ARDC Bushfires Data Challenge: BDC11 Australian Reference Genome Atlas project

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

taxonomy

ecolog
y

phenotyp
e

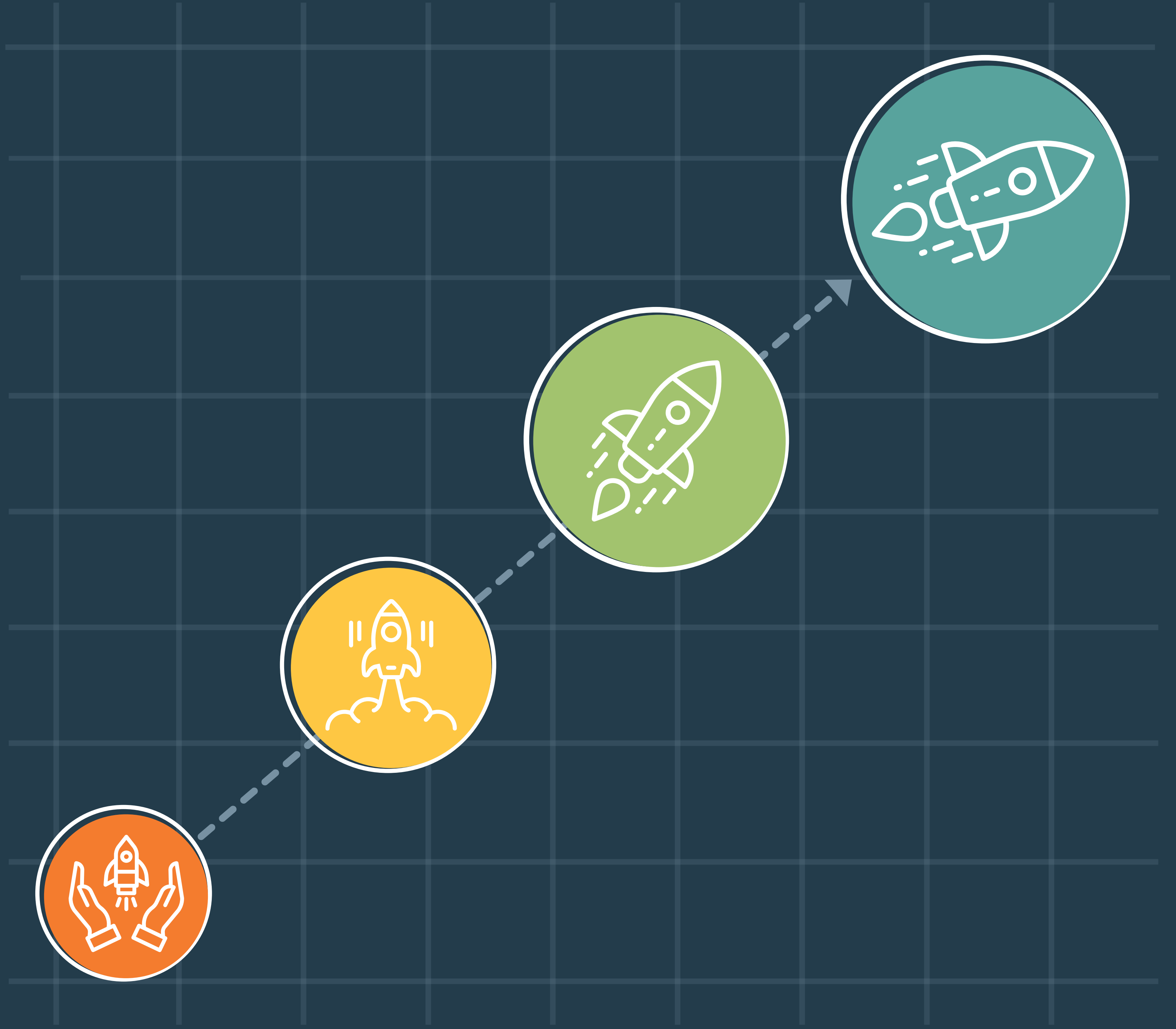
whole
genomes

DNA
sequences

variant
data

ARGA Project is building an indexing service for discovering, filtering and accessing complex life science data within biological contexts.

Project trajectory



ARGA app full release
December 2023

ARGA MVP early release
3 November 2023

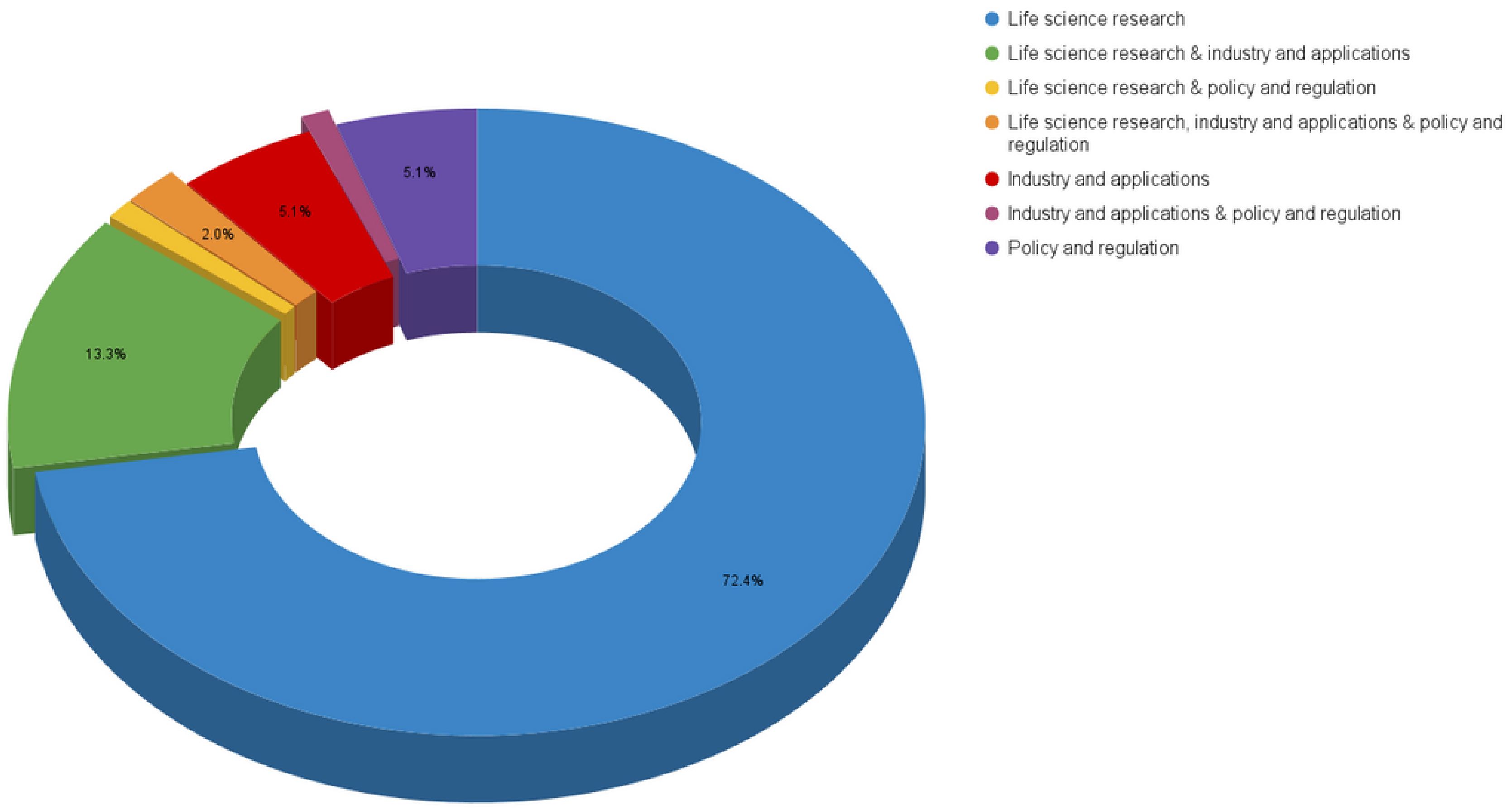
pre-launch testing
July 2023

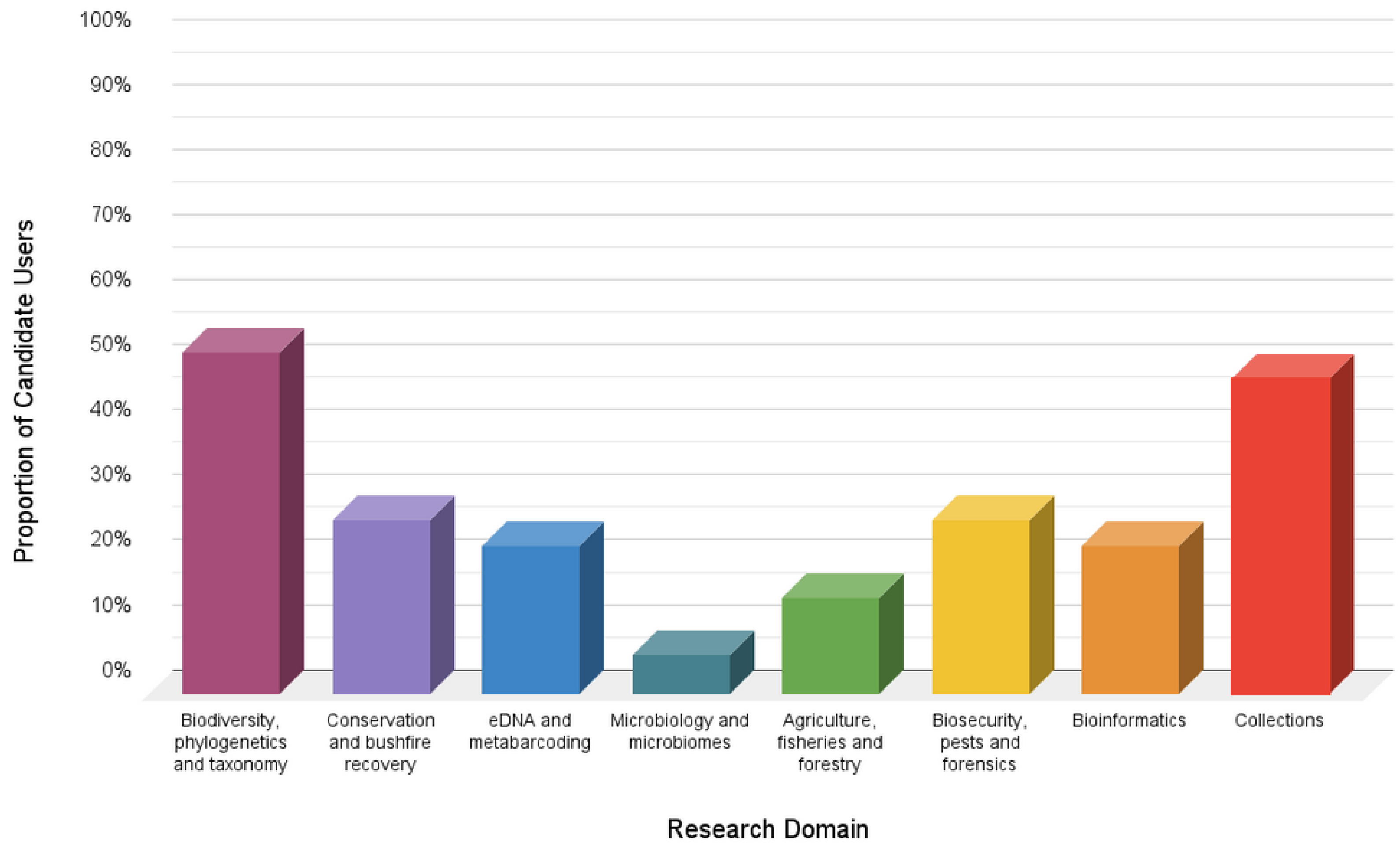
ARGA app design and build
May 2022 – June 2023

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

Community derived aims for the ARGA application

Users wanted to trust found data

- taxonomic certainty
- data quality
- metadata sufficiency

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

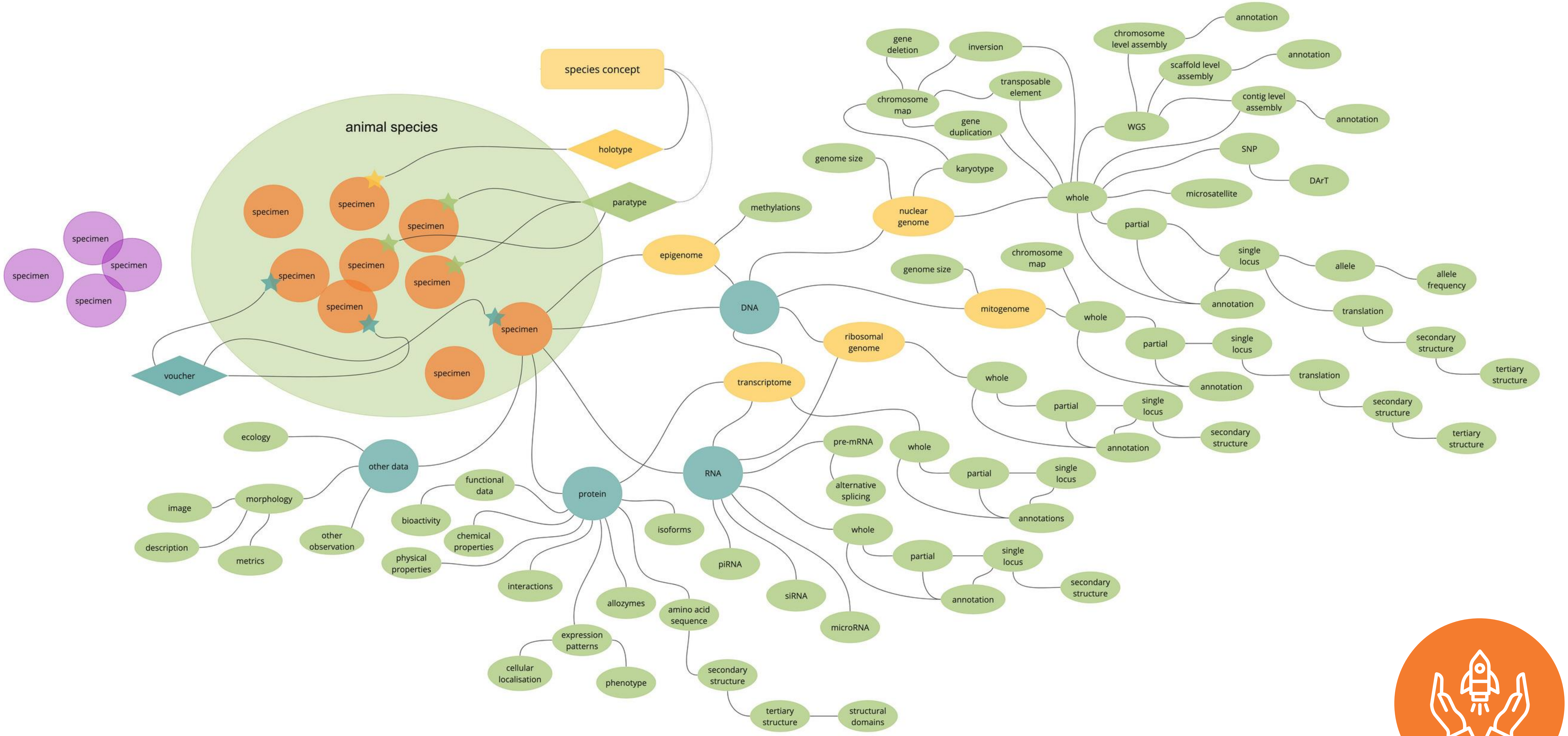
Data quality assessment

- visualisation tools for raw data

Metadata sufficiency

- new links with specimen records from ALA and specimen repositories

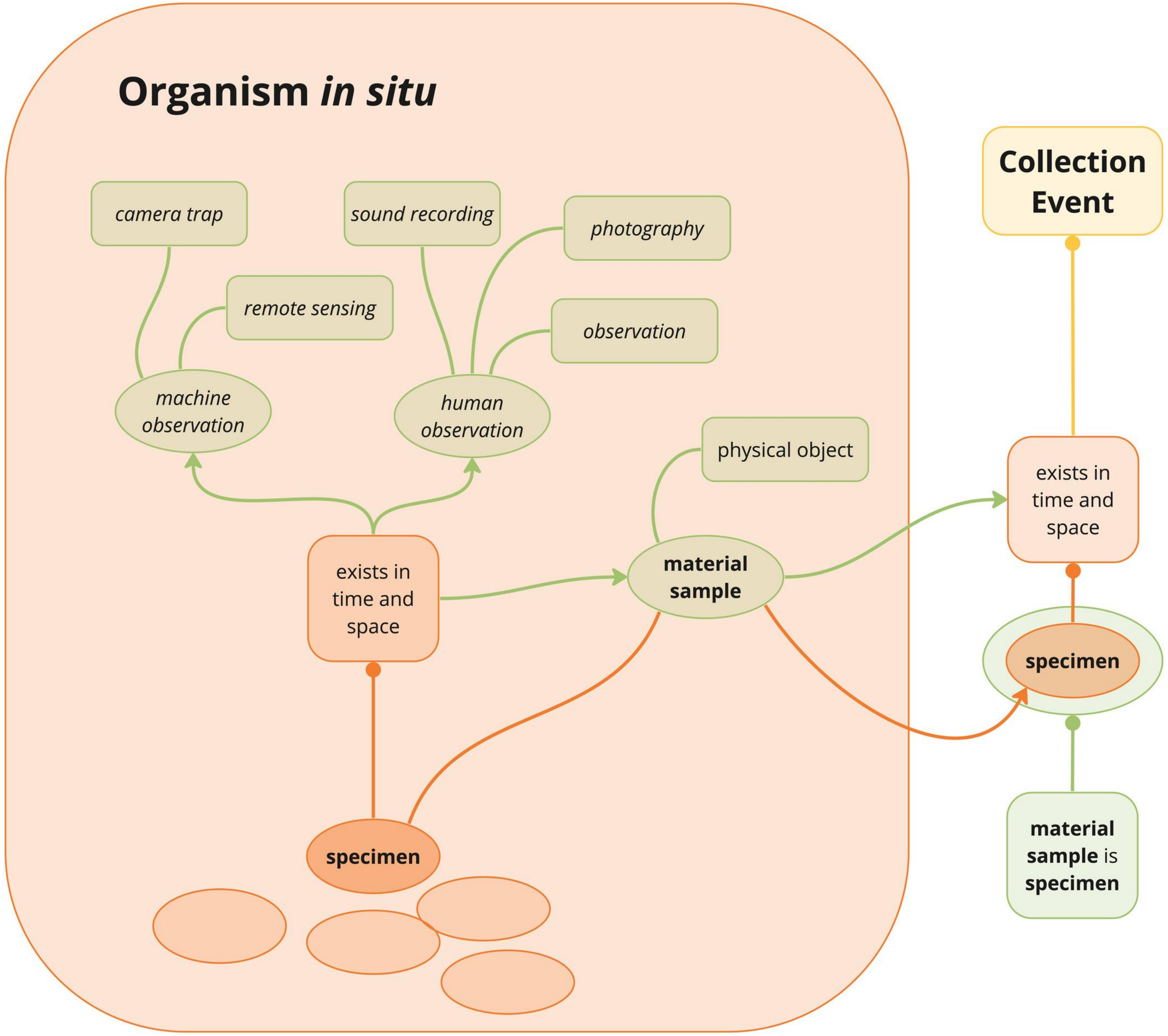
Understanding the data space



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

Material samples and specimens



Organism *in situ*

camera trap

sound recording

photography

remote sensing

observation

machine observation

human observation

physical object

exists in time and space

material sample

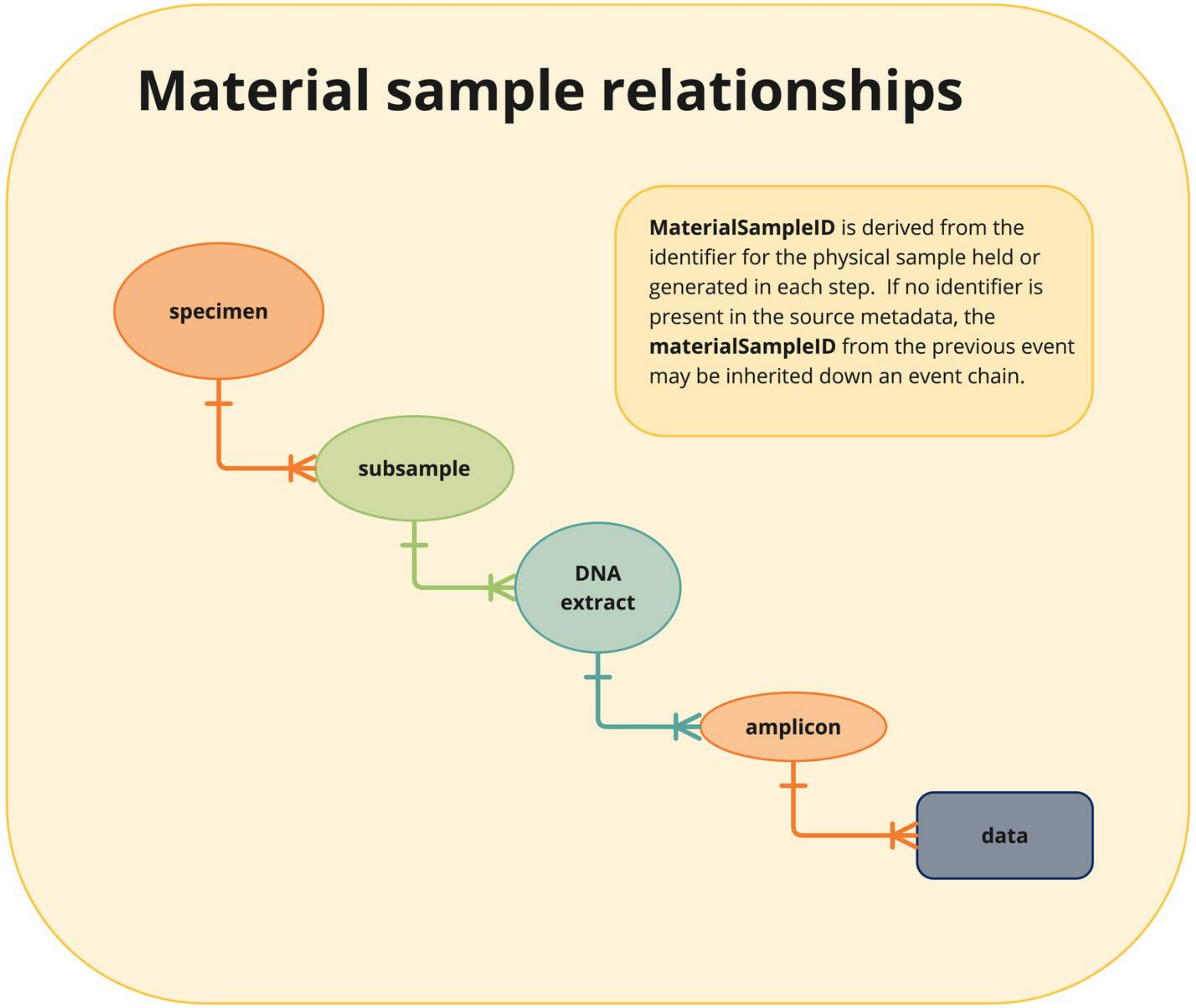
specimen

Collection Event

exists in time and space

specimen

material sample is **specimen**



Material sample relationships

MaterialSampleID is derived from the identifier for the physical sample held or generated in each step. If no identifier is present in the source metadata, the **materialSampleID** from the previous event may be inherited down an event chain.

specimen

subsample

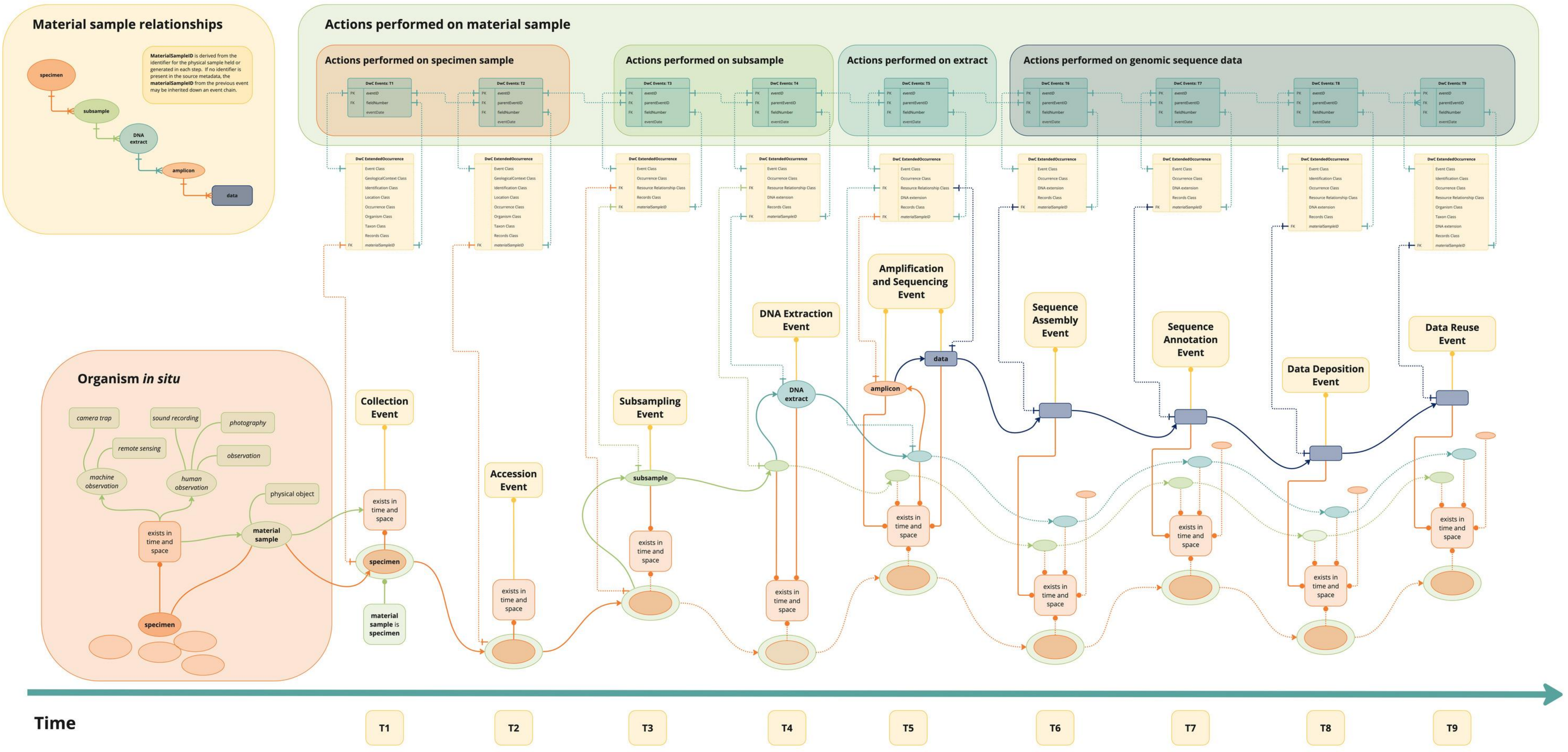
DNA extract

amplicon

data

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**



MaterialSampleID is derived from the identifier for the physical sample held or generated in each step. If no identifier is present in the source metadata, the **materialSampleID** from the previous event may be inherited down an event chain.

DwC Events: T1	
PK	*eventID*
FK	fieldNumber
	eventDate

DwC Events: T2	
PK	*eventID*
FK	parentEventID
FK	fieldNumber
	eventDate

DwC Events: T3	
PK	*eventID*
FK	parentEventID
FK	fieldNumber
	eventDate

DwC Events: T4	
PK	*eventID*
FK	parentEventID
FK	fieldNumber
	eventDate

DwC Events: T5	
PK	*eventID*
FK	parentEventID
FK	fieldNumber
	eventDate

DwC Events: T6	
PK	*eventID*
FK	parentEventID
FK	fieldNumber
	eventDate

DwC Events: T7	
PK	*eventID*
FK	parentEventID
FK	fieldNumber
	eventDate

DwC Events: T8	
PK	*eventID*
FK	parentEventID
FK	fieldNumber
	eventDate

DwC Events: T9	
PK	*eventID*
FK	parentEventID
FK	fieldNumber
	eventDate

DwC ExtendedOccurrence	
	Event Class
	GeologicalContext Class
	Identification Class
	Location Class
	Occurrence Class
	Organism Class
	Taxon Class
	Records Class
FK	*materialSampleID*

DwC ExtendedOccurrence	
	Event Class
	GeologicalContext Class
	Identification Class
	Location Class
	Occurrence Class
	Organism Class
	Taxon Class
	Records Class
FK	*materialSampleID*

DwC ExtendedOccurrence	
	Event Class
	Occurrence Class
FK	Resource Relationship Class
	Records Class
FK	*materialSampleID*

DwC ExtendedOccurrence	
	Event Class
	Occurrence Class
FK	Resource Relationship Class
	DNA extension
	Records Class
FK	*materialSampleID*

DwC ExtendedOccurrence	
	Event Class
	Occurrence Class
FK	Resource Relationship Class
	DNA extension
	Records Class
FK	*materialSampleID*

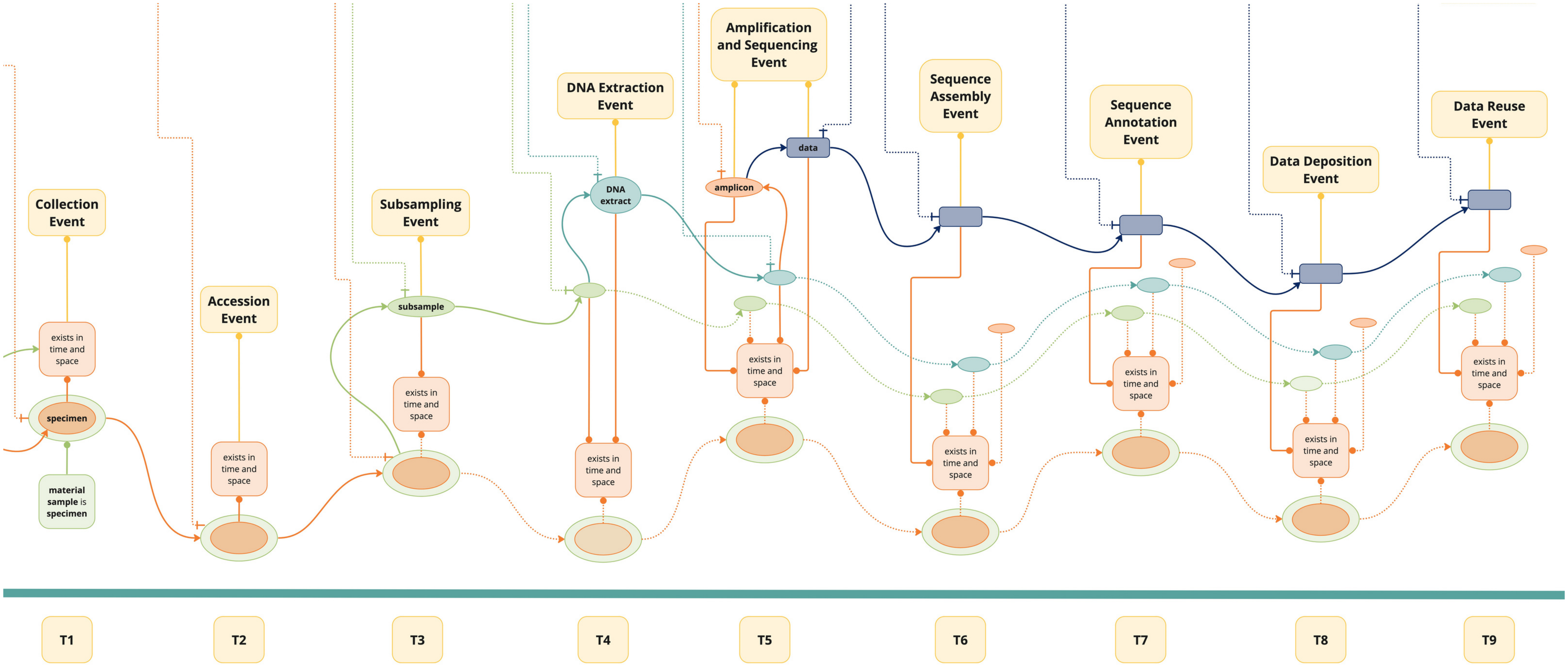
DwC ExtendedOccurrence	
	Event Class
	Occurrence Class
	DNA extension
	Records Class
FK	*materialSampleID*

DwC ExtendedOccurrence	
	Event Class
	Occurrence Class
	DNA extension
	Records Class
FK	*materialSampleID*

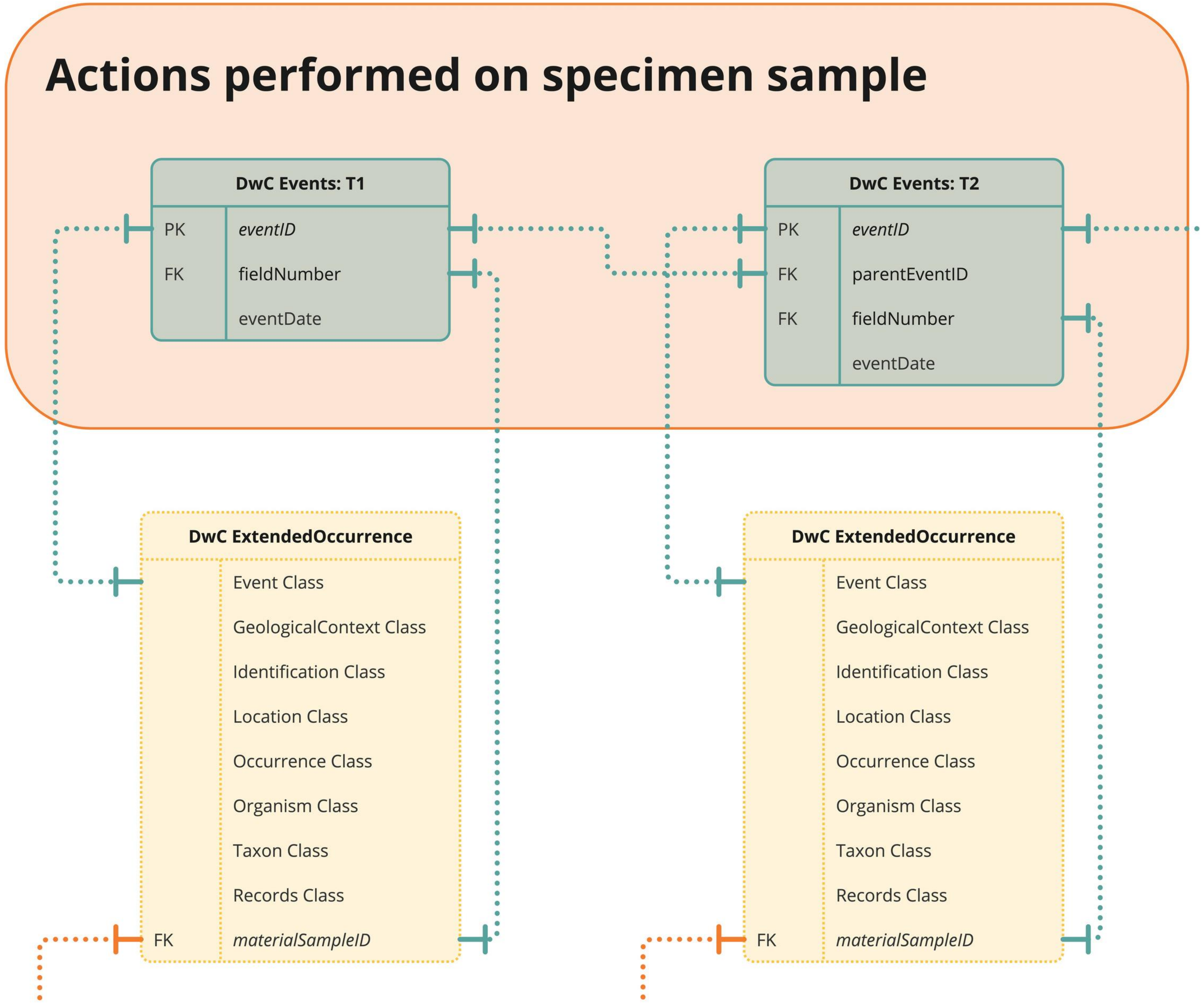
DwC ExtendedOccurrence	
	Event Class
	Identification Class
	Occurrence Class
	Resource Relationship Class
	DNA extension
	Records Class
FK	*materialSampleID*

DwC ExtendedOccurrence	
	Event Class
	Identification Class
	Occurrence Class
	Resource Relationship Class
	Organism Class
	Taxon Class
	DNA extension
	Records Class
FK	*materialSampleID*

Genomics data events

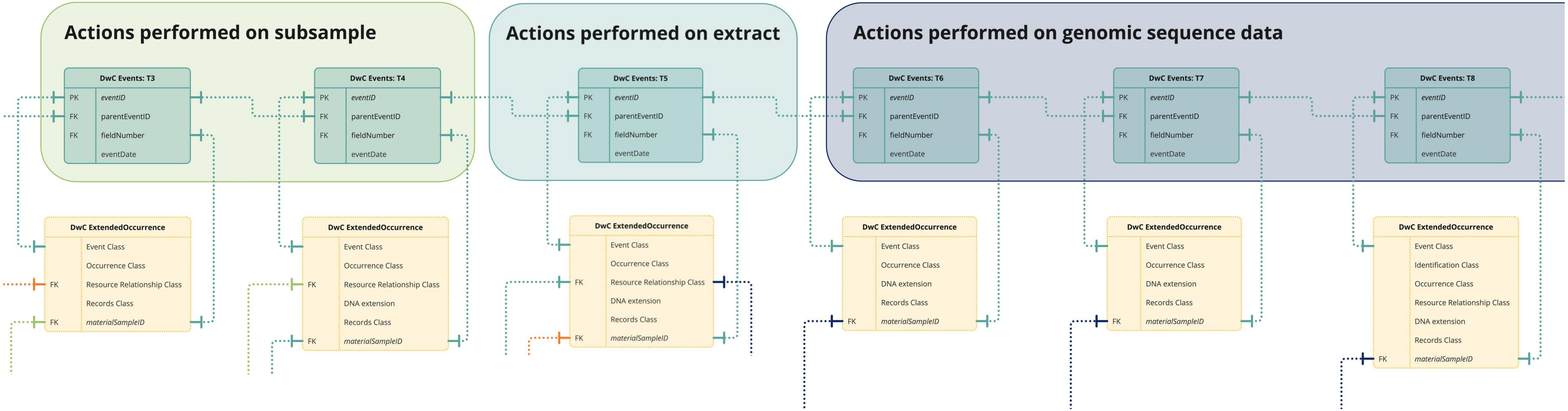


Modelling a single event



- 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



Actions performed on subsample

DwC Events: T3
PK eventID
FK parentEventID
FK fieldNumber
eventDate

DwC Events: T4
PK eventID
FK parentEventID
FK fieldNumber
eventDate

Actions performed on extract

DwC Events: T5
PK eventID
FK parentEventID
FK fieldNumber
eventDate

Actions performed on genomic sequence data

DwC Events: T6
PK eventID
FK parentEventID
FK fieldNumber
eventDate

DwC Events: T7
PK eventID
FK parentEventID
FK fieldNumber
eventDate

DwC Events: T8
PK eventID
FK parentEventID
FK fieldNumber
eventDate

DwC ExtendedOccurrence
Event Class
Occurrence Class
FK Resource Relationship Class
Records Class
FK materialSampleID

DwC ExtendedOccurrence
Event Class
Occurrence Class
FK Resource Relationship Class
DNA extension
Records Class
FK materialSampleID

DwC ExtendedOccurrence
Event Class
Occurrence Class
FK Resource Relationship Class
DNA extension
Records Class
FK materialSampleID

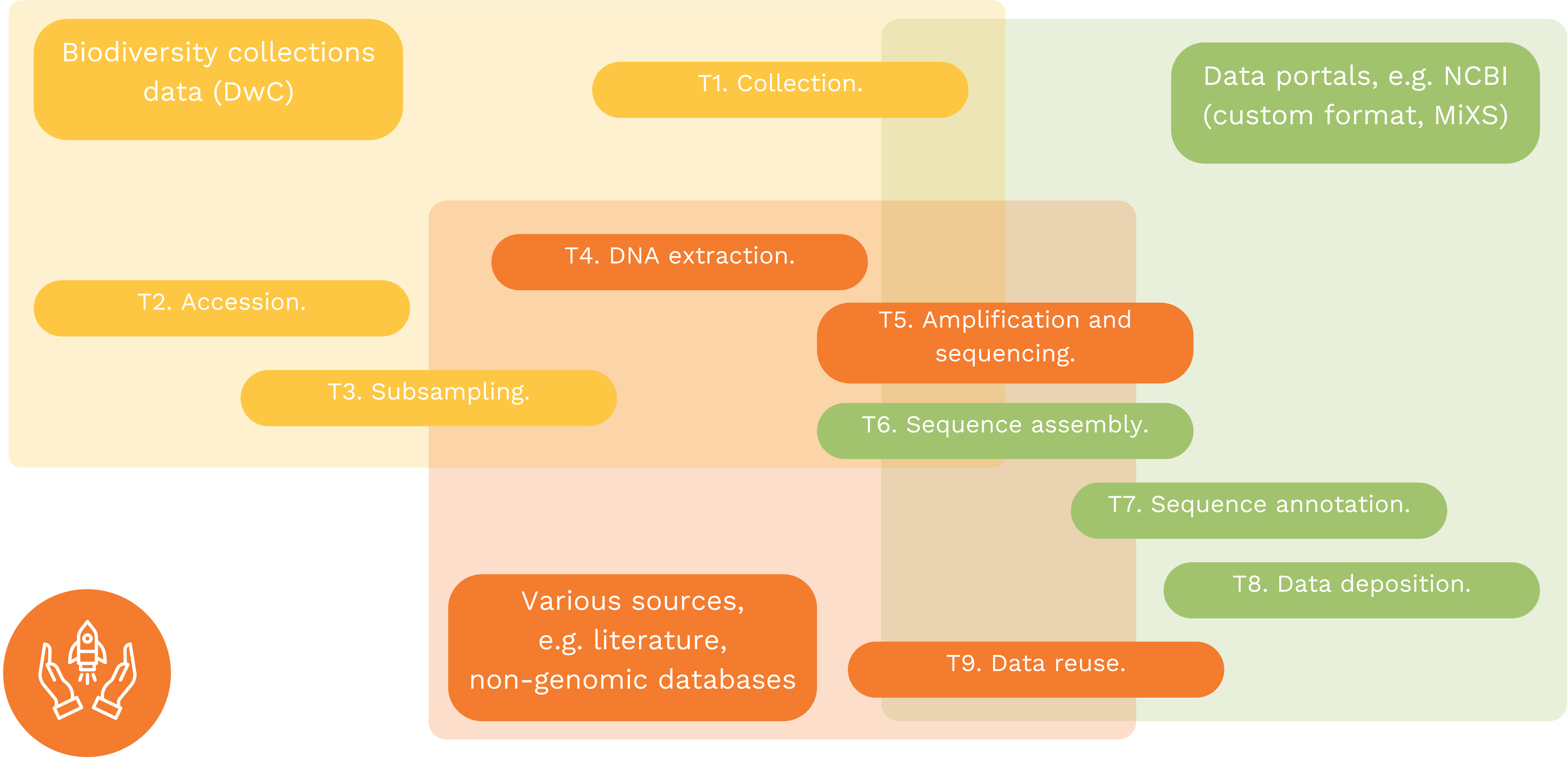
DwC ExtendedOccurrence
Event Class
Occurrence Class
DNA extension
Records Class
FK materialSampleID

DwC ExtendedOccurrence
Event Class
Occurrence Class
DNA extension
Records Class
FK materialSampleID

DwC ExtendedOccurrence
Event Class
Identification Class
Occurrence Class
Resource Relationship Class
DNA extension
Records Class
FK materialSampleID

- 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) sourced from NSL and AFD, with enrichment via GBIF and WoRMs

For aggregation via specimen:

- specimen numbers harmonised to Occurrences from ALA (DwC)

ARGA app UI testing

ARGA is solving a complex problem

- 74 potential databases identified and assessed
  - each has own unique structure
  - ingestion mapping devised for 16
- BPA, NCBI and BOLD now fully interoperable within ARGA
- biodiversity data are in Darwin Core format
  - genomic data normalised to modified DwC for interoperability
- MVP delivered

Bushfire susceptibility filters

ARGA MVP contains user-selectable filters

- 124 animal species:
  - 17 birds
  - 20 mammals
  - 23 reptiles
  - 16 amphibians
  - 21 fish
  - 22 crayfish
  - 5 other invertebrates
- 486 plants
- 191 high priority invertebrates
- 147 low-risk invertebrates

Sources:

- Rapid analysis of impacts of the 2019-20 fires on animal species, and prioritisation of species for management response [https://www.dcceew.gov.au/sites/default/files/env/pages/ef3f5ebd-faec-4c0c-9ea9-b7dfd9446cb1/files/assessments-species-vulnerability-fire-impacts-14032020.pdf].
- Bushfire impacts - Final priority list of plants [https://www.dcceew.gov.au/environment/biodiversity/bushfire-recovery/bushfire-impacts/priority-plants].
- Provisional list of invertebrates [https://www.dcceew.gov.au/sites/default/files/env/pages/9a6a5628-21b9-4a1b-9474-e953fc5138db/files/provisional-list-invertebrates-requiring-urgent-management-intervention.pdf].

App demonstration

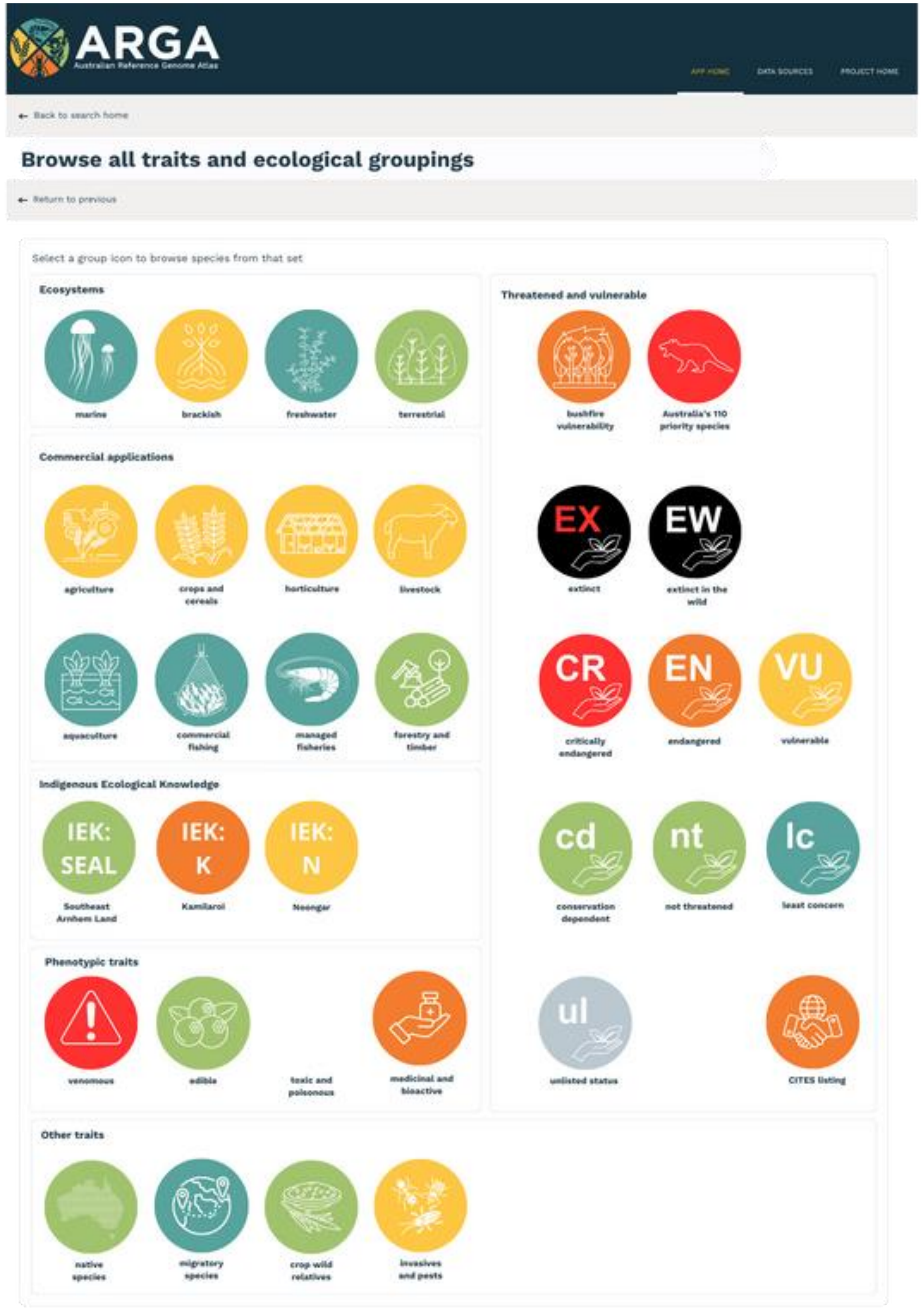
ARGA
Australian Reference Genome Atlas

Future directions

ARGA feature enhancements and integrations

- additional data source ingestion
- seamless integration with analysis pipelines
- traits and phenotype filter enhancements
- data visualisations and permutations
    - institutions/collections and researcher-based searching
    - geographical searching
    - user-customised reporting and statistics

Coming implementations



Browse all traits and ecological groupings

Select a group icon to browse species from that set

Ecosystems

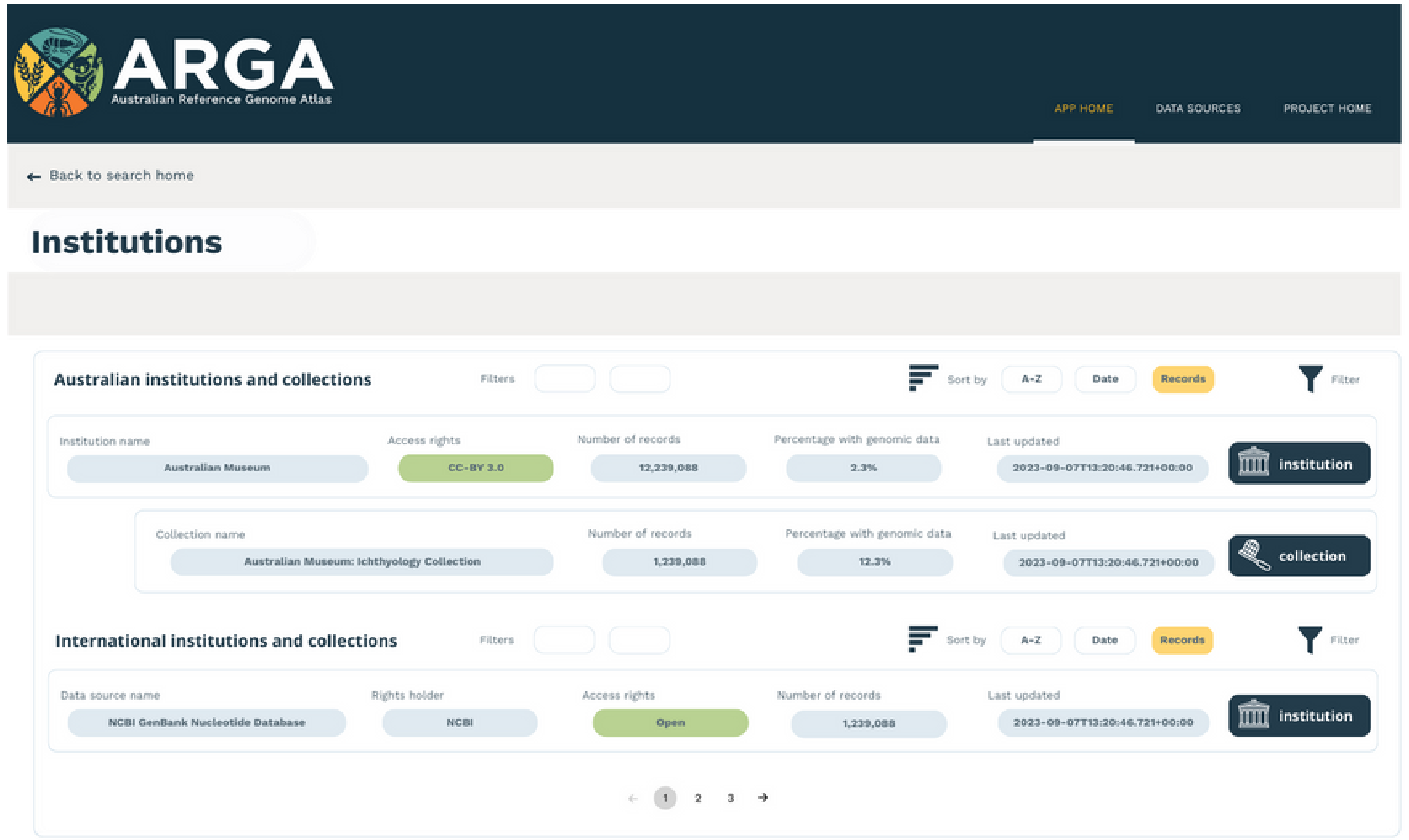
Threatened and vulnerable

Commercial applications

Indigenous Ecological Knowledge

Phenotypic traits

Other traits



Institutions

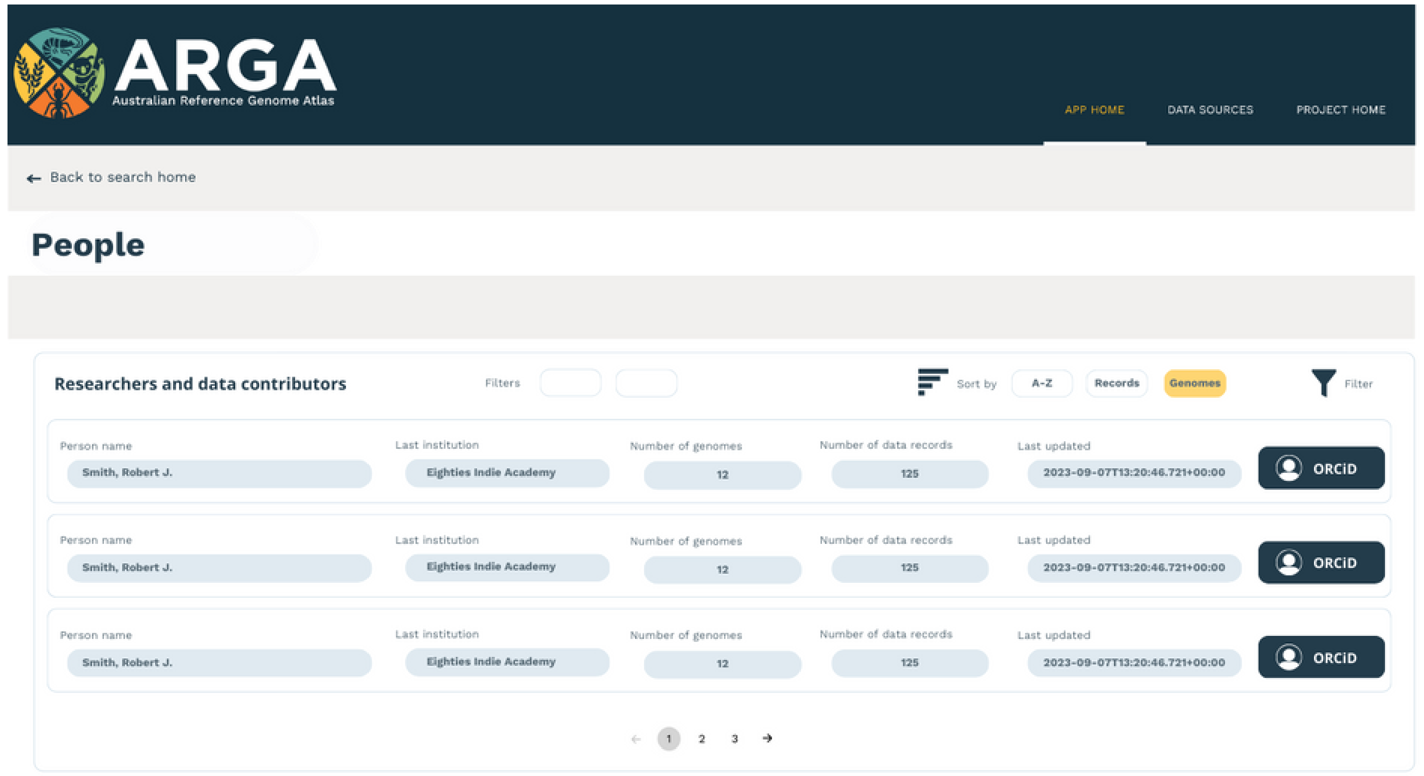
Australian institutions and collections

Institution name	Access rights	Number of records	Percentage with genomic data	Last updated	
Australian Museum	CC-BY 3.0	12,239,088	2.3%	2023-09-07T13:20:46.721+00:00	institution

Collection name	Number of records	Percentage with genomic data	Last updated	
Australian Museum: Ichthyology Collection	1,239,088	12.3%	2023-09-07T13:20:46.721+00:00	collection

International institutions and collections

Data source name	Rights holder	Access rights	Number of records	Last updated	
NCBI GenBank Nucleotide Database	NCBI	Open	1,239,088	2023-09-07T13:20:46.721+00:00	institution



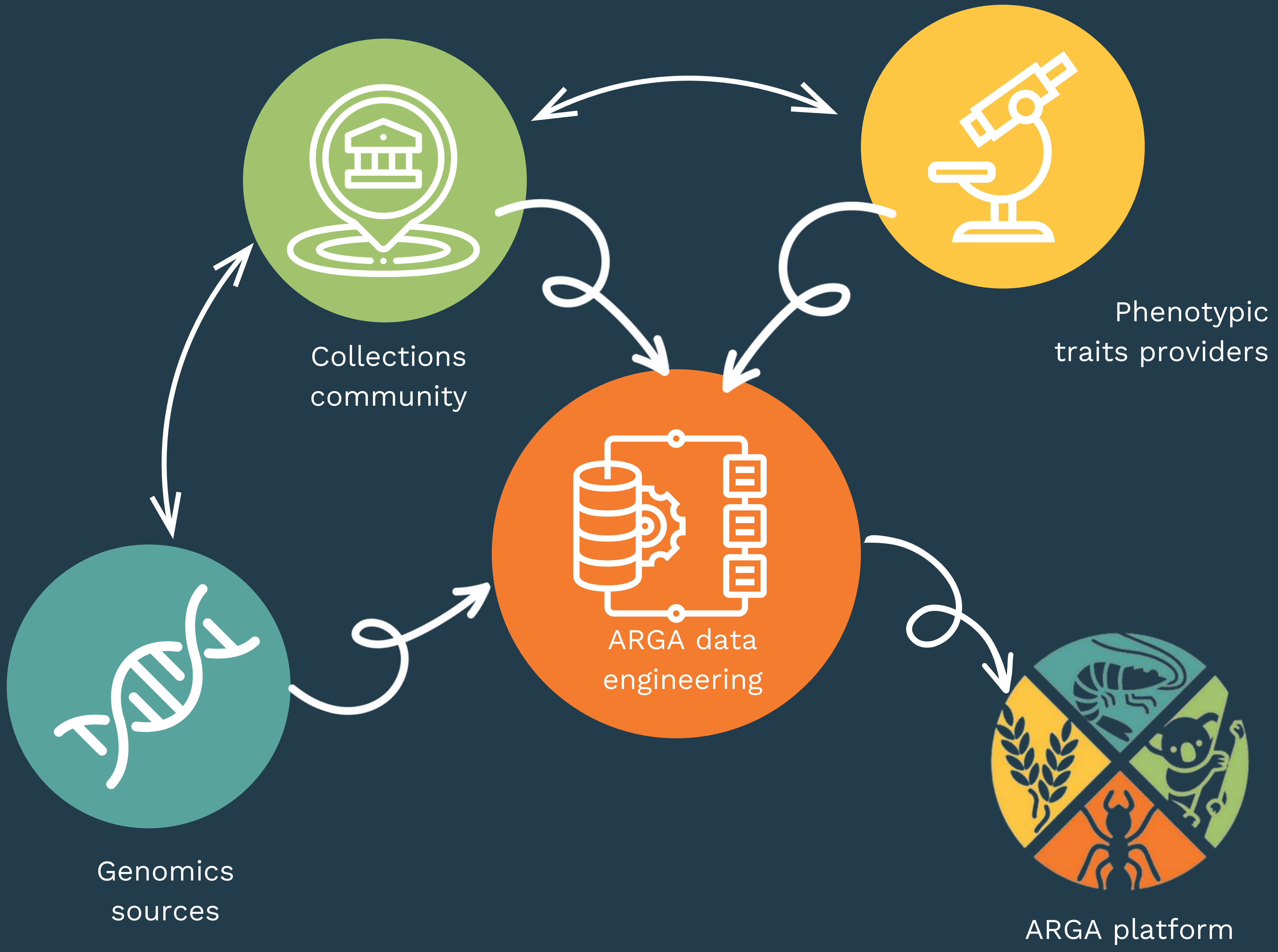
People

Researchers and data contributors

Person name	Last institution	Number of genomes	Number of data records	Last updated	
Smith, Robert J.	Eighties Indie Academy	12	125	2023-09-07T13:20:46.721+00:00	ORCiD
Smith, Robert J.	Eighties Indie Academy	12	125	2023-09-07T13:20:46.721+00:00	ORCiD
Smith, Robert J.	Eighties Indie Academy	12	125	2023-09-07T13:20:46.721+00:00	ORCiD

Community engagement

- data custodians from biosciences research domains
  - vocabulary focus
  - repository alignment
  - templates for metadata capture at point of data generation


Collections community

Phenotypic traits providers

Genomics sources

ARGA data engineering

ARGA platform

Key contacts

https://arga.org.au
https://arga.org.au/contact/

contact@arga.org.au

Keeva Connolly: Scientific Business Analyst

Kathryn Hall: Project Manager