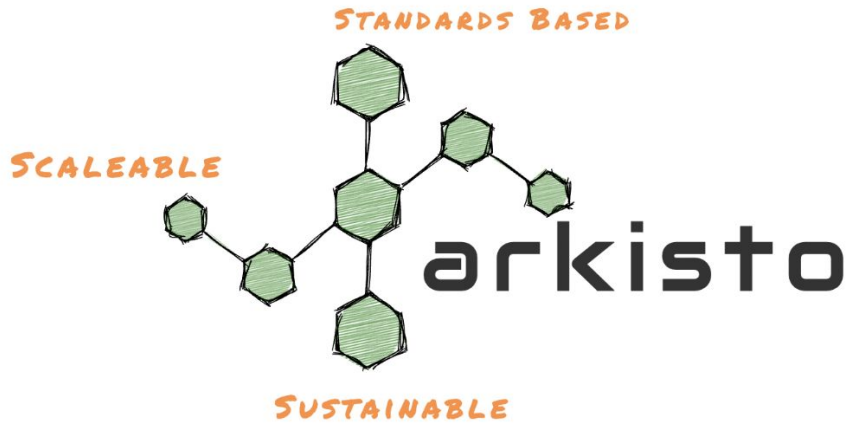


# HASS RDC Technical Advisory Group Meeting

LDaCA & ATAP

Intro

Peter Sefton - [p.sefton@uq.edu.au](mailto: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 always 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.

1.	<b>Conformance</b>
2.	<b>Terminology</b>
3.	<b>OCFL Object</b>
3.1	Object Structure
3.2	Object Conformance Declaration
3.3	Version Directories
3.3.1	Content Directory
3.4	Digests
3.5	Inventory
3.5.1	Basic Structure
3.5.2	Manifest
3.5.3	Versions
3.5.4	Fixity
3.6	Inventory Digest
3.7	Version Inventory and Inventory Digest
3.8	Logs Directory
3.9	Object Extensions
4.	<b>OCFL Storage Root</b>
4.1	Root Structure
4.2	Root Conformance Declaration
4.3	Storage Hierarchies
4.4	Storage Root Extensions
4.5	Filesystem features
5.	<b>Examples</b>
5.1	Minimal OCFL Object
5.2	Versioned OCFL Object
5.3	Different Logical and Content Paths in an OCFL Object
5.4	BagIt in an OCFL Object
5.5	Moab in an OCFL Object
5.6	Example Extended OCFL Storage Root

# Oxford Common File Layout Specification

Recommendation 07 July 2020

## This version:

<https://ocfl.io/1.0/spec/>

## Latest published version:

<https://ocfl.io/latest/spec/>

## Editors:

[Andrew Hankinson](#) ([Bodleian Libraries, University of Oxford](#))

[Neil Jefferies](#) ([Bodleian Libraries, University of Oxford](#))

[Rosalyn Metz](#) ([Emory University](#))

[Julian Morley](#) ([Stanford University](#))

[Simeon Warner](#) ([Cornell University](#))

[Andrew Woods](#) ([LYRASIS](#))

## Additional Documents:

[Implementation Notes](#)

[Validation Codes](#)

[Extensions](#)

## Previous version:

<https://ocfl.io/0.9/spec/>

## Repository:

[Github](#)

[Issues](#)

[Commits](#)

[Use Cases](#)

This document is licensed under a [Creative Commons Attribution 4.0 License](#).

[OCFL logo: hand-drive](#) by [Patrick Hochstenbach](#) is licensed under [CC BY 2.0](#).



## Introduction

*This section is non-normative*

Addressable  
resources



Local Data



<https://orcid.org/0000-0001-2345-6789>

ID? Title? Description?



Who created this data?



What parts does it have?



When?



What is it about?



How can it be reused?



As part of which project?



Who funded it?



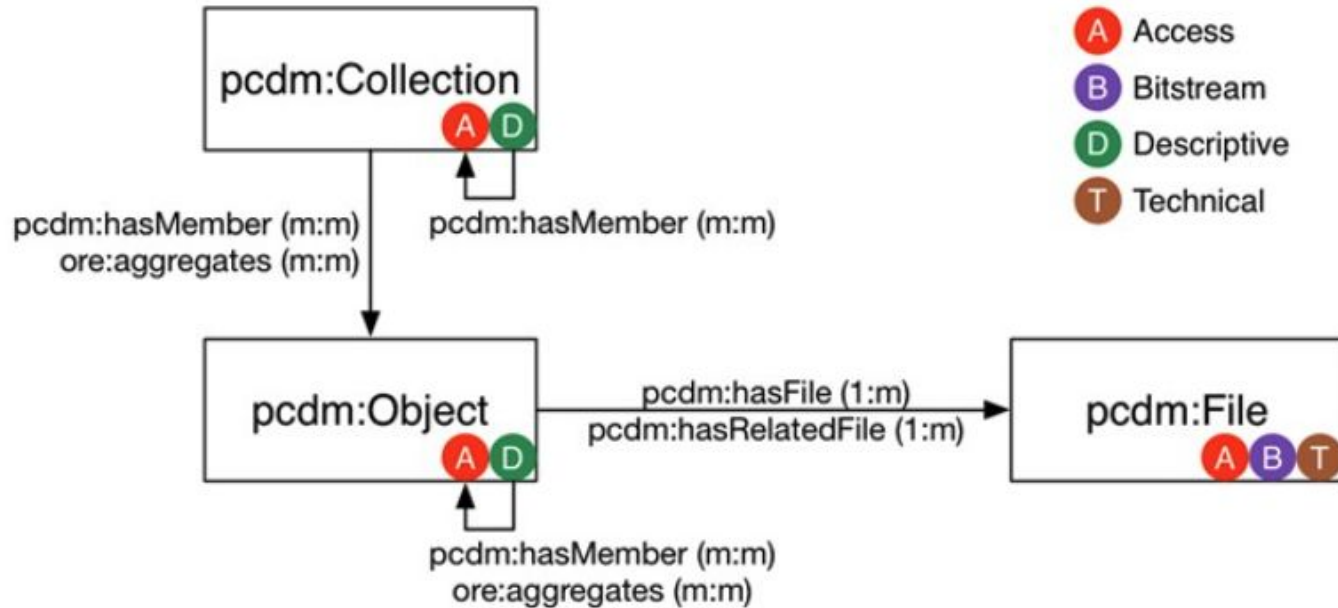
How was it made?



[https://en.wikipedia.org/wiki/Scanning\\_electron\\_microscope](https://en.wikipedia.org/wiki/Scanning_electron_microscope)



# Portland Common Data Model



## Collection of Australian Government Documents in six languages

 Download all the metadata for Collection of Australian Government Documents in six languages in JSON-LD format

[Check this crate](#)

This RO-Crate contains an entire PCDM collection

## Browse files Collection of Australian Government Documents in six languages

**@id** arcp://name,multilingual.gov.au/corpus/root

**name** [?] Collection of Australian Government Documents in six languages

**@type**

- Dataset
- RepositoryCollection

**description** [?] This dataset consists of documents created by government agencies in Australia which provide information to members of the Australian community. They originate as English text and also exist as translations into various other languages used in the Australian community. The initial dataset (12/2021) includes documents prepared by Services Australia (<https://www.servicesaustralia.gov.au/>) and by the Victorian Department of Health ( Health Translations) in English, Arabic, Farsi, Turkish, Vietnamese and Chinese.

**datePublished** [?] 2022-01-10

**author** [?]

- [Simon Musgrave](#)
- [Peter Sefton](#)

**hasFile** [?]

**copyrightNotice** Services Australia: Copyright in the documents in this collection is owned by the Commonwealth of Australia, represented by Services Australia. Content is licenced under a Creative Commons Attribution 3.0 Australia Licence, with the exception of: the Commonwealth Coat of Arms the Australian Government Services Australia logo any protected names and symbols under Commonwealth legislation any registered trade marks owned by the Commonwealth of Australia images content supplied by third parties, as identified. Full details of the licence terms are available on the Creative Commons website. The terms of use for the Coat of Arms are available from the Department of the Prime Minister and Cabinet website. Department of Health (Vic): Copyright in this website (including content and design) is owned by the State of Victoria or used under licence. You may make limited copies of this website in accordance with the Copyright Act 1968 (Cth), including copies for research, study, criticism, review or news reporting. You may not publish, reproduce, adapt, modify, communicate or otherwise use any part of this website (in particular for commercial purposes).

**hasPart** [?]

- ▶ [ChildCareSubsidy -to- .: Persian, Iranian \(PDF\) خوشونت خرداگي و خانگي](#)
- ▶ [.: Persian, Iranian\(TEXT\) -to- Parenting Payment: Chinese, Mandarin \(PDF\) خوشونت خرداگي و خانگي](#)
- ▶ [Parenting Payment: Chinese, Mandarin\(TEXT\) -to- Status Resolution Support Services payment: Persian, Iranian \(PDF\)](#)
- ▶ [Status Resolution Support Services payment: Persian, Iranian\(TEXT\) -to- Status Resolution Support Services payment: Chinese, Mandarin\(TEXT\)](#)

**license** [?] [Attribution 3.0 Australia \(CC BY 3.0 AU\)](#)

RO-Crates MUST have licence information that sets out conditions for use/reuse of the data



## Collection of Australian Government Documents in six languages

 Download all the metadata for Collection of Australian Government Documents in six languages in JSON-LD format

[Check this crate](#)

### DisabilitySupportPension

**@id** arcp://name,multilingual.gov.au/item/DisabilitySupportPension

**name** [\[?\]](#) DisabilitySupportPension

**@type**

- RepositoryObject
- ProceduralText

**author** [\[?\]](#) CommonwealthGovernment

**hasFile** [\[?\]](#)

- [معاش دعم العاقبة: Arabic, Standard \(PDF\)](#)
- [معاش دعم العاقبة: Arabic, Standard\(TEXT\)](#)
- [DisabilitySupportPension: ORIGINAL English \(TEXT\)](#)
- [برنامه مستمری حمایتی معوالن: Persian, Iranian \(PDF\)](#)
- [برنامه مستمری حمایتی معوالن: Persian, Iranian\(TEXT\)](#)
- [Engellilik Destek Emeklilik Maaşı: Turkish \(PDF\)](#)
- [Engellilik Destek Emeklilik Maaşı: Turkish\(TEXT\)](#)
- [Disability Support Pension: Vietnamese \(PDF\)](#)
- [Disability Support Pension: Vietnamese\(TEXT\)](#)
- [Disability Support Pension: Chinese, Mandarin \(PDF\)](#)
- [Disability Support Pension: Chinese, Mandarin\(TEXT\)](#)



Disability Support Pension - Arabic 1 / 4 100%

Australian Government  
Department of Human Services

ARABIC

### معاش دعم الإعاقة (Disability Support Pension)

يوفر Disability Support Pension (DSP) دعماً مالياً للأشخاص الذين لديهم حالة مرضية بدنية أو عقلية أو نفسية تعيقهم عن العمل، أو المصابين بالعمى الدائم.

#### الأهلية لـ Disability Support Pension

قد تكون مؤهلاً لـ DSP إذا:

- كان عمرك بين 16 و age pension
- كنت تفي بمتطلبات الإقامة
- كنت تفي بمتطلبات اختبارات الدخل والامتلاك المنطبقة على حالتك، و
- كنت مصاباً بالعمى الدائم، أو
- تم تقييم حالتك وثبت أن لديك ضعف بدني أو عقلي أو نفسي، و
- كنت غير قادر على العمل، أو تتدرب للعمل، لمدة 15 ساعة أو أكثر في الأسبوع وتتقاضى أجراً بنقص معدل الحد الأدنى للأجور ذي الصلة أو أعلى منه، خلال العامين القادمين، وذلك بسبب ضعفك، و
- شاركت بفعالية، أو أكملت Program of Support، إن كان مطلوباً.

#### المطالبة بـ Disability Support Pension

عندما تقدم مطالبته، سوف يُطلب منك تقديم إثبات طبي. وقد نحتاج إلى أخذ نسخ ضوئية عن مستنداتك وإرجاع الأصل إليك.

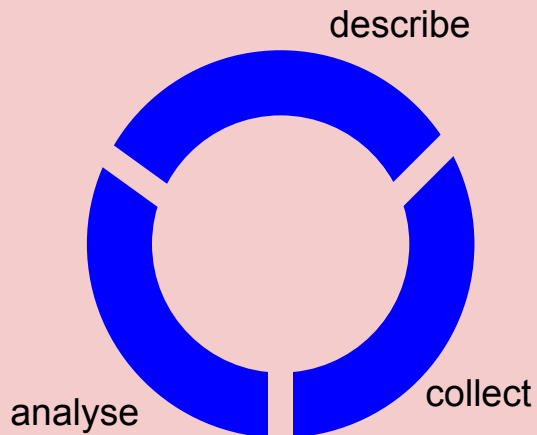
@id	DisabilitySupportPension/12626-1506ar.pdf
name [?]	معاش دعم الإعاقة: Arabic, Standard (PDF)
@type	File
encodingFormat [?]	<ul style="list-style-type: none"><li>• <a href="#">Acrobat PDF 1.5 - Portable Document Format</a></li><li>• application/pdf</li></ul>
contentSize [?]	404819
dateModified [?]	2022-01-05T12:01:37+11:00
language [?]	Arabic, Standard
translationOfWork [?]	DisabilitySupportPension
size [?]	404819

Link back to the container which has type RepositoryObject

Workspaces:

- working storage
- domain specific tools
- domain specific services

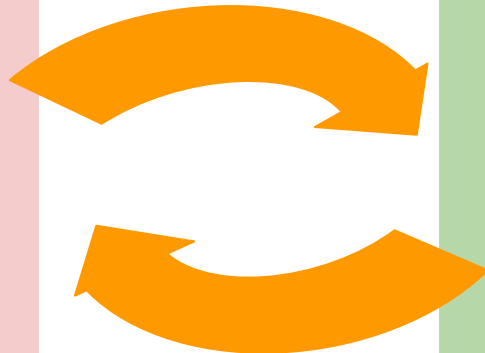
Research  
Data  
Management  
Plan



Active cleanup processes  
workspaces considered ephemeral

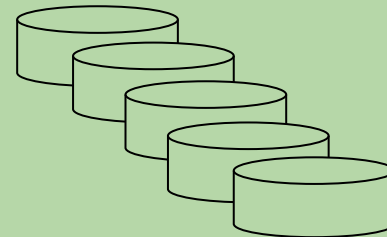
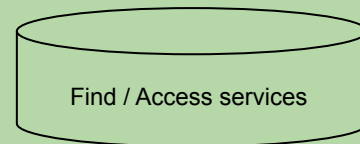
Reusable, Interoperable  
data objects

- 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

Data Curation  
& description

describe

analyse

collect

Data Cleaning

OCR / transcription  
format migration

BYOData



Identity Management

AAF / social media accounts

Archive & Preservation Repositories  
institutional, domain or both

Harvested

external

PARADISEC

AU Nat. Corpus

AusLan (sign)

Sydney Speaks

ATAP Corpus  
Reference, Training & BYO

ATAP Notebooks  
Apps, Code, Workflows

... etc

Lang. portal(s)

Corpus discovery  
Item discovery  
Authenticated API  
Create virtual corpora

Deposit / Publish

Licence  
Server

Reuse

Analytics  
Portal

Code discovery  
Launch / Rerun  
Data Discovery  
Authenticated API

STORAGE (including Cloudstor)



Workbench

Notebooks  
Data import by URL  
Export fully described pkg  
**Stretch goals:**  
Code gen / simple  
interfaces eg Discursis

Compute

HPC

Cloud

Desktop

Workspaces:

- working storage
- domain specific tools
- domain specific services



Active cleanup processes  
workspaces considered ephemeral

Data Curation  
& description

describe

Deposit / Publish

Archive & Preservation Repositories  
institutional, domain, or both

Harvested

external

PARADISEC

AU Nat. Corpus

AusLan (sign)

Sydney Speaks

ATAP Corpus  
Reference, Training & BYO

ATAP Notebooks  
Apps, Code, Workflows

etc

Lang. portal(s)

Corpus discovery  
Item discovery  
Authenticated API  
Create virtual corpora

Licence  
Server

Analytics  
Portal

Code discovery  
Launch / Reuse  
Data Discovery  
Authenticated API

Reuse

collect

Data Cleaning

OCR / transcription  
format migration

BYOData



STORAGE (including Cloudstor)

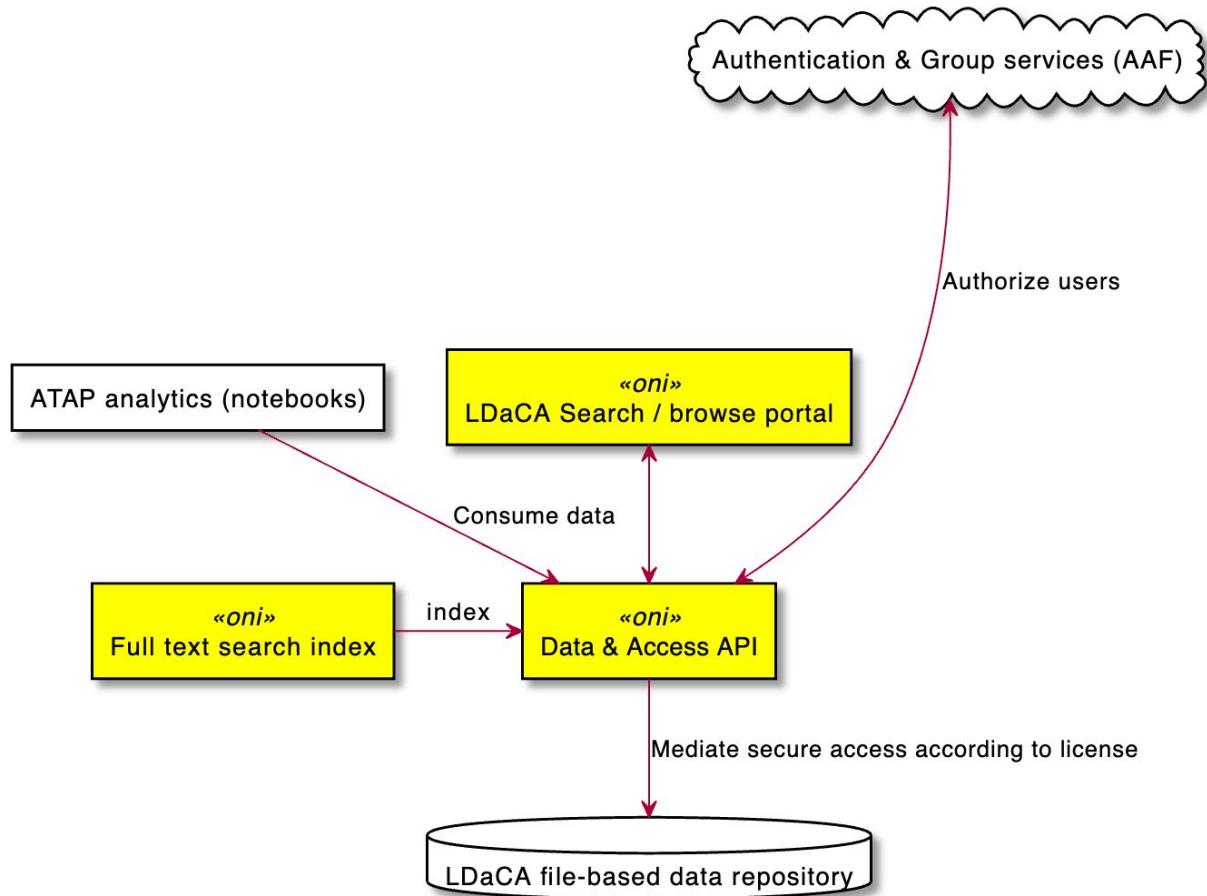
Talking mainly  
about this bit  
today



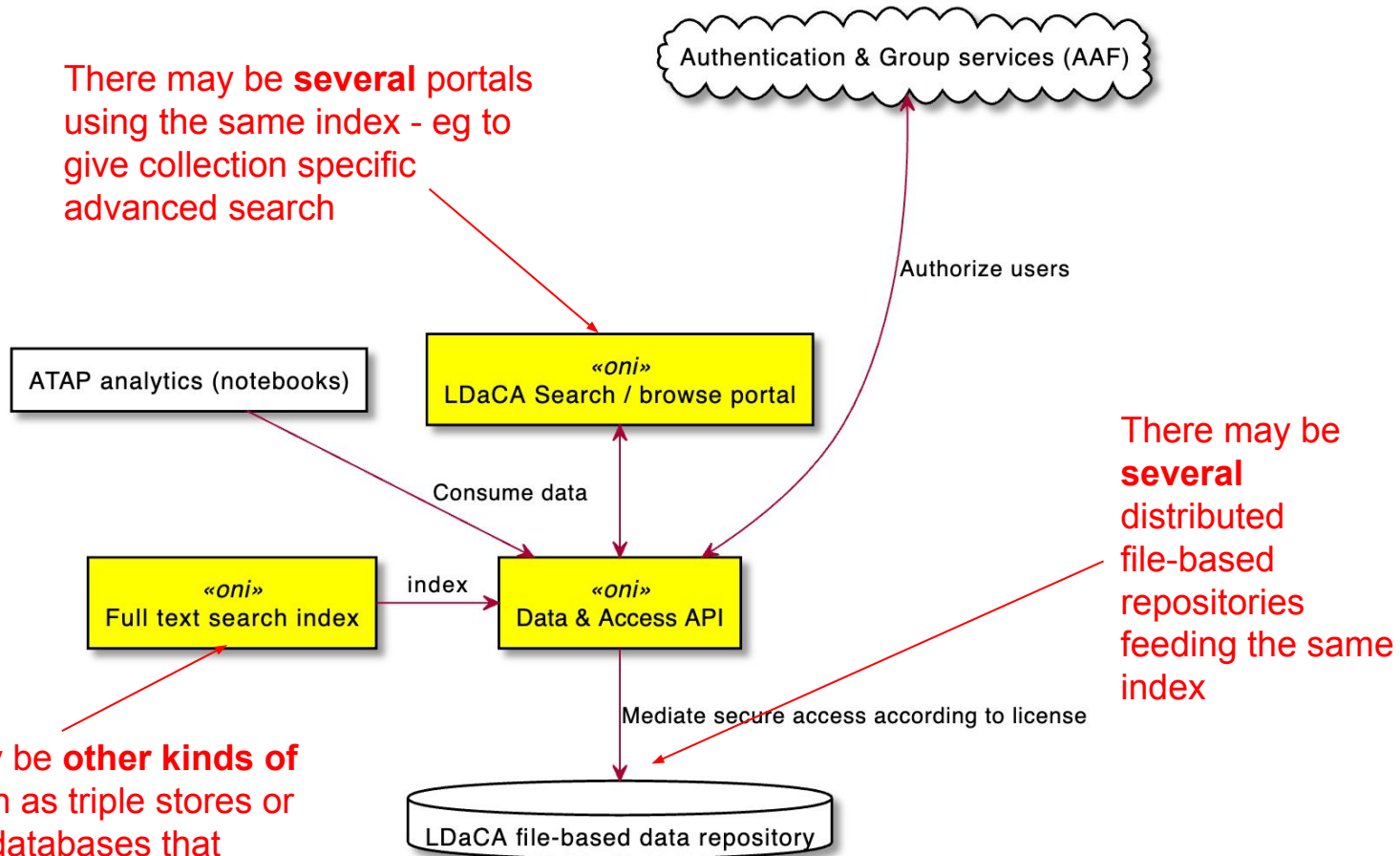
Workflow

Notebooks  
Data import by URL  
Export fully described pkg  
Stretch goals:  
Code generation  
interfaces eg Discursis

analyse



All data stored using the Research Object Crate metadata specification, with an OLAC-derived metadata schema and with re-use license information



# Oni

---

- Oni
  - [Start developing](#)
  - [Running the tests](#)
  - [Tech](#)
    - [Frontend - VueJS and friends](#)
    - [Backend - Restify and sequelize](#)
  - [Repo layout](#)
  - [Documentation](#)

Oni consists of a VueJS SPA (ui) and restify JS backend (api). This repo structure is shared with [Describo Online](#) and the [Nyingarn Workspace](#). Look there for more code.

## Start developing

---

To get started developing copy `configuration/example-configuration.json` to `configuration/development-configuration.json` and edit as required.

```
> docker-compose up
```

This will start the UI, API and db containers. It will automatically run `npm install` in both ui and api folders so you don't need to.

Saving UI and API code triggers auto reload.

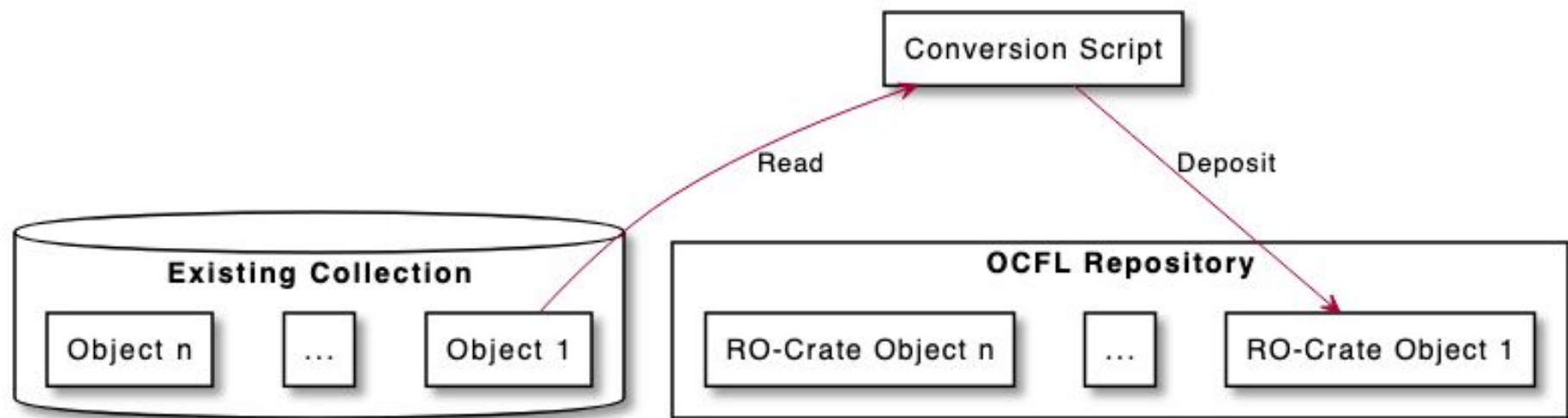
## Running the tests

---

- Find the api container ID: `dps | grep api | awk '{print $1}'`
- Exec into the container: `docker exec -it ${CONTAINER ID} bash`
- Run the Jest Testing environment: `npm run test:watch`

When you save a test file the tests will re-run automatically. Saving a changed code file (ie not a test file) does not re-run the tests.






Type ▾

Language ▾

Sort ▾

 New repository7 results for all repositories matching **corpus-tools** sorted by **last updated** Clear filter**corpus-tools-coeee** Private

Corpus prep tools for the COEEE corpus (using the spreadsheet that comes with the corpus)

● JavaScript  GPL-3.0  0  0  2  0 Updated now**corpus-tools-farms-to-freeways** Private

Node scripts to build an Arkisto-ready language data collection from the "From Farms to Freeways" history project data.

● JavaScript  GPL-3.0  0  0  0  0 Updated 2 minutes ago**corpus-tools-python3-soffice** Public

Docker container with LibreOffice and Python 3 for data massaging of word processing files, eg .doc -&gt; .docx, .docx -&gt; .pdf, or .pdf -&gt; .svg

 GPL-3.0  0  0  0  0 Updated 5 days ago**corpus-tools-gov.au-multilingual** Private

Node tools to prepare an Arkisto repository with parallel translations of Australian Government documents

● JavaScript  GPL-3.0  0  0  0  0 Updated 19 days ago**corpus-tools-sydney-speaks** Private

Node tools to prepare an Arkisto repository with language data from the Sydney Speaks project

● JavaScript  GPL-3.0  0  0  0  0 Updated on 16 Dec 2021**corpus-tools-monash-coe** Private

Corpus tools to prepare the Monash Corpus of English for ingest into RO-Crate and OCFL

● JavaScript  GPL-3.0  0  0  0  0 Updated on 8 Dec 2021**corpus-tools-alveo-export** Private

Node tools to export Alveo collections as Arkisto-ready repository objects

● JavaScript  GPL-3.0  0  0  0  0 Updated on 8 Sep 2021

Limit your search

cat

[Advanced search](#)

Search Metadata >

Collection **▼**

cooee 18

Created **▼**

[1820 - 1829](#) 2

[1830 - 1839](#) 1

[1840 - 1849](#) 2

[1850 - 1859](#) 2

[1870 - 1879](#) 2

[1880 - 1889](#) 4

[1890 - 1899](#) 4

[1900 - 1909](#) 1

Mode >

Speech Style >

Publication Status >

Written Mode >

Interactivity >

Communication Context >

Communication Medium >

Communication Setting >

Audience >

Discourse Type >

Language (ISO 639-3 Code) >

Type >

## You searched for:

cat x Collection > cooee x

Add Selected to list ▼ Add All to list ▼ 1 - 18 of 18

<input type="checkbox"/>	Identifier	Title	Created Date	Type(s)
<input type="checkbox"/>	<a href="#">cooee:4-281</a>	unspecified	1892	Original, Text
<input type="checkbox"/>	<a href="#">cooee:3-163</a>	unspecified	1858	Original, Text
<input type="checkbox"/>	<a href="#">cooee:4-008</a>	unspecified	1878	Original, Text
<input type="checkbox"/>	<a href="#">cooee:2-068</a>	unspecified	1831	Text, Original
<input type="checkbox"/>	<a href="#">cooee:4-045</a>	unspecified	1883	Text, Original
<input type="checkbox"/>	<a href="#">cooee:4-024</a>	unspecified	1882	Text, Original
<input type="checkbox"/>	<a href="#">cooee:1-259</a>	unspecified	1825	Original, Text
<input type="checkbox"/>	<a href="#">cooee:2-026</a>	unspecified	1829	Text, Original
<input type="checkbox"/>	<a href="#">cooee:3-090</a>	unspecified	1854	Text, Original
<input type="checkbox"/>	<a href="#">cooee:2-351</a>	unspecified	1849	Text, Original
<input type="checkbox"/>	<a href="#">cooee:4-177</a>	unspecified	1888	Text, Original
<input type="checkbox"/>	<a href="#">cooee:4-124</a>	unspecified	1887	Text, Original
<input type="checkbox"/>	<a href="#">cooee:2-239</a>	unspecified	1841	Text, Original
<input type="checkbox"/>	<a href="#">cooee:4-402</a>	unspecified	1900	Original, Text
<input type="checkbox"/>	<a href="#">cooee:3-273</a>	unspecified	1871	Original, Text
<input type="checkbox"/>	<a href="#">cooee:4-381</a>	unspecified	1897	Original, Text
<input type="checkbox"/>	<a href="#">cooee:4-397</a>	unspecified	1899	Text, Original
<input type="checkbox"/>	<a href="#">cooee:4-330</a>	unspecified	1896	Text, Original

Mode	>
Speech Style	▼
unspecified	18
Publication Status	▼
unspecified	18
Written Mode	▼
unspecified	18
Interactivity	▼
unspecified	18
Communication Context	▼
unspecified	18
Communication Medium	▼
unspecified	18

METHOD ▼ `{{HOST}}/object?conformsTo=https://github.com/Language-Research-Technology/ro-crate-profile%23Collection`Send ▼Parameters ● Authorization Headers (7) Body Pre-request Script Tests Settings


Cookies

Query Params

KEY	VALUE	DESCRIPTION	...	Bulk Edit
conformsTo	https://github.com/Language-Research-Technology/ro-c...			
Key	Value	Description		

:response ▼

Click Send to get a response

Python - Requests ▼ 

```
1 import requests
2
3 url = "https://oni-dev.text-commons.org/
  api/object?conformsTo=https://github.
  com/Language-Research-Technology/
  ro-crate-profile%23Collection"
4
5 payload={}
6 headers = {
7     'Authorization': 'Bearer
  sk8479-wt1-486-w2e-48277884'
8 }
9
10 response = requests.request("GET", url,
  headers=headers, data=payload)
11
12 print(response.text)
13
```

main ro-crate-metadata / ro-crate-metadata.ipynb Go to file ...

moisbo commented out ro-crate output Latest commit 6814c70 on 8 Dec 2021 History

2 contributors

4043 lines (4043 sloc) | 162 KB

<> Raw Blame

## Loading Farms to Freeways from the API and ro-crate metadata file

The Language Data Commons of Australia (LDAc) packages all their data collections in an [ro-crate](#). There is a metadata file called `ro-crate-metadata.json` that comes with every data collection and this is how we can obtain metadata on this collection of research objects.

The metadata file is in the json format, and so we'll be learning how to read a json file in this notebook.

**Skills**

- json file format (see <https://en.wikipedia.org/wiki/JSON>)
- working with dataframes, via pandas
- discovering and exploring metadata
- extracting ngrams, via textacy

**Skill level:** Intermediate

This notebook uses the library 'requests', as shown in the [Using APIs: Open Australia](#) notebook. If you haven't already familiarised yourself with that notebook, it might be a good idea to do so first.

```
In [3]: # Before we begin, let's make sure that we install all the requirements that we need
import sys
!{sys.executable} -m pip install -r requirements.txt
```

```
Collecting en_core_web_sm
Using cached en_core_web_sm-3.0.0-py3-none-any.whl
Collecting matplotlib==3.4.3
Using cached matplotlib-3.4.3-cp39-cp39-manylinux2014_aarch64.whl (9.0 MB)
Collecting requests==2.26
Using cached requests-2.26.0-py2.py3-none-any.whl (62 kB)
Collecting pandas==1.3.4
Using cached pandas-1.3.4-cp39-cp39-manylinux_2_17_aarch64.manylinux2014_aarch64.whl (10.9 MB)
Collecting spacy<4.0.0,>=3.0.0
Using cached spacy-3.2.1.tar.gz (1.1 MB)
```



# Language Data Commons of Australia

## Aggregations

CLEAR

### @type

- File **349**
- OrthographicTranscription **68**
- RepositoryObject **49**
- TextDialogue **34**
- ProceduralText **15**
- Dataset **2**
- RepositoryCollection **2**

### language.name.@value

- Arabic, Standard **26**

Found 401 Results

## Child Care Subsidy: Arabic, Standard (PDF)

Contains:

Languages:

## Child Care Subsidy: Arabic, Standard(TEXT)

Contains:

Languages:

## ChildCareSubsidy: ORIGINAL English (TEXT)

Contains:

Languages:

## Child Care Subsidy: Persian, Iranian (PDF)

Contains:





Compute

HPC

Cloud

Desktop

Workspaces:

- working storage
- domain specific tools
- domain specific services



Active cleanup processes  
workspaces considered ephemeral

Data Curation  
& description

describe

analyse



**Workbench**

Notebooks  
 Data import by URL  
 Export fully described pkg  
**Stretch goals:**  
 Code gen / simple  
 interfaces eg Discursis

**Identity Management**

AAF / social media accounts

*Our demo today looks at this part ...*

publish

reuse

training

description  
information

BYOData



**Archive & Preservation Repositories**  
institutional, domain or both

Harvested

external

PARADISEC

AU Nat. Corpus

AusLan (sign)

Sydney Speaks

ATAP Corpus  
Reference, Training & BYO

ATAP Notebooks  
Apps, Code, Workflows

... etc

**Lang. portal(s)**

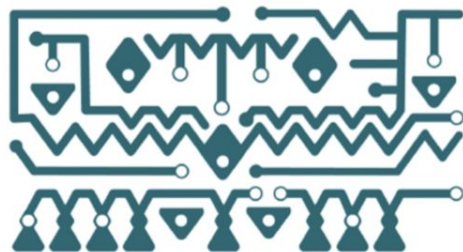
Corpus discovery  
 Item discovery  
 Authenticated API  
 create virtual corpora

Licence  
Server

**Analytics  
Portal**

Code discovery  
 Launch / Rerun  
 Data Discovery  
 Authenticated API

STORAGE (including Cloudstor)



# CARE Principles for Indigenous Data Governance

The CARE Principles for Indigenous Data Governance can be downloaded here in [summary](#) or [full](#)

The CARE Principles in Spanish - [CREA para la Gobernanza de Datos Indigenas](#)

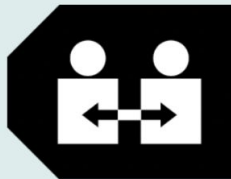
The CARE Principles in Vietnamese - [Các nguyên tắc CARE đối với quản trị dữ liệu bản địa](#)

## CARE Principles for Indigenous Data Governance

The current movement toward open data and open science does not fully engage with Indigenous Peoples rights and interests. Existing principles within the open data movement (e.g. FAIR: findable, accessible, interoperable, reusable) primarily focus on characteristics of data that will facilitate increased data sharing among entities while ignoring power differentials and historical contexts. The emphasis on greater data sharing alone creates a tension for Indigenous Peoples who are also asserting greater control over the application and use of Indigenous data and Indigenous Knowledge for collective benefit.

## Protocol Labels

Protocol Labels outline traditional protocols associated with access to this material and invite viewers to respect community protocols.



TK Men  
Restricted  
(TK MR)



TK Women  
Restricted  
(TK WR)



TK Culturally  
Sensitive  
(TK CS)



TK Secret /  
Sacred  
(TK SS)



TK Verified  
(TK V)



TK Non-Verified  
(TK NV)



TK Seasonal  
(TK S)



TK Women  
General  
(TK WG)



TK Men General  
(TK MG)



Sydney Speaks Project ▾

People +

Sydney Speakers

News & Events

Dissemination

Corpora

Sydney Speaks Apps +



## SYDNEY SPEAKS PROJECT

### Sydney Speaks: Language variation and change in a diverse society

This project seeks to document and explore Australian English, as spoken in Australia's largest and most ethnically and linguistically diverse city – Sydney. The title "Sydney Speaks" captures a key defining feature of the project: the data come from recorded conversations between Sydney siders, as they tell stories about their lives and experiences, their opinions and attitudes. This allows us to measure how their lived experiences impact their

# Sydney Speaks Licenses

- A Data can be used by other researchers and excerpts can be played in public settings
- B Data can be used by other researchers and excerpts cannot be played in public settings
- C No access (no consent for sharing with other researchers)
- D Data unavailable (not anonymised, transcription incomplete, no consent for including audio in web-based projects)

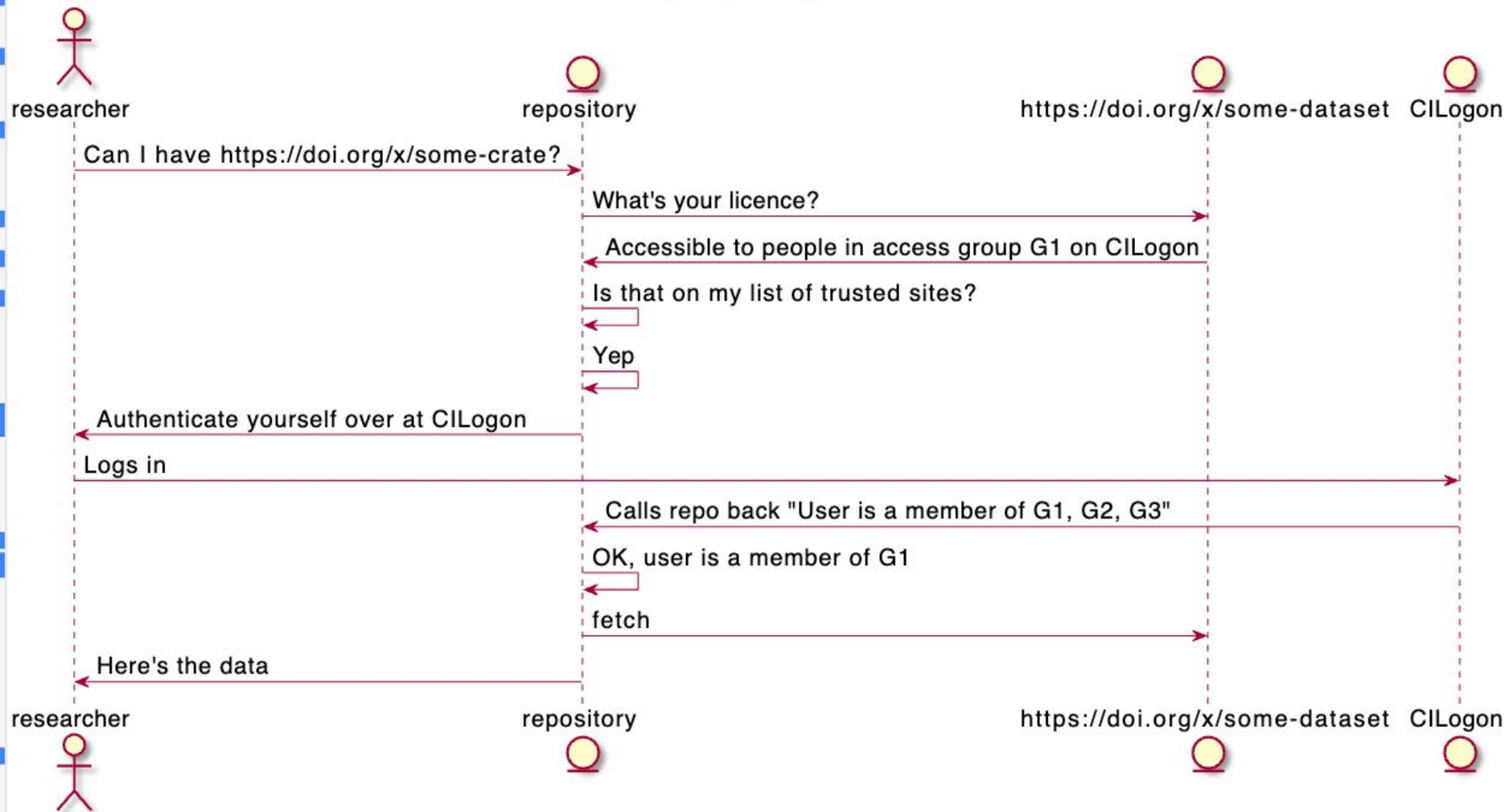
[Search](#)**Type**[CorpusItem](#) (1630)[Corpus](#) (5)[SubCorpus](#) (5)[Dataset](#) (2)[All...](#)**COOEE**[Corpus](#)**AustLit**

AustLit provides full-text access to hundreds of examples of out of copyright poetry, fiction and criticism ranging from 1795 to the 1930s. The collection includes literature intended for popular audiences as well as literature intended for audiences concerned with literary quality or the establishment of a national canon. The bibliographical information associated with these records enables researchers to investigate the relationships between texts and particular publishers or to track the first publication of each text in newspapers, magazines or journals. This provides indirect evidence of the original audience for each text and the evolution of reception over time if the texts were subsequently republished in other contexts.

[Corpus](#)**Text 1-002 1788 Phillip, Arthur**[CorpusItem](#)



# Simple repository access



# TODO

- Scope the infrastructure we need to support this (need more clarity on what data we will have and where it will be housed)
- Improve our testing for scale and implement Continuous Integration so we don't break things with every new Corpus that comes on board
- Pick our metadata terms we will probably build on the OLAC (Open Language Archives) vocabularies - but there are other options such as the CLARIN (Eu) vocabs
- Integrate better with the Australian Text Analytics Platform ATAP - eg fire up a notebook from the search portal to operate on a collection of interest