Chasing rabbits

How ARGA greatly expands provenance using CmRDTs and a Hybrid Logical Clock

Goran Sterjov
eResearch 2024, 31 Oct 2024

ARGA
Australian Reference Genome Atlas

What is ARGA?

The Australian Reference Genome Atlas aggregates:

- Genomic data
- Collections data
- Taxonomic data
- Traits data

What is ARGA?


Browse by data type
Genome assemblies
2.06M records
Single loci
4.01M records
Specimens
34.76M records
Browse by taxonomic group
Animals
126.16k records
Plants
32.41k records
Fungi
13.24k records
Protista
1.79k records
All species
174.78k records
Browse by functional or ecological group
Bushfire Recovery
738 records
Commercial
1.25k records
Venomous and Poisonous

What’s the problem?

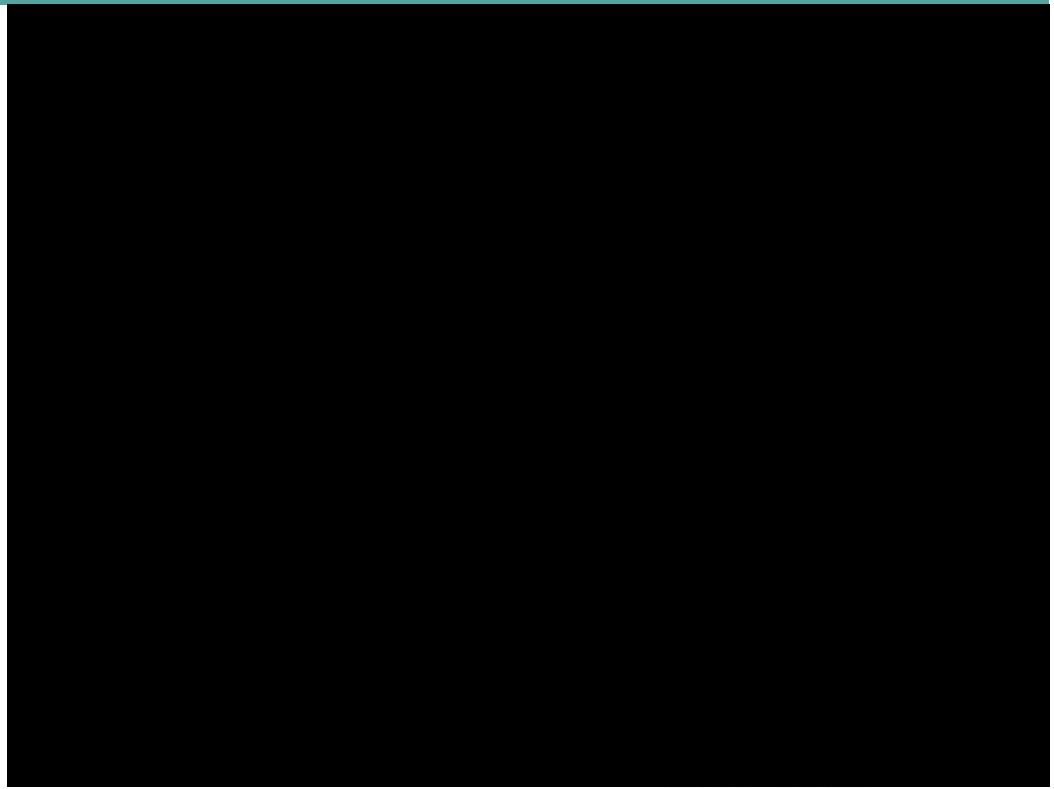
- Data quality varies greatly
  - From 10 columns to 900 columns
- Lack of vocabularies, lots of free text
  - Especially with genomics
- Shallow taxonomic systems
  - No history, inconsistent publications format, often lacks type specimens
- Data derived from other aggregated data
  - Eg. BOLD data created from NCBI data, both datasets we aggregate

History is key

- Data quality will improve over time from datasets
- Curation can fill in the gaps and fix data
- Knowing where the data is coming from when multiple datasets contribute to the same datum

CRDTs - Conflict-free replicated data types

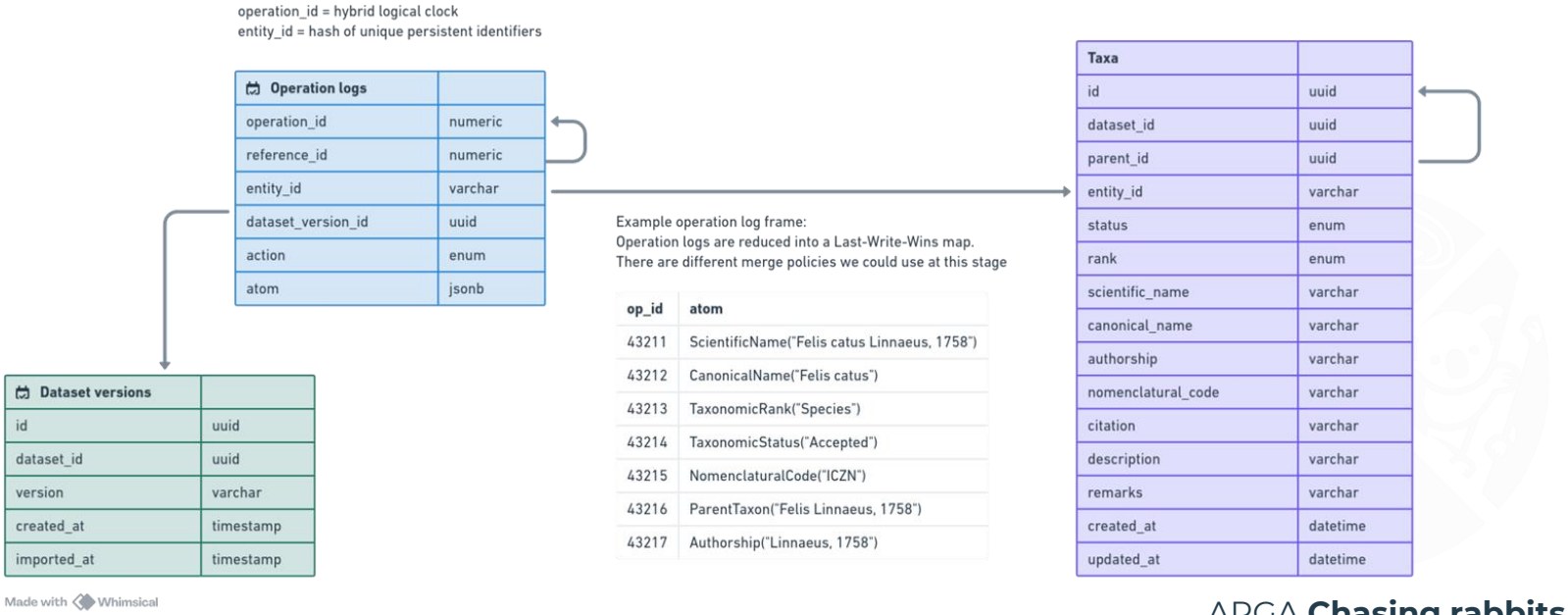
- Merge data edited concurrently
- Can retain diffs
- Eventual consistency



Video from Loro
https://loro.dev/docs/advanced/event_graph_walker

CmRDTs - Operation based

- Changes are recorded as operations on a record
- Verbose but very flexible



Record History - *Dendrolagus lumholtzi Collett, 1884*

Create AUSTRALIAN FAUNAL DIRECTORY
01/01/1970, 6:40:23 pm

Update AUSTRALIAN FAUNAL DIRECTORY
TAXON_ID → b99a5735-4e26-4e12-ac9b-d435b3a5f85e
01/01/1970, 6:40:23 pm

Update AUSTRALIAN FAUNAL DIRECTORY
SCIENTIFIC_NAME → Dendrolagus lumholtzi Collett, 1884
01/01/1970, 6:40:23 pm

Update AUSTRALIAN FAUNAL DIRECTORY
CANONICAL_NAME → Dendrolagus lumholtzi
01/01/1970, 6:40:23 pm

Hybrid logical clock

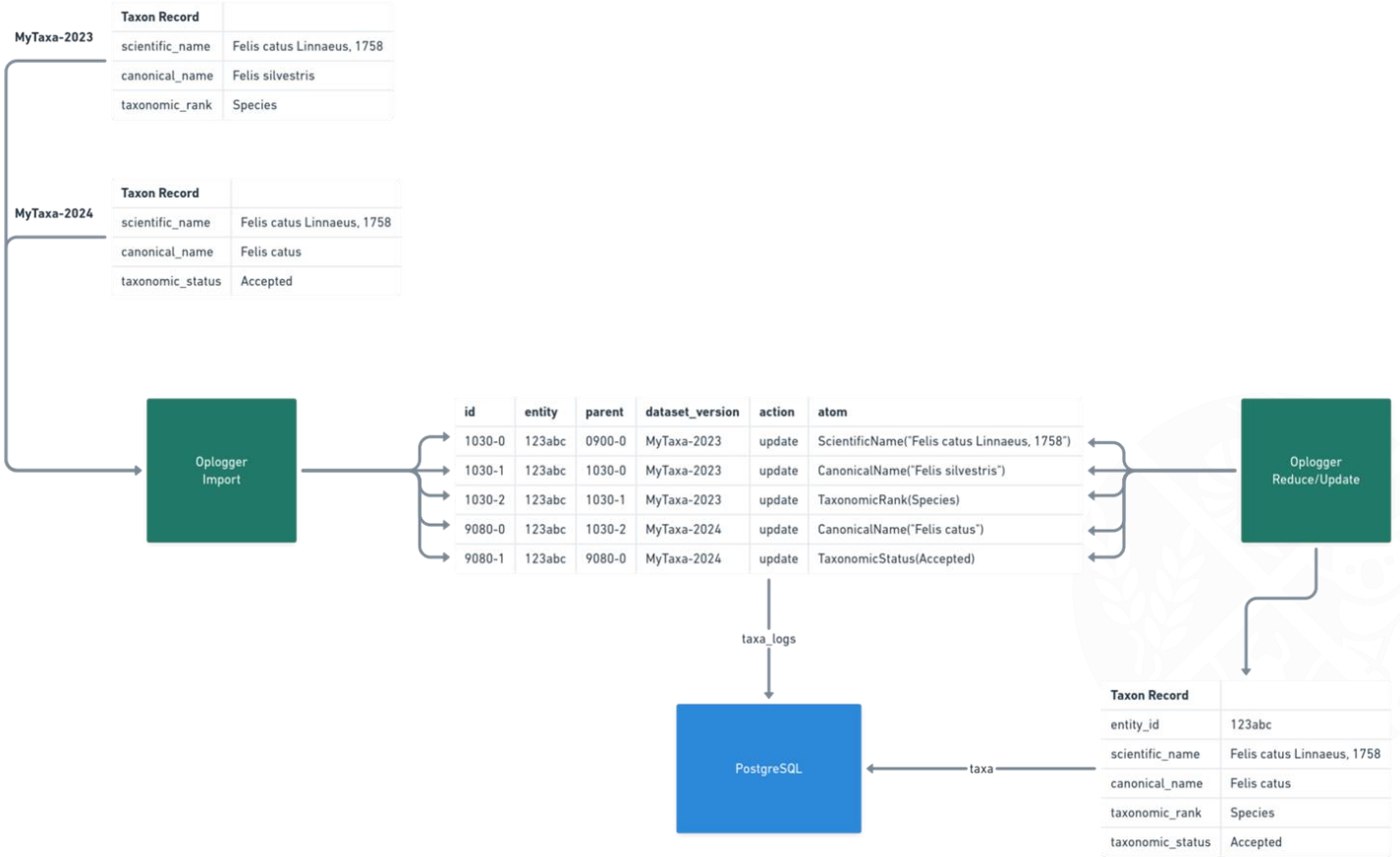
A custom timestamp combined with a logical clock and shoved into a 64bit unsigned integer.

It provides *causal* ordering

Used when time travelling

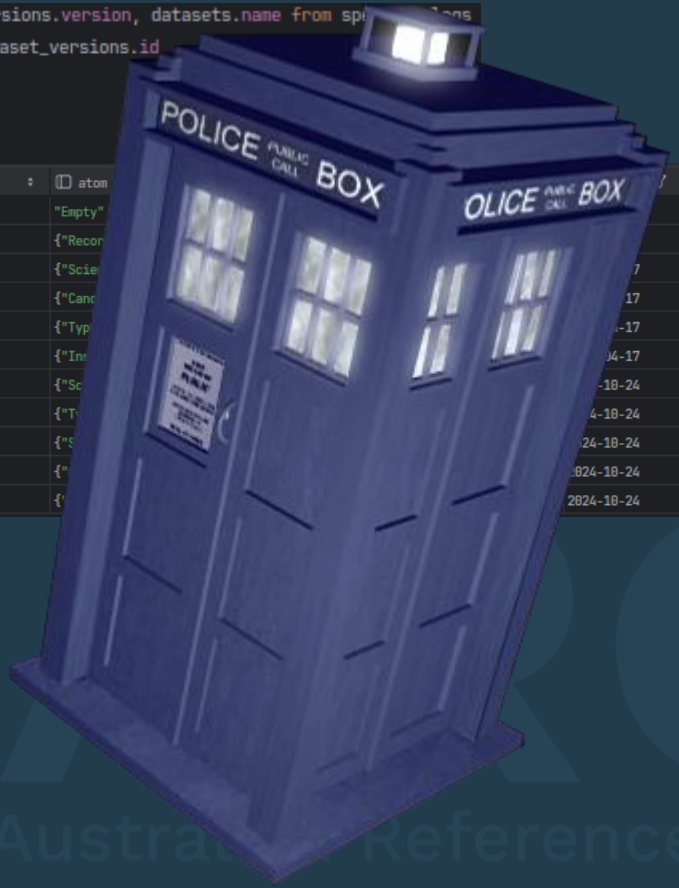
```
// MMDHmSssnn
// months = 16bit = 65536 (5461 years)
// days = 6bit = 64
// hours = 6bit = 64
// minutes = 6bit = 64
// seconds = 6bit = 64
// milliseconds = 10bit = 1024
// logical = 14bit = 16384
```

History and time travelling



```
select operation_id, parent_id, action, atom, dataset_versions.version, datasets.name from sp
join dataset_versions 1..n<->1: on dataset_version_id=dataset_versions.id
join datasets 1..n<->1: on dataset_id=datasets.id
where entity_id='13114533894762173098'
order by operation_id;
```

	operation_id	parent_id	action	atom	name
1	3791746164293632	3791746164129792	create	"Empty"	OZCAM
2	3791746164293633	3791746164293632	update	{"Recor	OZCAM
3	3791746164293634	3791746164293633	update	{"Scie	OZCAM
4	3791746164293635	3791746164293634	update	{"Canc	OZCAM
5	3791746164293636	3791746164293635	update	{"Typ	OZCAM
6	3791746164293637	3791746164293636	update	{"In	OZCAM
7	3791746956541994	3791746956541993	update	{"Sc	Australian Reference Genome Atlas (ARGA) manually curated
8	3791746956541996	3791746956541995	update	{"T	Australian Reference Genome Atlas (ARGA) manually curated
9	3791746956541999	3791746956541998	update	{"S	Australian Reference Genome Atlas (ARGA) manually curated
10	3791746956542000	3791746956541999	update	{"	Australian Reference Genome Atlas (ARGA) manually curated
11	3791746956542001	3791746956542000	update	{"	Australian Reference Genome Atlas (ARGA) manually curated



```
select operation_id, parent_id, action, atom, dataset_versions.version, datasets.name from specimen_logs
join dataset_versions 1..n<->1: on dataset_version_id=dataset_versions.id
join datasets 1..n<->1: on dataset_id=datasets.id
where entity_id='13114533894762173098'
and operation_id <= 3791746164293637
order by operation_id;
```

	operation_id	parent_id	action	atom	version	name
1	3791746164293632	3791746164129792	create	"Empty"	2024-04-17	OZCAM
2	3791746164293633	3791746164293632	update	{"RecordId": "QM J11296"}	2024-04-17	OZCAM
3	3791746164293634	3791746164293633	update	{"ScientificName": "Dendrolagus lumholtzi"}	2024-04-17	OZCAM
4	3791746164293635	3791746164293634	update	{"CanonicalName": "Dendrolagus lumholtzi"}	2024-04-17	OZCAM
5	3791746164293636	3791746164293635	update	{"TypeStatus": "HOLOTYPE"}	2024-04-17	OZCAM
6	3791746164293637	3791746164293636	update	{"InstitutionCode": "QM"}	2024-04-17	OZCAM

ARGA

Australian Reference Genome Atlas

Acknowledging 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

Thanks!

https://github.com/ARGA-Genomes
https://github.com/ARGA-Genomes/arga-oplogger/