

NeSI

New Zealand eScience
Infrastructure



**genomics
aotearoa**

Taonga: building a data repository for genomics research in New Zealand

eResearch NZ 2021

Presenter: Jun Huh, NeSI



Background

History

- NeSI has entered a partnership with Genomics Aotearoa in 2018
- Beginning of the data repository
 - Early 2019: Dr Maren Wellenreuther's snapper research data needing storage- Aarnet Cloudstor to a temp repository at University of Otago
 - Late 2019: DOC's kākāpō data - migrating from AWS cloud
- Supported using Globus- data transfer platform developed by Ian Foster's team in the University of Chicago
- By early 2020, hosting 6 data sets of Taonga species (around 7TB)
 - 4 listed on an html page <https://www.genomics-aotearoa.org.nz/data>
 - Black rat, Kōura, Manuka, and Snapper



Kākāpō¹



Kōkako³



Snapper²



Mānuka⁴

Data Repository project

- Later in 2020, NeSI and GA entered a new contract implementing a prototype data repository to host GA researchers' genomic data for taonga species and capture richer metadata
- Drivers
 - Many of the GA projects are focused on early phases of genomics research pipeline, which involve sequencing genomes and generating huge raw files
 - More researchers would soon be needing a place to host their research data
 - Data sovereignty of taonga species
 - FAIR
- GA considers the genomics data repository as a great opportunity and **potential to become a national treasure**
- Could live beyond the scope of NeSI/GA

The Team

- Genomics Aotearoa data repository group
 - Mik Black - University of Otago
 - Ben Te Aika- GA Vision Mātauranga Coordinator
 - Rudi Brauning- AgResearch
 - Libby Liggins- Massey University
 - Miles Benton - ESR
 - Ben Curran - University of Auckland
- NeSI development team
 - Eiran Perkins
 - Jun Huh
 - Brian Flaherty

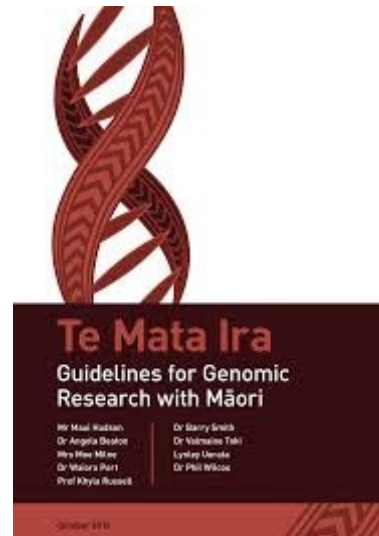




Implementation

What we did

- Guidelines - Te Mata Ira, FAIR/CARE principles
- Implementation using Gen3- open source genomic data repository solution developed by a team in University of Chicago together with NCI



Accessing data

Researchers accessing data



Filters

Project Method Data

[Collapse all](#)

▼ Instrume... 1 selected X Q

☒ Illumina HiSeq 2000 46

☐ Illumina HiSeq 2500 8

☐ no data 3

☐ AB SOLID 2 2

▼ Experimental Strategy Q

☐ RNA-Seq 40

☐ WGS 6

File Format

FASTQ

40
(87%)

FASTA

6
(13%)



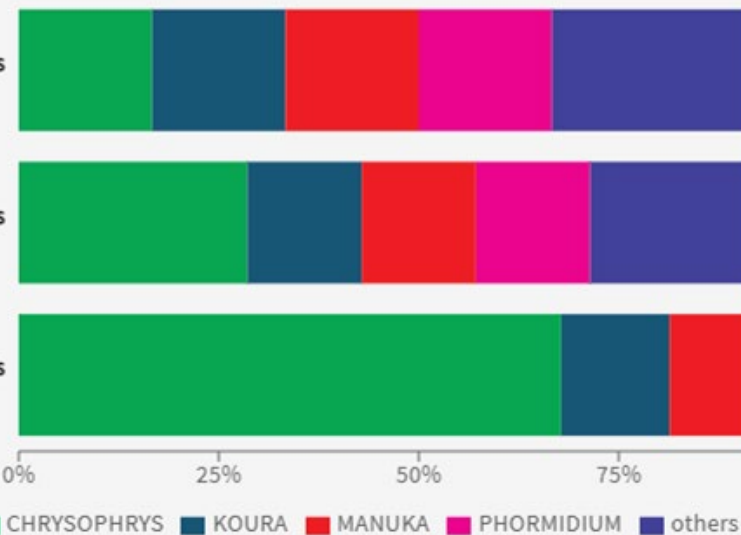
Showing 1 - 20 of 46 files

Project Id	Data Category	Data Format	File Name	File Size	Globus Url
TAONGA-CHRYSO	Sequencing Data	FASTQ	FCH7FYVBBXX-HKFISwsyEAADRAAPEI-204_L3_1.fq.gz	523.71 MB	https://trans origin_id=a6 02fcc9cdd752
TAONGA-CHRYSO	Sequencing Data	FASTQ	FCH7FY5BBXX-HKFISwsyEAAVRAAPEI-225_L3_1.fq.gz	534.64 MB	https://trans origin_id=a6 02fcc9cdd752

6 Projects

7 Experiments

59 Files



TAONGA-CHRYSOPHRYS [browse nodes](#)

Project name: Chrysophrys

Description: RNA-seq data for domesticated and wild type snapper (*Chrysophrys auratus*) individuals

Publications:

Wellenreuther M, Le Luyer J, Cook D, Ritchie PA, Bernatchez L

'Domestication and temperature modulate gene expression signatures and growth in the Australasian G3: Genes, Genomes, Genetics, January 2019, 9 (1), 105-116

<https://doi.org/10.1534/g3.118.200647>

Catanach A, Crowhurst R, Deng C, David C, Bernatchez L, Wellenreuther M

'The genomic pool of standing structural variation outnumbers single nucleotide polymorphism by the Molecular Ecology, 2019, 28 (6)

<https://doi.org/10.1111/mec.15051>

Number of files: 40

PI: Dr. Maren Wellenreuther

Access request form for Genomics Aotearoa data repository


* Required



Email address *



Please select the data set you are requesting access to *


Next



Page 1 of 4



File Manager







Panels  







Collection  



Path 


select all  


view 

NAME 	LAST MODIFIED	SIZE
For more information about this endpoint, please click here! 		
 Chrysophrys_auratus_scaffolds.assemblathon_sta...	11/15/2019 11:27am	8 KB
 Chrysophrys_auratus_scaffolds.fsa	11/15/2019 11:27am	745.87 MB
 md5sum.txt	11/15/2019 11:27am	204 B
 README.md	11/15/2019 11:27am	10.66 KB



select all  

view

Search for a collection to begin
 Get started by taking a short tour.

GEN3
DATA COMMONS

Dictionary

Exploration

Query

Workspace

Profile


Quit


Files


Running

Clusters

Select items to perform actions on them.

☐ 0 

 / pd

Name 


☐ ..


☐ data

☐ dockerHome

☐ lost+found

Upload

New 



Notebook:

Python 3

R

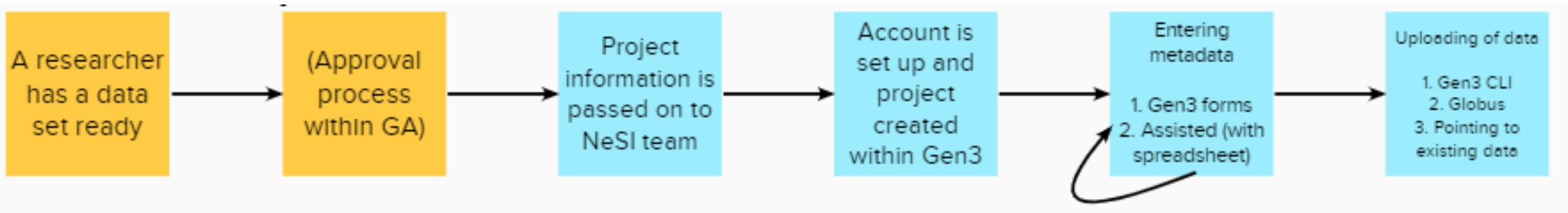
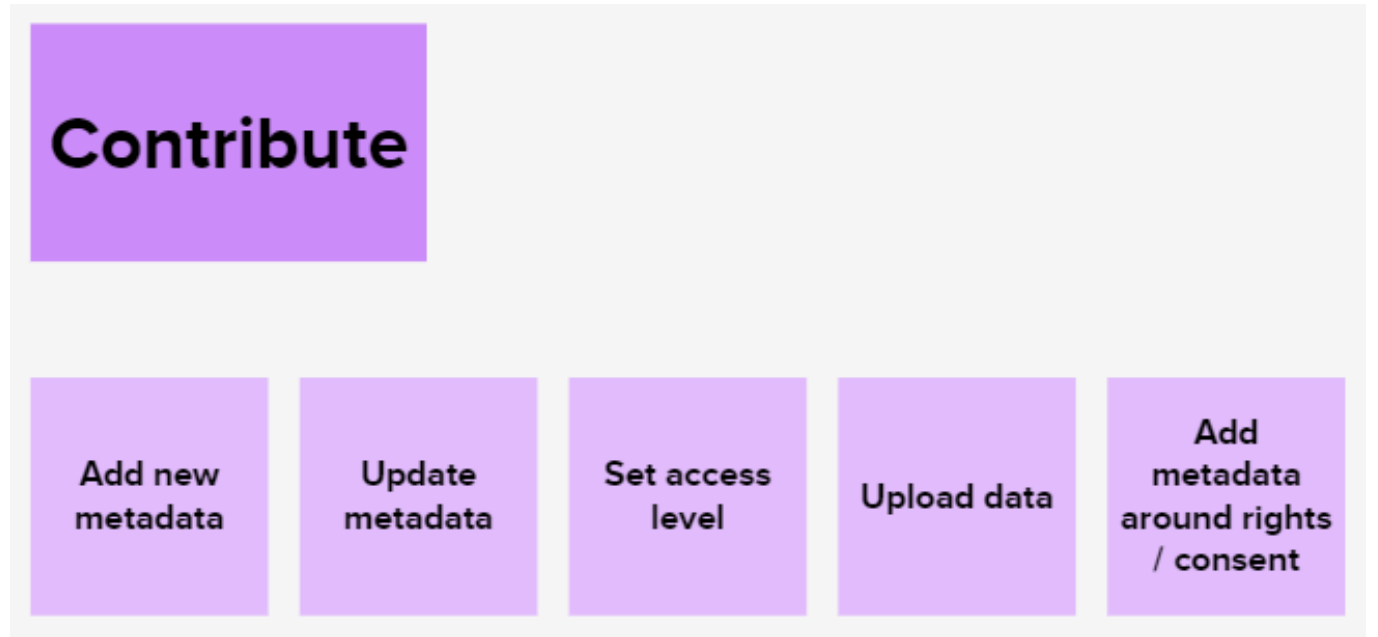
Other:

Text File

Folder

Terminal

Contributing data



Māori governance questions for depositing data

- Species name
- Sample area name
- Sample site GPS
- Māori written consent received - Y/N
- Māori Kaitiaki group name
- Māori Kaitiaki representative name
- Māori Kaitiaki email
- Māori Kaitiaki phone number
- Research institution
- Research institution officer name
- Research Institution officer contact email
- File size

Technology choices

- Gen3 - open source genomic data repository solution
 - Developed by a team in University of Chicago together with NCI
 - 20+ implementations - NCI, The Anvil, Chicagoland Covid19
 - Inheriting a lot of valuable domain specific knowledge from work done by UoC and NCI
 - Open source / active community support
 - Can support federated approach- e.g. pointing to data on NCBI
- Globus ⇒ Object storage
- Workflow support
 - Cloud solutions for application forms for now - Google Forms, Zendesk
 - Policy evaluation of form data ownership was done

Data ownership and policy evaluation

	Zendesk	Google Forms	Survey Monkey	Typeform	Locally Hosted / NeSI
Privacy Policy	https://www.zendesk.com/company/customers-partners/privacy-policy/	https://policies.google.com/privacy	https://www.surveymonkey.com/mp/legal/privacy-policy/	https://admin.typeform.com/to/dwk6gt	https://www.nesi.org.nz/privacy
Information Collected	Cookies, Logs, Info from 3rd party providers, Account & registration info, device	Account info; apps, browsers, devices, activity, inc. views & interaction with content, location information.	Usage, device/browser, log data, referral info, Cookies, Contact Information (from inquiries)	Browser profiling, cookies, sign-up (contact) information	
Security	Compliance with high security standards, such as encryption of data in motion over public networks, auditing standards (SOC 2, ISO 27001, ISO 27018), Distributed Denial of Service ("DDoS") mitigations, and a Support team that is on-call 24/7. Zendesk servers are	Certifications: ISO/IEC 27001 for the systems, technology, processes, and data centers; ISO/IEC 27017 information security controls for cloud services, SO/IEC 27018:2014 international privacy and data protection standards;	Encryption to keep data private while in transit; Safe Browsing, Security Checkup, and 2 Step Verification; SOC 2 accredited data centers; ISO/IEC 27001, IS 705333	Compliant with security and privacy standards, including Privacy Shield. data in-transit (end-to-end, including within the virtual private cloud at AWS) is encrypted using secure TLS cryptographic protocols (TLS 1.2)	
Service Data Ownership	The customer retains ownership of and control over Service Data in its account.	The customer retains ownership of and control over Service Data in its account.	The customer retains ownership of and control over Service Data in its account.	The customer retains ownership of and control over Service Data in its account.	
HIPAA Compliance	Advanced Compliance: This add-on helps customers fulfill obligations under the Health Insurance Portability and Accountability Act (HIPAA). With this add-on, customers have the ability to enter into a Business Associate Agreement (BAA) with Zendesk. The Advanced Compliance add-on is available for Enterprise plans.	Gmail, Google Drive and Docs all support HIPAA-compliant access to patient information. customers who are subject to HIPAA and wish to use G Suite or Cloud Identity with PHI (Protected Health Information) must sign a Business Associate Agreement (BAA) with Google.	HIPAA-compliant" means that we offer a service that enables covered entities to collect and manage PHI through surveys in a manner compliant with HIPAA. SurveyMonkey only permits PHI to be collected by regulated entities if it is done through a "HIPAA-enabled account" with a business	Typeform is currently not HIPAA compliant.	



Learning and Challenges

Understanding the needs

- 3 workshops
 - Indigenous governance
 - Metadata dictionary
 - End user workshop
- Weekly review with the working group
- Case studies- in progress
 - Learning about metadata requirements on different types of research

Workshop outcomes 1

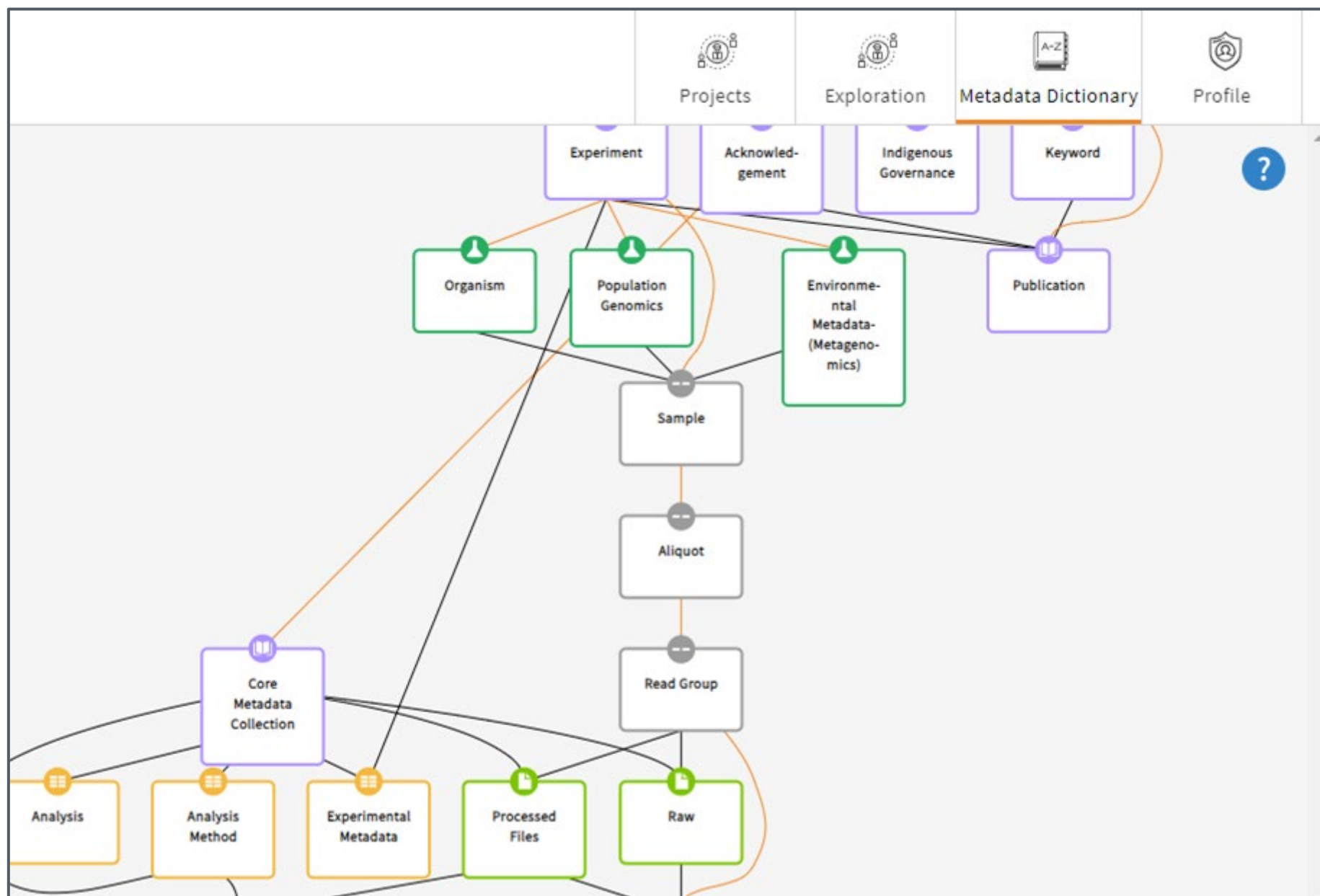
Purpose of the repository

- Taonga species as a starting point
- Mix of open/restricted data
 - Public data can go to NCBI (National Center for Biotechnology Information)
- Non-indigenous & non-public data (commercial data for example) → could be hosted and granted access to non-commercial research
- Mix of active and archival data

Workshop outcomes 2

Metadata dictionary

- Initially using Gen3's default dictionary - built by Gen3 team with genomics researchers at NCI
 - Focus on human health - cancer data
- Needed to capture non-human biospecimen (organism, metagenome, and population)
 - Shaped by existing data sets, NCBI templates, case studies, inputs from workshops - Libby Liggins's examples from Ira moana/GEOME project, Kim Handley's metagenome sample data
- Instruments and methods metadata have been kept
- Added a dictionary definition to capture Māori governance information



Workshop outcomes 3

Māori governance

- Support multiple levels of permissions
- Capturing metadata around consent
 - context of the consent, such as expiry date
- TK Labels
- Māori data management plans - pre-repository phase
- Visibility of every stage of the process
 - Reporting
 - Governance process dictated by kāhui - based on Ben Te Aika's recommendations

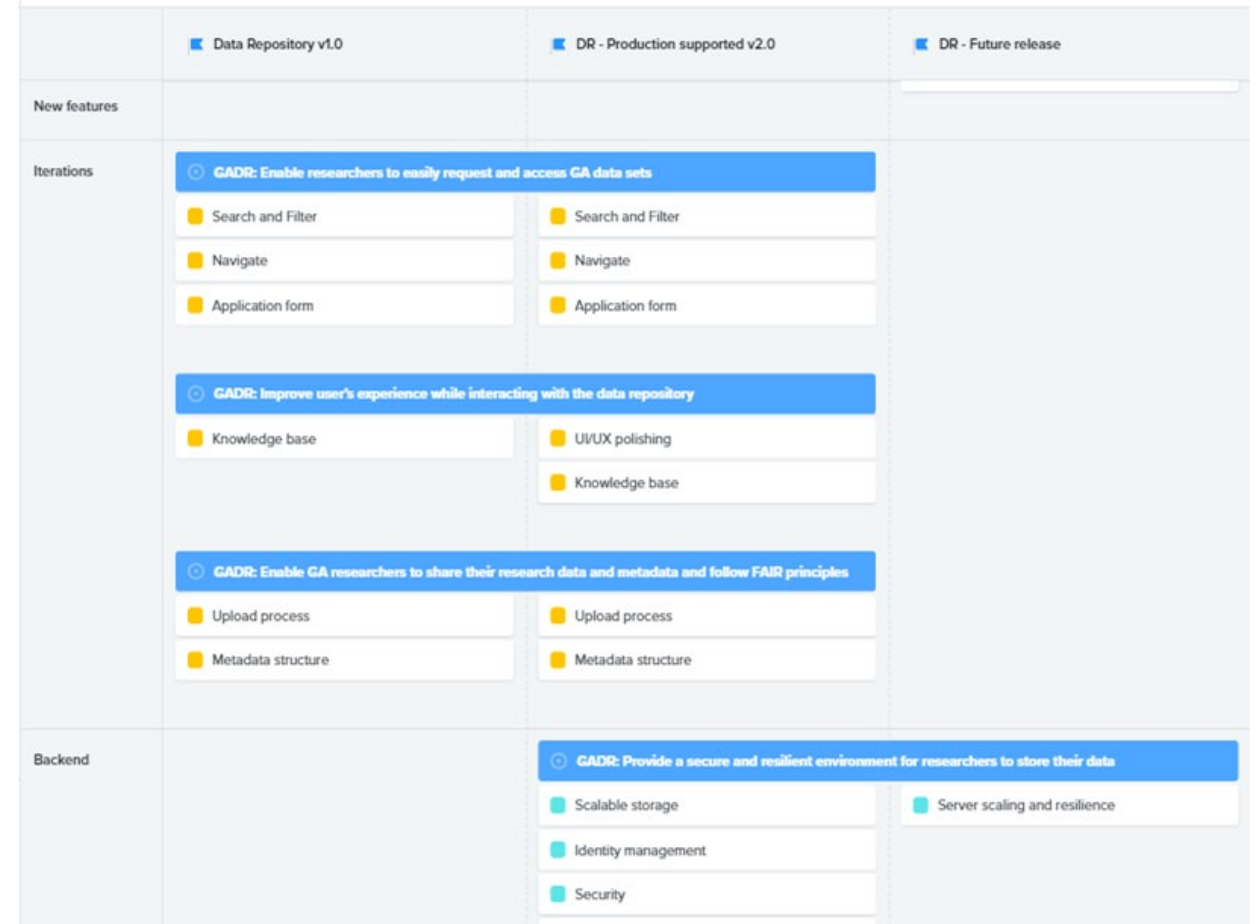
	Metadata	Data
Public visibility		
Logged-in users		
Approved access	Kākāpō	



What's next

Development roadmap

- Timeframe with NeSI's new infrastructure and NeSI IdM project
 - Flexibility with storage, activeness of data, and resilience
- Going to production
- Continued iterations and polishing



Bigger picture

- Genomics Aotearoa
 - Time to reach out to a wider audience
 - NeSI
 - Considering researcher data lifecycle
 - Te Ao Māori in eResearch BoF
-
- Genomics Aotearoa Data Repository BETA is at:
 - <https://repo.data.nesi.org.nz/>
 - Feedback portal
<https://portal.productboard.com/grfyksgghs3jpxnlu5rexvp2s>



Thank you