eResearch Australasia
18 October 2018

The Bioplatforms Australia Data Portal

Adam Hunter[1], **Grahame Bowland[2]**, Samuel Chang[3], Tamas Szabo[4], **Kathryn Napier[5]**, Mabel Lum[6], Anna MacDonald[7], Jason Koval[8], Sophie Mazard[9], Anna Fitzgerald[10], Matthew Bellgard[11]

[1]*Centre for Comparative Genomics, Murdoch University , ahunter@ccg.murdoch.edu.au*
[2]*Centre for Comparative Genomics, Murdoch University, gbowland@ccg.murdoch.edu.au*
[3]*Centre for Comparative Genomics, Murdoch University, schang@ccg.murdoch.edu.au*
[4]*Centre for Comparative Genomics, Murdoch University, tszabo@ccg.murdoch.edu.au*
[5]*Centre for Comparative Genomics, Murdoch University, now: Curtin Institute for Computation, Curtin University, kathryn.napier@curtin.edu.au*
[6]*Bioplatforms Australia, Sydney, Australia, mlum@bioplatforms.com*
[7]*John Curtin School of Medical Research, The Australian National University, anna.macdonald@anu.edu.au*
[8]*Ramaciotti Centre for Genomics, University of New South Wales, j.koval@unsw.edu.au*
[9]*Bioplatforms Australia, Sydney, smazard@bioplatforms.com*
[10]*Bioplatforms Australia, Sydney, afitzgerald@bioplatforms.com*
[11]*Office of eResearch, Queensland University of Technology, matthew.bellgard@qut.edu.au*

Innovative life science research requires access to **state of the art infrastructure**.

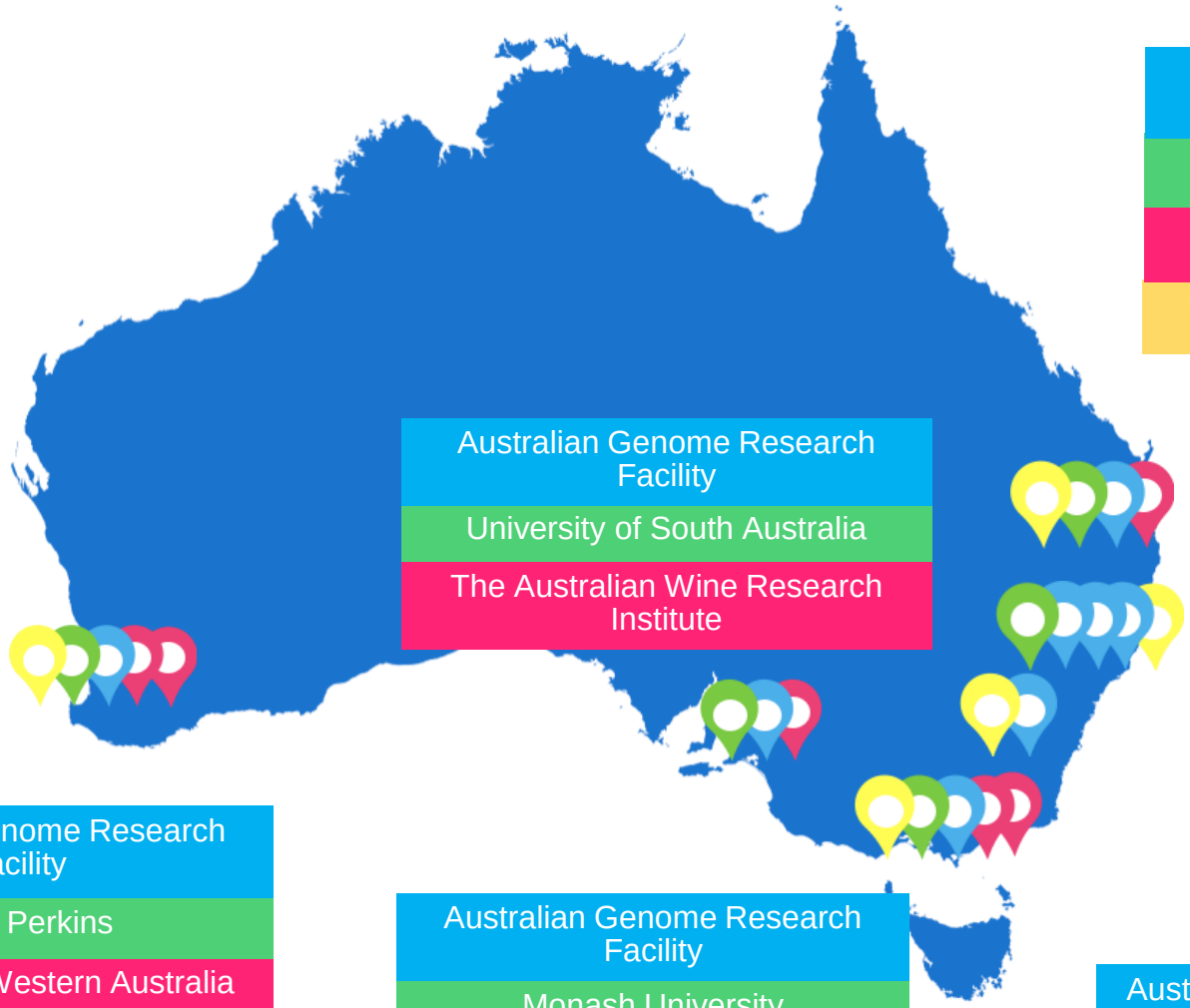
BIOPLATFORMS AUSTRALIA

CENTRE FOR COMPARATIVE GENOMICS
Western Australia

Bioplatforms Australia enables **innovation and collaboration in life science research** by investing in world class infrastructure and associated expertise in molecular platforms and informatics.

NCRIS
National Research Infrastructure for Australia
An Australian Government Initiative

Genomics | Proteomics | Metabolomics | Bioinformatics



Australian Genome Research Facility
University of Queensland
University of Queensland
University of Queensland

Australian Genome Research Facility
University of South Australia
The Australian Wine Research Institute

Ramaciotti Centre
Kinghorn Centre for Clinical Genomics
Australian Genome Research Facility
Australian Genome Research Facility
University of New South Wales

Australian Genome Research Facility
Harry Perkins
University of Western Australia
Murdoch University
Murdoch University

Australian Genome Research Facility
Monash University
University of Melbourne BioScience, Bio21
University of Melbourne Melbourne Bioinformatics

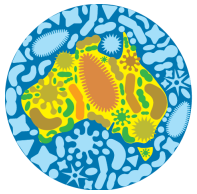
Australian National University
National Computational Infrastructure

Bioplatforms Australia invests in **collaborative open-data** projects

BIOPLATFORMS AUSTRALIA

CENTRE FOR COMPARATIVE GENOMICS Western Australia

Bioplatforms Australia invests in **collaborative open-data** projects



Australian Microbiome Database

The Oz Mammals Genomics Initiative

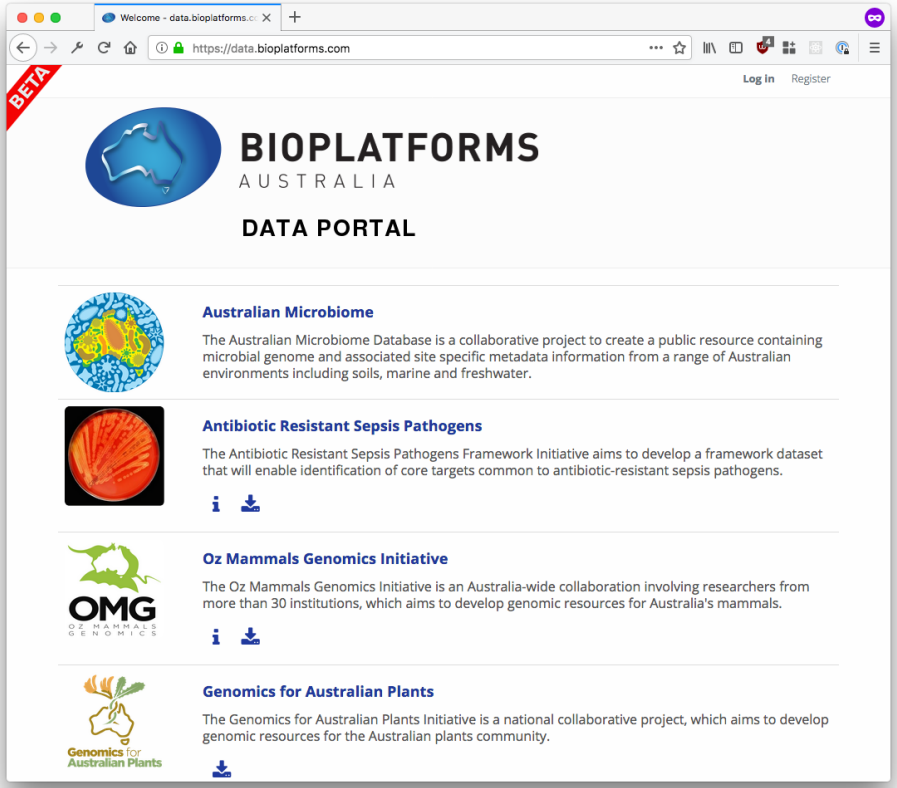
Antibiotic Resistant Sepsis Pathogens

Genomics for Australian Plants

Bioplatforms Australia Data Portal

A data archive repository housing:

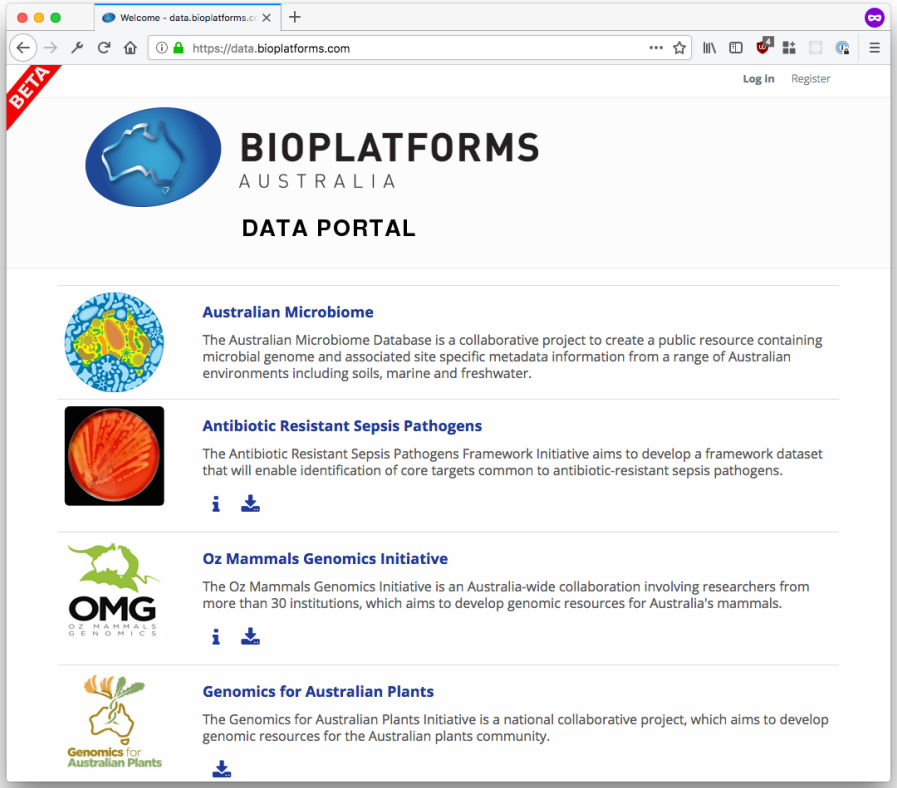
- Raw sequence data
- Analysed data
- Associated metadata



https://data.bioplatforms.com/

Bioplatforms Australia Data Portal

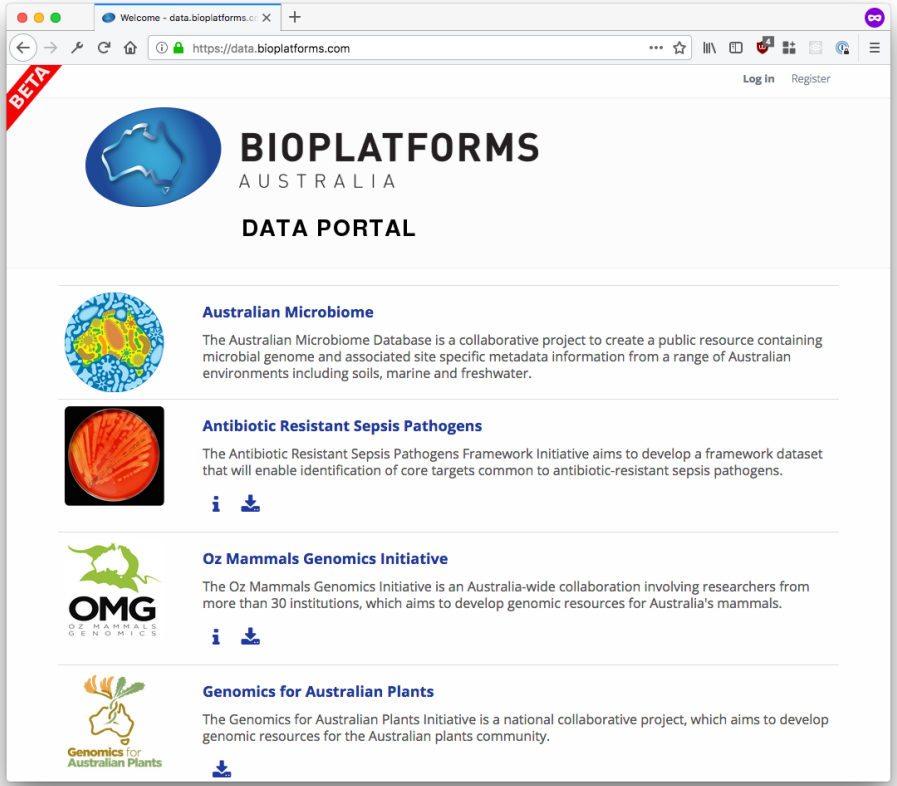
- Built upon open source technology (CKAN) - which also powers data.gov, data.gov.au, data.wa.gov.au, ….



https://data.bioplatforms.com/

Bioplatforms Australia Data Portal

- Hosted on the cloud (AWS)
- Data files are stored in Simple Storage Service (S3)
- Backup flat-file archive at QCIF



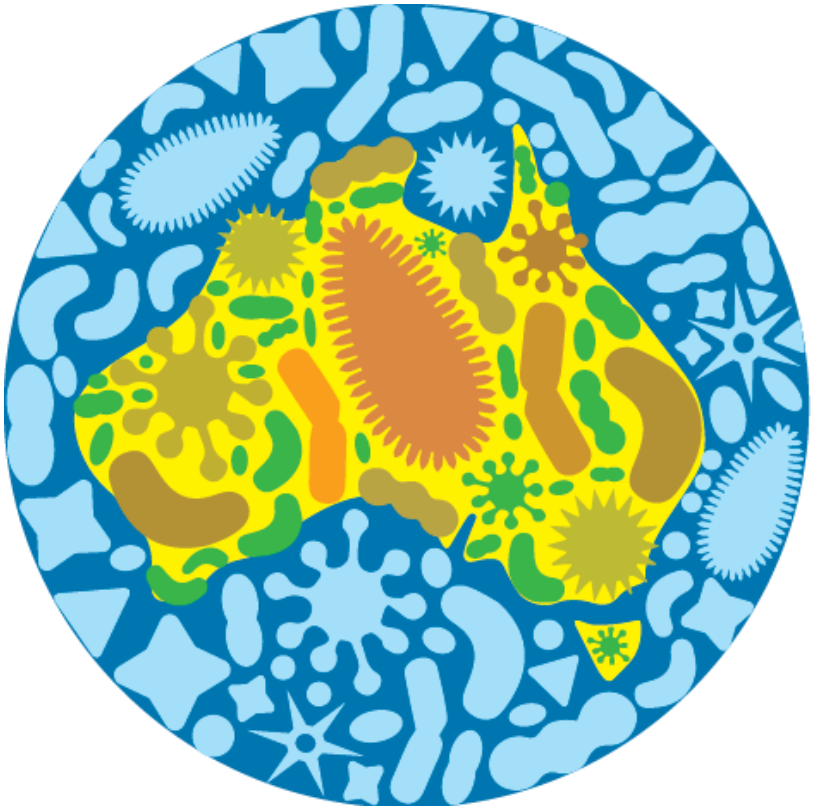
https://data.bioplatforms.com/

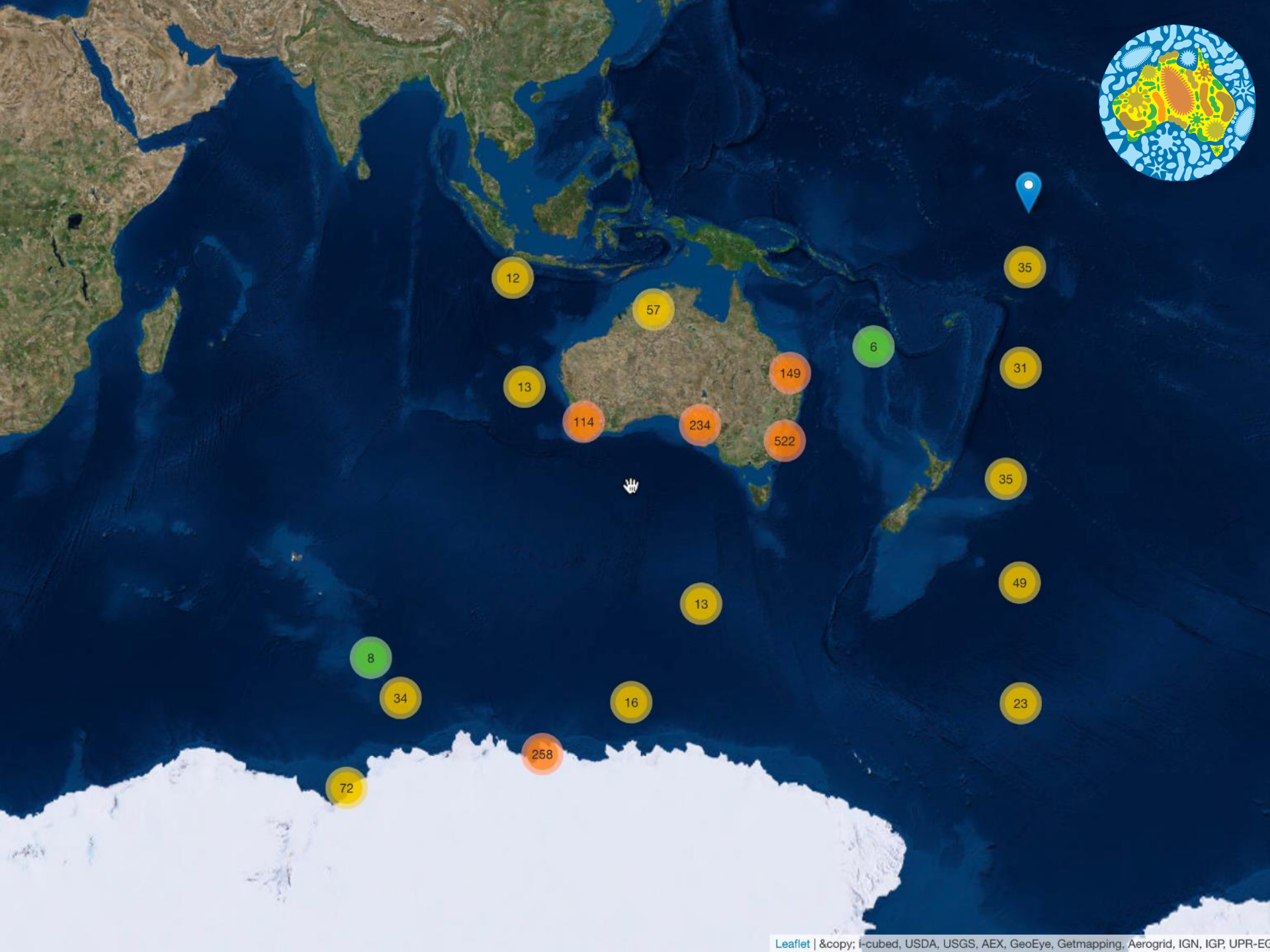
Data Portal metadata sources

- Sample contextual metadata
- Sequencing metadata (from sequencing providers)
- Taxonomic metadata
- Voyage metadata
- High terabytes of data, hundreds of thousands of raw files

Metadata sources become available over **differing timeframes** and have **different sources of truth**

Diving into a project
The Australian Microbiome




12
57
35
6
31
149
13
114
234
522
35
49
13
8
34
16
23
258
72
Leaflet | © i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EC

Data Portal metadata sources (revisited)

- Sample contextual metadata
- Sequencing metadata (from sequencing providers)
- Taxonomic metadata
- Voyage metadata
- Tens of terabytes of data, hundreds of thousands of raw files

Metadata sources become available over **differing timeframes** and have **different sources of truth**

Designed along FAIR Principles

Findable

ANDS-issued handles (DOIs being investigated)

Datasets (collections of files) have a stable ID based on these elements:

- the -omics (e.g. genomics, metabolomics, transcriptomics, …)
- the technology (hiseq, miseq, exon capture, …)
- the ANDS issued handled (102.100.100/X)

Designed along FAIR Principles

Accessible

- Access data using well documented, open-source CKAN API (REST, R/Python bindings available)
- Bulk download of large datasets, via a generated bash/Powershell script - ckanext-bulk
- Integrations with external systems (e.g. ARDC Research Data Cloud project - see Dr Jeff Christiansen's talk at 2pm)

Designed along FAIR Principles

Interoperable

- Software enforced data vocabularies (e.g. Australian Soil Classification, …)
- Industry standard file formats (FASTQ, ….)
- Samples, libraries derived from samples, issued identifiers and cross-linked

Designed along FAIR Principles

Reusable

- Raw data from sequencing facilities available for future processing
- Largely automated upload to international repositories (NCBI Sequence Read Archive [SRA])
- Analysed data available, with processing pipelines documented and versioned (processing run at BPA facilities / via research partners, e.g. CSIRO)

How we deliver this

Reproducible ingest

Python program “bpa-ingest” which consumes metadata and data sources, builds a target state for the data portal, and then works to assert that state on the portal by:

- Generating CKAN schemas
- Mutating metadata (via the CKAN API)
- Uploading data files (via the CKAN API)

User-accessible data is never manually uploaded.

How we deliver this

Reproducible ingest

“bpa-ingest” enforces metadata standards, including data vocabularies.

Can perform data and metadata integrity checks.

Issues are escalated to program managers.

The Data Portal has directly managed the ingestion of **tens of thousands of samples** constituting over **100 terabytes of data**.

Over **three hundred researchers** make use of the portal.

Bioplatforms Australia enables a broad scope of research endeavours through **investment in nationally collaborative programs** that fund the building of new datasets, ultimately offering them as a public resource.

Thank you

Murdoch UNIVERSITY

CENTRE FOR COMPARATIVE GENOMICS

BIOPLATFORMS AUSTRALIA

NCRIS National Research Infrastructure for Australia
An Australian Government Initiative