LDaCA Technical Architecture update 2025

PT Sefton, Moises Sacal Bonequi, Ben Foley

What’s in this presentation?

- News
- Refresh memories of the distinction between Workspaces vs Archival Repositories
- Explore the architecture of our work on Archival Repositories
  - Decentralised approach: multiple Data Stores under appropriate governance
  - Standards and specifications:
    - RO-Crate for describing data objects
    - RO-Crate Metadata Profiles for data interchange within a discipline or domain (like language data)
  - Open source tools

News!

- John Ferlito (PARADISEC) has created a new version of the LDaCA portal using a simpler API that can be used for PARADISEC and LDaCA (and potentially Nyingarn and many other repositories)
- New API is “An RO-Crate API” - AROCAPI
  - Generic API for collections of Objects/Items
  - Objects are described using RO-Crates
- Working together on a new Oni-stack using the new API
- New stack can be used for RAPID and other data portals

More news

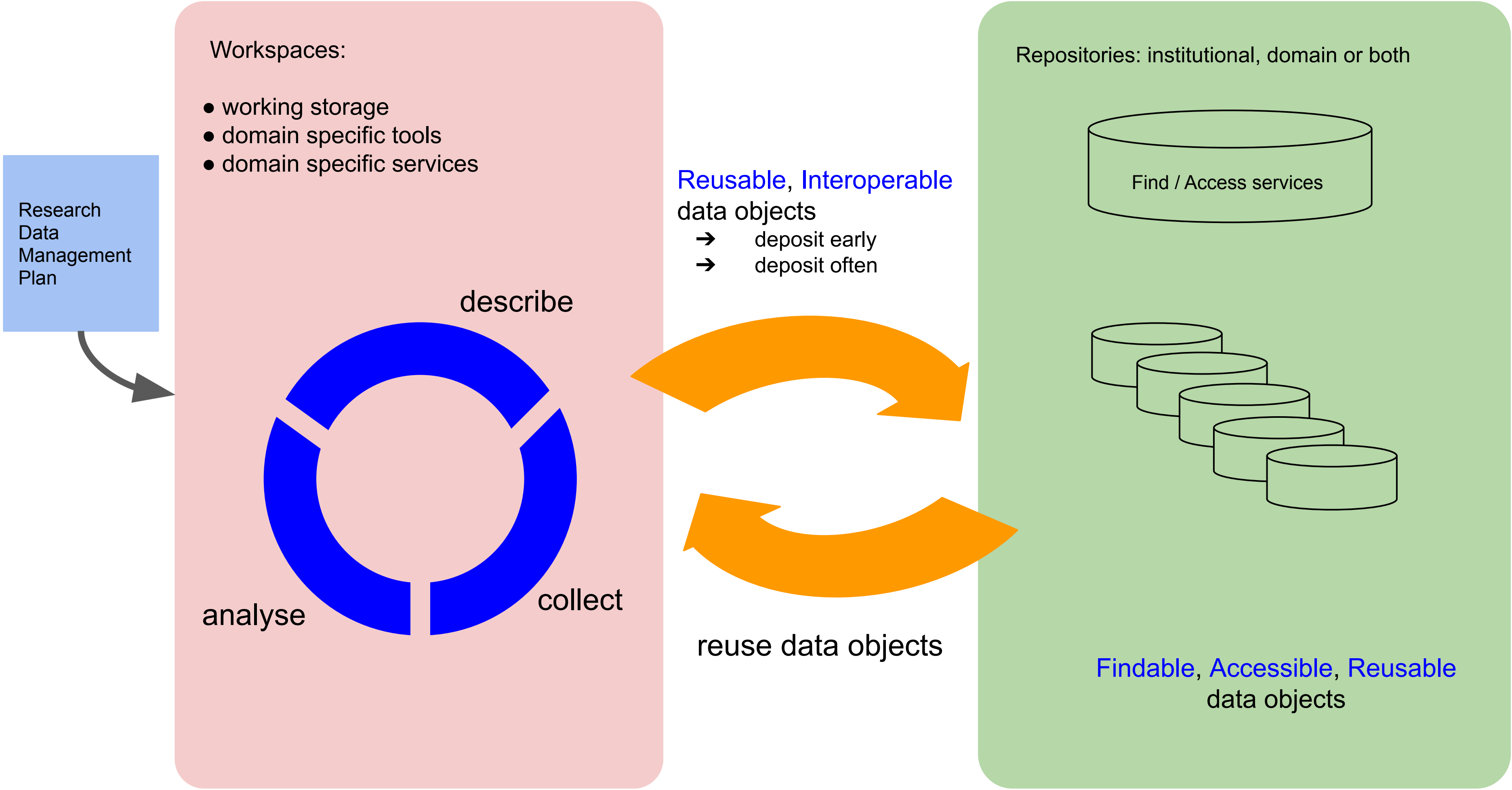
Version: 2025-07-31

LDaCA Execution Strategy Overview

	Starting state (2021)	Desired state (2028)
collect & organise	Language data is rarely organised or described in reusable ways, if it's described at all	Standards and tools are available and being applied by data stewards
conserve	A lot of language data is at risk of being lost forever	Good governance and standardised, distributed storage of data helps preserve and return data
find	It's difficult to know what language data exists and where to find it	Discovering and locating language data is easy via linked portals
access	Processes for granting permissions and getting access to data are either absent or aren't easy to understand or apply	Access controls are in place and easy to use, so that data access can be given to the right people
analyse > analysis overview	Ad hoc tools, analysis and annotation methods are used, lacking reproducibility	Shared tools can process, analyse, reuse, repurpose, annotate, visualise and enhance data at scale
guide	Guidance and training for collecting, handling, using and analysing data are scattered and hard to find	Best practice advice and training for working with language data is available from a single source which is easy to find

Activities

- Strengthen the data management skills of language worker communities
- Develop shared tools, standards and technical infrastructure to help data stewards care for data for the long term
- Build data portals with useful search functions and lightweight technical structures
- Create guidance for data stewards to document and grant access and reuse rights
- Support language communities to gain greater control over their language data
- Develop tools for data and metadata conversion, processing, analysis, annotation, visualisation, and enrichment
- Develop and guide the implementation of local and national policy and governance toolkits
- Provide examples and training for research at scale



Research Data Management Plan

Workspaces:

- working storage
- domain specific tools
- domain specific services

describe

collect

analyse

Reusable, Interoperable data objects

➔ deposit early

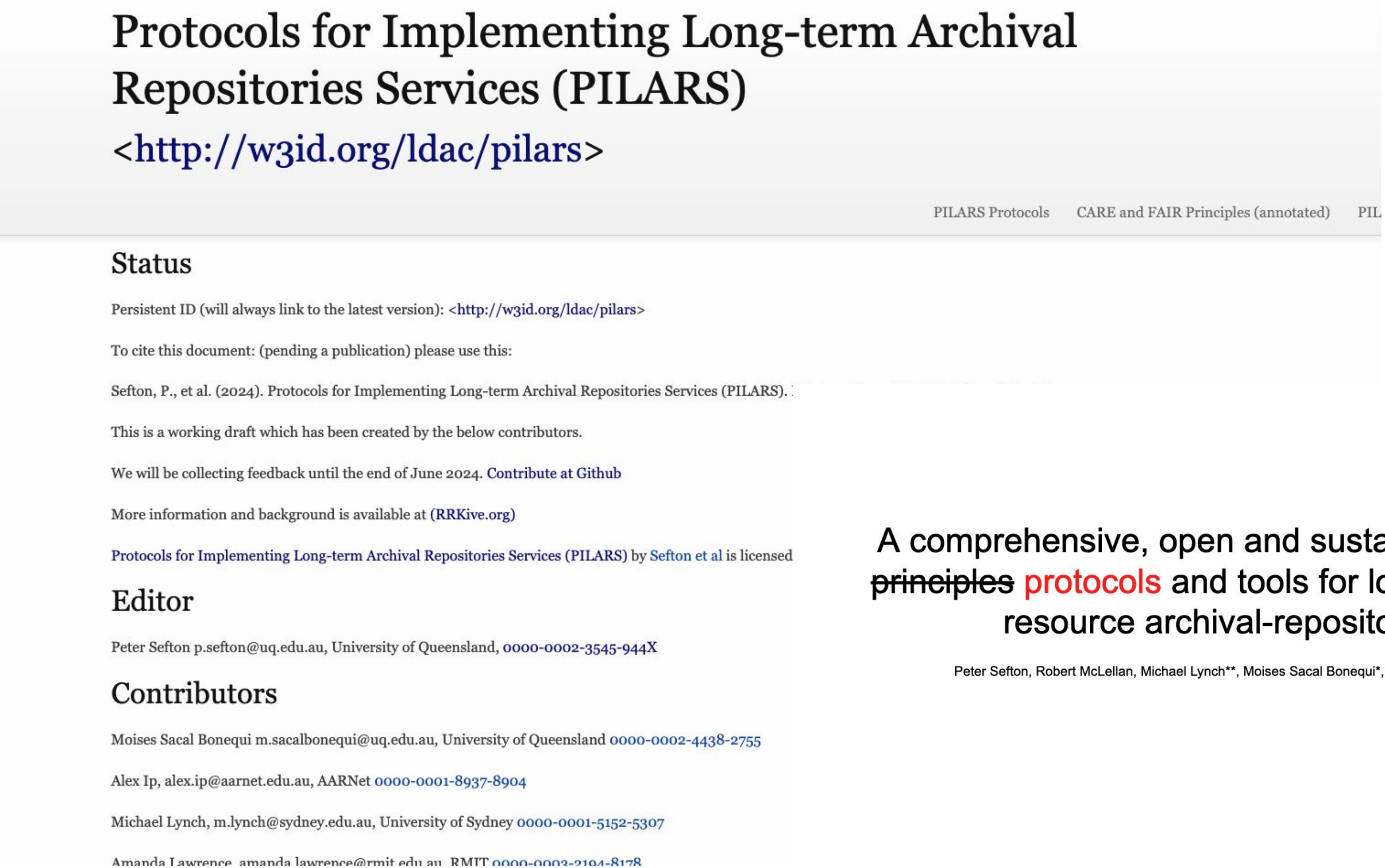
➔ deposit often

reuse data objects

Repositories: institutional, domain or both

Find / Access services

Findable, Accessible, Reusable data objects



Protocols for Implementing Long-term Archival Repositories Services (PILARS)

<http://w3id.org/ldac/pilars>

Status

Persistent ID (will always link to the latest version): <http://w3id.org/ldac/pilars>

To cite this document: (pending a publication) please use this:

Sefton, P., et al. (2024). Protocols for Implementing Long-term Archival Repositories Services (PILARS).

This is a working draft which has been created by the below contributors.

We will be collecting feedback until the end of June 2024. Contribute at Github

More information and background is available at (RRKive.org)

Protocols for Implementing Long-term Archival Repositories Services (PILARS) by Sefton et al is licensed

Editor

Peter Sefton p.sefton@uq.edu.au, University of Queensland, 0000-0002-3545-944X

Contributors

Moises Sacal Bonequi m.sacalbonequi@uq.edu.au, University of Queensland 0000-0002-4438-2755

Alex Ip, alex.ip@aarnet.edu.au, AARNet 0000-0001-8937-8904

Michael Lynch, m.lynch@sydney.edu.au, University of Sydney 0000-0001-5152-5307

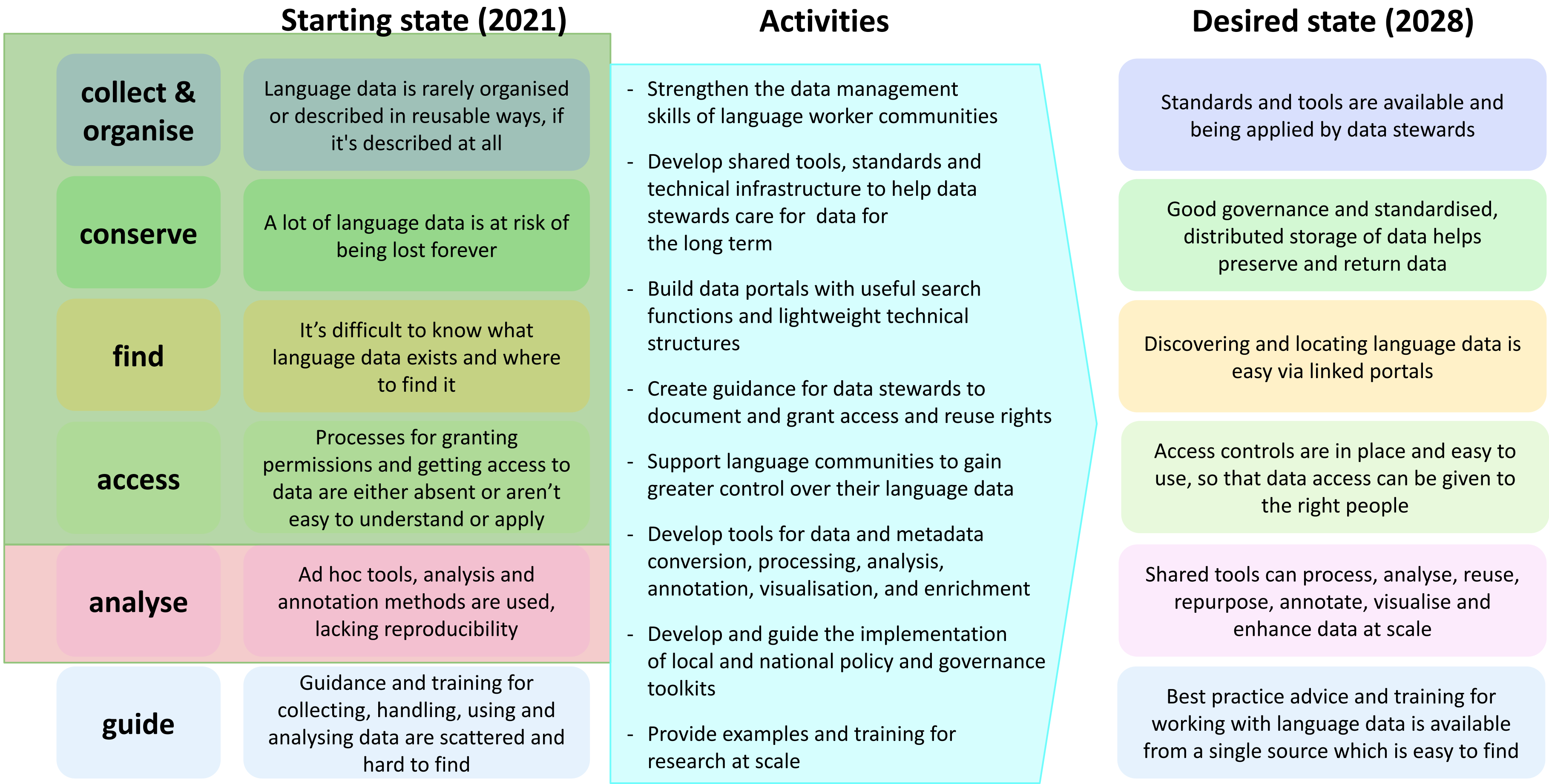
Amanda Lawrence, amanda.lawrence@rmit.edu.au, RMIT 0000-0003-2194-8178

A comprehensive, open and sustainable set of ~~principles~~ protocols and tools for low (and high) resource archival-repositories

Peter Sefton, Robert McLellan, Michael Lynch**, Moises Sacal Bonequi*, Nick Thieberger***

Version: 2025-07-31

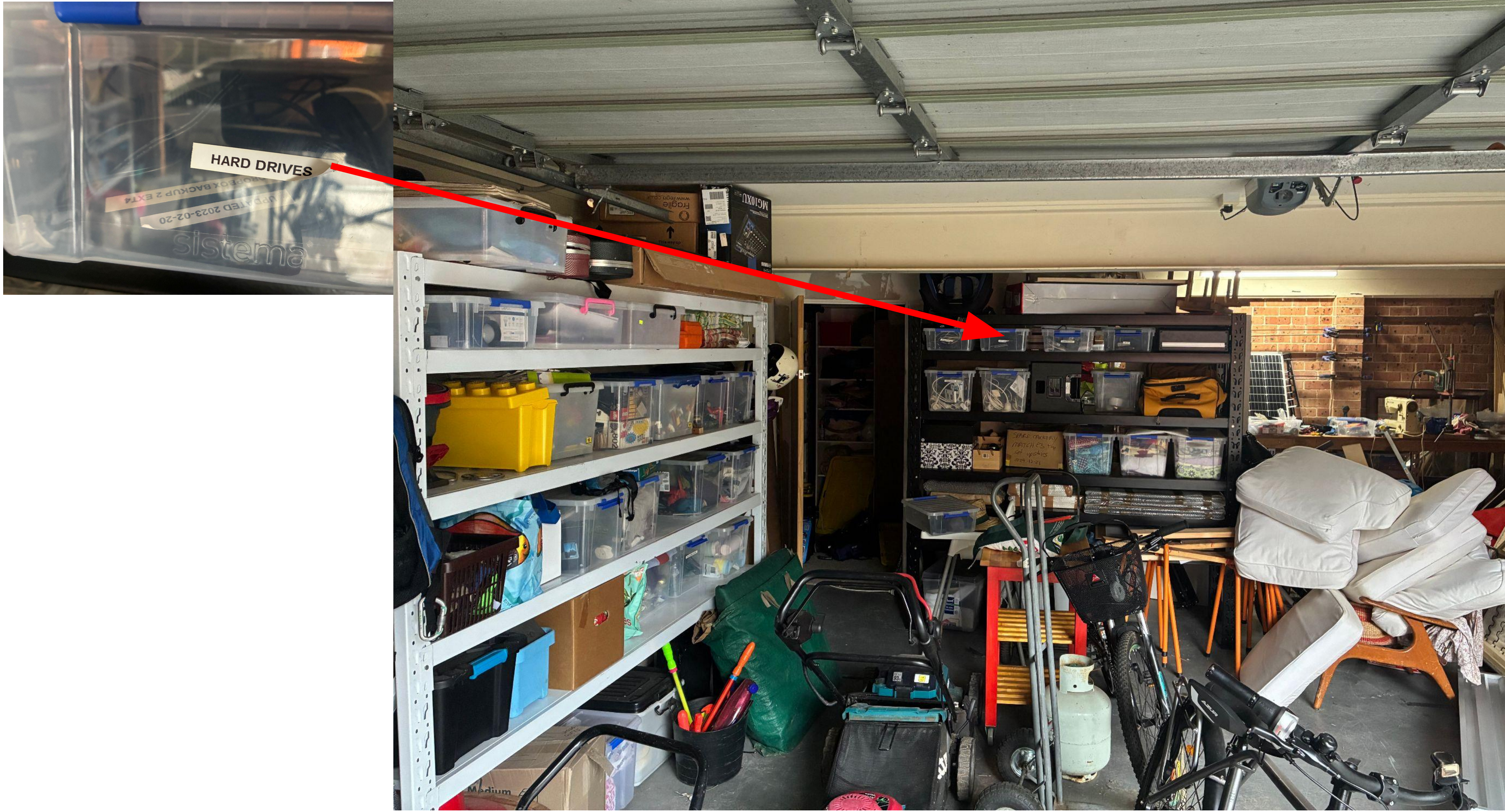
LDaCA Execution Strategy Overview



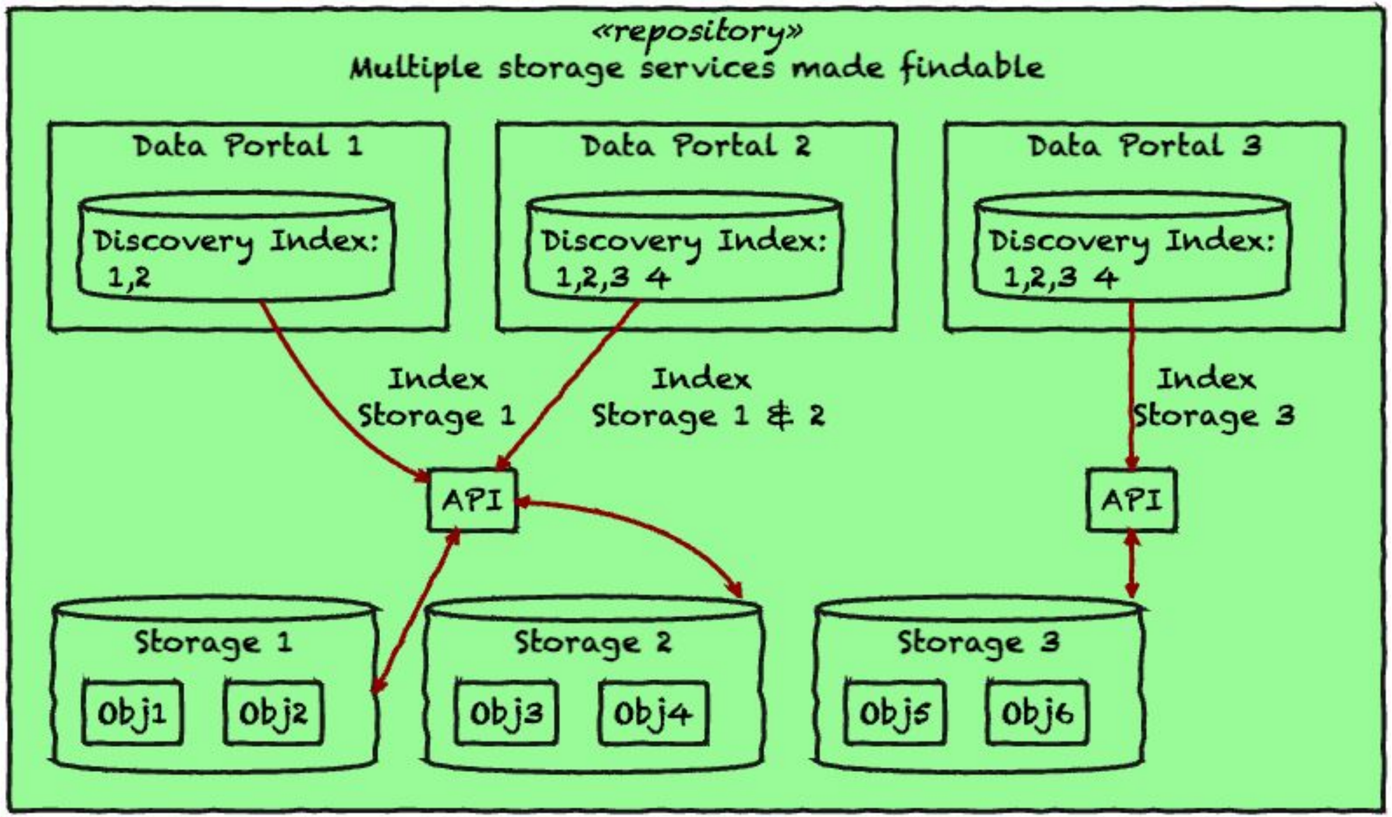
	Starting state (2021)	Desired state (2028)
collect & organise	Language data is rarely organised or described in reusable ways, if it's described at all	Standards and tools are available and being applied by data stewards
conserve	A lot of language data is at risk of being lost forever	Good governance and standardised, distributed storage of data helps preserve and return data
find	It’s difficult to know what language data exists and where to find it	Discovering and locating language data is easy via linked portals
access	Processes for granting permissions and getting access to data are either absent or aren’t easy to understand or apply	Access controls are in place and easy to use, so that data access can be given to the right people
analyse	Ad hoc tools, analysis and annotation methods are used, lacking reproducibility	Shared tools can process, analyse, reuse, repurpose, annotate, visualise and enhance data at scale
guide	Guidance and training for collecting, handling, using and analysing data are scattered and hard to find	Best practice advice and training for working with language data is available from a single source which is easy to find

Activities

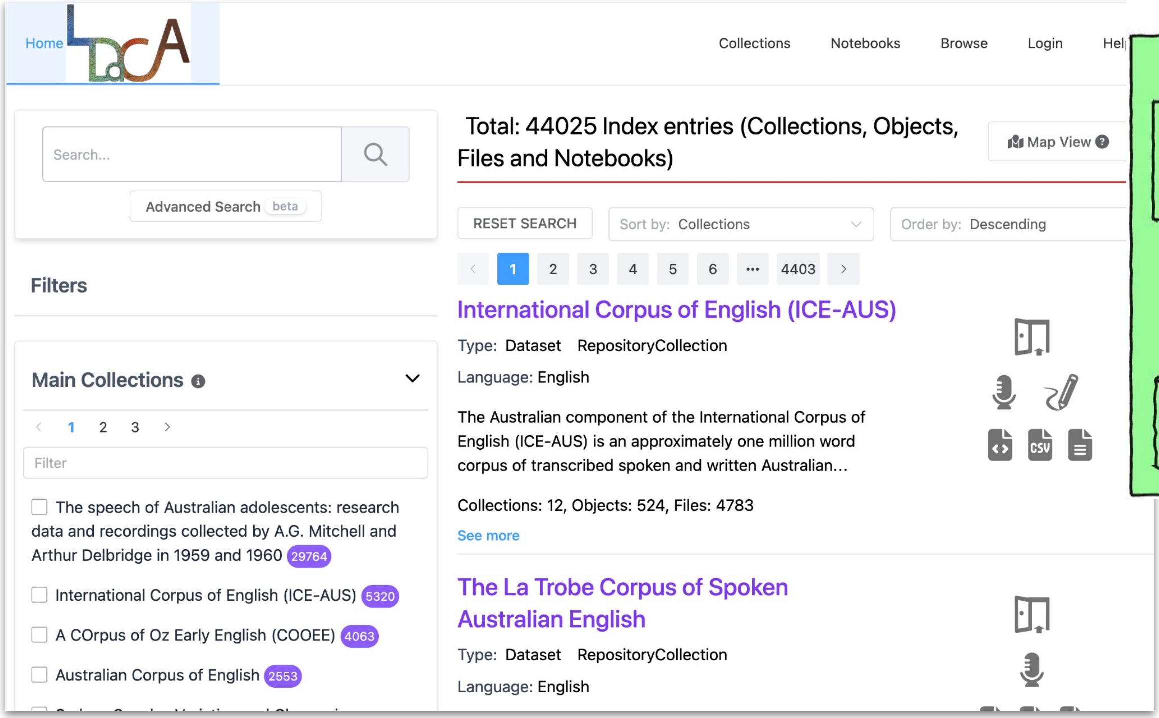
- Strengthen the data management skills of language worker communities
- Develop shared tools, standards and technical infrastructure to help data stewards care for data for the long term
- Build data portals with useful search functions and lightweight technical structures
- Create guidance for data stewards to document and grant access and reuse rights
- Support language communities to gain greater control over their language data
- Develop tools for data and metadata conversion, processing, analysis, annotation, visualisation, and enrichment
- Develop and guide the implementation of local and national policy and governance toolkits
- Provide examples and training for research at scale


HARD DRIVES
DROP-BOX BACKUP 2 EXT4
UPDATED 2023-02-20
sistema
SPARE CROCKERY MATCHES the set upstairs 2024-12-21

PILARS 1: Data is Portable: assets are not locked-in to a particular mode of storage, interface or service



One storage service ↔ One API ↔ One Portal



Home

Collections Notebooks Browse Login Hel

Search...

Advanced Search beta

Total: 44025 Index entries (Collections, Objects, Files and Notebooks)

Map View

RESET SEARCH Sort by: Collections Order by: Descending

1 2 3 4 5 6 ... 4403

Filters

Main Collections

1 2 3

Filter

- □ The speech of Australian adolescents: research data and recordings collected by A.G. Mitchell and Arthur Delbridge in 1959 and 1960 29764
- □ International Corpus of English (ICE-AUS) 5320
- □ A COrpus of Oz Early English (COOEE) 4063
- □ Australian Corpus of English 2553

International Corpus of English (ICE-AUS)

Type: Dataset RepositoryCollection

Language: English

The Australian component of the International Corpus of English (ICE-AUS) is an approximately one million word corpus of transcribed spoken and written Australian...

Collections: 12, Objects: 524, Files: 4783

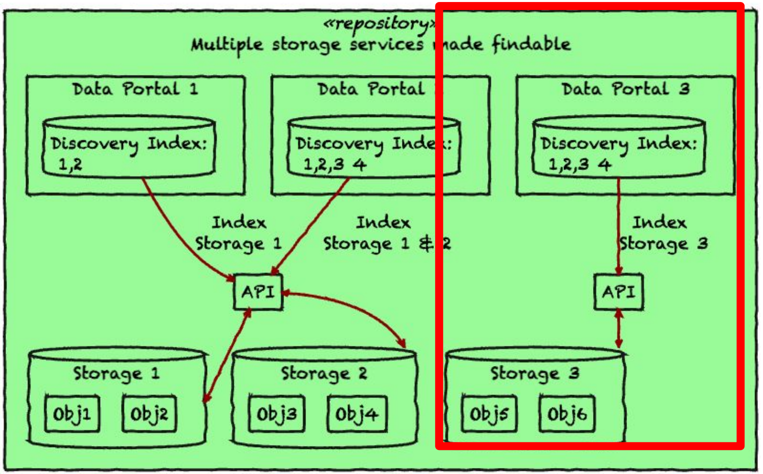
See more

The La Trobe Corpus of Spoken Australian English

Type: Dataset RepositoryCollection

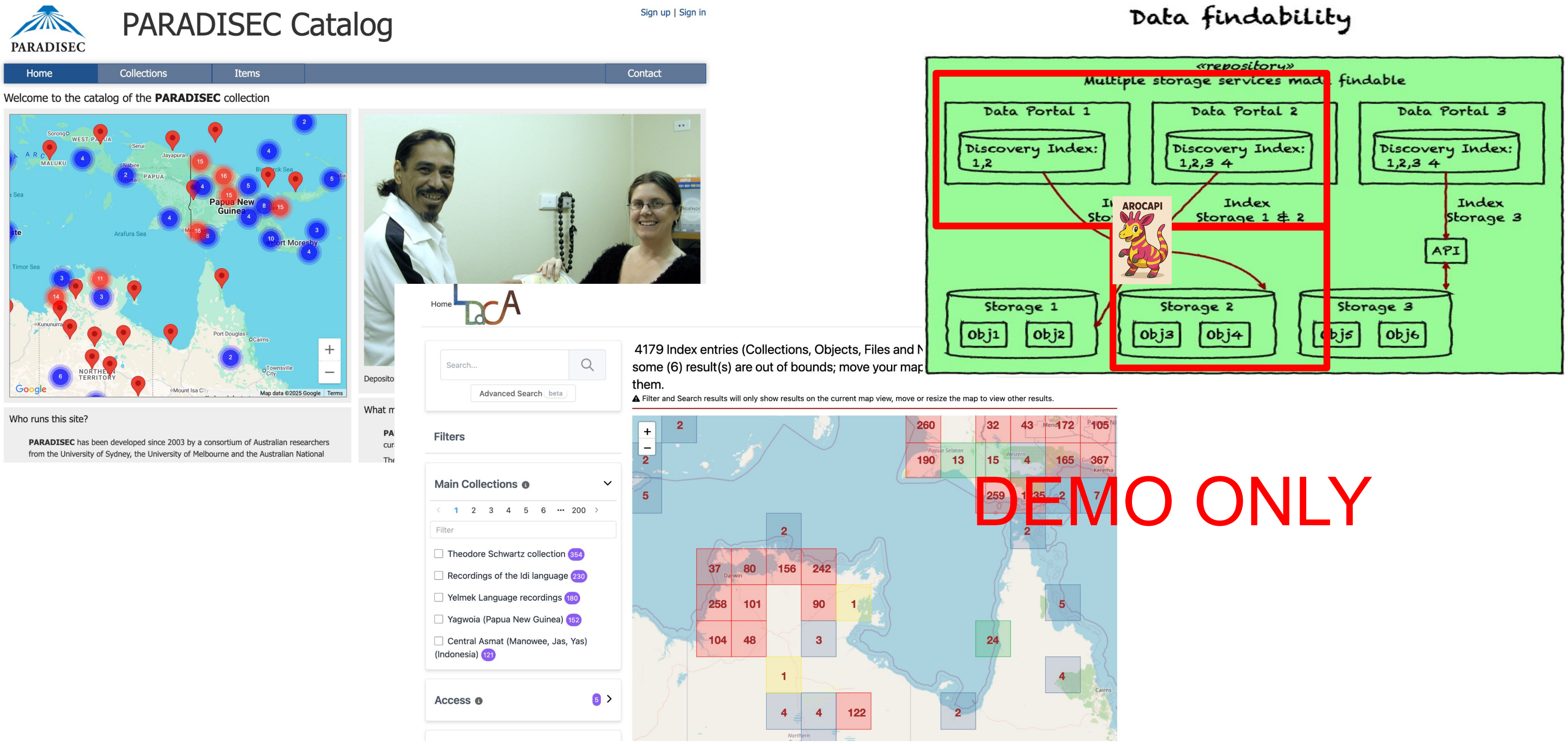
Language: English

Data findability



This pattern will be used for LDaCA, the Batchelor CALL Collection, RAPID (Hansard) the UTS Research Data Repository and other major collections

One data store ↔ One API ↔ 2 portals (demo)



PARADISEC Catalog

Sign up | Sign in

Home | Collections | Items | Contact

Welcome to the catalog of the **PARADISEC** collection

Who runs this site?

PARADISEC has been developed since 2003 by a consortium of Australian researchers from the University of Sydney, the University of Melbourne and the Australian National

Deposito

What m

PA
cur
The

Home

Search...

Advanced Search beta

Filters

Main Collections

1 2 3 4 5 6 ... 200

Filter

- □ Theodore Schwartz collection 354
- □ Recordings of the Idi language 230
- □ Yelmek Language recordings 180
- □ Yagwoia (Papua New Guinea) 152
- □ Central Asmat (Manowee, Jas, Yas) (Indonesia) 121

Access 5

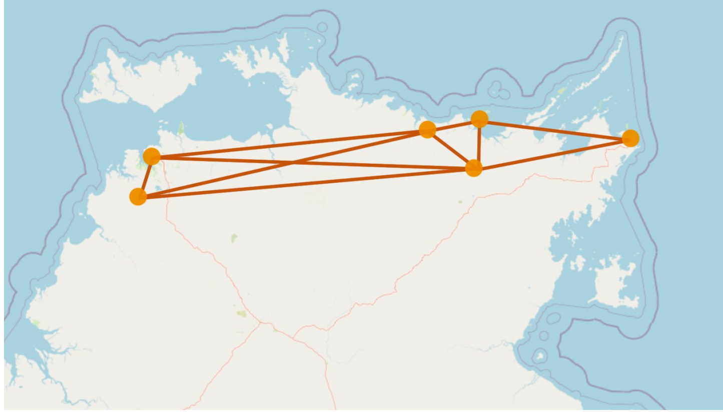
4179 Index entries (Collections, Objects, Files and N
some (6) result(s) are out of bounds; move your map
them.

⚠ Filter and Search results will only show results on the current map view, move or resize the map to view other results.

DEMO ONLY

Other deployment options

- Set up a stand-alone service for a specific archive (Batchelor CALL Collection work in progress)
- Automation of deployment of portals on demand for testing or show and tell
- Distributed regional archival repositories, local orgs share infrastructure, avoiding cloud services



- Put part (or all) of a collection on a tiny computer (Raspberry Pi) for distribution

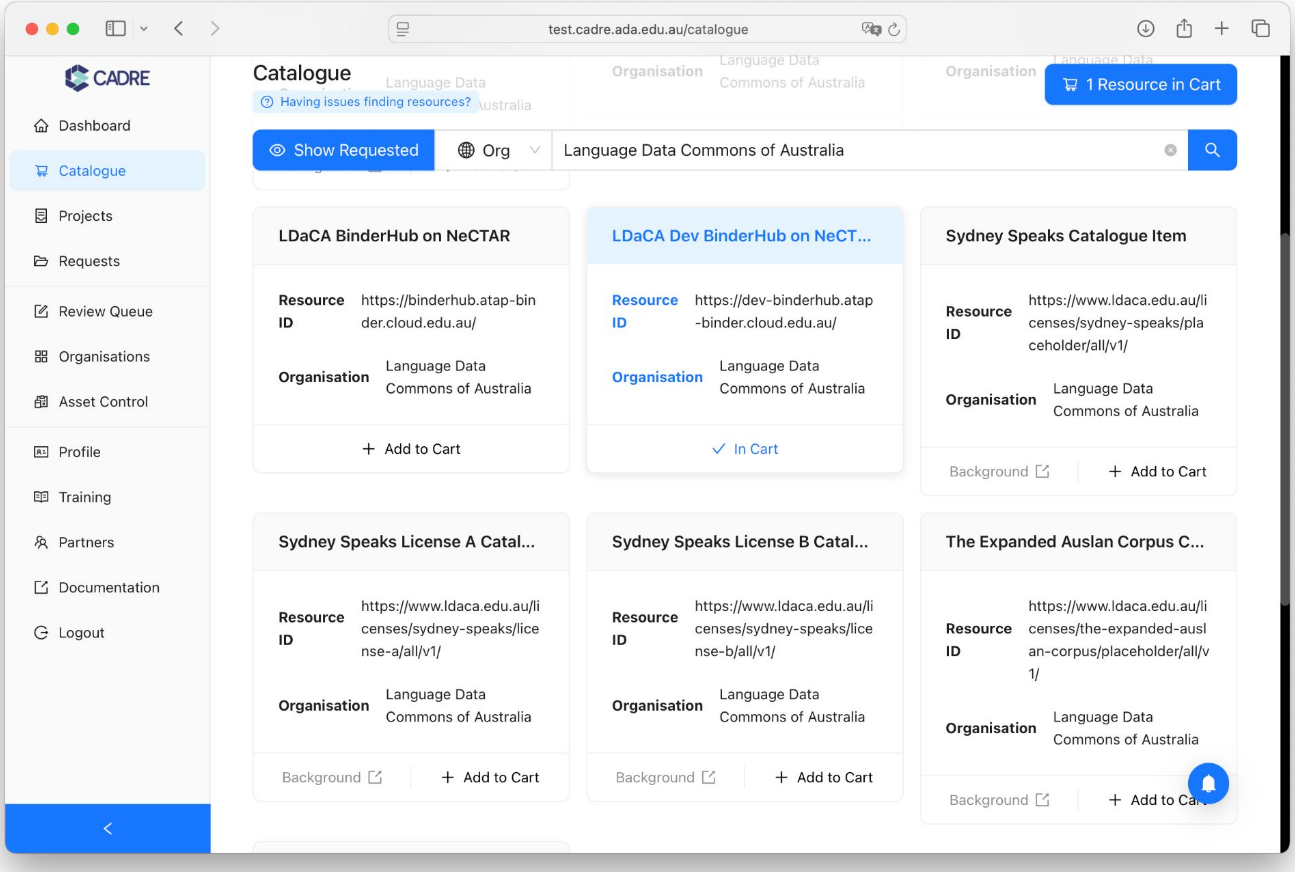
↑Raspberry Pi containing a collection

←Access on mobile via wifi

Services, software, standards and guides

Services	Standards / specs	Guides	Open Source Software
LDaCA Data Portal (oni) PARADISEC Catalogue Nyingarn Workspace RO-Crate Playground	RO-Crate specification LDaC Schema (OLAC plus) RO-Crate Profile(s)	Persistent ID policy Collections Strategy Data onboarding materials (Crate-O, spreadsheets)	Oni repo & portal Corpus tools (~30) RO-Crate libraries and tooling for editing and displaying crates and maintaining Schemas
CADRE (authorization) – ANU	OLAC, CLDF		

Tools



Catalogue

LDaCA BinderHub on NeCTAR

Resource ID https://binderhub.atap-binder.cloud.edu.au/

Organisation Language Data Commons of Australia

LDaCA Dev BinderHub on NeCT...

Resource ID https://dev-binderhub.atap-binder.cloud.edu.au/

Organisation Language Data Commons of Australia

Sydney Speaks Catalogue Item

Resource ID https://www.ldaca.edu.au/licenses/sydney-speaks/placeholder/all/v1/

Organisation Language Data Commons of Australia

Sydney Speaks License A Catal...

Resource ID https://www.ldaca.edu.au/licenses/sydney-speaks/license-a/all/v1/

Organisation Language Data Commons of Australia

Sydney Speaks License B Catal...

Resource ID https://www.ldaca.edu.au/licenses/sydney-speaks/license-b/all/v1/

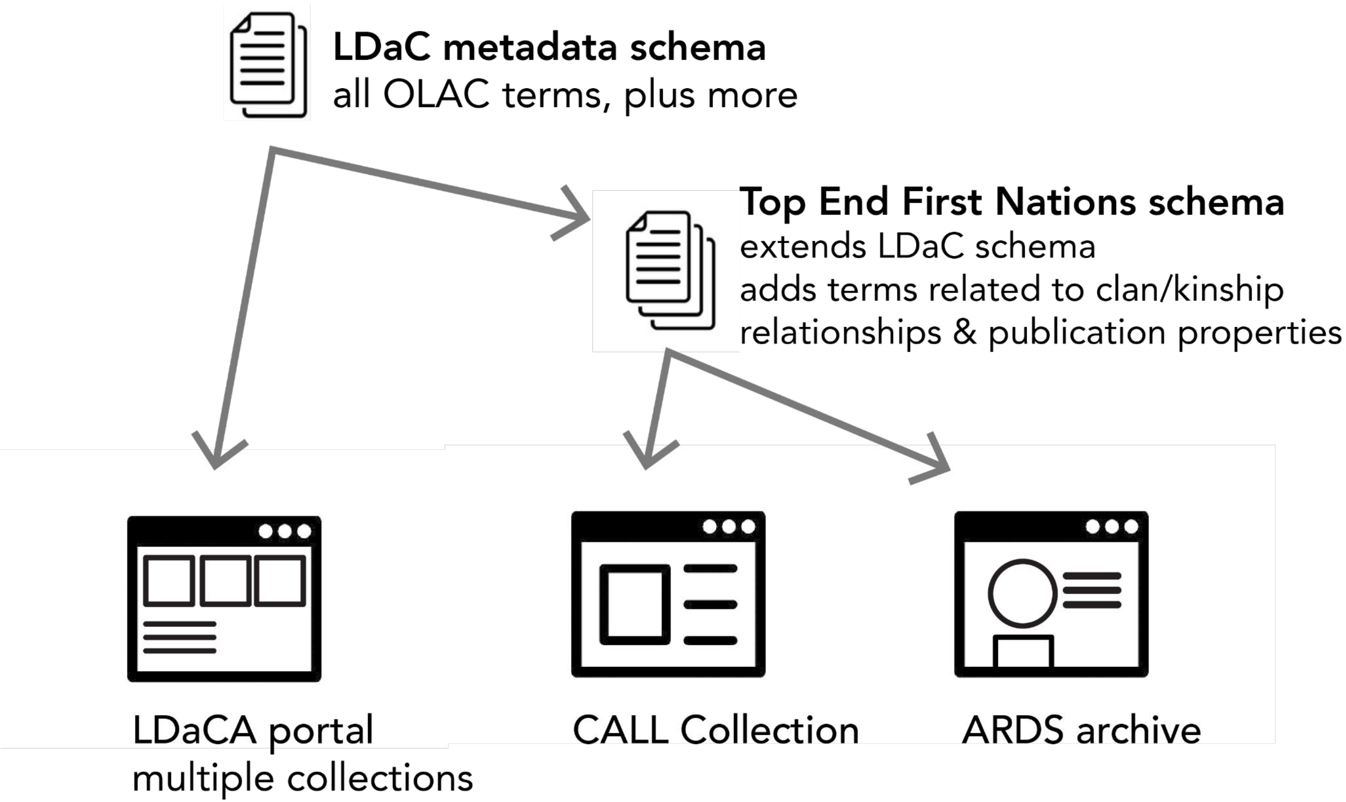
Organisation Language Data Commons of Australia

The Expanded Auslan Corpus C...

Resource ID https://www.ldaca.edu.au/licenses/the-expanded-auslan-corpus/placeholder/all/v1/

Organisation Language Data Commons of Australia

Example of standards/specs used in multiple sites



LDaC metadata schema
all OLAC terms, plus more

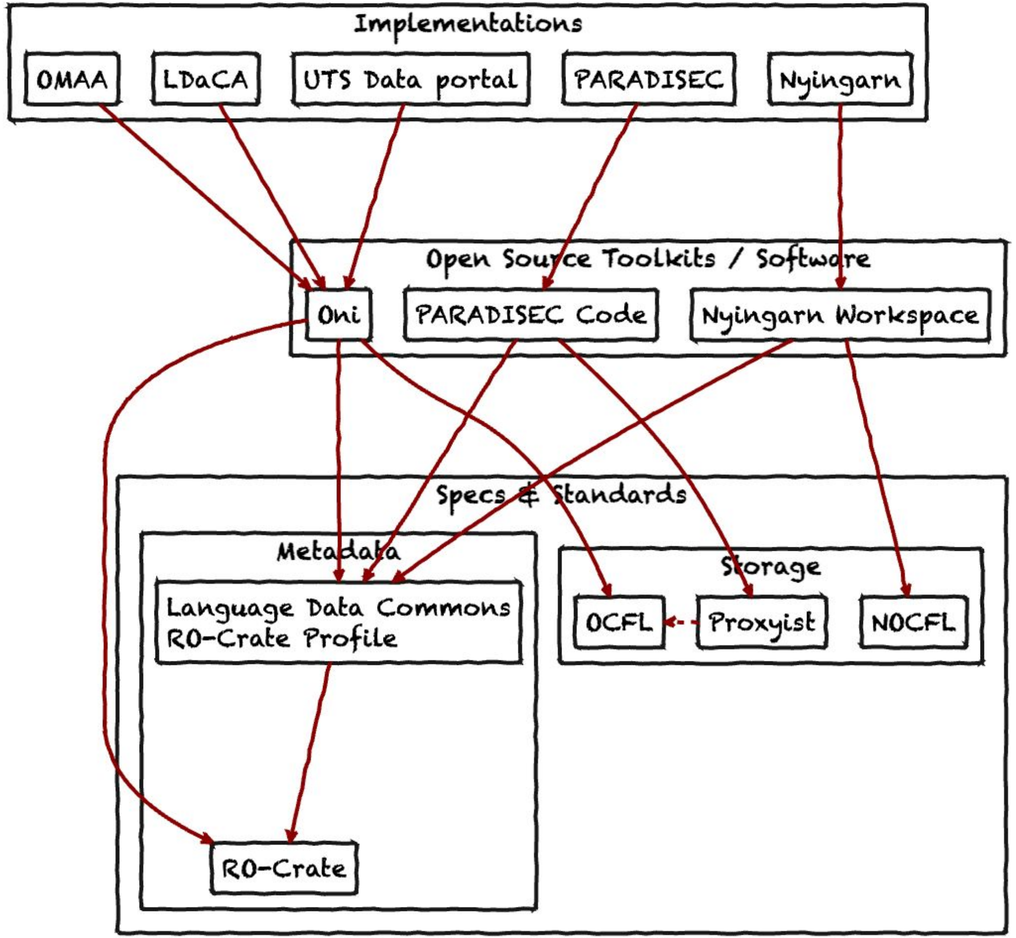
Top End First Nations schema
extends LDaC schema
adds terms related to clan/kinship
relationships & publication properties

LDaCA portal
multiple collections

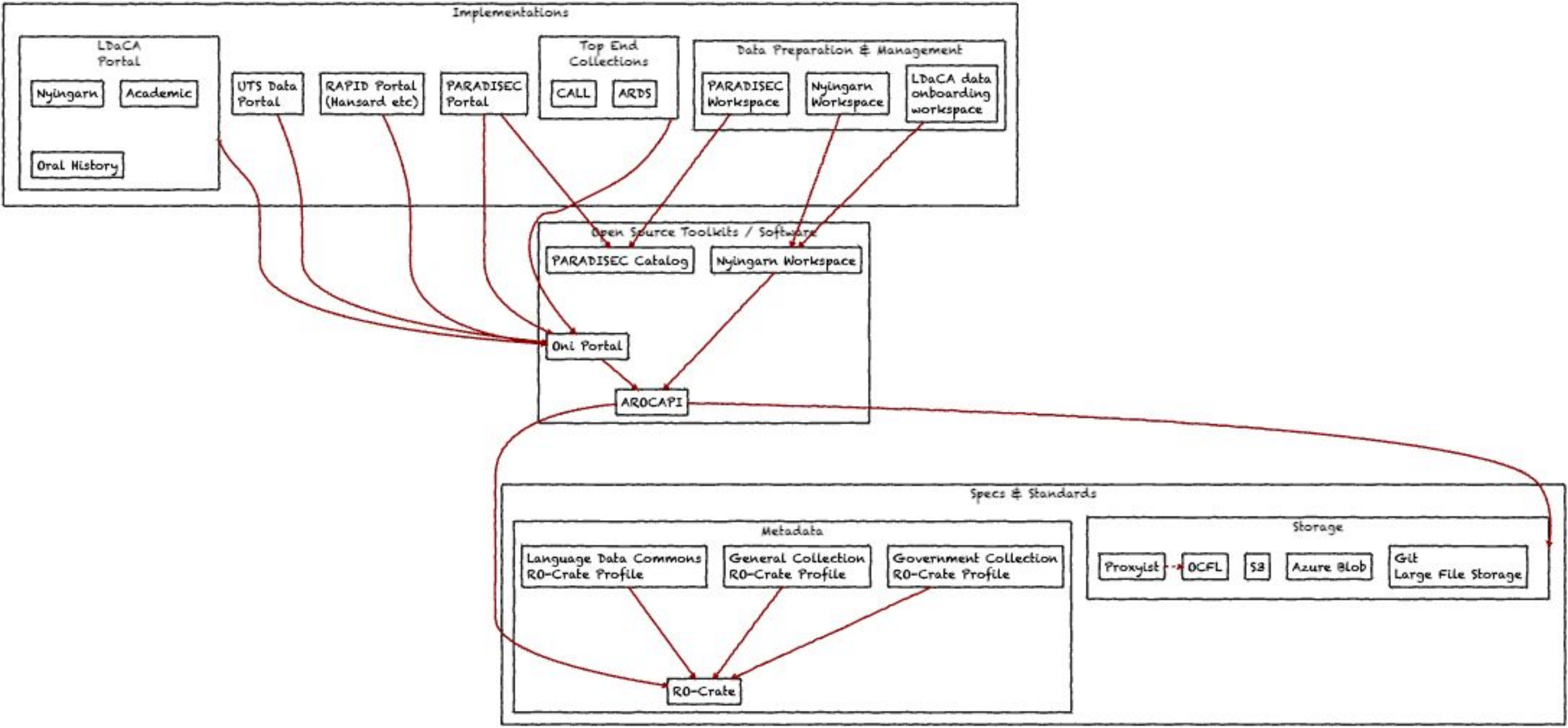
CALL Collection

ARDS archive

PILARS Implementations - mid 2024



PILARS Implementations - mid 2026?



Implementations

LDaCA Portal

Nyingarn

Academic

Oral History

UTS Data Portal

RAPID Portal (Hansard etc)

PARADISEC Portal

Top End Collections

CALL

ARDS

Data Preparation & Management

PARADISEC Workspace

Nyingarn Workspace

LDaCA data onboarding workspace

Open Source Toolkits / Software

PARADISEC Catalog

Nyingarn Workspace

Oni Portal

AROCAPI

Specs & Standards

Metadata

Language Data Commons RO-Crate Profile

General Collection RO-Crate Profile

Government Collection RO-Crate Profile

RO-Crate

Storage

Proxyist

OCFL

S3

Azure Blob

Git Large File Storage