

# Low friction FAIR interoperability using RO-Crate metadata in text analytics pipelines

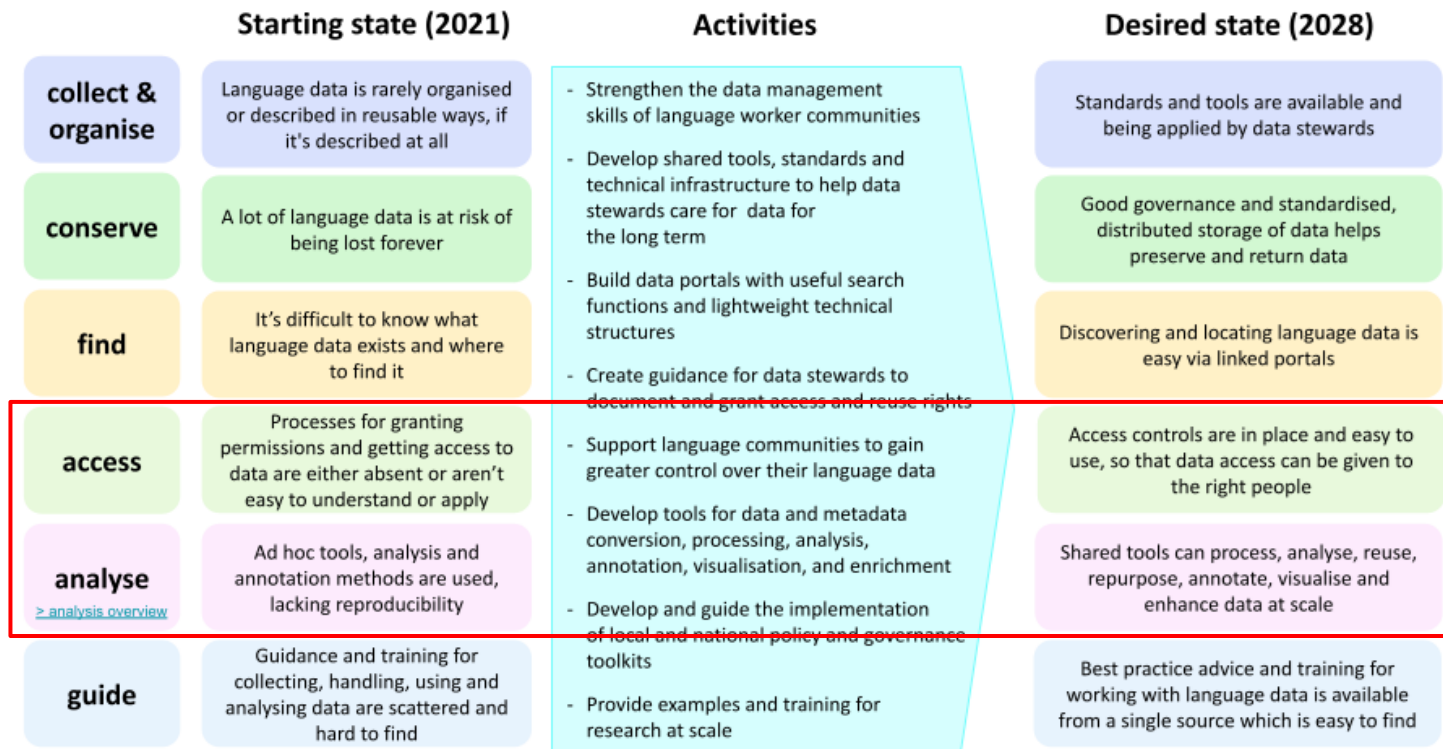
Rosanna Smith, Mike Lynch, Peter Sefton, Simon Musgrave,  
River Tae Smith

# About Us

The Language Data Commons of Australia (LDaCA) is part of the Humanities and Social Sciences and Indigenous Research Data Commons which is led by the Australian Research Data Commons (ARDC).

Version: 2025-07-31

## LDaCA Execution Strategy Overview



# LDaCA Analyse - Strategic Overview

## Starting state (2021)

## Activities

## Desired state (2028)

**transparent**

Analytical workflows are typically not published, re-runnable or reusable

- Document, demonstrate and teach methods for publishing findable, (re)usable and readable research code

Researchers can use tools and processes to publish, find, (re)use and adapt computational methods to new contexts

**documented**

(Meta)Data formats, tools, research workflows are varied, and under-documented

- Train researchers in data management, standardised data formats, preparation, transformation and wrangling of data for analysis

Documentation and training programs available to help researchers adopt appropriate standards

**findable**

Appropriate implementations of analytical methods are hard to find

- Document methods and develop toolkits to transform BYO (meta)data to standard formats without compromising data integrity

LDaCA infrastructure is interconnected, with suitable interfaces, data formats and guidance on appropriate usage

**adaptable**

Methods are specialized to particular studies or research cohorts

- Train researchers in computational methods application and development
- Develop guidance and train researchers on how to choose appropriate analytical approaches eg ethical and appropriate use of AI - raise awareness of computational methods

Key methods and workflows are adaptable to work in different research contexts with documentation of their uses and limitations

**contextually appropriate**

It is unclear when and how methods can and should be (re)used in different research contexts

- Identify promising methods, practices and workflows, including emerging methods (AI) and adapt them across research contexts ethically and appropriately

Researchers are more aware of computational methods, and can use LDaCA guidance to match appropriate methods to their and others' data.

**connected**

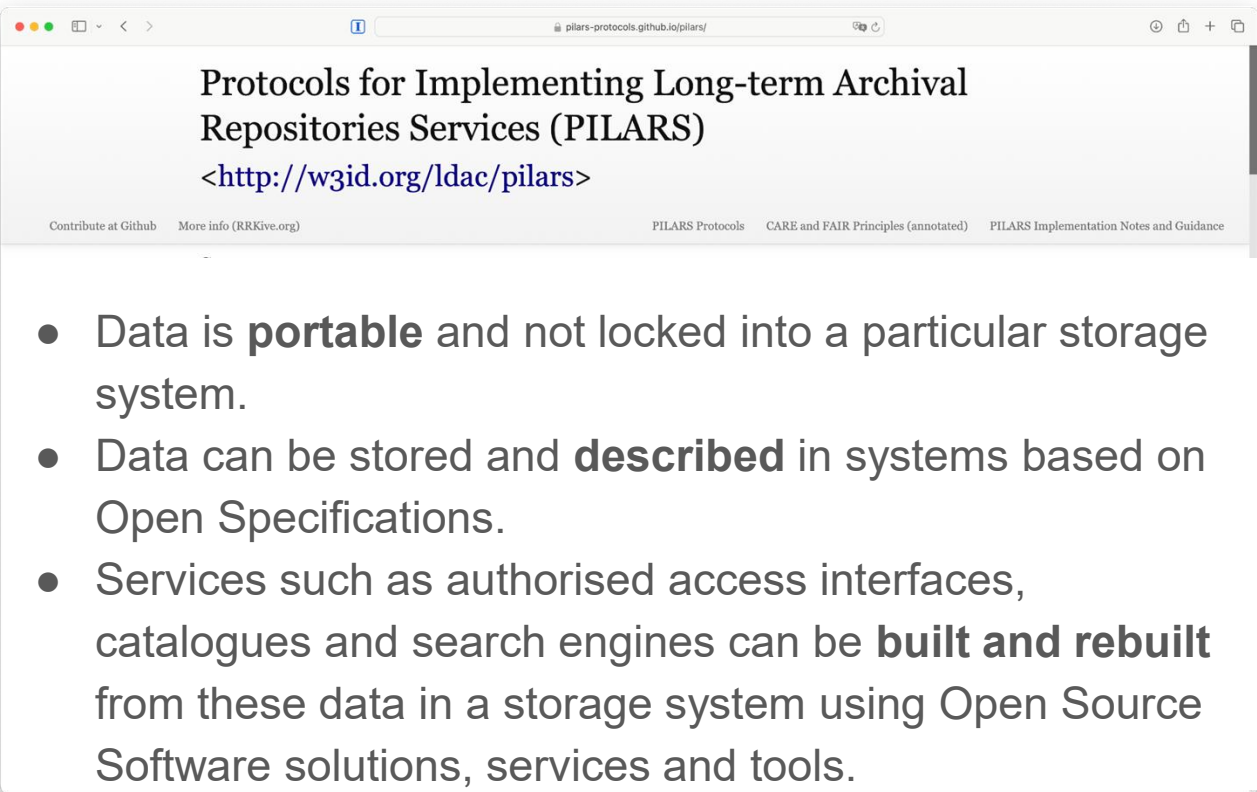
There are many analytical tools available but they require different input formats

- Develop and demonstrate end-to-end best-practice workflows to connect researchers, data and computational tools

There are readily accessible, self documenting connectors making it easy to apply analytical methods

# Implementation

The LDaCA architecture is implemented using the Protocols for Implementing Long-Term Archival Repository Services (PILARS)



The screenshot shows a web browser window with the URL <https://w3id.org/ldac/pilars/>. The page title is "Protocols for Implementing Long-term Archival Repositories Services (PILARS)". Below the title is a navigation bar with links: "Contribute at Github", "More info (RRKive.org)", "PILARS Protocols", "CARE and FAIR Principles (annotated)", and "PILARS Implementation Notes and Guidance". The main content area features a bulleted list of three key principles:

- Data is **portable** and not locked into a particular storage system.
- Data can be stored and **described** in systems based on Open Specifications.
- Services such as authorised access interfaces, catalogues and search engines can be **built and rebuilt** from these data in a storage system using Open Source Software solutions, services and tools.

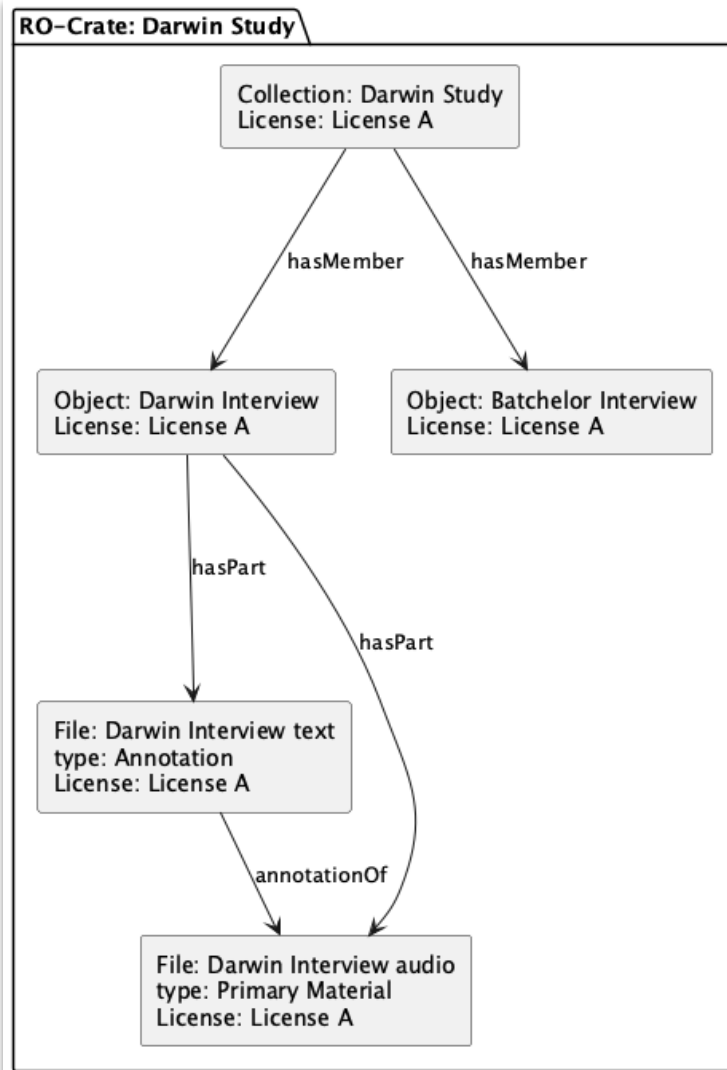
<https://w3id.org/ldac/pilars>

# Storage

Storage Objects are deposited in a repository. In LDaCA each storage object is an RO-Crate.

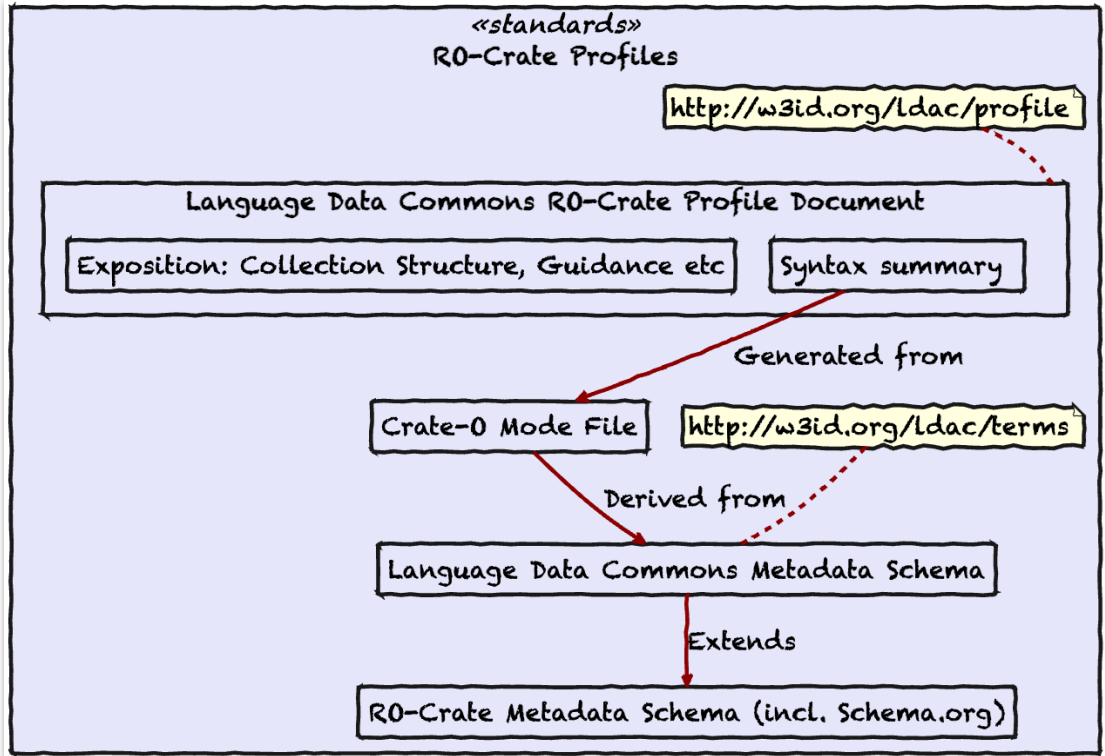
An RO-Crate is a Research Object (or RO) formed of a collection of data (a crate), a special `ro-crate-metadata.json` file which describes the collection and its license information.

The `ro-crate-metadata.json` file is a JSON-LD metadata file at the root of an RO-Crate that describes the crate, its contents, and their relationships in a machine-readable way.



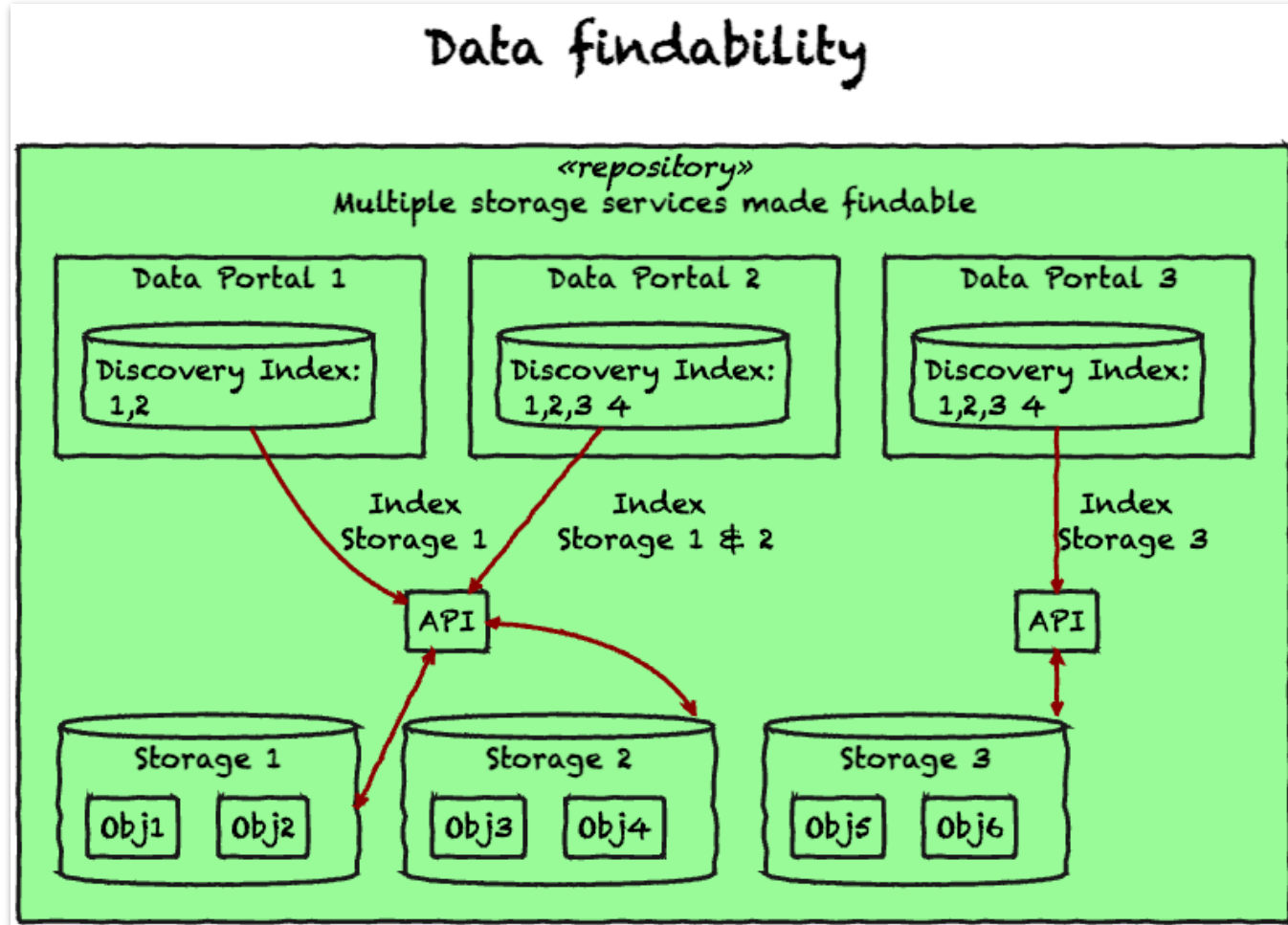
# Data Annotation

1. A Metadata Profile describes how storage objects should be modelled, and how files should be described.  
<https://w3id.org/ldac/profile>
2. This profile draws on the Language Data Commons Schema – a Schema.org Style set of terms for describing language data in an Archival Repository and for data interchange.  
<https://w3id.org/ldac/terms>
3. LDaCA uses Research Object Crate (RO-Crate) metadata to describe each storage object. Each OCFL object has an RO-Crate metadata document (ro-crate-metadata.json), making it an RO-Crate.



# Index

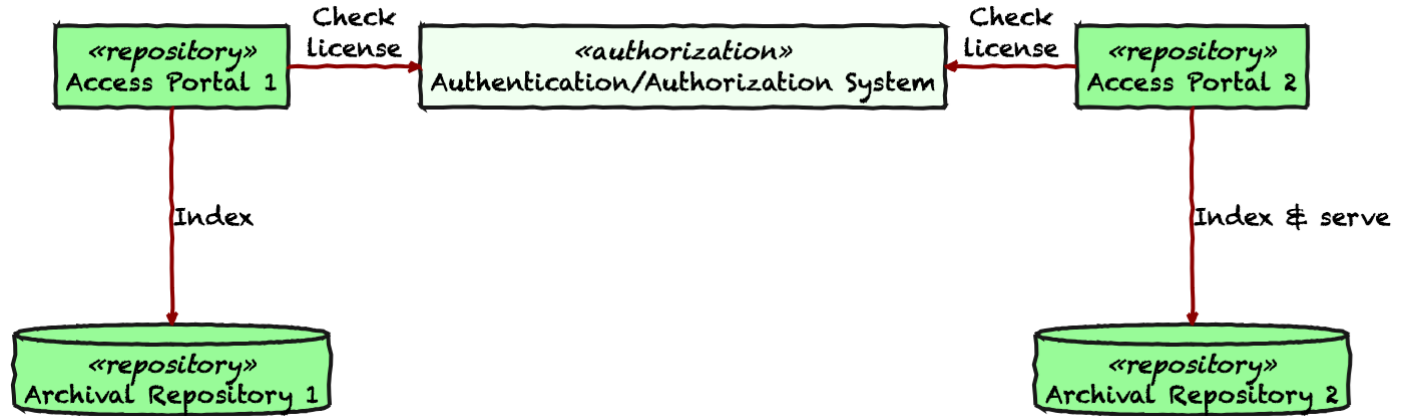
Portals can be indexed from the storage to make them findable.



# Distributed Access Control

## Motivation

- FAIR (Findable, Accessible, Interoperable, Reusable) data principles require not just openness but **controlled access** in many contexts.
- Traditional centralized access control solutions struggle with scalability, cross-institutional trust, privacy, and fine-grained permissions.



# Portal(s)

Main portal:  
[data.ldaca.edu.au](http://data.ldaca.edu.au)

Terraform  
automation allows  
for additional  
portals to be built  
on demand.

The screenshot displays the LDCA (Language Data Collection Australia) portal interface. At the top left, there is a navigation menu with 'Home' and the LDCA logo. The top right contains links for 'Collections', 'Notebooks', 'Browse', 'Login', and 'Help'. A search bar is located on the left side of the main content area, with a search icon and an 'Advanced Search beta' button. Below the search bar, there are several filter panels: 'Main Collections' with a list of collections and their counts, 'Access' with a count of 10, 'Record Type' with 'RepositoryCollection' selected (count 15), and 'Language' with a count of 31. The main content area shows search results for 'RepositoryCollection' with 15 index entries. The first result is 'A Corpus of Oz Early English (COOEE)', a Dataset in English, with 1354 objects and 2708 files. The second result is 'AusReddit aggregated data - Collection', also a Dataset in English, with 6 objects. The third result is 'Australian Corpus of English', a Dataset in English, with 845 objects and 1707 files. Each result includes a brief description and a 'See more' link. On the right side of the results, there are icons for a document, a microphone, a pencil, and a document with a checkmark.

# Analysis

## COOEE Notebook

<b>Name</b>	COOEE Notebook
<b>Description</b>	A topic modeling notebook for the cooee collection
<b>Date Published</b>	Not Defined
<b>ID</b>	<a href="#">cooee.ipynb</a>
<b>Author</b>	Smith, Rosanna Musgrave, Simon Smith, River Tae
<b>Base64</b>	

### Notebook Viewer

#### Corpus of Oz Early English (COOEE)

COOEE is a collection of texts produced in Australia between 1788 and 1900. For each of four time periods (1788-1825, 1826-1850, 1851-1875, 1876-1900), the number of tokens included in the corpus is approximately equal. The corpus is also divided into four genres of material (Private Written, Public Written, Government English, Speech-Based) and the proportions of these type of materials is consistent for each time period. This organisation means that the corpus can be stratified into 16 sections to see whether linguistic features vary according to either or both of the variables.

This notebook illustrates how the corpus can be accessed via the Language Data Commons of Australia API and then how the downloaded data can be reconfigured as a flat tabular structure which makes the metadata variables easy to access. The notebook also demonstrates one way to split the data into 16 stratified sub-corpora which are then used as the basis for topic modeling. The final result is that we can make a visualisation showing what topics are more or less strongly associated with particular sub-corpora.

### Downloads

This item does not have a direct download link.

[Show All Related Downloads](#)

### Citation

[View citation details for this item](#)

### Try this Notebook

LDaCA-ATAP BinderHub   

MyBinder Public  
BinderHub

### Takedown Request

If you see an item on this page that you think should not be made public, you can request that it be taken down:

[Takedown Request Form](#)

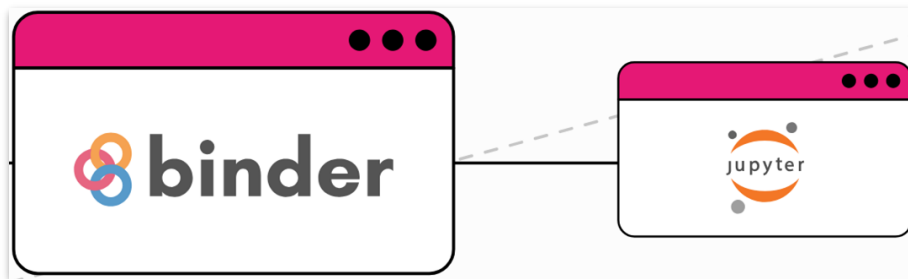
# Reproducible Analysis

Jupyter notebooks can break due to:

- Library upgrades
- Version changes
- Missing credentials
- Undocumented requirements

Mitigate this with BinderHub:

- Launch pre-configured notebooks as interactive computing environments
- Explicitly defined hardware and software requirements



# Example: A COrpus of Oz Early English (COOEE)

A collection of texts written in Australia between 1788 and 1900.

Divided into four time periods:

- Period 1: 1788-1825
- Period 2: 1826-1850
- Period 3: 1851-1875
- Period 4: 1876-1900

Contains material from four registers:

- Speech-based (SB)
- Private written (PrW)
- Public written (PcW)
- Government English (GE)

The screenshot shows the project page for 'A COrpus of Oz Early English (COOEE)' on the LdCA website. The page features a header with the LdCA logo and navigation links for Collections, Notebooks, Browse, and Login. The main content area is divided into sections for Name, Description, Date Published, ID, Accountable Person, Author, Related Works, and Conforms To. The Description section provides a detailed overview of the corpus, including its stratification into four time periods and registers. The Access section states that the material is free to share and adapt under Attribution 4.0 International. The Content section lists the language as English and the communication mode as WrittenLanguage and SpokenLanguage. The Downloads section provides information about the corpus files, including the number of files (2716), size (84.13 MB), and a download button.

Home **LdCA** Collections Notebooks Browse Login

## A COrpus of Oz Early English (COOEE)

<b>Name</b>	A COrpus of Oz Early English (COOEE)
<b>Description</b>	Material to be included had to meet with a regional and a temporal criterion. The latter required texts to have been produced between 1788 and 1900 in order to become eligible for COOEE. It was mandatory for a text to have been written in Australia, New Zealand or Norfolk Island. But in a few cases, other localities were allowed. For example, if a person who was a native Australian or who had lived in Australia for a considerable time, wrote a shipboard diary or travelled in other countries. Contains: Letters, published materials in book form, historical texts. The collection is stratified in two ways: Time period - The corpus is divided into four time periods: Period 1: 1788-1825 Period 2: 1826-1850 Period 3: 1851-1875 Period 4: 1876-1900 The initial numeral of each file name indicates the period from which the document comes. Register - The corpus contains material from four registers: Speech-based (sb) Private written (prw) Public written (pcw) Government English (ge) The register to which a file belongs is specified in the metadata at the start of each file in the form <=[register]> using the abbreviations above. This collection was previously accessible online via the Australian National Corpus (AusNC), an initiative managed by Griffith University between 2012 and 2023. Some texts that were included in COOEE do not have digital version, records for these texts are included here for completeness
<b>Date Published</b>	2012
<b>ID</b>	<a href="arcp://name,hdl10.26180-23961609">arcp://name,hdl10.26180-23961609</a>
<b>Accountable Person</b>	<a href="#">Clemens W. A. Fritz</a>
<b>Author</b>	<a href="#">Clemens W. A. Fritz</a>
<b>Related Works</b>	<a href="#">From English in Australia to Australian English</a>
<b>Conforms To</b>	<a href="#">con forms 1b</a>

### Access

You are free to share (copy and redistribute the material in any medium or format) and adapt (remix, transform and build upon the material for any purpose, even commercially).

[Attribution 4.0 International \(CC BY 4.0\)](#)

Public Metadata and Text is Searchable

### Content

Language  
English

Communication Mode  
[WrittenLanguage](#)  
[SpokenLanguage](#)

File Formats  
text/plain

### Downloads

A COrpus of Oz Early English (COOEE).zip  
Files: 2716, Size: 84.13 MB  
[Attribution 4.0 International \(CC BY 4.0\)](#)

[Download](#)

[Show All Related Downloads](#)

# COOEE Notebook - Metadata Standardisation

Stratifying COOEE across time period and register makes 16 sub-corpora, allowing for comparative analysis such as topic modeling.

Metadata standardisation required to reduce friction, increase interoperability:

- Mapping of collection-specific terms to [schema.org](https://schema.org) standard:
  - **Birth** → **birthDate**
  - **Gender** → **gender**
  - **Nr** → **identifier**
- Identifying the main text of the collection for analysis with the [Language Data Commons Schema](#) term **ldac:mainText**, e.g.

```
"ldac:mainText": {  
  "@id": "data/1-001-plain.txt"  
}
```

# COOEE Notebook - Downloading and preparing the data

- Download collection from LDaCA Portal.
- Use RO-Crate tabulator to create a table including both text data and metadata from the **RepositoryObject** entities, focussing on **Idac:mainText**.
- Convert to Pandas DataFrame
- Slice dataframe by register and time period to create 16 documents.

# Analysis

Because of the nature of the linked data format, we have built tools to convert linked data to tabular forms.

## RO-Crate tabulator

Python library to turn an RO-Crate into tabular formats.

### Installation

Install [uv](#), then

```
> git clone git@github.com:Sydney-Informatics-Hub/rocrate-ta
> cd rocrate-tabular
> uv run tabulator --help
```

`uv run` should create a local venv and install the dependencies




### Usage

First pass: this will scan an RO-Crate directory, build a `properties` table in the sqlite database `crate.db`, and generate a config file with a list of all available tables in the `potential_tables` section

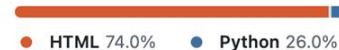
```
> uv run tabulator -c config.json ./path/to/crate crate.db
```

You can then edit the config file and move the tables you want to create in the database and/or csv to the `tables` section, and re-run the tabulator

```
> uv run tabulator -c config.json ./path/to/crate crate.db
```

-  **spikelynch** Mike Lynch
-  **r-tae** River
-  **moisbo** Moises Sacal
-  **h-croser** Hamish Croser

### Languages



# COOEE Notebook - Tokenization

- Natural Language Toolkit (NLTK)
- Documents converted to lists of words
- Removed punctuation marks, numbers, new line symbols and common words (e.g. *the*, *and*, *of*, etc.)

```
\nI have not much news since  
I wrote, except that the  
weather is now beautiful,  
and in consequence the gold  
frenzy has burst forth now  
in full force,
```

```
'much',  
'consequence',  
'news',  
'gold',  
'since',  
'frenzy',  
'write',  
'burst',  
'except',  
'forth'.
```

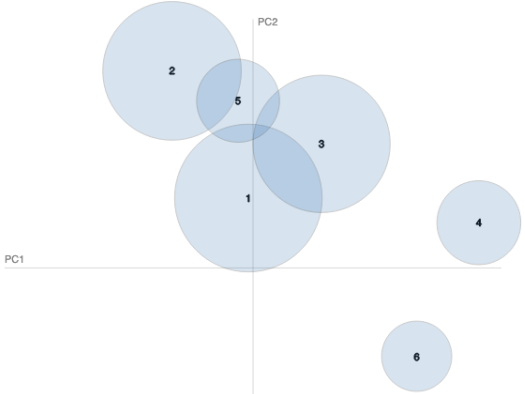
# COOEE Notebook - Visualising the data

```
pyLDAvis.gensim.prepare(lda_model, corpus, dictionary, mds='mmds')
```

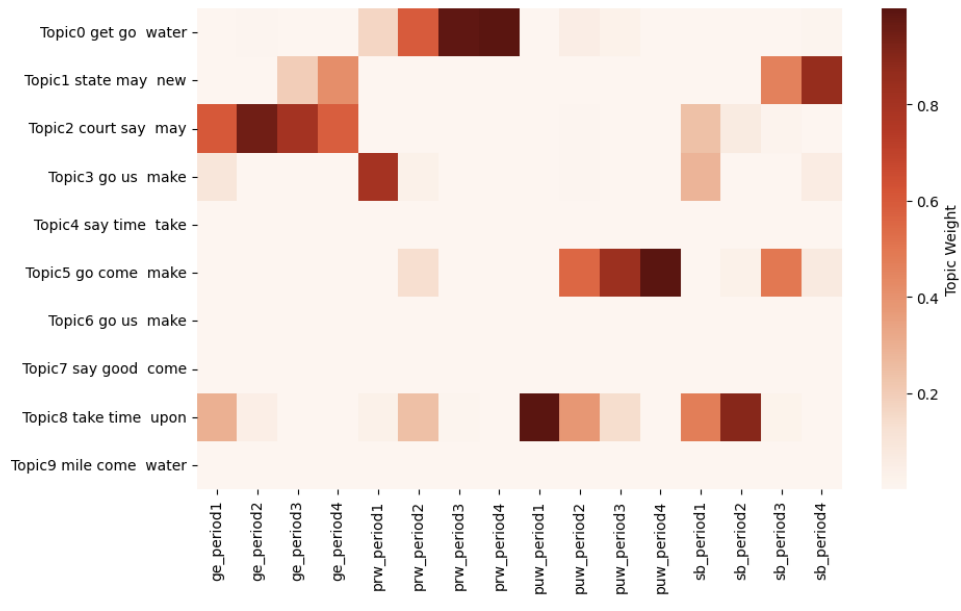
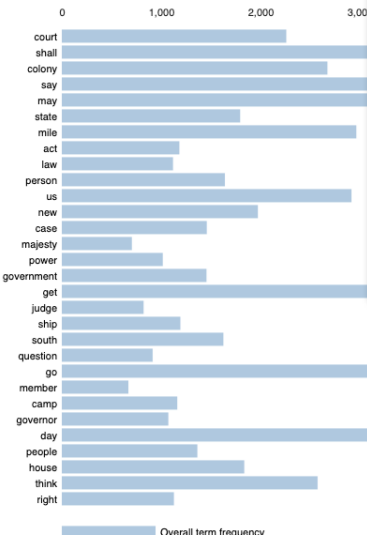
Selected Topic:

Slide to adjust relevance metric:<sup>(2)</sup>   $\lambda = 1$

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Salient Terms



Marginal topic distribution

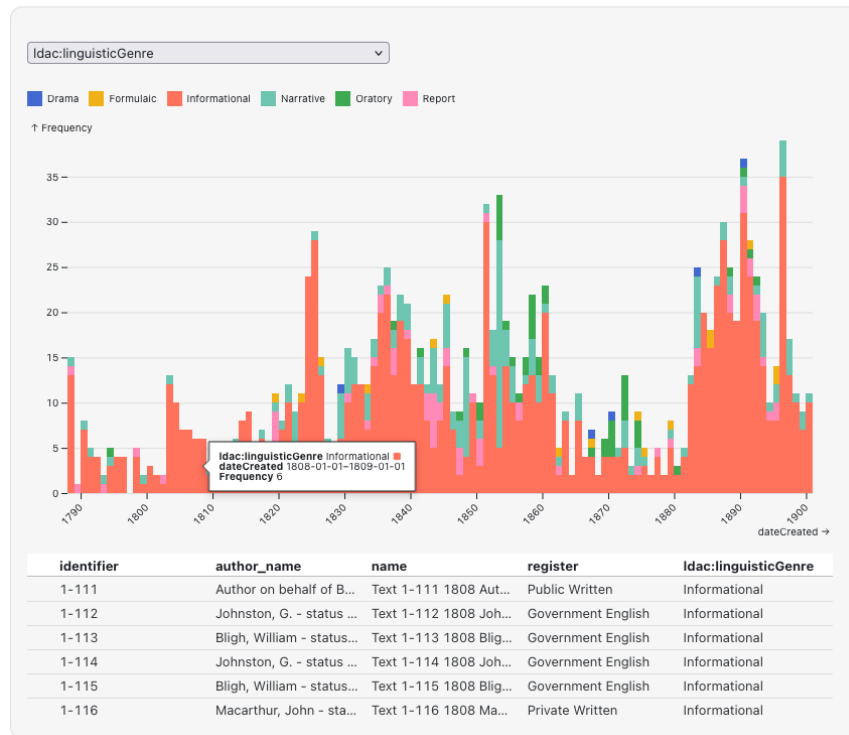


# Tabulator

- Observable v RO-Crates?
- Needed a tool to turn a JSON-LD graph into tables

```
"Tables": {  
  "RepositoryObject": [...],  
  "Person" [...],  
  ..  
}
```

## COOEE



# Annotated CSV exports

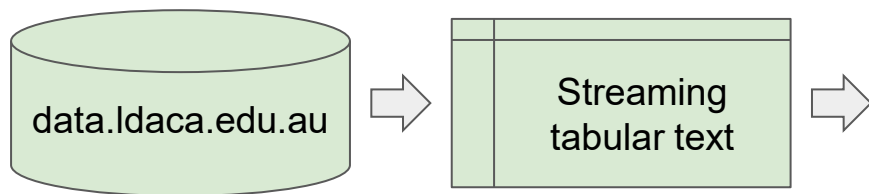
	A	B	C	D	E	F	G	H	I
1	entity_id	@type	name	birthDate	local:birthDateEstimateStart	local:birthDateEstimateEnd	birthPlace	birthPlace_id	gender
2	https://www.pe	Person	Clemens W. A. Fritz						
3	arcpc://name,hd	Person	Philip, Arthur - status 1788 text #1-001	1738	1738	1738	Great Britain	#place_GB	m
4	arcpc://name,hd	Person	Philip, Arthur	1738	1738	1738	Great Britain	#place_GB	m



```
{  
  "name": "author_local:birthDateEstimateStart",  
  "label": "local:birthDateEstimateStart",  
  "propertyUrl": "arcpc://name,hd10.26180~23961609/terms#birthDateEstimateStart",  
  "description": "The start of the range of possible birth dates for a person - this is  
used when the birth date field was specified to the decade like 188x",  
  "@id": "#COLUMN_documents.csv_author_local:birthDateEstimateStart",  
  "@type": "csvw:Column"  
}
```

# Getting specific

- Tabulator: general
- Tabulator-LDaCA: takes advantage of common features



The screenshot displays the "LDaCA Corpus Analysis" web interface. The top navigation bar includes the logo, the title "LDaCA Corpus Analysis", and a "Logout" link. Below the navigation bar is a sidebar menu with the following items: "Views", "Data Loader", "Data Preprocessing", "Token Frequency", "Concordance", "Timeline", "Topic Modeling", "Quotation" (which is highlighted), and "Export". The main content area is titled "Quotation Extraction" and contains the instruction: "Load quotations for a single node and highlight speaker, quote, and verb spans." Below this, there is a section for "Selected Nodes (1/1)" with a lock icon. It shows a selected node: "sample\_data/ADO/qdelection2020\_candidate\_tweets" with a shape of "2380 x 15" and a "Text Column:" dropdown menu set to "text". At the bottom of the main area, there is a message: "Analysis locked to the last request. Clear results to unlock and resync node choices." and two buttons: "Load Quotations" and "Clear Results".

## LDaCA Execution Strategy Overview

	Starting state (2021)	Activities	Desired state (2028)
<b>collect &amp; organise</b>	Language data is rarely organised or described in reusable ways, if it's described at all	<ul style="list-style-type: none"> <li>- Strengthen the data management skills of language worker communities</li> <li>- Develop shared tools, standards and technical infrastructure to help data stewards care for data for the long term</li> <li>- Build data portals with useful search functions and lightweight technical structures</li> <li>- Create guidance for data stewards to document and grant access and reuse rights</li> <li>- Support language communities to gain</li> </ul>	Standards and tools are available and being applied by data stewards
<b>conserve</b>	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
<b>find</b>	It's difficult to know what language data exists and where to find it		Discovering and locating language data is easy via linked portals
<b>access</b>	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
<b>analyse</b>	Ad hoc tools, analysis and annotation methods are used, lacking reproducibility		
<a href="#">&gt; analysis overview</a>			
<b>guide</b>	Guidance and training for collecting, handling, using and analysing data are scattered and hard to find		

Version: 2025-07-31

## LDaCA Analyse - Strategic Overview

	Starting state (2021)	Activities	Desired state (2028)
<b>transparent</b>	Analytical workflows are typically not published, re-runnable or reusable	<ul style="list-style-type: none"> <li>- Document, demonstrate and teach methods for publishing findable, (re)usable and readable research code</li> </ul>	Researchers can use tools and processes to publish, find, (re)use and adapt computational methods to new contexts
<b>documented</b>	(Meta)Data formats, tools, research workflows are varied, and under-documented	<ul style="list-style-type: none"> <li>- Train researchers in data management, standardised data formats, preparation, transformation and wrangling of data for analysis</li> <li>- Document methods and develop toolkits to transform BYO (meta)data to standard formats without compromising data integrity</li> </ul>	Documentation and training programs available to help researchers adopt appropriate standards
<b>findable</b>	Appropriate implementations of analytical methods are hard to find	<ul style="list-style-type: none"> <li>- Train researchers in computational methods application and development</li> <li>- <del>Develop guidance and train researchers on how to</del></li> </ul>	LDaCA infrastructure is interconnected, with suitable interfaces, data formats and guidance on appropriate usage
<b>adaptable</b>	Methods are specialized to particular studies or research cohorts	<ul style="list-style-type: none"> <li>- choose appropriate analytical approaches eg ethical and appropriate use of AI - raise awareness of computational methods</li> <li>- Identify promising methods, practices and workflows, including emerging methods (AI) and adapt them across research contexts ethically and appropriately</li> </ul>	Key methods and workflows are adaptable to work in different research contexts with documentation of their uses and limitations
<b>contextually appropriate</b>	It is unclear when and how methods can and should be (re)used in different research contexts	<ul style="list-style-type: none"> <li>- Develop data connections that link Language Data Commons compliant data - in portal and BYO - to analytical tools</li> </ul>	Researchers are more aware of computational methods, and can use LDaCA guidance to match appropriate methods to their and others' data.
<b>connected</b>	There are many analytical tools available but they require different input formats	<ul style="list-style-type: none"> <li>- Develop and demonstrate end-to-end best-practice workflows to connect researchers, data and computational tools</li> </ul>	There are readily accessible, self documenting connectors making it easy to apply analytical methods