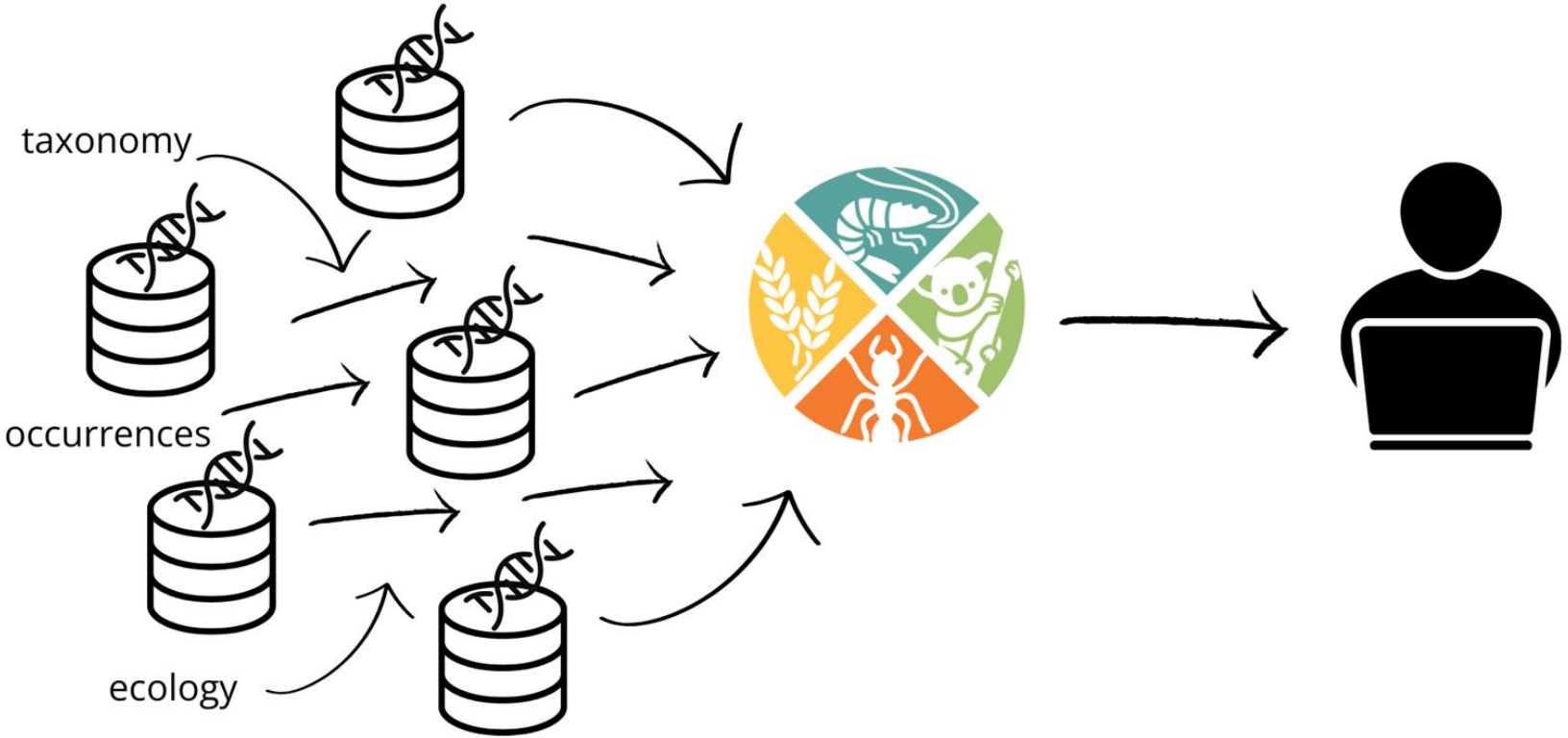
- Researchers with different research goals are interested in different facets of data and metadata
- Databases will select or develop a schema depending on the scope of their content
    - Schema may or may not implement existing data standards (of which there are many)

Disparate data are not easily integrated

- Without a common standard or standards which are interoperable, it can be difficult to:
  - Integrate data into one dataset
  - Exchange data between databases and resources

Case study 1: Australian Reference Genome Atlas (ARGA)

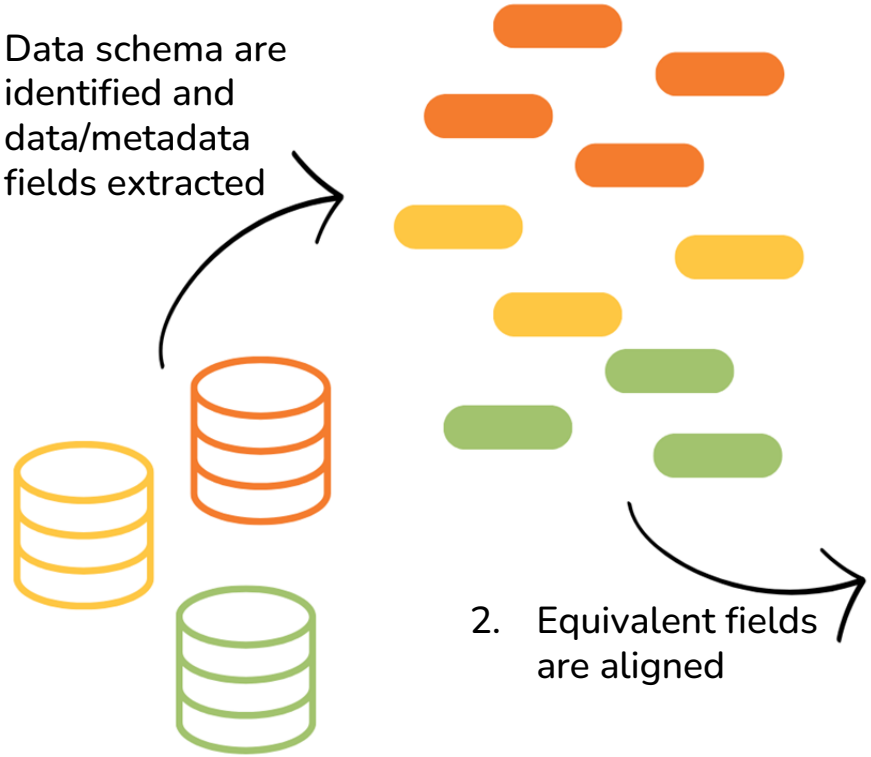
ARGA is a platform for genomic data discovery



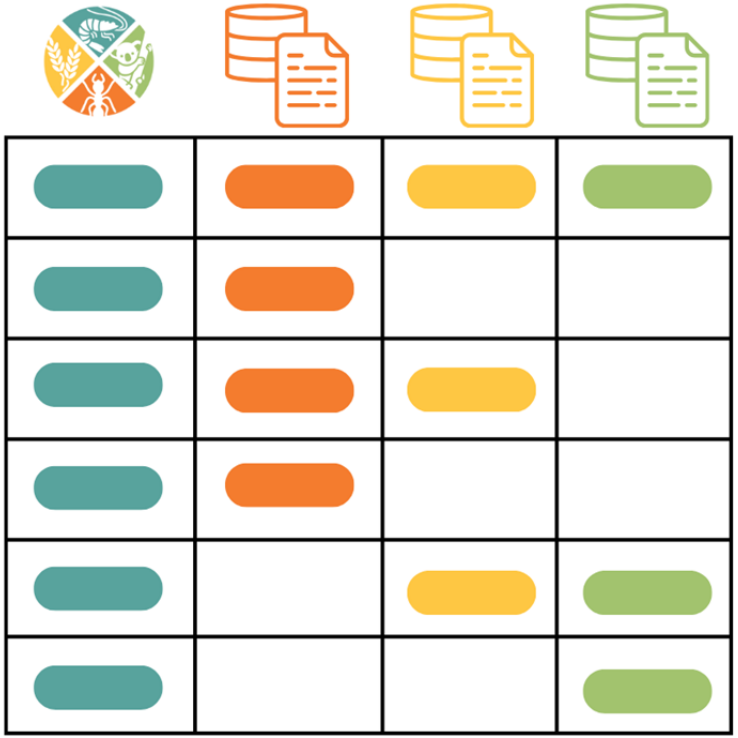
Australian BioCommons

Mapping data fields to a harmonising schema

1. Data schema are identified and data/metadata fields extracted



2. Equivalent fields are aligned



Australian BioCommons

Mapping data fields to a harmonising schema

ARGA schema	Darwin Core	Bioplatforms Australia	NCBI Nucleotide
collected_by	recordedBy	collector; collected_by; collectors; voucher_herbarium_collector_id	collected_by
collection_method	samplingProtocol	collection_method; samp_mat_process	-
identified_by	identifiedBy	identified_by	identified_by
identification_method	-	identification_method	-

Mapping data fields to a harmonising schema

Collection event

ARGA schema	Darwin Core	Bioplatforms Australia	NCBI Nucleotide
collected_by	recordedBy	collector; collected_by; collectors; voucher_herbarium_collector_id	collected_by
collection_method	samplingProtocol	collection_method; samp_mat_process	-
identified_by	identifiedBy	identified_by	identified_by
identification_method	-	identification_method	-

Accommodating complexity

- The original harmonising schema was relatively flat, which did not suit many-to-one and one-to-many relationships
- The revised ARGA format includes:
  - Additional record level metadata for each event
  - Unique entity identifiers for each event

Harmonisation facilitates integration

- Once data are harmonised to a common schema, records can be:
  - Subject to catalogue-wide **searching**
  - **Filtered** according to the metadata features
  - Combined in **summary statistics**
  - Made available to **download** in a standardised format
  - Combined across multiple sources to **supplement metadata** and contextual information

Case study 2: The Australian Tree of Life (AToL)

AToL assembles and brokers genome sequences

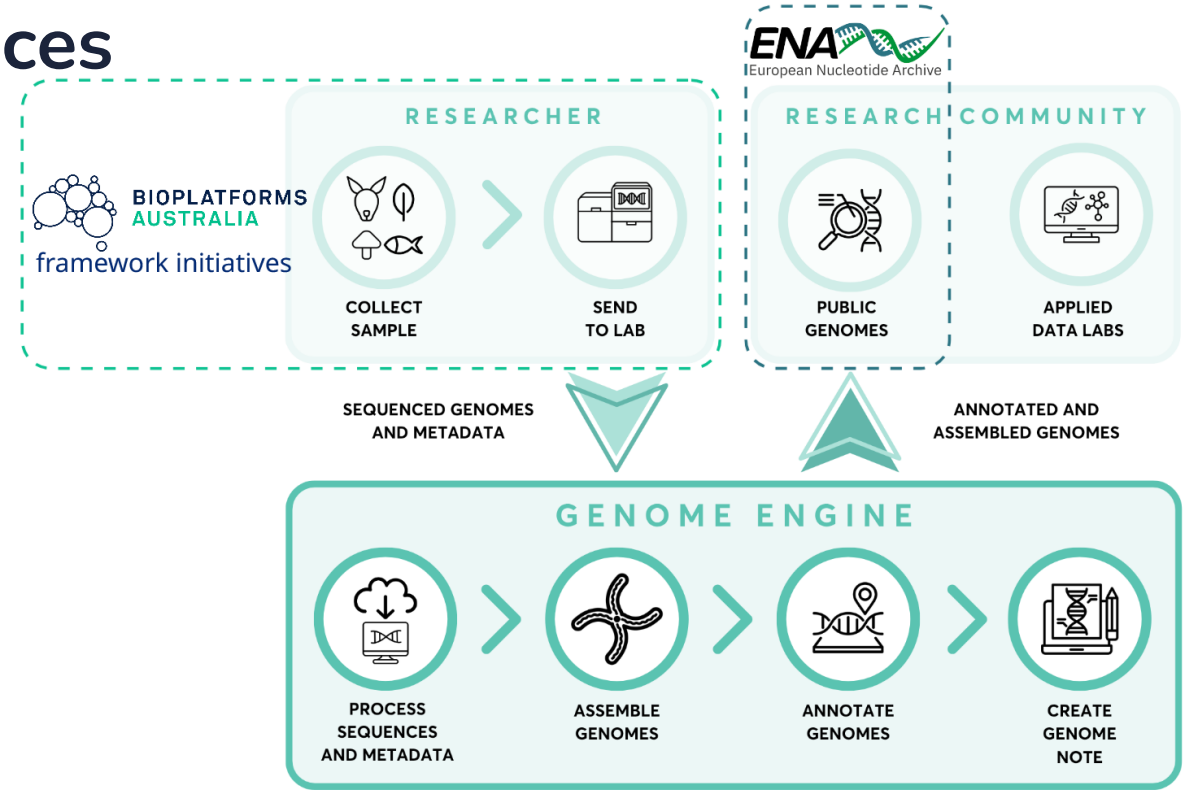


Image credit: Rahul Ratwatte & Christina Hall

Australian BioCommons

Data mapping and processing

1. Bioplatforms data records are **filtered**
2. Metadata are **mapped** from the Bioplatforms schema to the AToL schema
3. Metadata are **transformed**

Data mapping and processing

1. Bioplatforms data records are **filtered**, according to:
    a. Relevance
        i. Data generated by a biodiversity initiative
        ii. Data generated for the purposes of assembly and annotation
    b. Minimum metadata requirements (for brokering to ENA)

Data mapping and processing

2. Metadata are **mapped**:
    a. Values are mapped from Bioplatforms schema to the AToL schema
    b. Values are sanitised and normalised to controlled vocabularies (for applicable fields)
    c. Taxonomic information is aligned to the NCBI taxonomy

Data mapping and processing

3. Metadata are **transformed** to:
    a. Extract unique organisms (taxa)
    b. Extract unique samples

Mapped data are compliant with ENA standards

- The AToL schema has also been mapped to:
  - The ENA Biosample Tree of Life sample checklist
    - The Tree of Life checklist is also used by other international sequencing efforts - e.g. the Darwin Tree of Life
  - The ENA experiment/read database
- Processed and mapped metadata can be automatically formatted in XML for programmatic submission
  - Metadata should pass ENA validations

Harmonisation facilitates exchange and reuse

- After metadata mapping and processing:
  - Sequence read data and downstream assemblies can be **brokered** to ENA automatically
  - Metadata can be used to **populate genome note templates**, to fast-track publishable outputs
  - Researchers will be able to **discover, access and use assemblies** in their research

Acknowledgements

Australian Reference Genome Atlas

- Kathryn Hall
- Matt Andrews
- Jack Brinkman
- Christopher Mangion
- Goran Sterjov

Australian Tree of Life

- Tom Harrop
- Emily Marshall
- Amy Timms
- Jane Tung

BIOPLATFORMS
AUSTRALIA

Australian Research Data Commons

QCIF
Digital Research³