HASS RDC Technical Advisory Group Meeting

LDaCA & ATAP

Intro

Peter Sefton - p.sefton@uq.edu.au
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.
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

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://orcid.org/0000-0001-2345-6789

https://en.wikipedia.org/wiki/Scanning_electron_microscope
Portland Common Data Model

Diagram of data model:
- pcdm:Collection
  - pcdm:hasMember (m:m)
  - ore:aggregates (m:m)
- pcdm:Object
  - pcdm:hasFile (1:m)
  - pcdm:hasRelatedFile (1:m)
  - pcdm:hasMember (m:m)
  - ore:aggregates (m:m)
- pcdm:File
  - Access
  - Bitstream
  - Descriptive
  - Technical
RO-Crates MUST have licence information that sets out conditions for use/reuse of the data.
Collection of Australian Government Documents in six languages

DisabilitySupportPension

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

name: DisabilitySupportPension

@type: [RepositoryObject, ProceduralText]

author: CommonwealthGovernment

hasFile: [Arabic, Standard (PDF), Arabic, Standard (TEXT), DisabilitySupportPension: ORIGINAL English (TEXT), Persian, Iranian (PDF), Persian, Iranian (TEXT), Engelilik Destek Emeklliik Maasi: Turkish (PDF), Engelilik Destek Emeklliik Maasi: 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)

Disability Support Pension (DSP) هو نوع من معاشات الدعم الاحترافية التي تقدمها الدولة لمن يعانون من مشاكل صحية قوية أو عقلي أو نفسية.

**المواصفات**

- يجب أن يكون المتقاعد من العمر 65 عامًا على الأقل.
- عند وفاة المتقاعد في سن ما بعد 65، يتم الاستمرار في الدعم للمعاقين المتفقين.
- البرنامج يحتوي على برامج دعم مالية ورعاية صحية.

**البرامج**

- برامج الدعم المالي:
  - قسط مالي:
  - دعم جزئي:
- برامج الرعاية الصحية:
  - الرعاية المنزلية:
  - الرعاية المكملة:

**الدوام**

- يجب أن يتم تقديم الدعم بشكل دوري لفترة معينة.
- إذا لم يتم تقديم الدعم بشكل دوري، فقد يسمح برفع مبلغ الدعم.

**العوامل**

- أناقش بشكل عام عن فوائد البرامج وشركات الرعاية الصحية ومتطلبات الدعم.

**التعويضات**

- تعويضات مالية:
  - دعم مالي:
  - دعم جزئي:
- تعويضات صحية:
  - الرعاية المنزلية:
  - الرعاية المكملة:

**الملاحظات**

- إذا كنت متقاعدًا، يُرجى تقديم طلب لتقديم الدعم.
- إذا كنت متقاعدًا، يُرجى تقديم طلب لتقديم الدعم.
- إذا كنت متقاعدًا، يُرجى تقديم طلب لتقديم الدعم.

**نقرة على الملف**

إليك رابط مباشر لتنزيل الملف PDF:

```
@id
DisabilitySupportPension/12626-1506ar.pdf
```

**نوع الملف**

File

**هيكل الملف**

- Acrobat PDF 1.5 - Portable Document Format
- application/pdf

**المحتوى**

404819

**التاريخ**

2022-01-05T12:01:37+11:00

**اللغة**

Arabic, Standard

**ترجمة العمل**

DisabilitySupportPension

**الحجم**

404819
Repositories: institutional, domain or both

Find / Access services

Research
Data
Management
Plan

Workspaces:
- working storage
- domain specific tools
- domain specific services

● working storage
● domain specific tools
● domain specific services

collect
describe
analyse

Reusable, Interoperable data objects
➔ deposit early
➔ deposit often

Findable, Accessible, Reusable data objects

reuse data objects

Active cleanup processes
workspaces considered ephemeral

Policy based data management

V1.1 © Marco La Rosa, Peter Sefton 2021 https://creativecommons.org/licenses/by-sa/4.0/
Active cleanup processes: Workspaces considered ephemeral

Data Curation & description

Data Cleaning
OCR / transcription
format migration

BYODaTa 🥂

Identity Management
AAF / social media accounts

Analytical Portal
Code discovery
Launch / Rerun
Data Discovery
Authenticated API
Create virtual corpora

Analytics Server

Archive & Preservation Repositories
institutional, domain or both

STORAGE (including Cloudstor)

Compute
HPC
Cloud
Desktop

Apps, Code, Workflows
Deposit / Publish

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

Workspaces:
• working storage
• domain specific tools
• domain specific services

Active cleanup processes

Lang. portal(s)
Corpus discovery
Item discovery
Authenticated API
Create virtual corpora

Licence Server

Harvested
PARADISEC

AU Nat. Corpus

AusLan (sign)

Sydney Speaks

ATAP Corpus
Reference, Training & BYO

ATAP Notebooks
Apps, Code, Workflows

... etc

Data Discovery
Authenticated API
Create virtual corpora

Data Curation & description

Reuse

Identity Management
AAF / social media accounts

STORAGE (including Cloudstor)
Talking mainly about this bit today
ATAP analytics (notebooks) ➔ «oni» Full text search index ➔ «oni» LDaCA Search / browse portal ➔ «oni» Data & Access API ➔ 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.
There may be several portals using the same index - e.g., to give collection-specific advanced search.

There may be several distributed file-based repositories feeding the same index.

There may be other kinds of index such as triple stores or relational databases that index tabular data.

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.
<table>
<thead>
<tr>
<th>Repository Name</th>
<th>Visibility</th>
<th>Description</th>
<th>Language</th>
<th>License</th>
<th>Stars</th>
<th>forks</th>
<th>watchers</th>
<th>Updated</th>
</tr>
</thead>
<tbody>
<tr>
<td>corpus-tools-cooe</td>
<td>Private</td>
<td>Corpus prep tools for the COEE corpus (using the spreadsheet that comes with the corpus)</td>
<td>JavaScript</td>
<td>GPL-3.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Updated now</td>
</tr>
<tr>
<td>corpus-tools-farms-to-freeways</td>
<td>Private</td>
<td>Node scripts to build an Arkisto-ready language data collection from the &quot;From Farms to Freeways&quot; history project data.</td>
<td>JavaScript</td>
<td>GPL-3.0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>Updated 2 minutes ago</td>
</tr>
<tr>
<td>corpus-tools-python3-soffice</td>
<td>Public</td>
<td>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</td>
<td>Python3</td>
<td>GPL-3.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Updated 5 days ago</td>
</tr>
<tr>
<td>corpus-tools-gov.au-multilingual</td>
<td>Private</td>
<td>Node tools to prepare an Arkisto repository with parallel translations of Australian Government documents</td>
<td>JavaScript</td>
<td>GPL-3.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Updated 19 days ago</td>
</tr>
<tr>
<td>corpus-tools-sydney-speaks</td>
<td>Private</td>
<td>Node tools to prepare an Arkisto repository with language data from the Sydney Speaks project</td>
<td>JavaScript</td>
<td>GPL-3.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Updated on 16 Dec 2021</td>
</tr>
<tr>
<td>corpus-tools-monash-coe</td>
<td>Private</td>
<td>Corpus tools to prepare the Monash Corpus of English for ingest into RO-Crate and OCFL</td>
<td>JavaScript</td>
<td>GPL-3.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Updated on 8 Dec 2021</td>
</tr>
<tr>
<td>corpus-tools-alveo-export</td>
<td>Private</td>
<td>Node tools to export Alveo collections as Arkisto-ready repository objects</td>
<td>JavaScript</td>
<td>GPL-3.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Updated on 8 Sep 2021</td>
</tr>
<tr>
<td>Identifier</td>
<td>Title</td>
<td>Created Date</td>
<td>Type(s)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>-------</td>
<td>--------------</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coose:4-281</td>
<td>unspecified</td>
<td>1892</td>
<td>Original, Text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coose:3-163</td>
<td>unspecified</td>
<td>1858</td>
<td>Original, Text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coose:4-008</td>
<td>unspecified</td>
<td>1878</td>
<td>Original, Text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coose:2-068</td>
<td>unspecified</td>
<td>1831</td>
<td>Text, Original</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coose:4-045</td>
<td>unspecified</td>
<td>1883</td>
<td>Text, Original</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coose:4-024</td>
<td>unspecified</td>
<td>1882</td>
<td>Text, Original</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coose:1-259</td>
<td>unspecified</td>
<td>1825</td>
<td>Original, Text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coose:2-026</td>
<td>unspecified</td>
<td>1829</td>
<td>Text, Original</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coose:3-090</td>
<td>unspecified</td>
<td>1854</td>
<td>Text, Original</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coose:2-351</td>
<td>unspecified</td>
<td>1849</td>
<td>Text, Original</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coose:4-177</td>
<td>unspecified</td>
<td>1888</td>
<td>Text, Original</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coose:4-124</td>
<td>unspecified</td>
<td>1887</td>
<td>Text, Original</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coose:2-239</td>
<td>unspecified</td>
<td>1841</td>
<td>Text, Original</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coose:4-402</td>
<td>unspecified</td>
<td>1900</td>
<td>Original, Text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coose:3-273</td>
<td>unspecified</td>
<td>1871</td>
<td>Original, Text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coose:4-381</td>
<td>unspecified</td>
<td>1897</td>
<td>Original, Text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coose:4-397</td>
<td>unspecified</td>
<td>1899</td>
<td>Text, Original</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coose:4-330</td>
<td>unspecified</td>
<td>1896</td>
<td>Text, Original</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mode</td>
<td>Value</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------</td>
<td>----------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speech Style</td>
<td>unspecified</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Publication Status</td>
<td>unspecified</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Written Mode</td>
<td>unspecified</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interactivity</td>
<td>unspecified</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Communication Context</td>
<td>unspecified</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Communication Medium</td>
<td>unspecified</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Click Send to get a response
Loading Farms to Freeways from the API and ro-crate metadata file

The Language Data Commons of Australia (LDaCo) 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.

```
# Before we begin, let's make sure that we install all the requirements that we need
!pip install -r requirements.txt
```

Collecting en_core_web_sm
  Using cached en_core_web_sm-3.0.0-py3-none-any.whl
Collecting spacy-2.2.0
  Using cached spacy-2.2.0-py2.py3-none-any.whl (2.4 MB)
Language Data Commons of Australia

Found 401 Results

- Child Care Subsidy: Arabic, Standard (PDF)
  - Contains: File
  - Languages: Arabic, Standard

- Child Care Subsidy: Arabic, Standard (TEXT)
  - Contains: File
  - Languages: Arabic, Standard

- Child Care Subsidy: ORIGINAL English (TEXT)
  - Contains: File
  - Languages:

- Child Care Subsidy: Persian, Iranian (PDF)
  - Contains: File
  - Languages:
Our demo today looks at this part...
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 Indígenas

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

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 language use.
Sydney Speaks Licenses

Data can be used by other researchers and excerpts can be played in public settings
A public settings

Data can be used by other researchers and excerpts cannot be played in public settings
B played in public settings

No access (no consent for sharing with other researchers)
C No access (no consent for sharing with other researchers)

Data unavailable (not anonymised, transcription incomplete, no consent for including audio in web-based projects)
D consent for including audio in web-based projects
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.
Simple repository access

Can I have https://doi.org/x/some-crate?

What's your licence?

Accessible to people in access group G1 on CILogon

Is that on my list of trusted sites?

Yep

Authenticate yourself over at CILogon

Logs in

Calls repo back "User is a member of G1, G2, G3"

OK, user is a member of G1

fetch

Here's the data
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