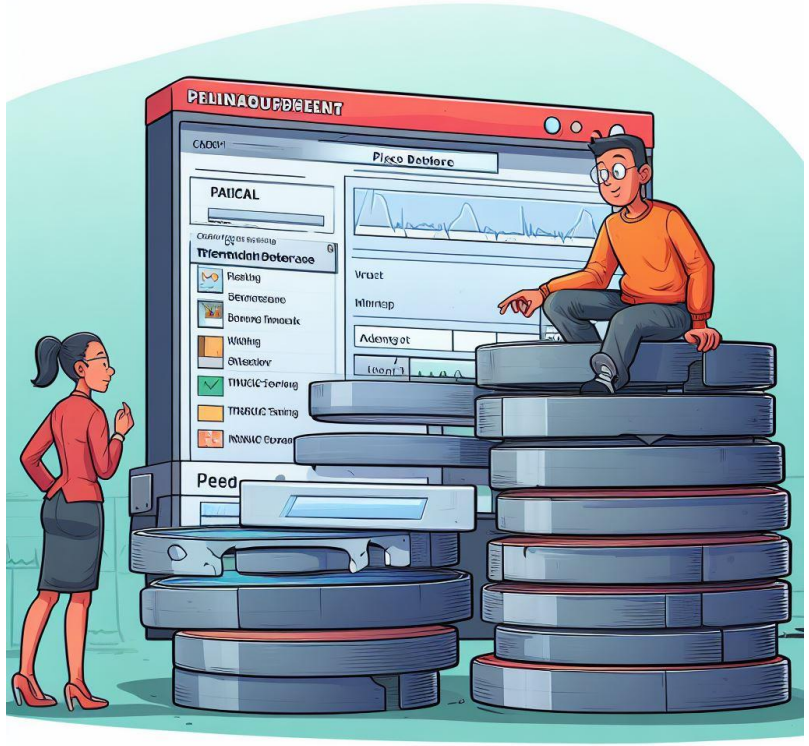
# Trialling Collective Access as a data curation and hosting platform



Mike Laverick

Centre for eResearch, University of Auckland

17 October 2023

# Setting the scene

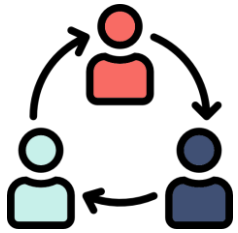
We help researchers with storage, compute, training, and bespoke research solutions

Approached in 2022 by an archaeology research group to help trial a data curation platform solution. This scenario is becoming more common:

Research group have a collection of historical data

Ongoing collection of more data from new projects

Need for central storage and to explore data



Need for collaborative curation of data

Need for fine grain access control

# Researchers suggested

What is Collective Access (CA) ?

- Open-source software designed to help catalogue and publish data collections
- Used by a number of museums and galleries around the world
- Highly customisable – can be tailored to specific collections/schemas/vocabulary

- Providence: a cataloguing portal backend to upload and curate data collections
- Pawtucket: a content-publishing front-end to present collections to public audience

# CA meets the needs of research group

We checked Collective Access could meet the groups specific needs:

- Store and display a variety of metadata and media of artefacts
- Ability to define and show hierarchical data and data relationships
- Capable of searching on arbitrary metadata of any database entries
- Fine-grain, role-based access controls to protect culturally-sensitive artefacts
- Ability to share data, for viewing and/or editing, with a wider audience
  i.e. engagement with relevant hapū/iwi/experts

# We agreed to help trial CA: Providence

Set up an MoU for 100 hours of work over 9 months (~2.5 weeks condensed)

Set up infrastructure and hosting on Nectar Research Cloud

Server maintenance and documentation of above processes

Help to customise Collective Access for research group needs

Help to scope out data ingestion workflow process

# Infrastructure and Hosting

Set up infrastructure and hosting on Nectar Research Cloud

Server maintenance and documentation of above processes

# Setting up Collective Access

The easiest aspects of this case study

- Routine set up of database and web server on Nectar VM (MySQL, PHP)
- We did not set up to production level, but CA can handle this
  (i.e. authentication via university systems, redundancy, automatic backups, etc)
- CA is fairly well documented on set up and configuration process
- Documentation on advanced configuration is hidden in config files themselves

# Maintaining Collective Access

The easiest aspects of this case study

- Trialled an in-place update to newest CA version midway through the case study
- Reasonably simple to do, and with minimal downtime (as usual make backups, etc)
- Ongoing maintenance requirements are quite light from technical support POV
  (VM patching, CA updates, storage and DB support, etc)
- Full case study documented on GitHub including:
  various CA config files, example CA schema mappings, ingest workflows, etc

# Customising to their data

Help to customise Collective Access for research group needs

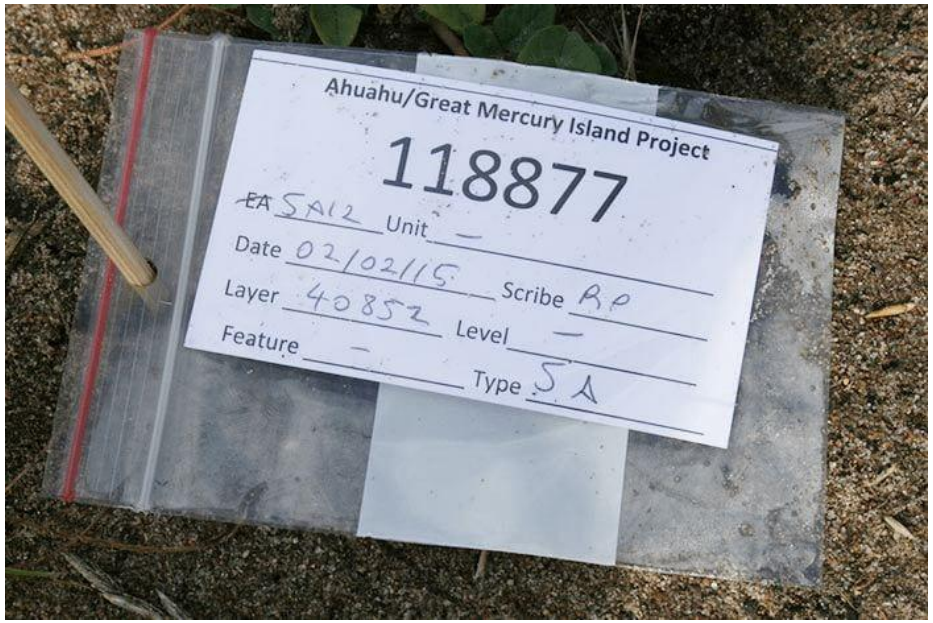
# Their archaeology data

The data is not “big” (order of GB’s, not TB’s) but is complex in relations and hierarchy

Trial with a historical, 6 years long excavation project

Over 200K artefact entries | 650+ locations within dig sites | 8500+ images from excavation

CSV files (various consistency), images, shapefiles

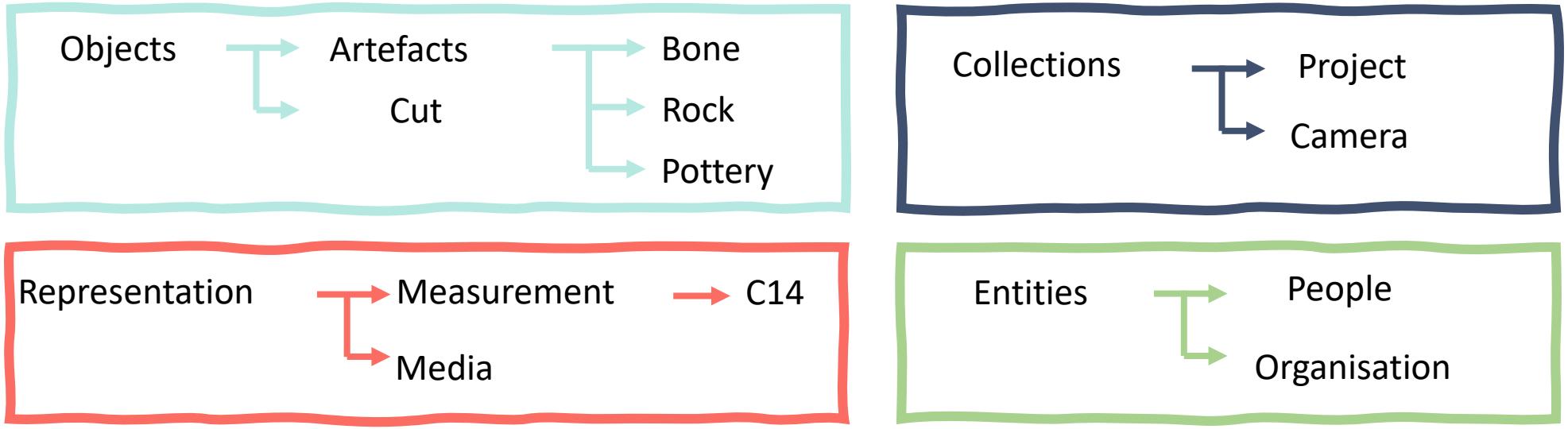



# How can CA be customised?

CA has its own intrinsic set of Primary tables so that it knows how to function and how to handle different types of content

i.e. Objects, Entities, Places, Collections, Occurrences, Representations

You can create custom subclasses as needed:



# How can CA be customised?

CA also allows you to create custom metadata fields and assign them to any of the tables in the database.

Similarly, you can create relationships and even User Interfaces (UIs) for specific tables and to display specific metadata

Finally, you can introduce entire vocabulary/label changes and multiple language support at every aspect of the tables/metadata/relations/UIs

All these customisations can be done via CA(providence)’s own backend UI or directly via the profile XML

# Mapping data onto the CA schema

This lengthiest aspect of this case study

- Flexibility inevitably requires complexity – Primary table types have quirks/nuances
- Mapping the existing archaeology schemas into an equivalent CA schema takes time (requires knowledge of both research data and CA software)
- CA documentation is extensive in places, and non-existent in other places

Main tables and relationships all created. Made examples of custom metadata & UIs. Documented the process for creating customisations, and various quirks/nuances

# Ingesting data into CA

Help to scope out data ingestion workflow process

# Ingesting data into CA

The trickiest aspect of this case study to do correctly

- CA has extensive import mechanisms, but the documentation is limited in places
- Bulk uploads were done using CSV data files, but required xlsx mapping files
- Creating the mapping files is once again very iterative and nuanced, and upload process is very fragile (poor exception handling, suggests “backup” before uploads!)

Uploaded all media and ~30K artefacts into CA. Examples of import/export mappings and documentation of process + more quirks/nuances

# Takeaway and outcome

Collective Access is a powerful and flexible platform that meets a lot of research requirements

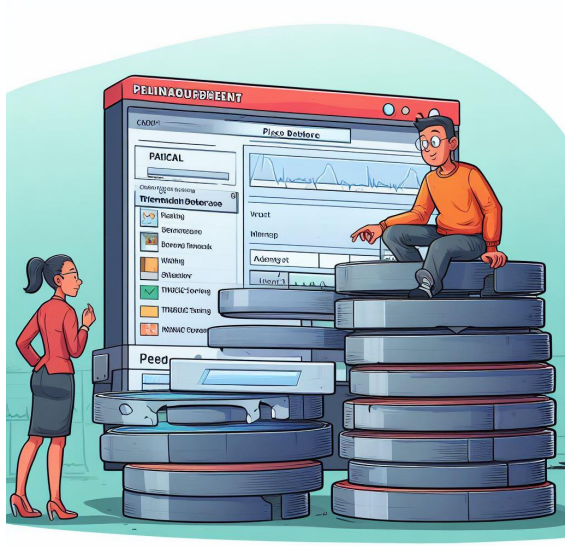
The trade off is that it requires substantial time and effort to map **new** data into CA
time + effort required from both Researchers and IT-support

Successfully created and trialled a Collective Access instance tailored for the research group
(30K artefacts, 8500+images, 650+ dig sites)

Our case study: the research champion left academia shortly after project finish
so next steps are left in limbo

**Collective Access is a powerful tool**
**but requires substantial effort from all involved**

# Thanks for listening!



Questions?

mike.laverick@auckland.ac.nz

https://github.com/UoA-eResearch/ahuahu_great_mercury