Genomic data - improving discovery and access management

Wednesday 14 June 2023

biocommons.org.au AustralianBioCommons @AusBiocommons

Australian BioCommons

Acknowledgement of Country

We acknowledge the Traditional Owners and their custodianship of the lands on which we meet today.

We pay our respects to their Ancestors and their descendants, who continue cultural and spiritual connections to Country.

We recognise their valuable contributions to Australian and global society.

Australian BioCommons

Australian BioCommons

Actively supporting Australian life sciences research through bioinformatics and bioscience data infrastructure

biocommons.org.au  AustralianBioCommonsChannel  @AusBiocommons

Introductions

Jess Holliday

Australian BioCommons

Mustafa Syed

Children's Cancer Institute

Dr Andrew Patterson

University of Melbourne

A/Prof Sarah Kummerfeld

Garvan Institute of Medical Research

Images from LinkedIn and organisation webpages

Australian BioCommons

Australian Context

Current state in Australia

- Huge amounts of human genomic data being generated
- No national solution for genomic data archiving, querying & sharing
- Data siloed within institutions and not discoverable
- Some data sent to European Genome-Phenome Archive (EGA) but:
  - It is challenging
  - Contextual metadata is lacking

https://www.biocommons.org.au/human-genome-informatics-initiative

images from https://www.australiansiloarttrail.com/

Australian BioCommons

Human Genomes Platform Project

The Human Genomes Platform Project

THE OPPORTUNITY:

- Enhance capability for **securely**, **responsibly** & **FAIRly sharing** human genomics research data nationally and internationally
- Ensure **maximum value** can be derived from Australian research outputs

THE SOLUTION:

- **HGPP**
- Nationally-funded **collaborative** project
- Running from Jan 2021- Nov 2023
- Deliver a **services toolbox** based on emerging **global standards**

https://www.biocommons.org.au/hgpp

Australian BioCommons

HUMAN GENOMES PLATFORM PROJECT

Virtual Cohorts

DAC Automation

Federated Identity & Access Management

Data & Metadata Archiving

Communications, Documentation & Training

Project Partners

- Australian BioCommons
- Australian Access Federation
- Bioplatforms Australia
- Garvan Institute of Medical Research
- NCI Australia
- QIMR Berghofer Medical Research Institute
- The University of Melbourne
- zero Childhood Cancer

What's in our toolbox?

virtual cohorts

Beacon Network

DAC approval

REMS

CSC

Identity & access

CILogon

AUSTRALIAN ACCESS FEDERATION

Data archives

FEDERATE EUROPEAN GENOME-PHENOME ARCHIVE

Documentation & training

zenodo

YouTube zoom

Principles:

- open source - global standards - build on what exists - document

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

elixir

Outputs so far...

- Knowledge discovery reports and documentation published to Zenodo, GitHub
- Posters and oral presentations delivered at multiple national and international conferences
- Prototype implementations of Beacon, Beacon network, CILogon, REMS, LocalEGA
- Connections established to international initiatives - ELIXIR, GA4GH, EGA, CILogon
- Relationships built between major players in human genomics and beyond
- Foundation for future national-scale projects and initiatives

Today’s webinar

Identity & access

Andrew Patterson

virtual cohorts

Mustafa Syed

DAC approval

Sarah Kummerfeld

Australian
BioCommons

IAM (Identity Access Management)

Identity management a key infrastructure plank in any federation of systems (i.e. cross institute)

There is a broadly agreed base technology set that achieves this

- Open ID Connect (OIDC)
- OAuth
- SAML
- (and to some extent GA4GH Passports/Visas on top)

IAM (Identity Access Management)

TLDR; What all this means is that users can use their own institute credentials to login to collaborative websites.

The same technology as “login via Google” / “login via Github”.

But across all Australian universities and a lot of research institutes.

The “user” has a consistent identifier across the entire network.

IAM in Practice

In practice then, what were the goals of the IAM stream of HGPP?

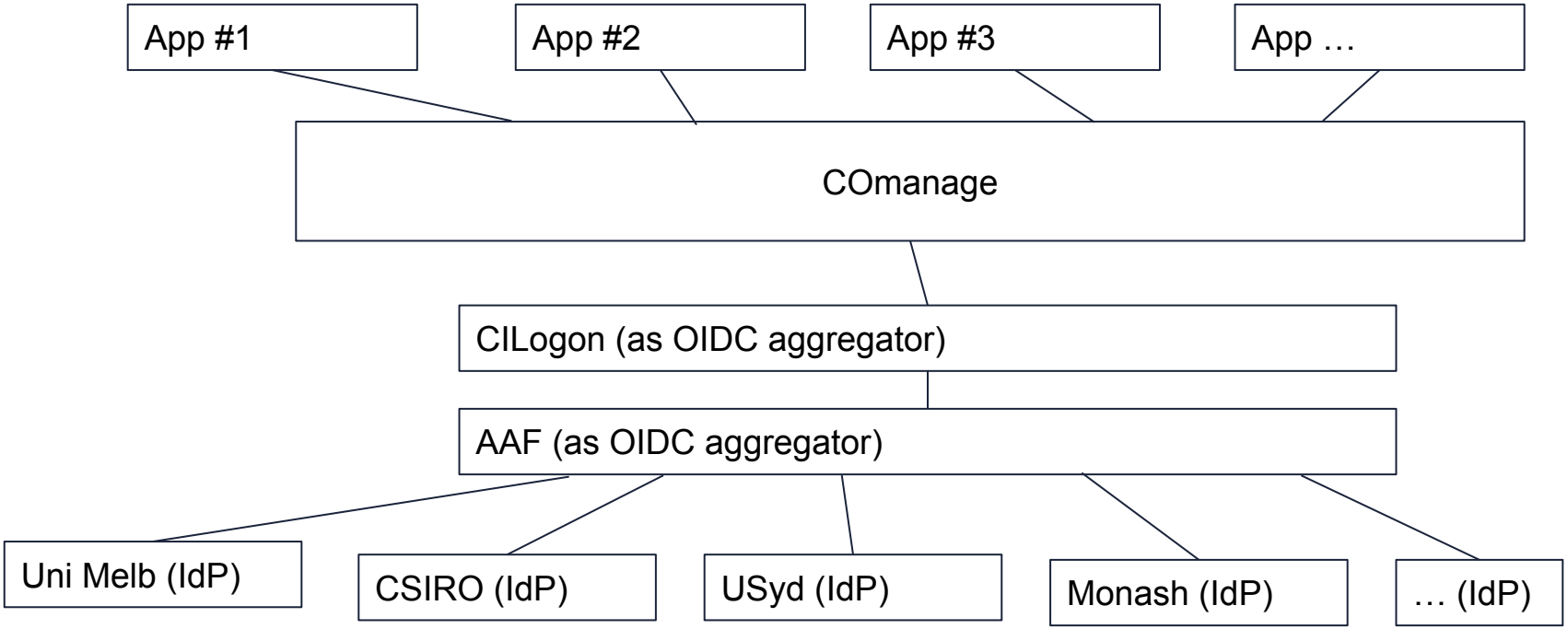
- Demonstrate the technology in practice
- Investigate gaps and recommend solutions to address these gaps
- Pilot solutions

IAM in Practice

Main gap identified in the base technology is “Community Management” - the ability of groups of users to self organise into communities and control rights etc

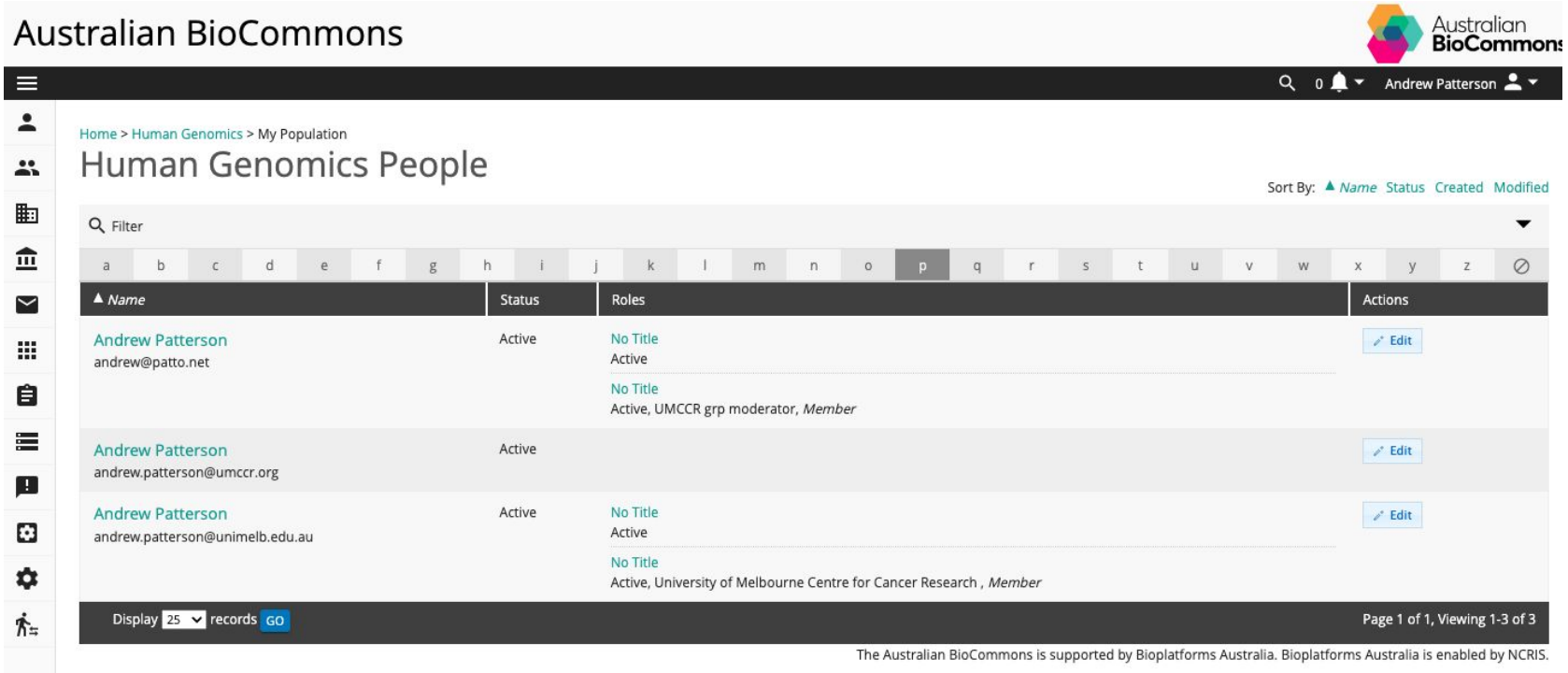
After some scans of products - chose CILogon from University of Illinois - National Center for Supercomputing Applications (https://www.cilogon.org/)

CILogon / COmanage



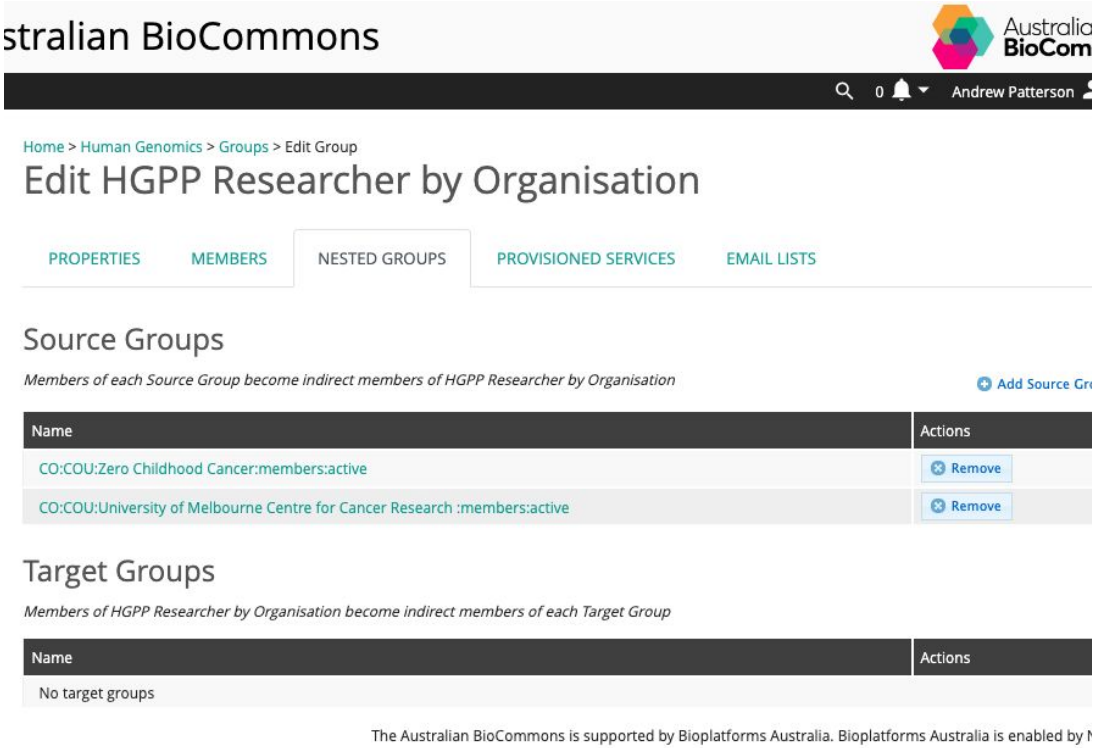
Australian BioCommons

CILogon / COmanage



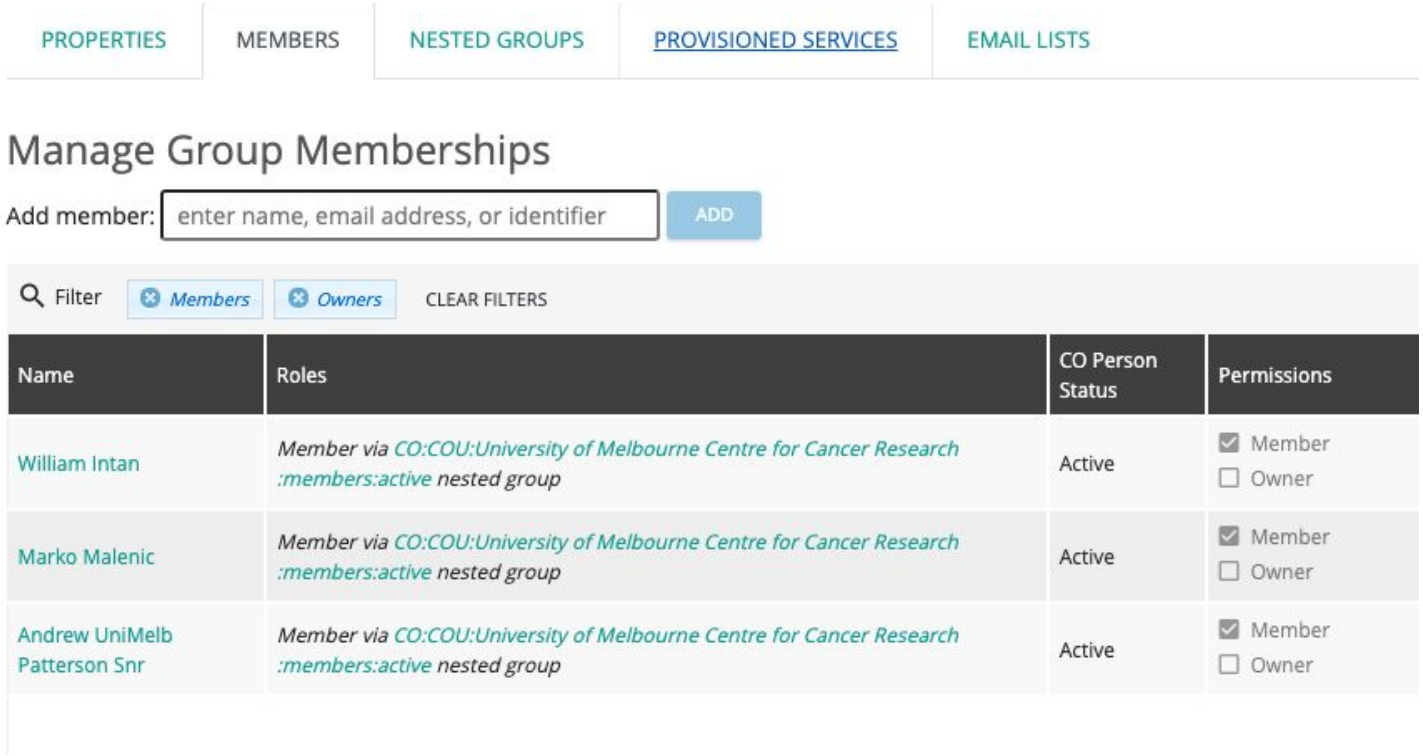
Australian BioCommons

CILogon / COmanage



Australian BioCommons

CILogon / COmanage



Australian BioCommons

IAM in Practice (app #1)

We used CILogon in a particularly boring way - as a gateway into logging in to an application (Elsa Data)

The “group” was the blocker to being able to login (only users in the group could do that).

IAM in Practice (app #2)

In a more sophisticated usage - we wanted to establish the concept of a “bona fide researcher”. A bona fide researcher was similarly someone who was registered in a group signifying their trust within a community - but this information needed to be passed out to the Beacon network federation.

We will now hand over to Mustafa for a demonstration of Beacon network. The CILogon IAM underpins the authentication in this cross institute system.

Mustafa

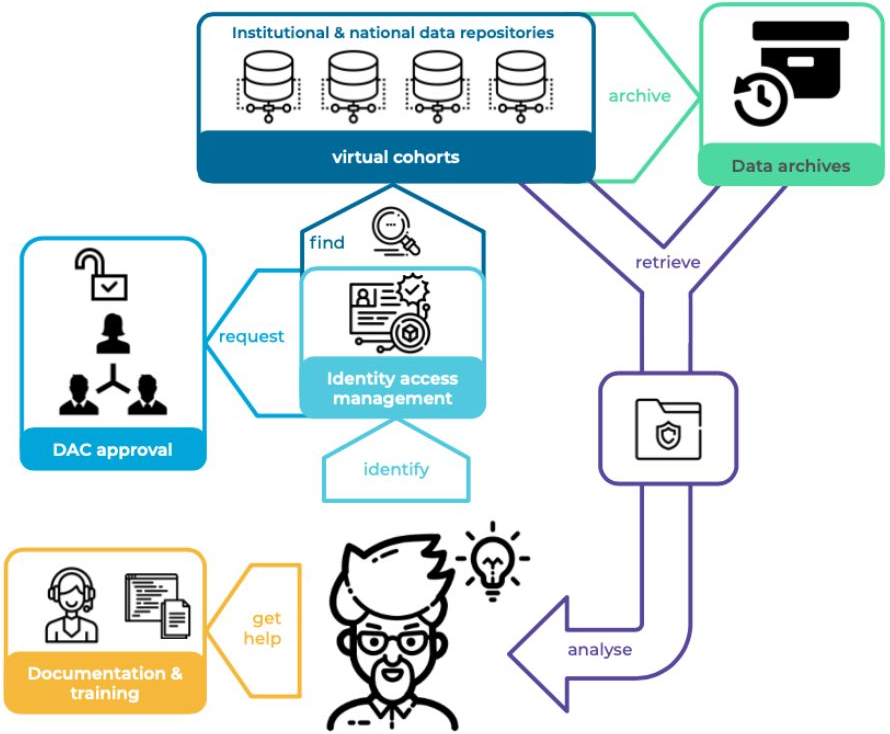
Demo

Goal

Enhance capability for securely and responsibly sharing human genomics research data nationally

Australian BioCommons

Concept



Australian BioCommons

Current State

- We have identify set of problems that need to be addressed
- The current state of cross-institutional human genomic data querying in Australia
- The current state of processes and tools for virtual cohort querying
- Recommendations on preferred technology and proposed implementation architecture
- Candidate solutions to enable cross-institutional virtual cohort querying

Requirements

- Common data description (ontology) to annotate data from different sources
- A minimal common data model
- Implementation of distributed search/query protocol or framework
- User interface - browse, search data, request access, download data
- Integration with other sub-project - IAM, DAC, Data Metadata Arhieve

User Stories

Rank	QIMR Berghofer MRI	Garvan	NCI (JCSMR)	UMCCR	ZERO
U.S.1	*As a research user:* I want to know who holds sequencing data for PDAC cases ***So that:*** We can build a virtual cohort of cases for discovery	*As a research user:* I want to identify all individuals with a particular set of clinical characteristics and obtain primary data ***So that:*** We can build a virtual cohort of cases for discovery	*As a research user:* I have an interest in research topic X. What datasets have the required consents for me to use to address research topic X? ***So that:*** We can build a virtual cohort of cases for discovery	*As a curator:* I want to find variant information for cancer samples of a given subtype ***So that:*** We can assess a novel variant	*As a research user:* I want to find all medulloblastoma samples, get access and download th ***So that:*** We can utilise them for rese
U.S.2	*As a research user:* I want to know who holds sequencing data for PDAC cases, from fresh-frozen tissue ***So that:*** We can build a virtual cohort of cases for discovery	*As a research user:* I want to identify all individuals with a particular set of variants and/or clinical characteristics and obtain primary data ***So that:*** We can build a virtual cohort of cases for discovery	*As a research user:* I want to know what restrictions I have on the use of data? ***So that:*** We can build a virtual cohort of cases for discovery	*As a research user:* I want to find primary / read level data for published cancer cohorts stored in Australia ***So that:*** So I can re-process / harmonise data	*As a research user:* I don't have access to large where can I run my analysis samples ***So that:*** I can perform my analyses i cohort
U.S.3	*As a research user:* I want to know who holds sequencing data for PDAC cases, from fresh-frozen tissue, with survival timepoints ***So that:*** We can build a virtual cohort of cases for discovery	*As a research user:* I want to run analyses on my virtual cohorts in situ (i.e. bringing compute to the data) ***So that:*** We can analyse the data in the virtual cohort	*As a research user:* Can I download the data and share it with my collaborators in Australia and/or overseas? ***So that:*** We can establish allowed uses of the data	*As a research user:* I want to find primary / read level data for published cancer cohorts stored in Australia of a given phenotype / with minimal metadata requirements ***So that:*** So I can re-process / harmonise data	*As a research user:* How can I find all paediatric < 21 yrs) ***So that:*** I can consolidate my data wi
U.S.4	*As a research user:* I want to know how frequently a particular germline variant occurs in cases of healthy normal/never diagnosed ***So that:*** We can better understand variant distribution in the Australian population	*As a research user:* I want to share data and analyses on my virtual cohorts in situ (i.e. bringing compute to the data) ***So that:*** We can analyse the data in the virtual cohort	*As a research user:* Where can I perform computation on data once I have identified all required samples to comply with DAC requirements? ***So that:*** We can analyse the data in the virtual cohort	*As a research user:* I want to find primary / read level data for published cancer cohorts stored in Australia of a given phenotype / with minimal metadata requirements and with data access control requirements matching my research plan ***So that:*** So I can re-process / harmonise data	*As a research user:* I want to be able to access various cohorts and studies ***So that:*** I know how to normalise my data with the virtual cohort(s)
U.S.5	*As a clinician researcher user:* I want to know who holds clinical data including treatment regime and survival timepoints, for PDAC cases with KRAS G12D mutation ***So that:*** We can build a virtual cohort of cases for analysis	*As a research user:* *I want to identify samples with a particular set of clinical characteristics and/or variants that have available tissue for follow up studies* ***So that:*** *We can perform follow up research*	*As a research user:* [How] can I reconnect with participants for follow up sample, additional information, or return results of incidental findings? ***So that:*** We can perform follow-up research and potentially return results	*As a research user:* I want to *share* primary / read level and secondary / variant level data for our own research cohorts alongside agreed-upon phenotype and minimal metadata annotation restricted by DUO codes ***So that:*** others can use our data	*As a clinician researcher:* I want to build a cohort where after last follow-up the patient has sta ***So that:*** I can build a virtual cohort fo analysis
U.S.6	*As a data custodian:* I want to limit which users can view which information — e.g. public access for catalogue type data (what do we hold) plus possibly somatic variants ***So that:*** Access to data is restricted or exposed as appropriate				*As a research user:* I want to identify all Neurobl patients with ALK fusions, w disease status at most rece ***So that:*** I can determine the prognostic impact of this driver mutation (SNPs and AMP are well established biomarkers in this disease).

Data Model

Sample types: blood, fresh frozen, nail, skin biopsies

Sequencing Data types: primary, secondary, WGS

Phenotype: cancer diagnosis, or a patient's genetic disorder

Mutations: SNVs, indels, and gene fusions

Data Access: Consent, data location, data use requirements

Clinical metadata: survival time, age, follow-up, disease status

Cohort level data: Number of patients of a certain type

Australian BioCommons

Candidate Solutions

CINECA - GECKO

GA4GH - Data Connect

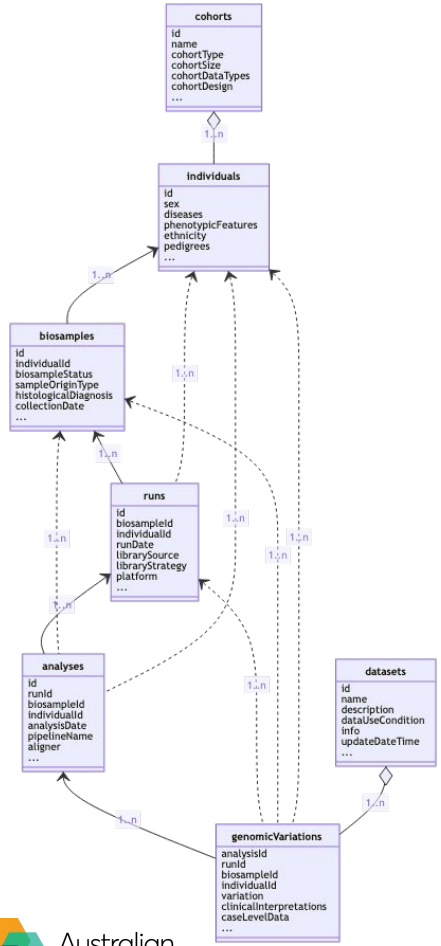
GA4GH - Phenopacket

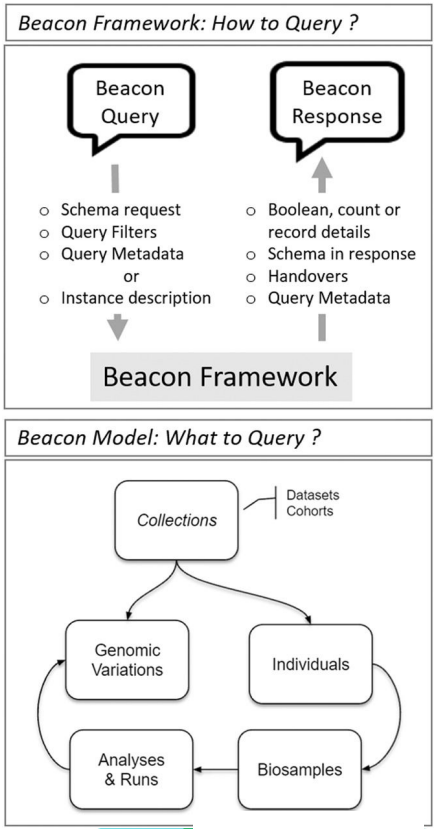
Gen3 Data Commons

OMOP - CDM

Beacon v2

Beacon v2





Rambla, Jordi, et al.

SIMPLE

Beacons can be implemented on top of any underlying variant or variant annotation datastore

FEDERATED

Beacons can be lit and maintained by individual organizations, and assembled into a distributed network

AGGREGATIVE

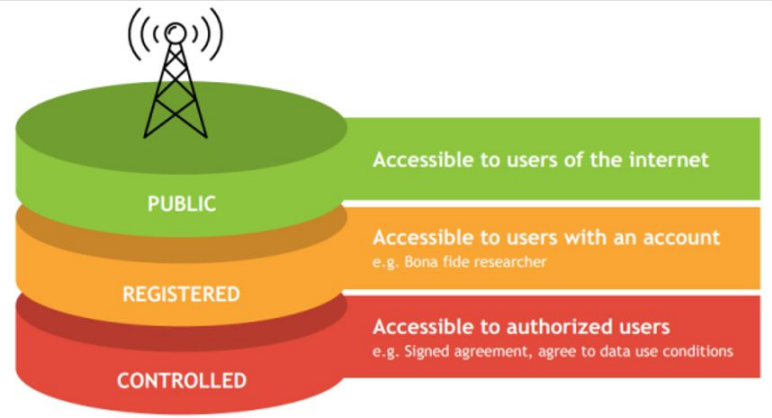
Beacon results can be aggregated across an entire populations to support de-identification.

SECURABLE

Beacon access can be restricted using institutional security protocols, and authorization schemes can be implemented to respect consent conditions.

Australian BioCommons

Data Access Control

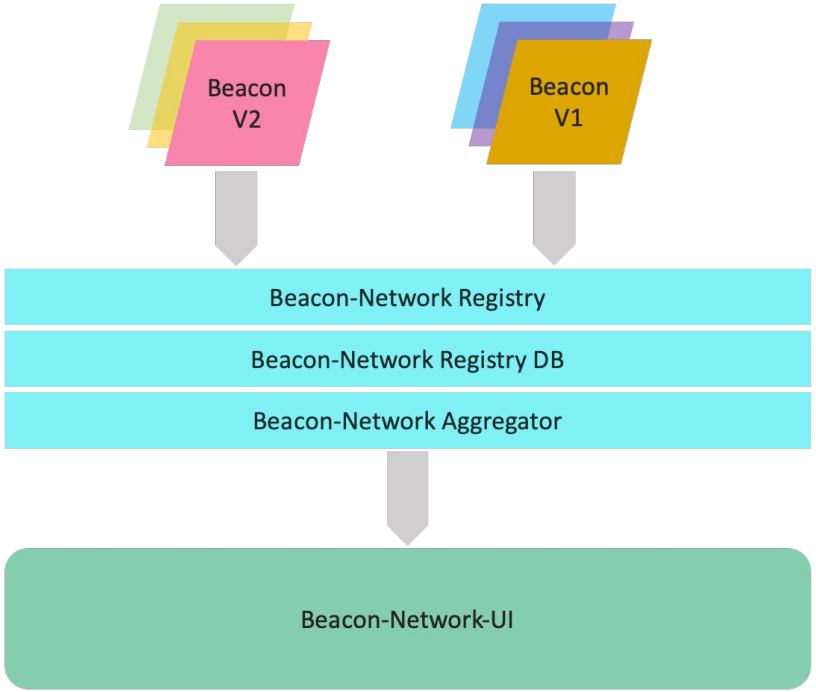


Boolean | Count | Resultset


```
"datasets": [
  {
    "id": "urn:hg:1000genome",
    "name": "1000 genomoe",
    "externalUrl": "ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/",
    "description": "Data from the 1000 Genomes Project. The 1000 Genomes Project ran between 2008 and 2015
the Wellcome Trust to maintain and expand the resource.",
    "assemblyId": "GRCh38",
    "variantCount": 4242,
    "callCount": 3892,
    "sampleCount": 2504,
    "version": "v0.4",
    "info": {
      "accessType": "PUBLIC"
    },
    "createDateTime": "2013-05-02T12:00:00Z",
    "updateDateTime": "2013-05-02T12:00:00Z"
  },
  {
    "id": "urn:hg:example-controlled",
    "name": "Example CONTROLLED dataset",
    "externalUrl": null,
    "description": "Example dataset to showcase CONTROLLED access level.",
    "assemblyId": "GRCh38",
    "variantCount": 1,
    "callCount": 2,
    "sampleCount": 1,
    "version": "1.0.0",
    "createDateTime": null,
    "updateDateTime": null,
    "info": {
      "accessType": "CONTROLLED"
    }
  },
  {
    "id": "urn:hg:example-registered",
    "name": "Example REGISTERED dataset",
    "externalUrl": null,
    "description": "Example dataset to showcase REGISTERED access level for ELIXIR Bona Fide status.",
    "assemblyId": "GRCh38",
    "variantCount": 1,
    "callCount": 2,
    "sampleCount": 1,
    "version": "1.0.0",
    "createDateTime": null,
    "updateDateTime": null,
    "info": {
      "accessType": "REGISTERED"
    }
  }
],
```


Australian BioCommons

Architecture Diagram



Beacon-v1: Genomic
Beacon-v2: Genomic, Phenotypic, Clinical
Query API

Registry: Register beacons
Registry-DB: Save beacon details
Aggregator: Query multiple beacons

Show registered beacons
Run Queries

Australian
BioCommons

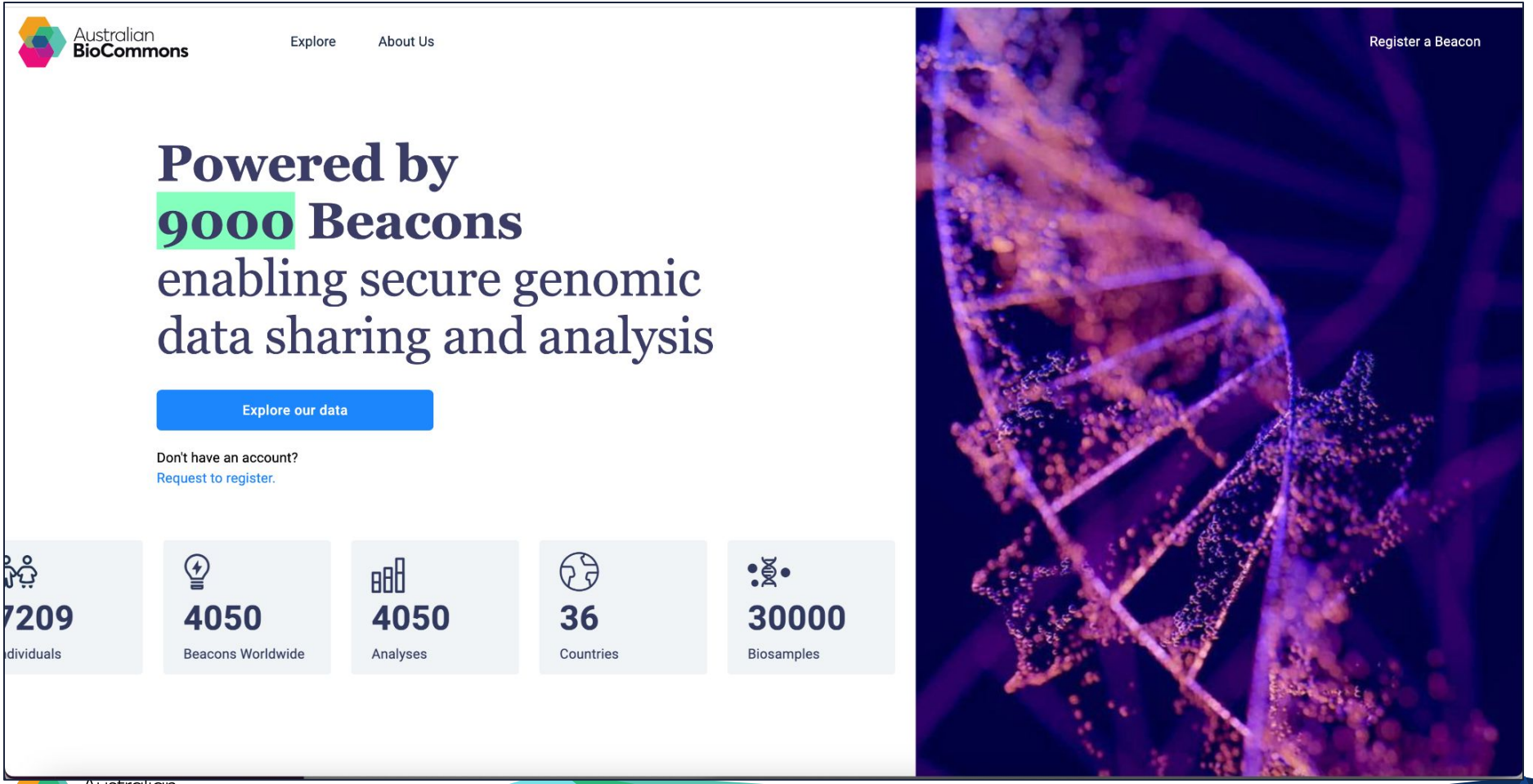
Beacon Network v2 UI

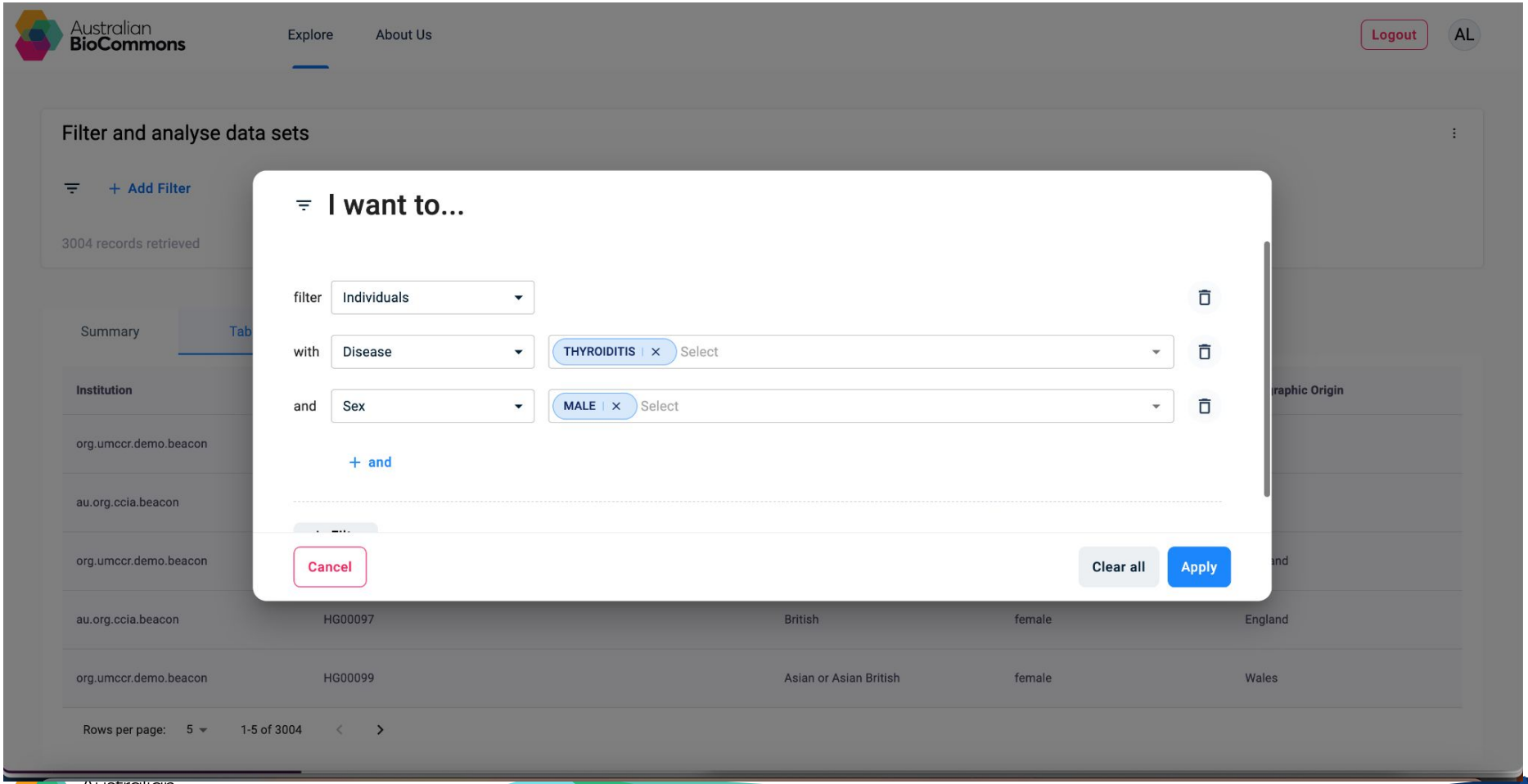
We have designed and implemented a UI to support Beacon v2.

The purpose of the UI is to enable querying the full catalogue of data by all members of the Beacon network.

Demo: https://beacon-network.dev.umccr.org/

Demo


Australian BioCommons
Explore
About Us
Register a Beacon
Powered by
9000 Beacons
enabling secure genomic
data sharing and analysis
Explore our data
Don't have an account?
Request to register.
7209
dividuals
4050
Beacons Worldwide
4050
Analyses
36
Countries
30000
Biosamples


Australian BioCommons
Explore
About Us
Logout
AL
Filter and analyse data sets
Add Filter
3004 records retrieved
I want to...
filter
Individuals
with
Disease
THYROIDITIS
Select
and
Sex
MALE
Select
and
Cancel
Clear all
Apply
Summary
Institution
org.umccr.demo.beacon
au.org.ccia.beacon
org.umccr.demo.beacon
au.org.ccia.beacon
HG00097
British
female
England
org.umccr.demo.beacon
HG00099
Asian or Asian British
female
Wales
Rows per page: 5
1-5 of 3004

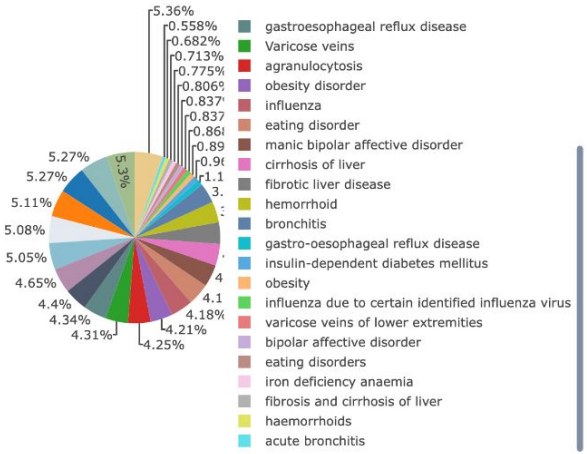
Filter and analyse data sets

+ Add Filter

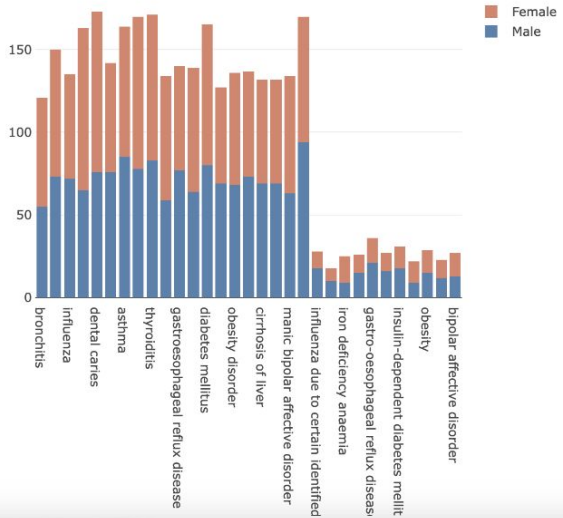
3004 records retrieved

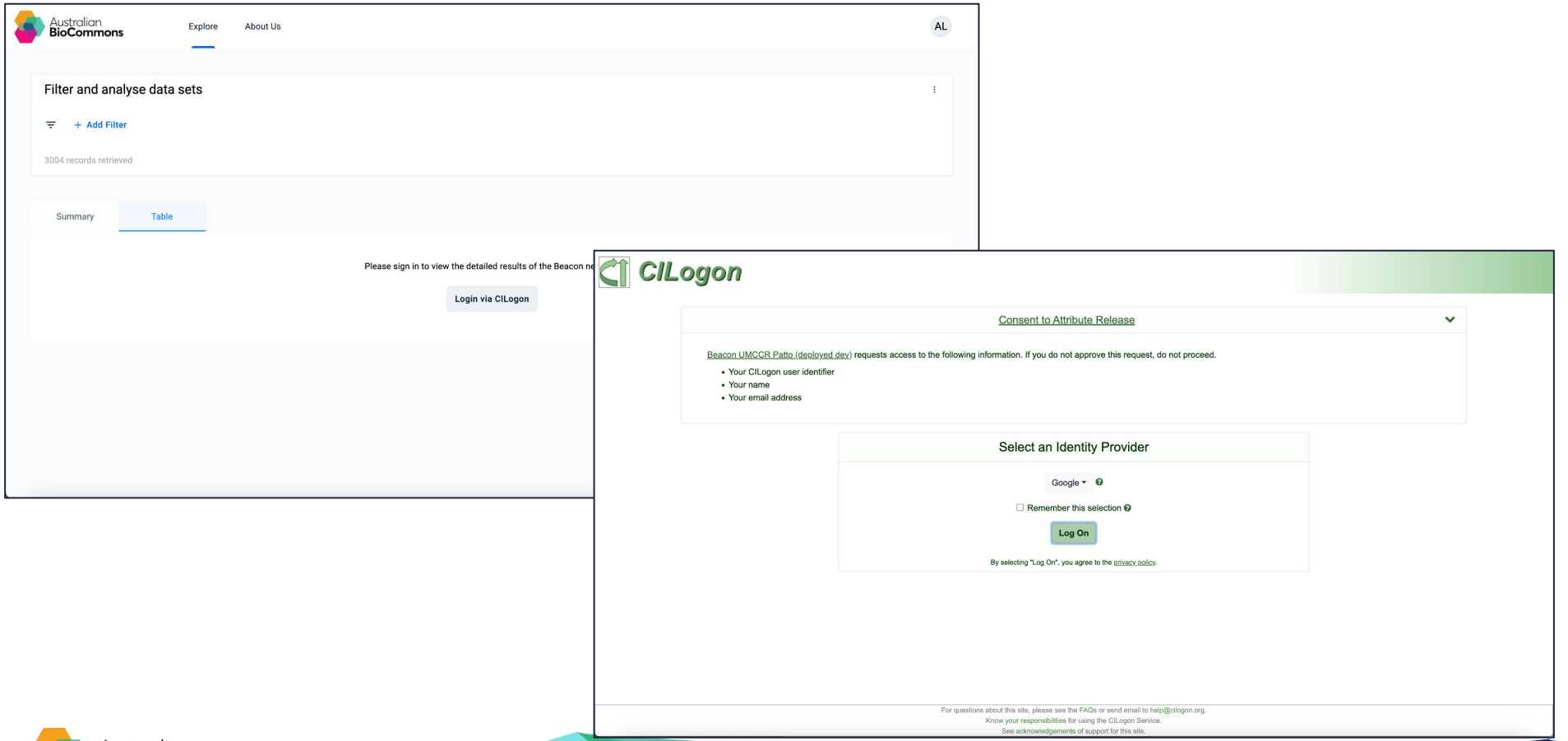
Summary | Table

Disease distribution



Sex distribution across diseases





Australian BioCommons

Australian BioCommons

Explore About Us

Logout AL

Filter and analyse data sets

+ Add Filter

3004 records retrieved

Summary Table

Institution	Id	Disease	Ethnicity	Sex	Geographic Origin
org.umccr.demo.beacon	HG00096		African	male	-
au.org.ccia.beacon	HG00096		African	male	-
org.umccr.demo.beacon	HG00097		British	female	England
au.org.ccia.beacon	HG00097		British	female	England
org.umccr.demo.beacon	HG00099		Asian or Asian British	female	Wales

Rows per page: 5 1-5 of 3004

Australian BioCommons

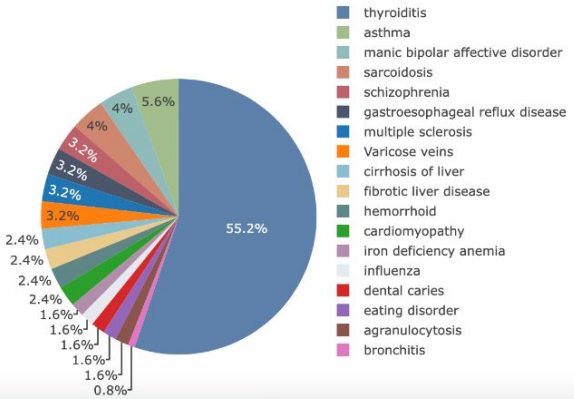
Filter and analyse data sets

+ Add Filter

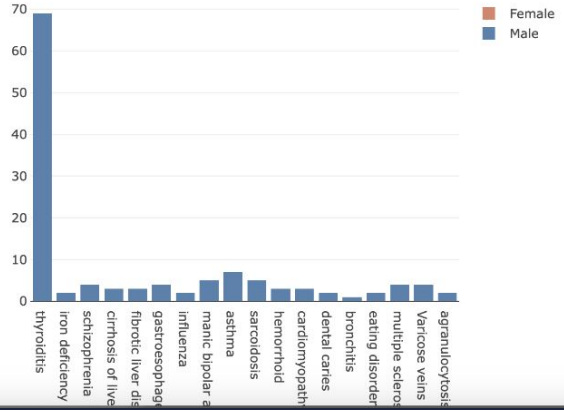
69 records retrieved

Summary | Table

Disease distribution



Sex distribution across diseases

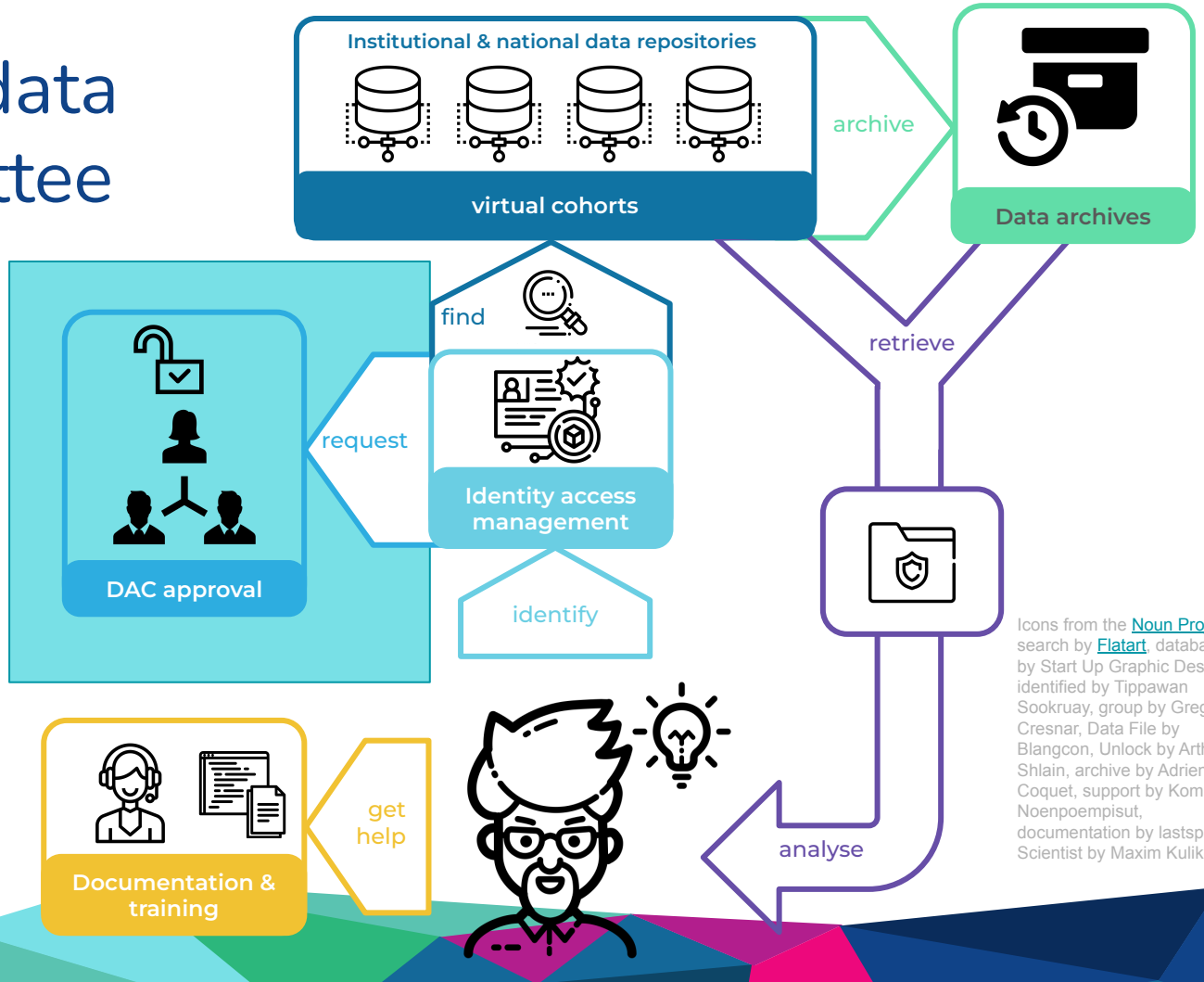


Australian BioCommons

Data Access Committee Automation

Australian BioCommons

Streamlining data access committee management



Icons from the Noun Project: search by Flatart, database by Start Up Graphic Design, identified by Tippawan Sookruay, group by Gregor Cresnar, Data File by Blangcon, Unlock by Arthur Shlain, archive by Adrien Coquet, support by Komkrit Noenpoempisut, documentation by lastspark, Scientist by Maxim Kulikov.

Australian BioCommons

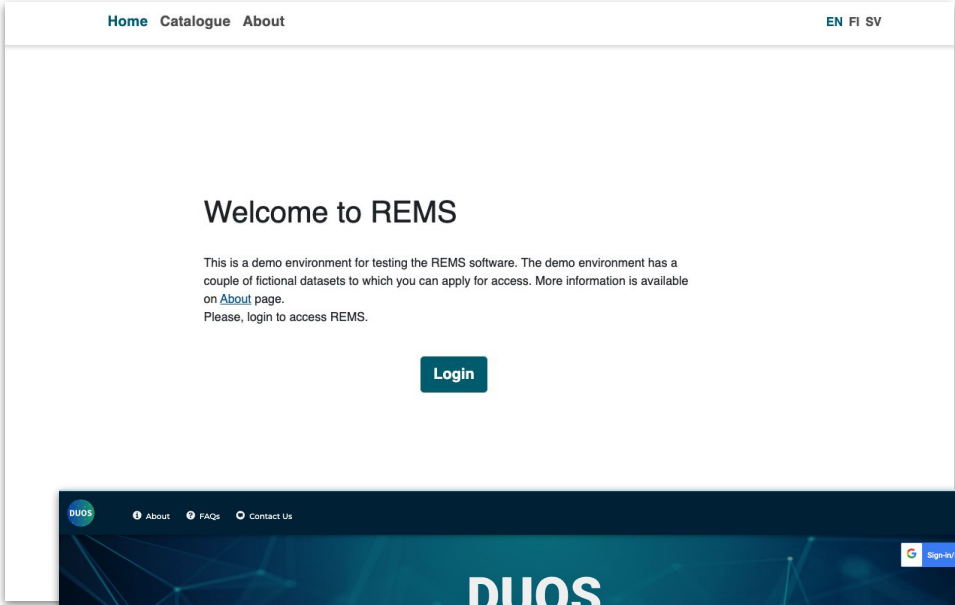
Data Access Committee Support

?

Australian BioCommons

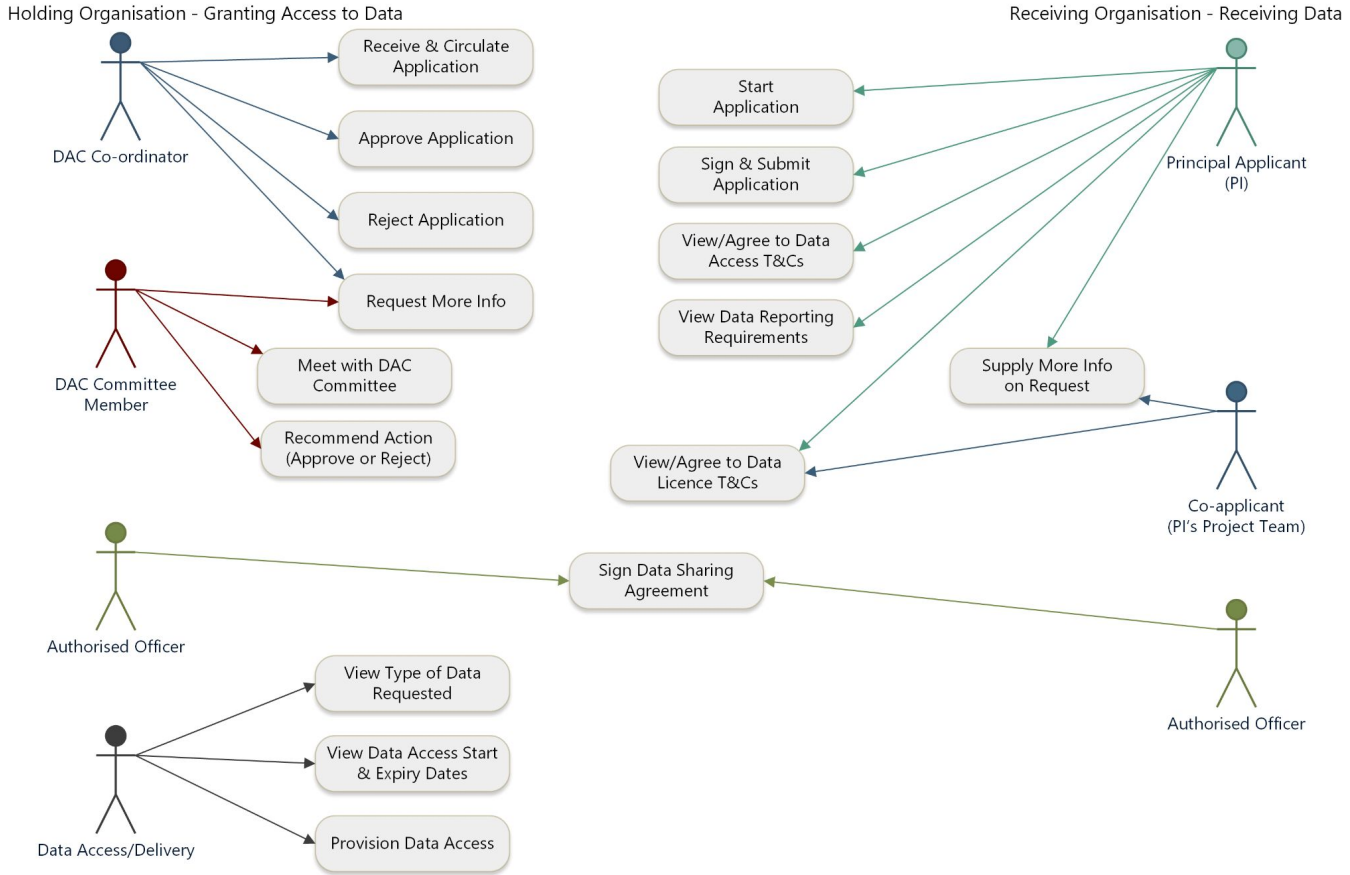
Subproject process

- Compile user stories
- Document requirements and test scenarios
- Identify candidate solutions: REMS and DUOS
- REMS installed at Garvan and UMCCR
- Business analysis of REMS
- Interaction with REMS team in Finland to add features (allow use of AWS email servers) was quick and effective
- Assess REMS against requirements using test scenarios



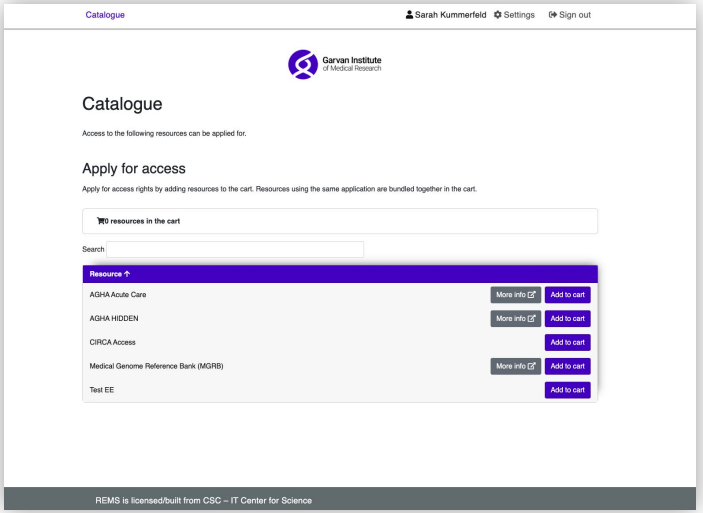
Australian BioCommons

System Use Cases | Submit Data Access Application and Approve, Reject, Sign, Provision



Australian BioCommons

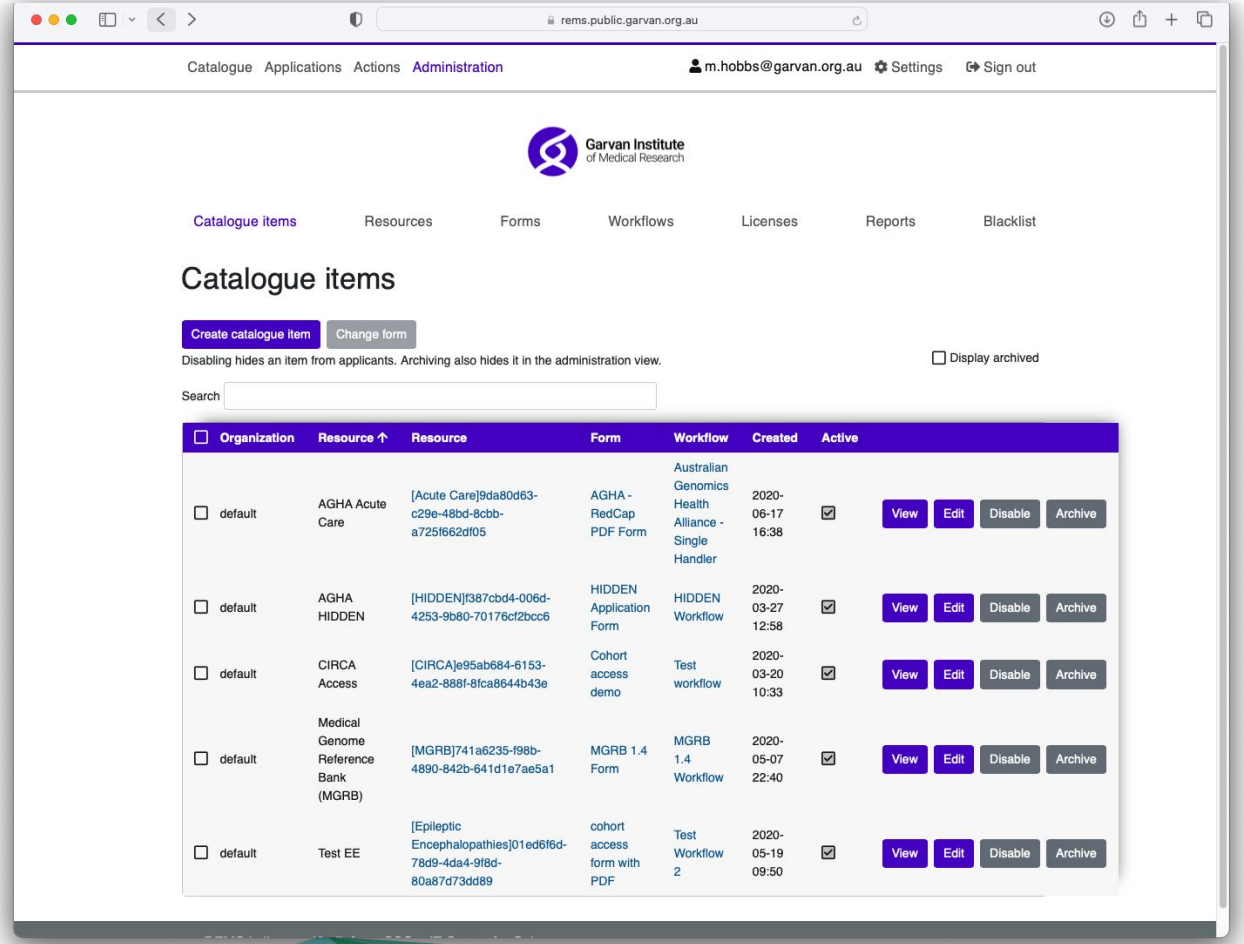
Case Study: Garvan Data Access Committee Automation with REMS



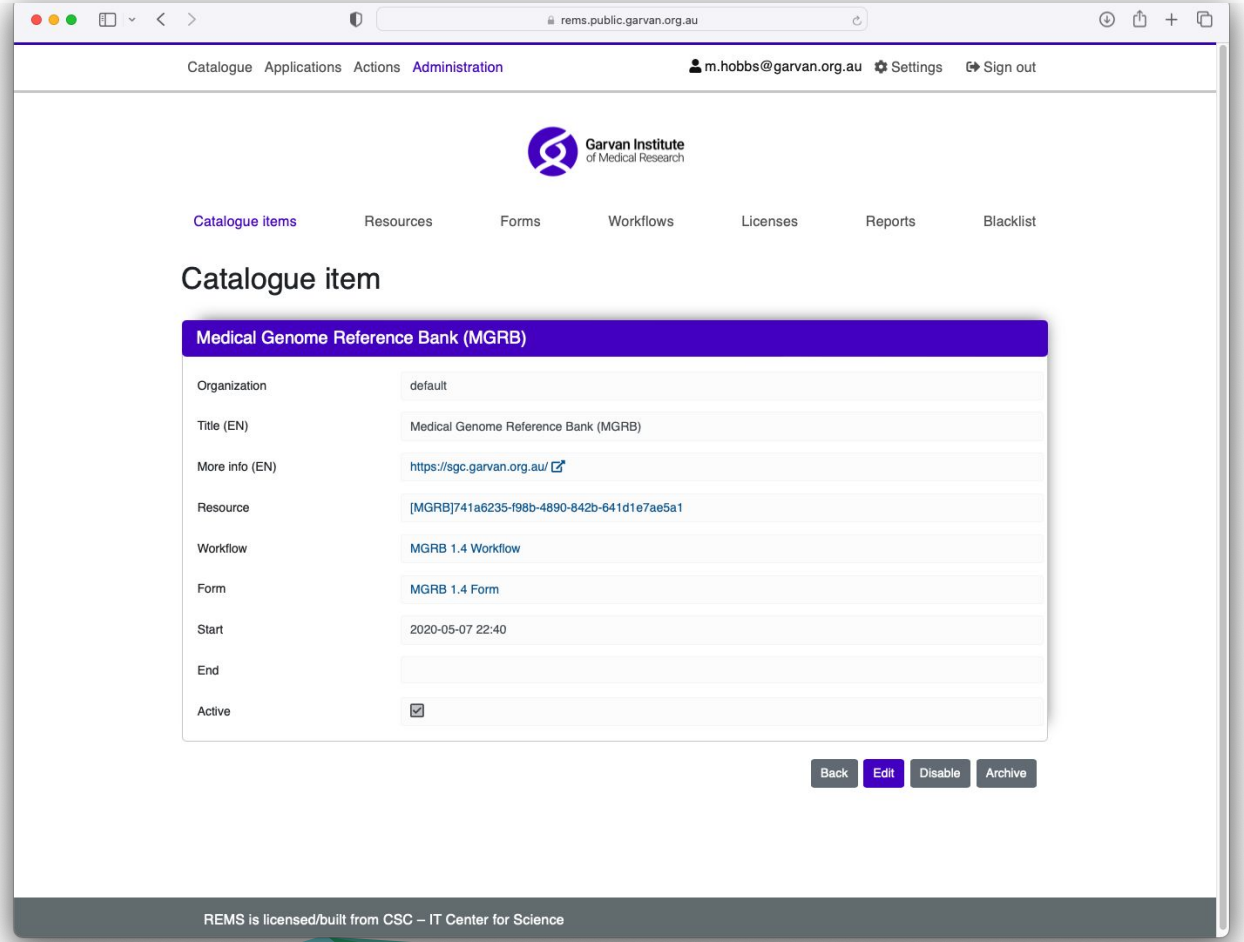
Australian BioCommons

Streamlining data access

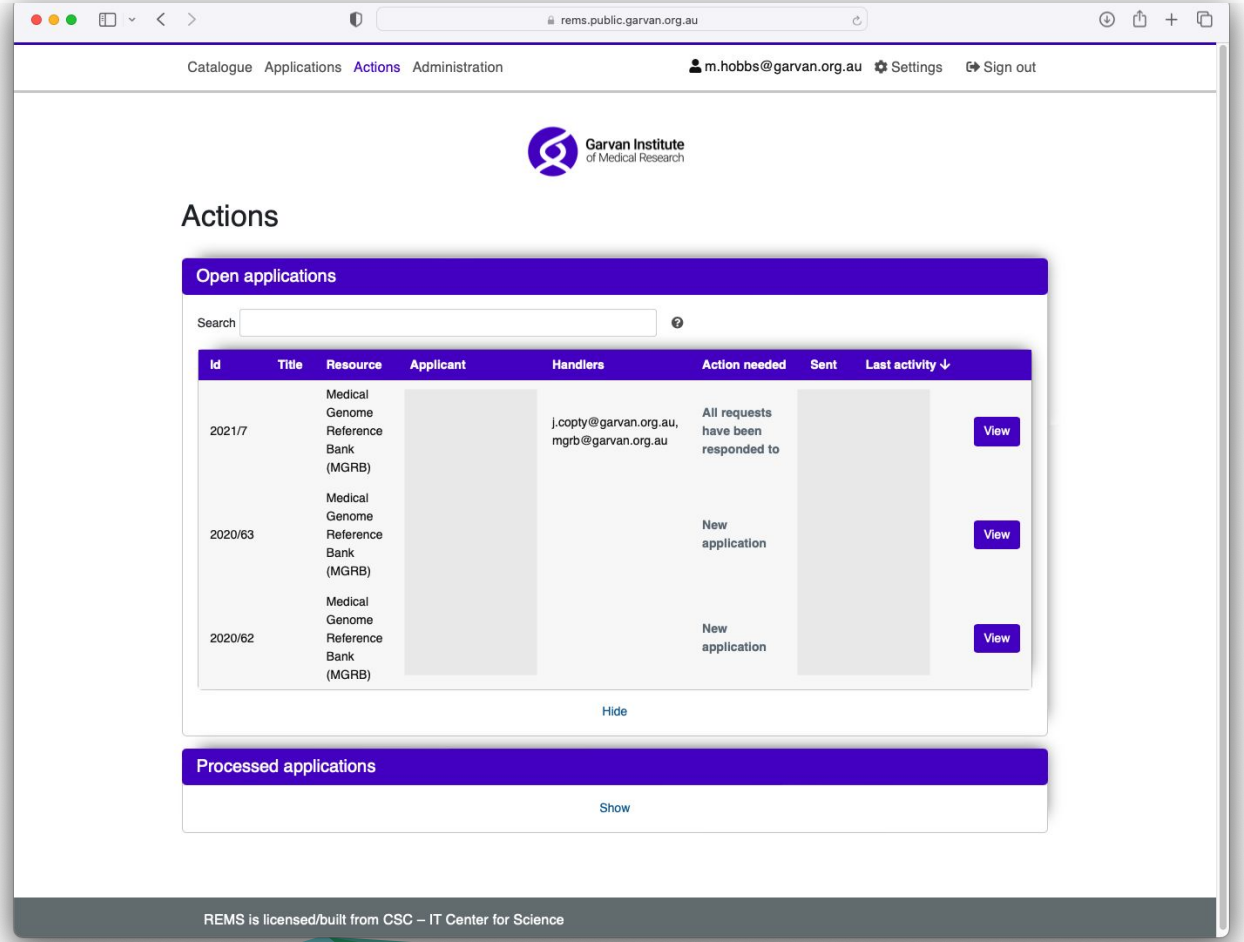
- High value, access controlled datasets
- Each cohort has a separate data access committee
- Users from around the world request access
- Garvan's data science platform are the data custodians



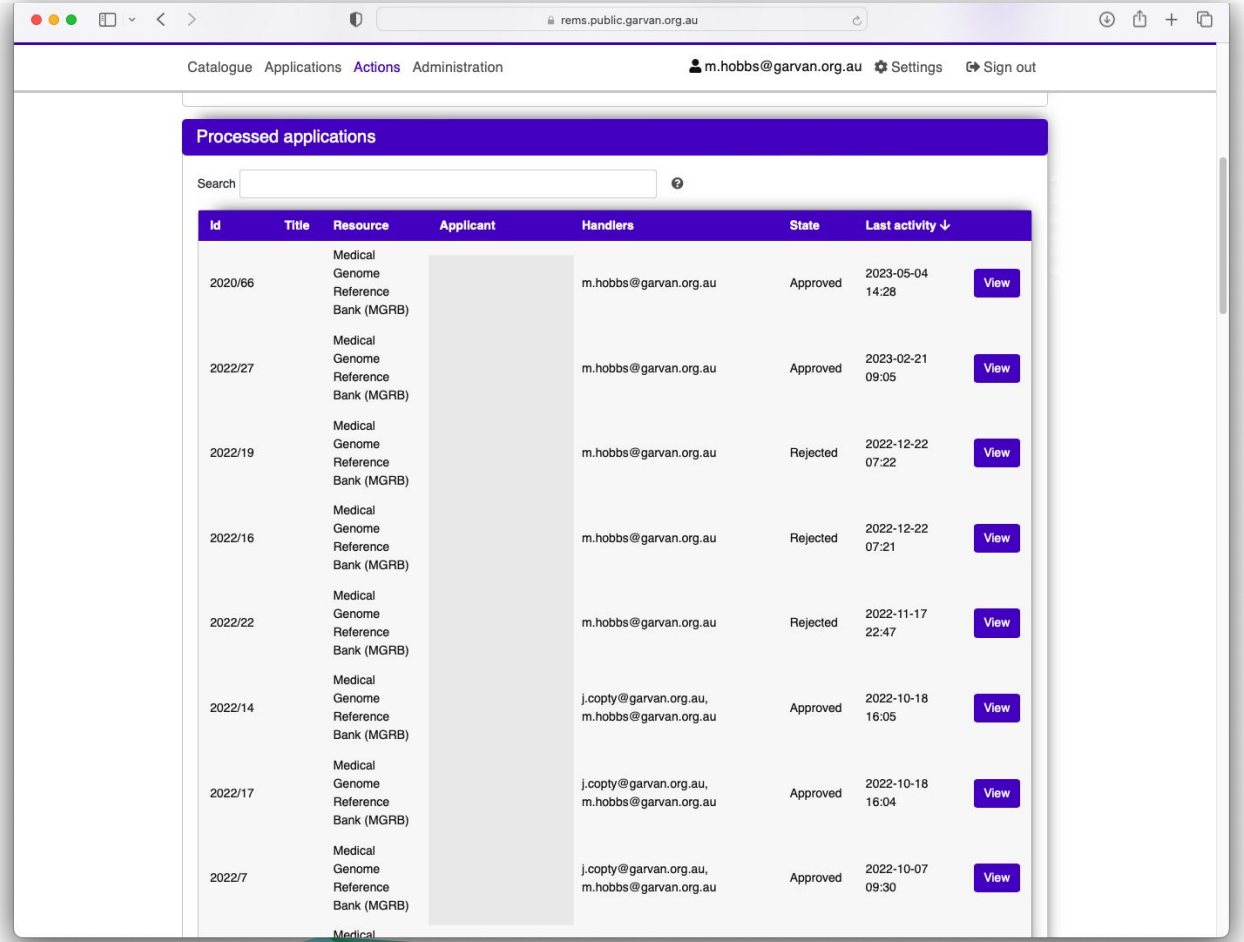
Australian BioCommons



Australian BioCommons



Australian BioCommons



Australian BioCommons

Opportunities

- Expanded reporting (in discussion with REMS team)
- After applications are approved: Data access automation
- REMS as a service

Acknowledgements

The **HGPP** receives investment from the **NCRIS**-enabled **ARDC** infrastructure under investment identifier https://doi.org/10.47486/PL032, as well as being funded through **BioPlatforms Australia**. Contributions are also made from each **partner organisation.**

Australian BioCommons

Thanks!

Any questions?

You can email the Australian BioCommons team at:
contact@biocommons.org.au

@AusBiocommons

AustralianBioCommonsChannel

zenodo

Australian BioCommons Community

Australian BioCommons

Questions?

Australian BioCommons

NEXT ...

International Congress of Genetics

16-21 July 2023, Melbourne Convention and Exhibition Centre

Upcoming training events

https://www.biocommons.org.au/webinars-workshops

Australian BioCommons

Thanks for joining us!

The Australian BioCommons is enabled by NCRIS via Bioplatforms Australia funding

Australian
BioCommons