Knitting jumpers from steel wool and spaghetti: implementing a modified Darwin Core Event model for the Australian Reference Genome Atlas (ARGA) to increase trust through provenance

Kathryn Hall*, Matt Andrews, Keeva Connolly, Nick dos Remedios, Yasima Kankanamge, Christopher Mangion, Winnie Mok, Vikas Nagaraju, Lars Nauheimer, Sarah Richmond, Goran Sterjov, Nigel Ward, and Peter Brenton

Atlas of Living Australia
Australian BioCommons
Bioplatforms Australia

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

ARDC
Australian Research Data Commons

NCRIS
National Research Infrastructure for Australia
An Australian Government Initiative

Why build ARGA?

Data sources are
obtuse
complex
different
scattered
disconnected

Genomics can
improve outcomes
for livestock
breeding and
primary industries
research

Bushfires
(and another
environmental
catastrophe)
responses can be
proactive,
not reactive

15,000 life science
researchers in Australia
can supercharge their
searches for relevant
data using occurence
records and curated
traits filters

Community derived aims for the ARGA application

Users wanted to trust found data

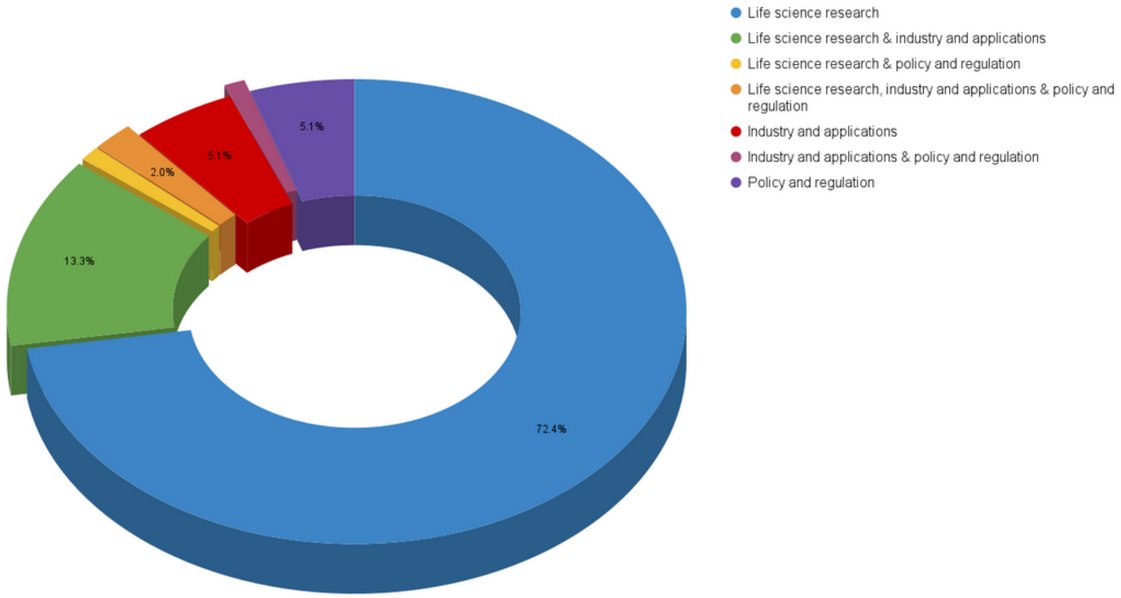
- data quality
- taxonomic certainty
- metadata sufficiency

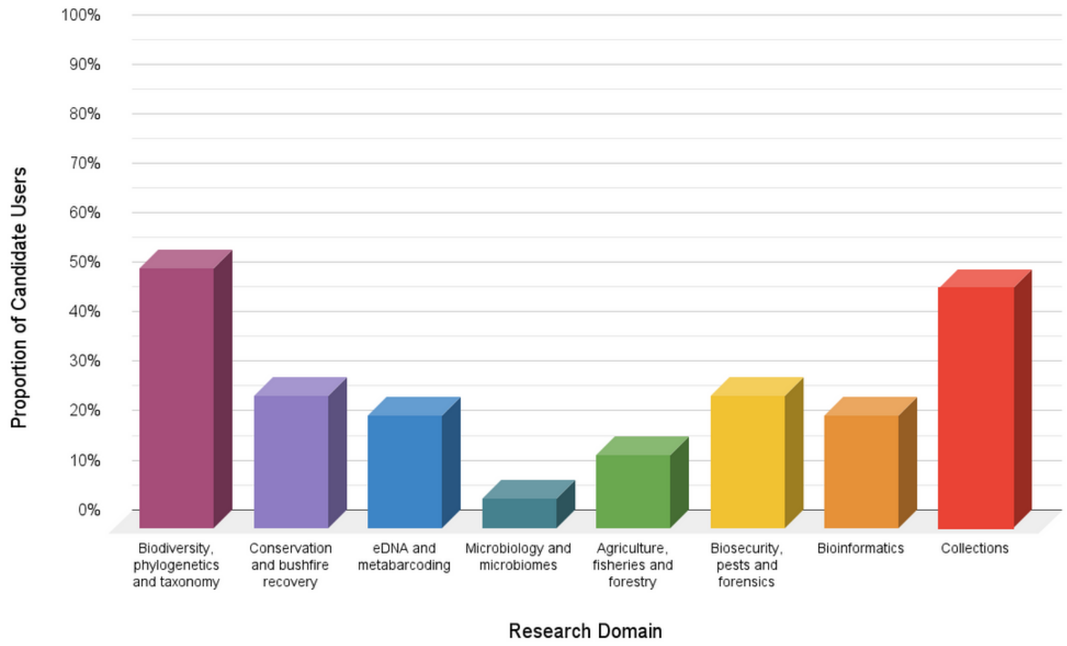
Connolly, K., & Hall, K. A. (2023, May 31). ARGA Project Community Engagement Report. https://doi.org/10.17605/OSF.IO/YRZ72

Consultation cohort

ARGA consulted with

- 98 people from
- 38 institutions around Australia





Connolly, K., & Hall, K. A. (2023, May 31). ARGA Project Community Engagement Report. https://doi.org/10.17605/OSF.IO/YRZ72

	PROPOSED SOLUTION	ISSUES ADDRESSED			SOLUTION ARCHITECTURE	TECHNICAL COMPLEXITY
		Taxonomic certainty	Metadata sufficiency	Data quality		
LITERATURE AND SECONDARY SOURCES	Provide link to source publication or PubMed page	✓	✓	✓	ARGA to provide data enriched by this source	1
	Provide link to relevant ALA page or other relevant database page	✓	✓		ARGA to provide data enriched by this source	2
	Provide citation count for source publication	✓		✓	ARGA to provide data enriched by this source	3
SPECIMEN METADATA	Generate de novo taxonomic confidence scores	✓			Build custom algorithm within ARGA	2
	Provide specimen accession number	✓			Build systems to access and ingest collection data	3
	Provide voucher/registration status	✓			Build systems to access and ingest collection data	3
	Provide specimen photo	✓			Build systems to access and ingest collection data	3
	Provide contact information for specimen identifier/collector	✓	✓		Build systems to access and ingest collection data; integrate with ORCiD	4
SEQUENCING HISTORY AND ANALYSIS	Provide metadata regarding data provenance (e.g. author names, institution, sequencing project)	✓		✓	Build access path to data from original source	1
	Provide original sequence chromatograms and other QC data generated during sequencing	✓		✓	Build access path to data from original source; create raw data preview functionality	1
	Provide contact information for data depositor		✓		Build systems to access and ingest collection data; integrate with ORCiD	2
	Generate metadata completeness scores (based on a series of assertions checking presence or absence of relevant metadata)		✓		Build custom algorithm within ARGA	2
	Implement systems to package data and ship to Galaxy Australia for QC analyses	✓		✓	Build systems integrating ARGA into the BioCommons ecosystem; access via CILogon	4
USER-BASED TOOLS	Implement customisable metadata filters		✓		ARGA to provide data enrichment via additional original sources	1
	Provide view and download counts for each datum	✓		✓	Build custom system within ARGA	2
	Implement user-based "add to favourites"/"up-vote" function for each datum	✓		✓	Build custom system within ARGA	2
	Implement ticket-based system to register and respond to user feedback or queries	✓		✓	Build systems to facilitate community-based data curation	2

Building ARGA

Resolving taxonomic uncertainty

- NSL and AFD as primary name sources
- inclusion of informal taxa

Data quality assessment

- visualisation tools for raw data

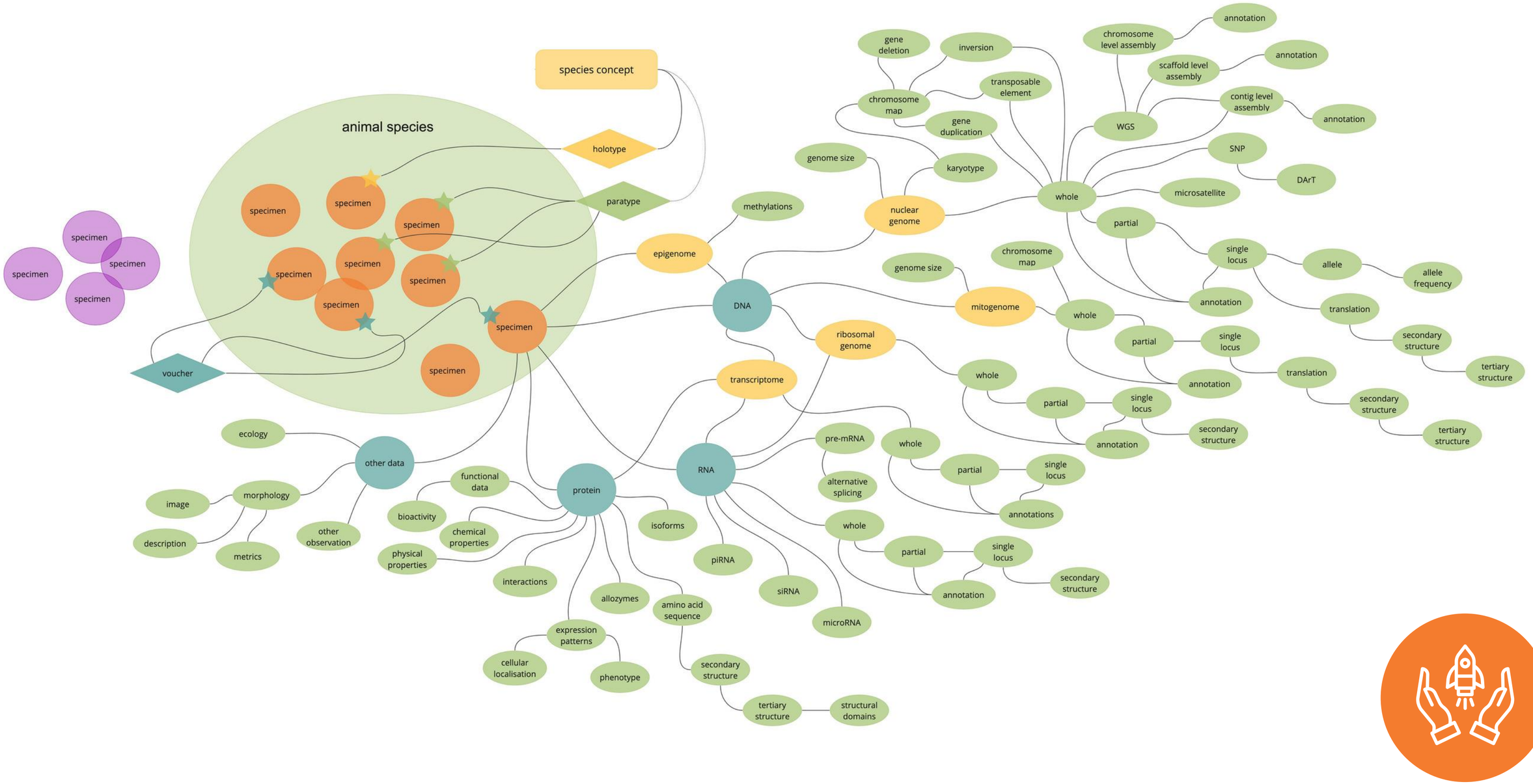
Metadata sufficiency

- new links with specimen records from ALA and specimen repositories

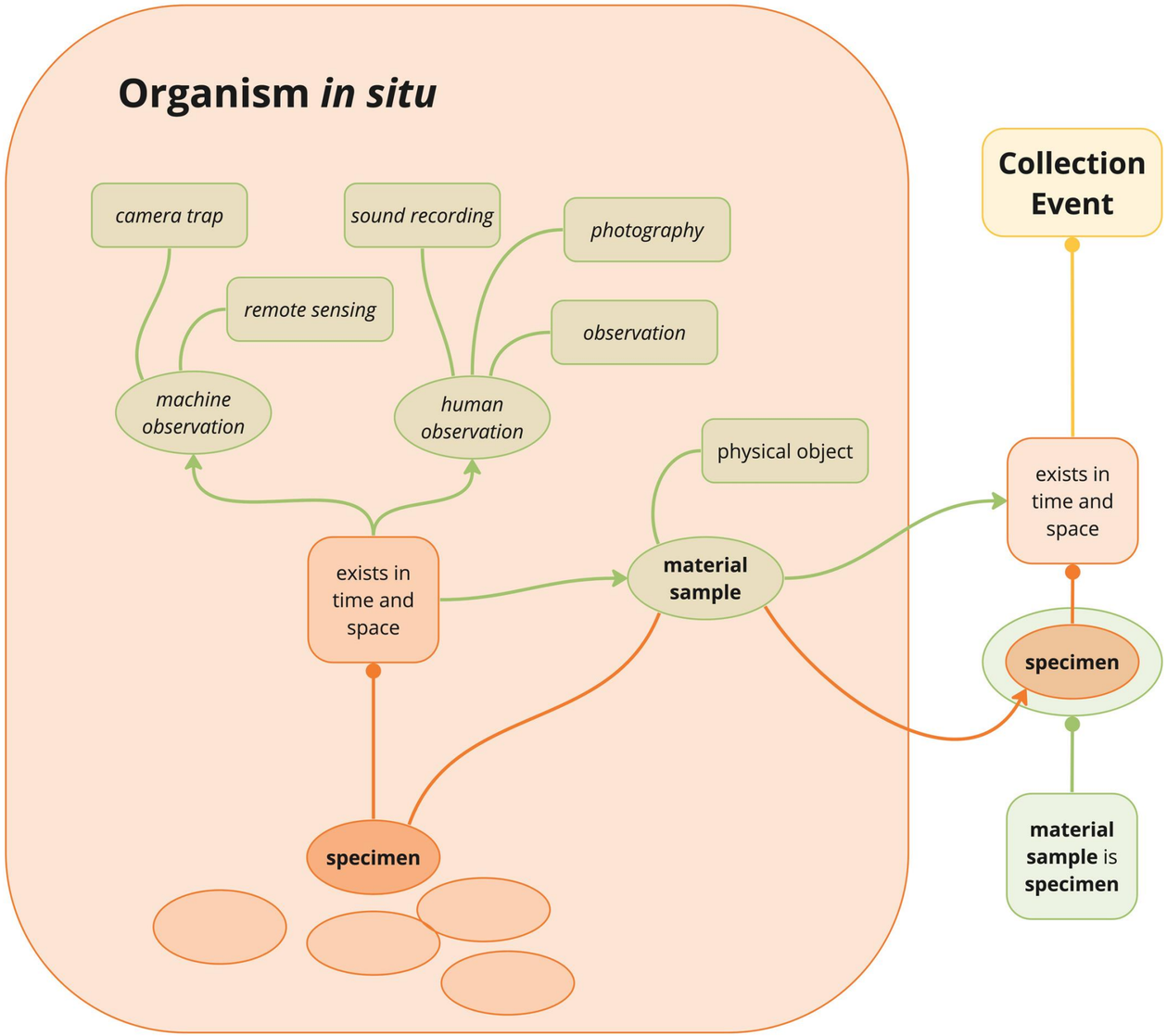
ARGA data model

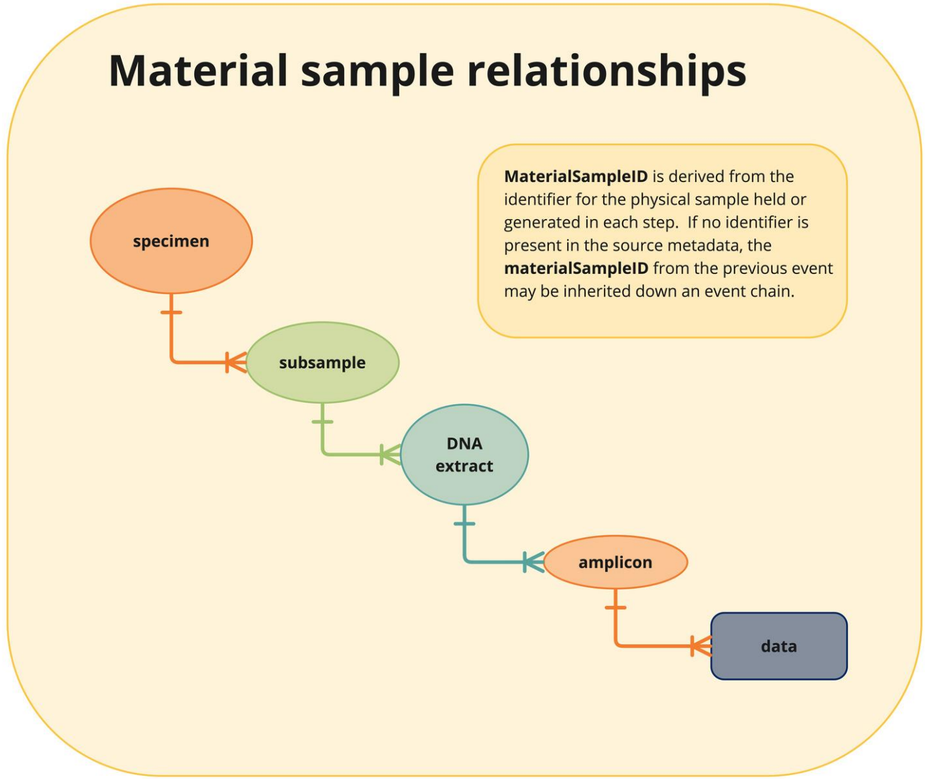
- respect hierarchical nature of data
- integrate multiple source genomic repository structures
- interoperate with biodiversity data
- standardise and unify data under a Darwin Core formatted schema

Understanding the data space



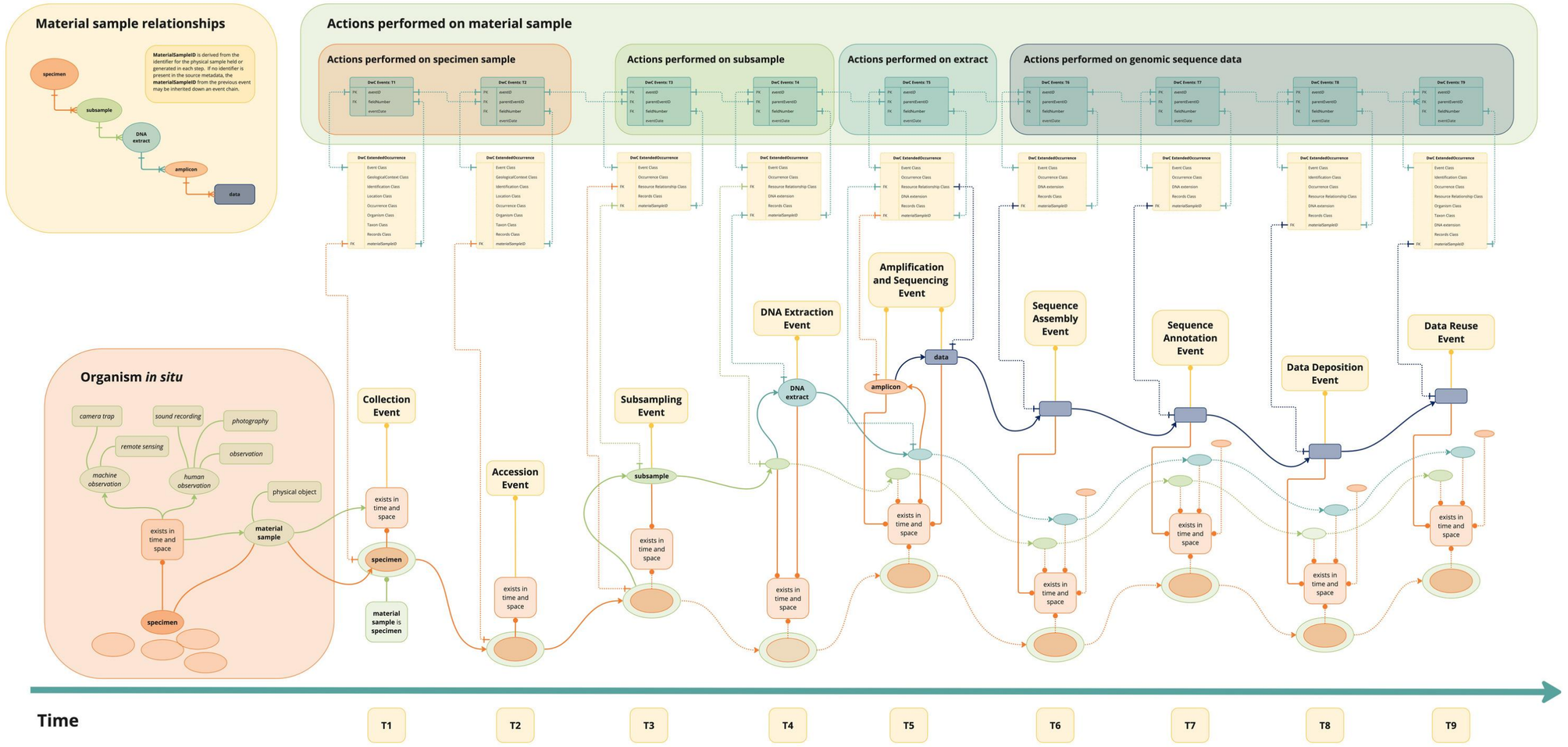
Material samples and specimens



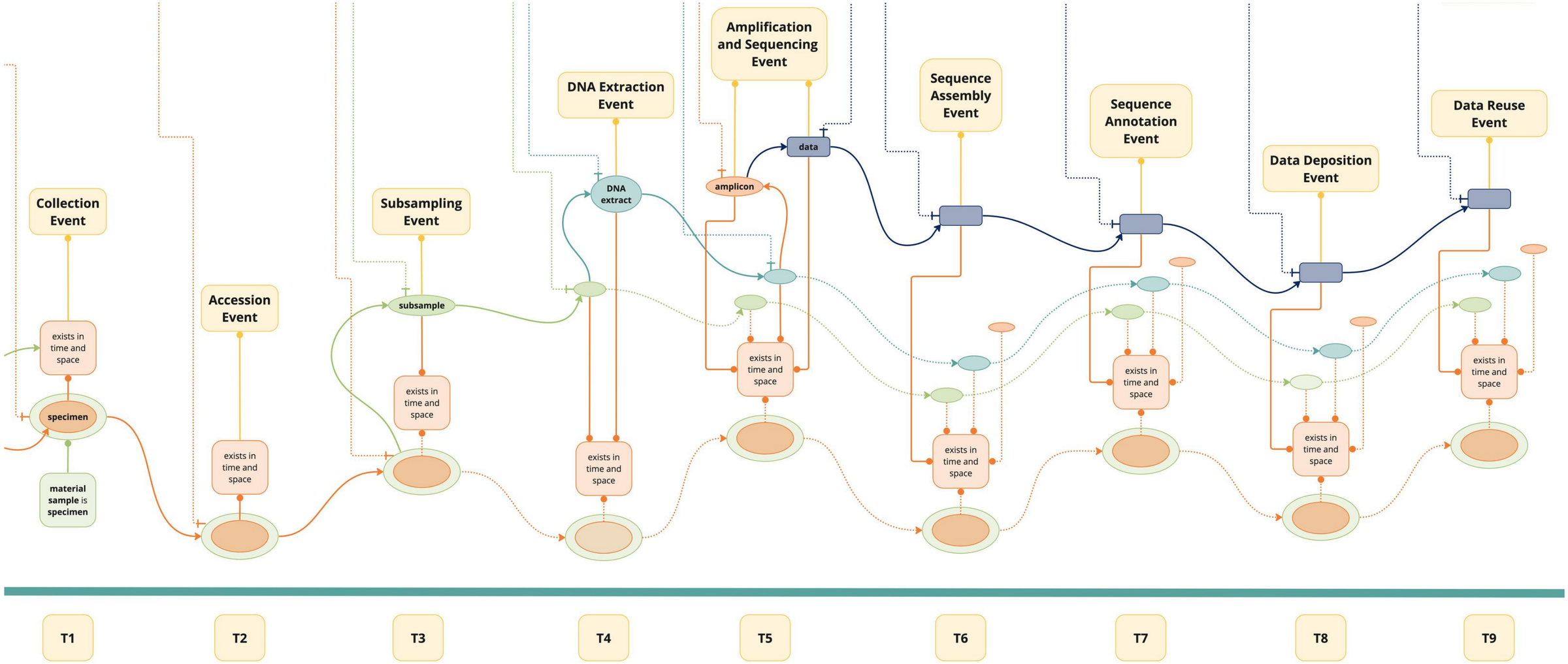


Modelling an organism from environment through to genomic data generation and deposition

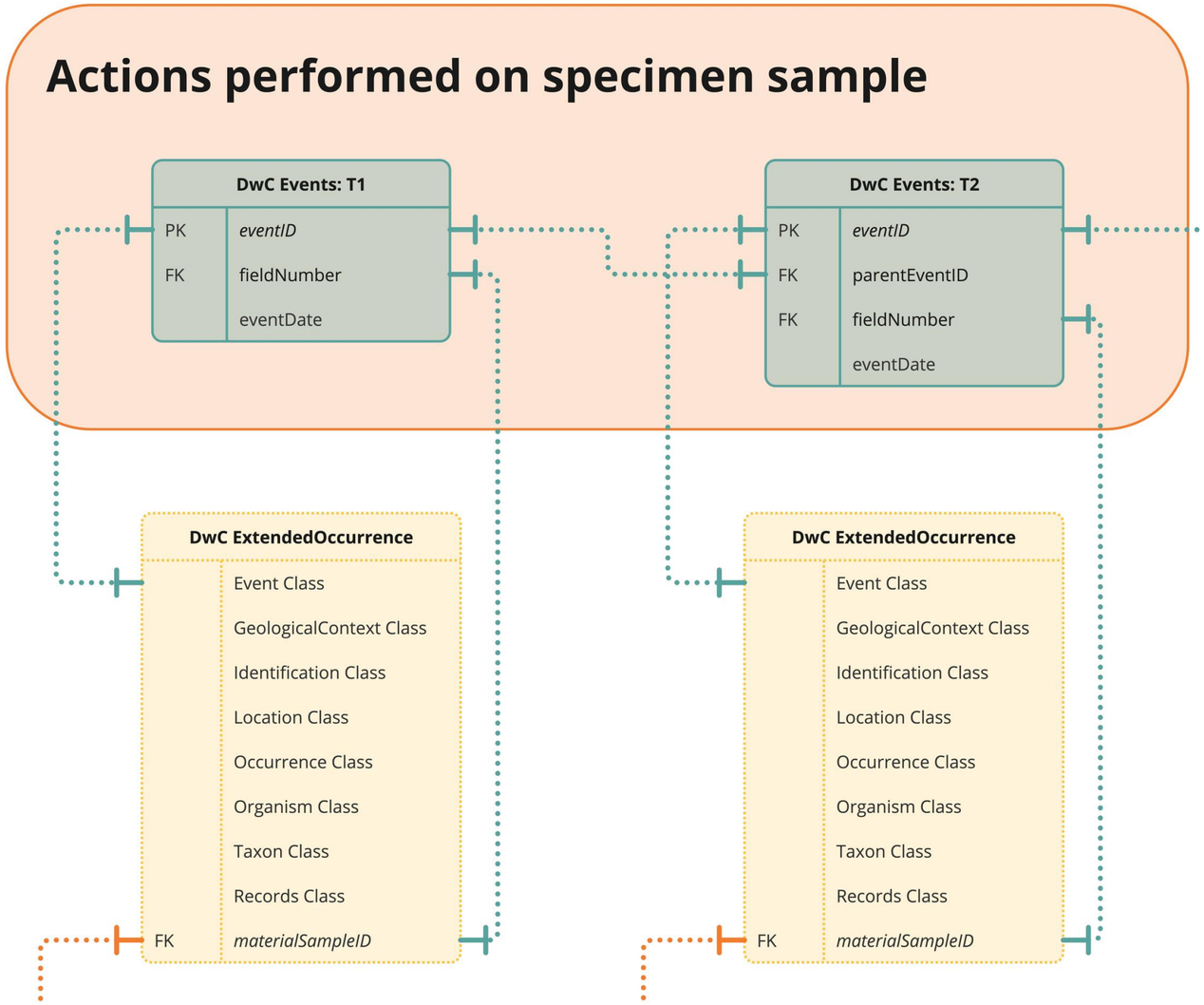
- **material samples** are hierarchical
- **data derived** from material samples are hierarchical
- **different actions** are performed on different types of material samples at various times
- **actions** are hierarchically related via **Events**



Genomics data events

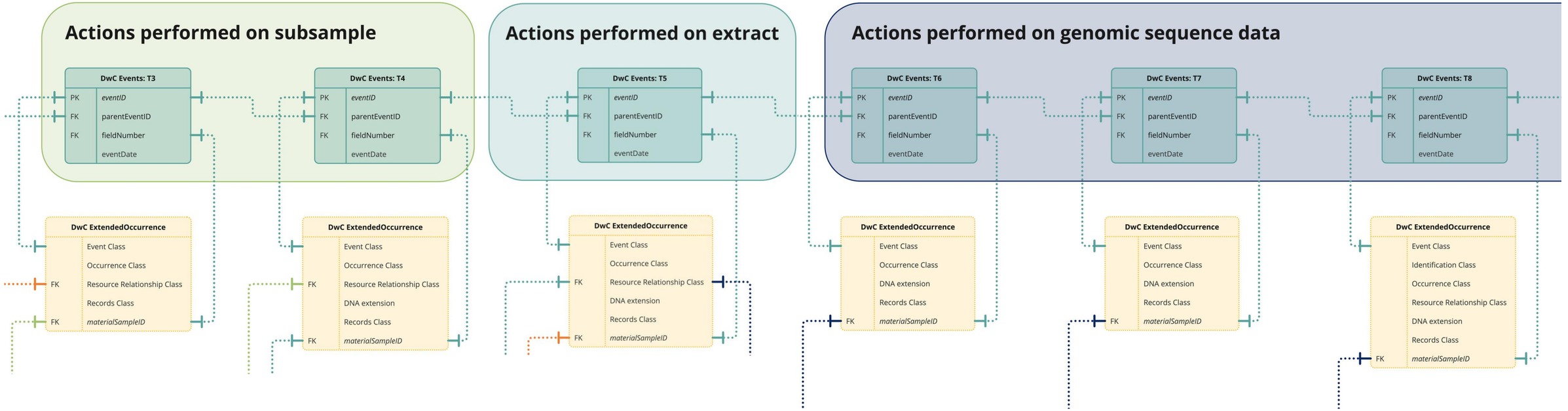


Modelling a single event



- events are created for each action
- each event connects to an **Extended Occurrence** (Darwin Core format)
- events nest to form provenance chains for each datum in the ARGA index

Modelling different event types



- the **Extended Occurrence** block can be tailored to capture only relevant metadata for each individual event

Key challenges when aligning data among events

Biodiversity collections data (DwC)

T1. Collection.

Data portals, e.g. NCBI (custom format, MiXS)

T4. DNA extraction.

T2. Accession.

T5. Amplification and sequencing.

T3. Subsampling.

T6. Sequence assembly.

T7. Sequence annotation.

T8. Data deposition.

Various sources, e.g. literature, non-genomic databases

T9. Data reuse.

Steel wool and spaghetti

- genomics data mapped to Darwin Core using **GBIF DNA derived data extension**

https://rs.gbif.org/extension/gbif/1.0/dna_derived_data_2022-02-23.xml

- unique mappings for each genomics data repository
- data preprocessed to field maps prior to ingestion to ARGA to create DwC-A
- unmapped fields retained as verbatim fields

For aggregation via taxonomy:

- canonical name matching to backbone taxonomy (DwC)

For aggregation via specimen:

- specimen numbers harmonised to Occurrences from ALA (DwC)

ARGA app released for UI testing

ARGA is solving a complex problem

- BPA and NCBI are not the same
- biodiversity data are in Darwin Core format
- molecular data and biodiversity data are stored flat
  - in reality neither is
- animal taxonomy in Australia is not yet on the NSL

app.arga.org.au

ARGA

Australian Reference Genome Atlas

ARGA app launch

ARGA team is perfecting the UI

- new mapping features
- streamlining data pages

Launching on 3 November 2023

- https://www.biocommons.org.au/events/arga-launch

Key contacts

https://arga.org.au

https://app.arga.org.au

keeva.connolly@qcif.edu.au

kathryn.hall@csiro.au

ARGA

Australian Reference Genome Atlas