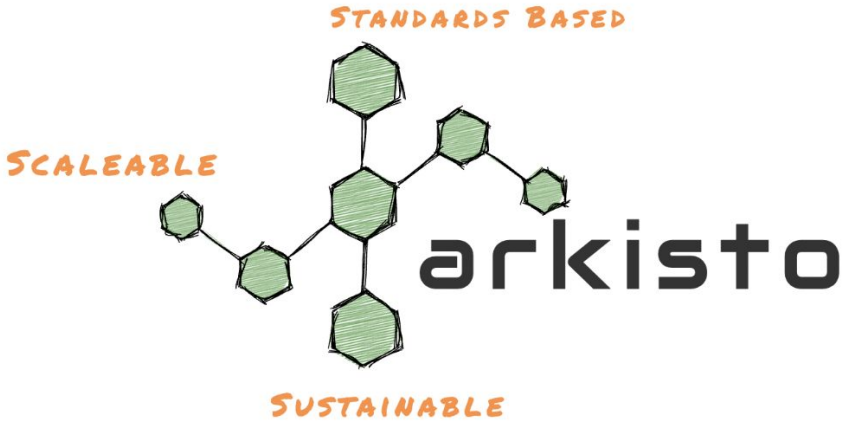
HASS RDC Technical Advisory Group Meeting

ATAP Architecture Discussion

Intro

Moises Sacal Bonequi - m.sacalbonequi@uq.edu.au
Ben Foley - b.foley@uq.edu.au
Peter Sefton - p.sefton@uq.edu.au



A scaleable, standards based platform for sustainable data.

The basis of Arkisto is that the long-term preservability of well-described data is *always* the first consideration.

Data on an Arkisto deployment is alway available on disc (or object storage) with a complete description *independently* of any services such as websites or APIs. Once the data is safe and well described, Arkisto has a flexible model for how data can be accessed using a variety of services.

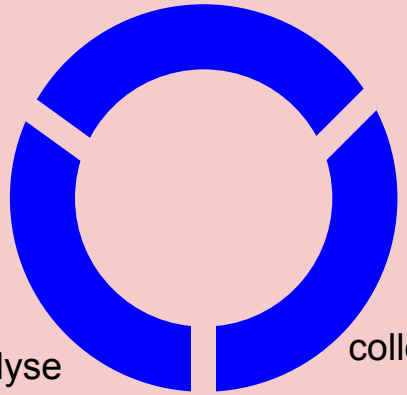
Arkisto is built on top of Research Object Crate (RO-Crate) and the Oxford Common File System Layout (OCFL).

With Arkisto there is no messy data migration.

Research Data Management Plan

Workspaces:

- working storage
- domain specific tools
- domain specific services



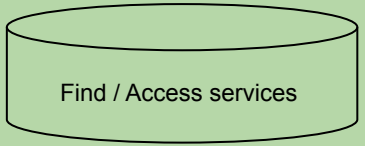
Active cleanup processes
workspaces considered ephemeral

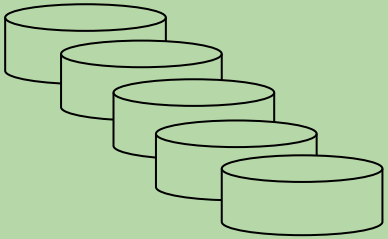
Reusable, Interoperable
data objects

➔ deposit early
➔ deposit often

reuse data objects

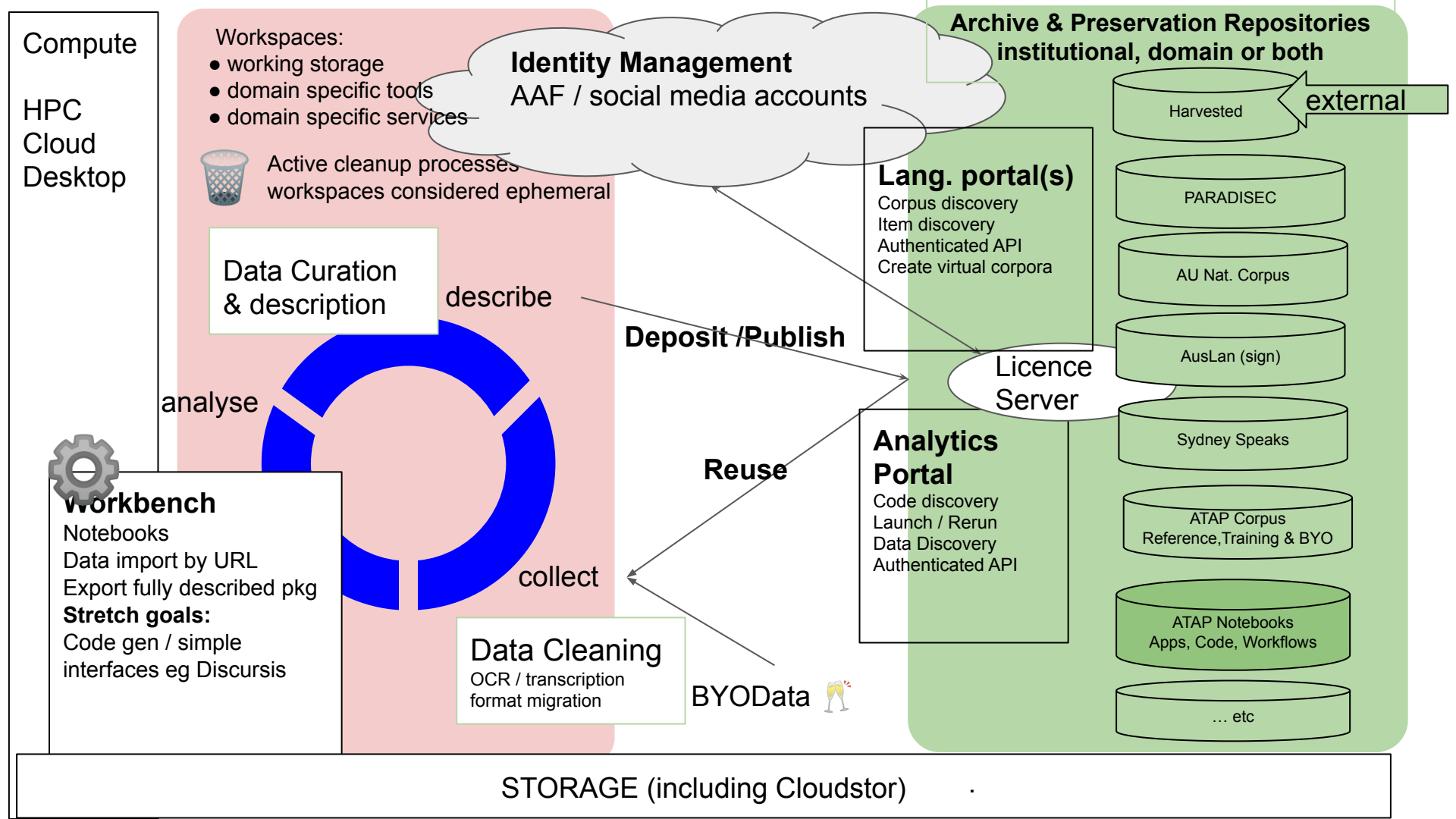
Repositories: institutional, domain or both

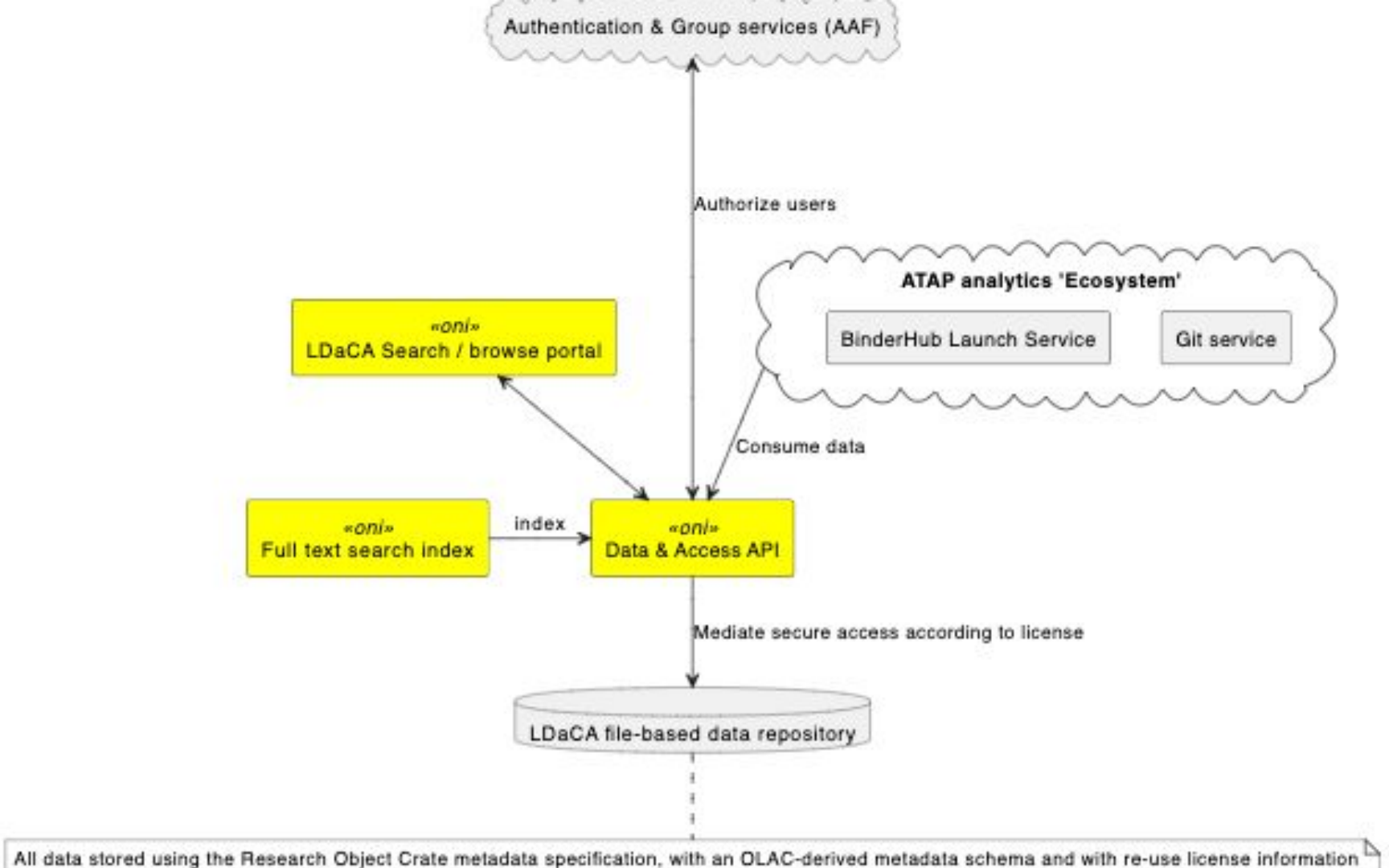


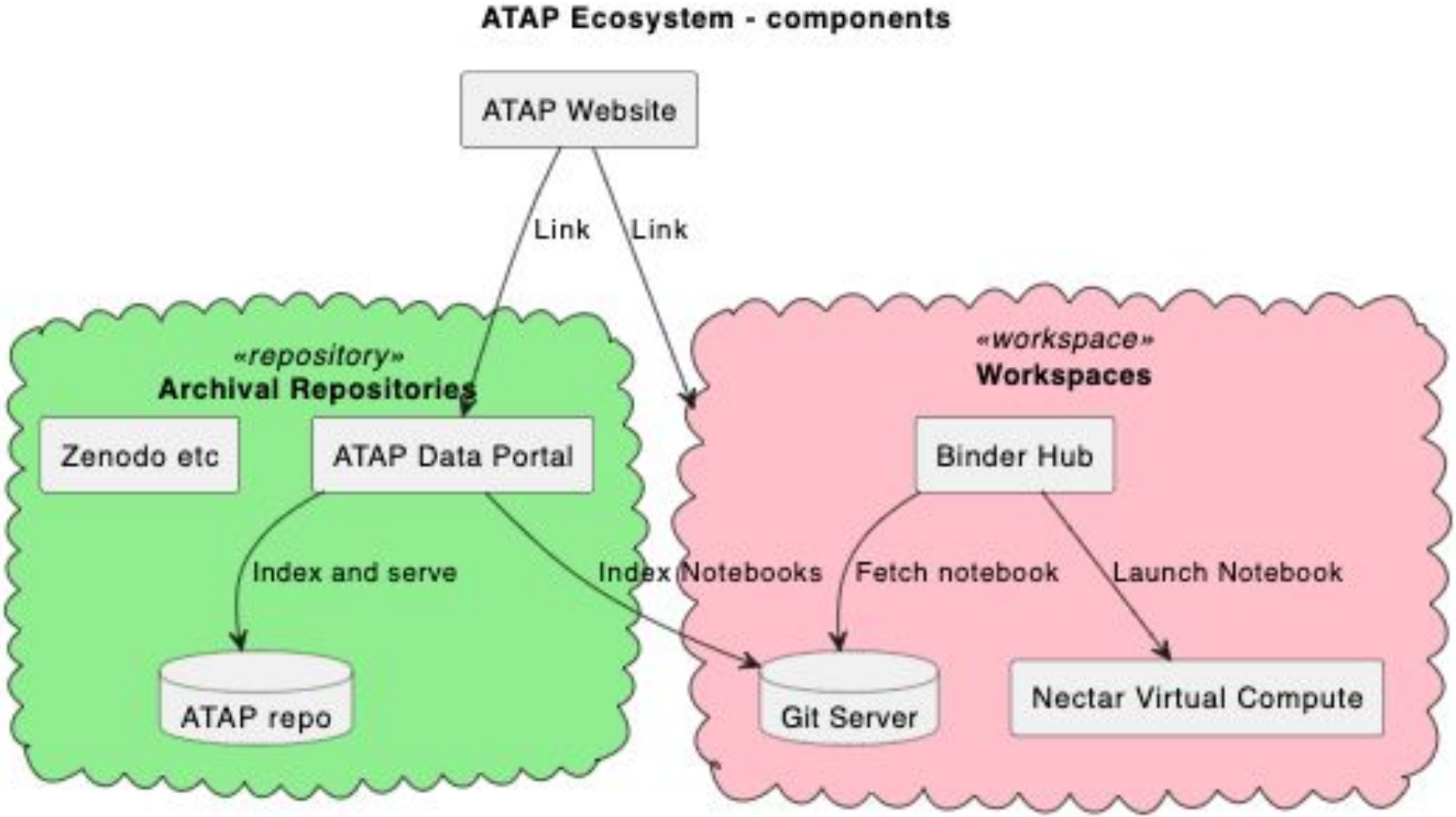


Findable, Accessible, Reusable
data objects

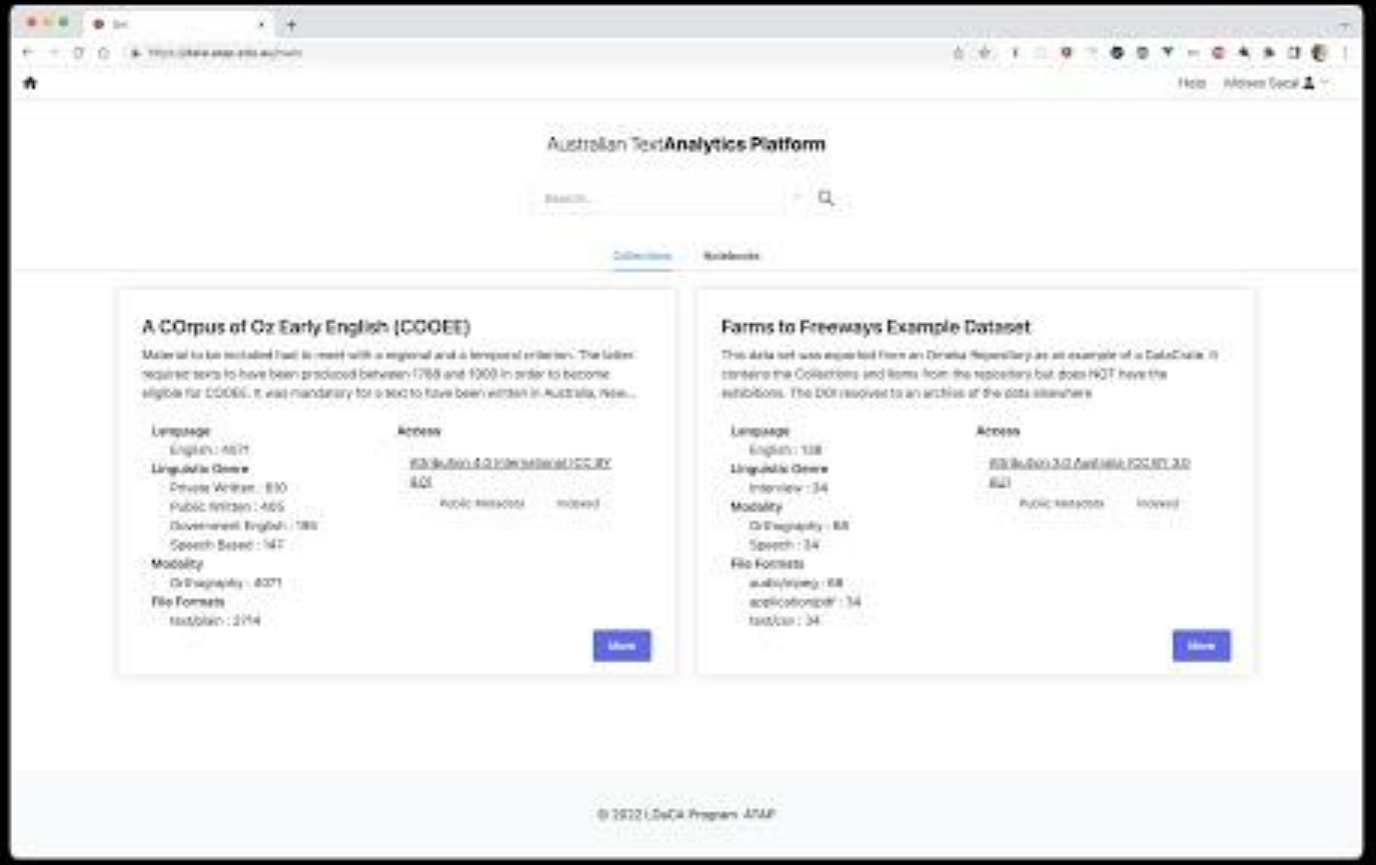
Policy based data management


Compute
HPC
Cloud
Desktop
Workspaces:
• working storage
• domain specific tools
• domain specific services
Active cleanup processes
workspaces considered ephemeral
Identity Management
AAF / social media accounts
Archive & Preservation Repositories
institutional, domain or both
Harvested
external
Lang. portal(s)
Corpus discovery
Item discovery
Authenticated API
Create virtual corpora
PARADISEC
AU Nat. Corpus
AusLan (sign)
Sydney Speaks
ATAP Corpus
Reference,Training & BYO
ATAP Notebooks
Apps, Code, Workflows
… etc
Data Curation
& description
describe
Deposit /Publish
Licence
Server
analyse
Reuse
collect
Workbench
Notebooks
Data import by URL
Export fully described pkg
Stretch goals:
Code gen / simple
interfaces eg Discursis
Analytics
Portal
Code discovery
Launch / Rerun
Data Discovery
Authenticated API
Data Cleaning
OCR / transcription
format migration
BYOData
STORAGE (including Cloudstor)


Authentication & Group services (AAF)
Authorize users
ATAP analytics 'Ecosystem'
BinderHub Launch Service
Git service
«oni»
LDaCA Search / browse portal
Consume data
«oni»
Full text search index
index
«oni»
Data & Access API
Mediate secure access according to license
LDaCA file-based data repository
All data stored using the Research Object Crate metadata specification, with an OLAC-derived metadata schema and with re-use license information


ATAP Ecosystem - components
ATAP Website
Link
Link
«repository»
Archival Repositories
Zenodo etc
ATAP Data Portal
Index and serve
ATAP repo
«workspace»
Workspaces
Binder Hub
Index Notebooks
Fetch notebook
Launch Notebook
Git Server
Nectar Virtual Compute

Demo


Australian TextAnalytics Platform
A COrpus of Oz Early English (COOEE)
Farms to Freeways Example Dataset