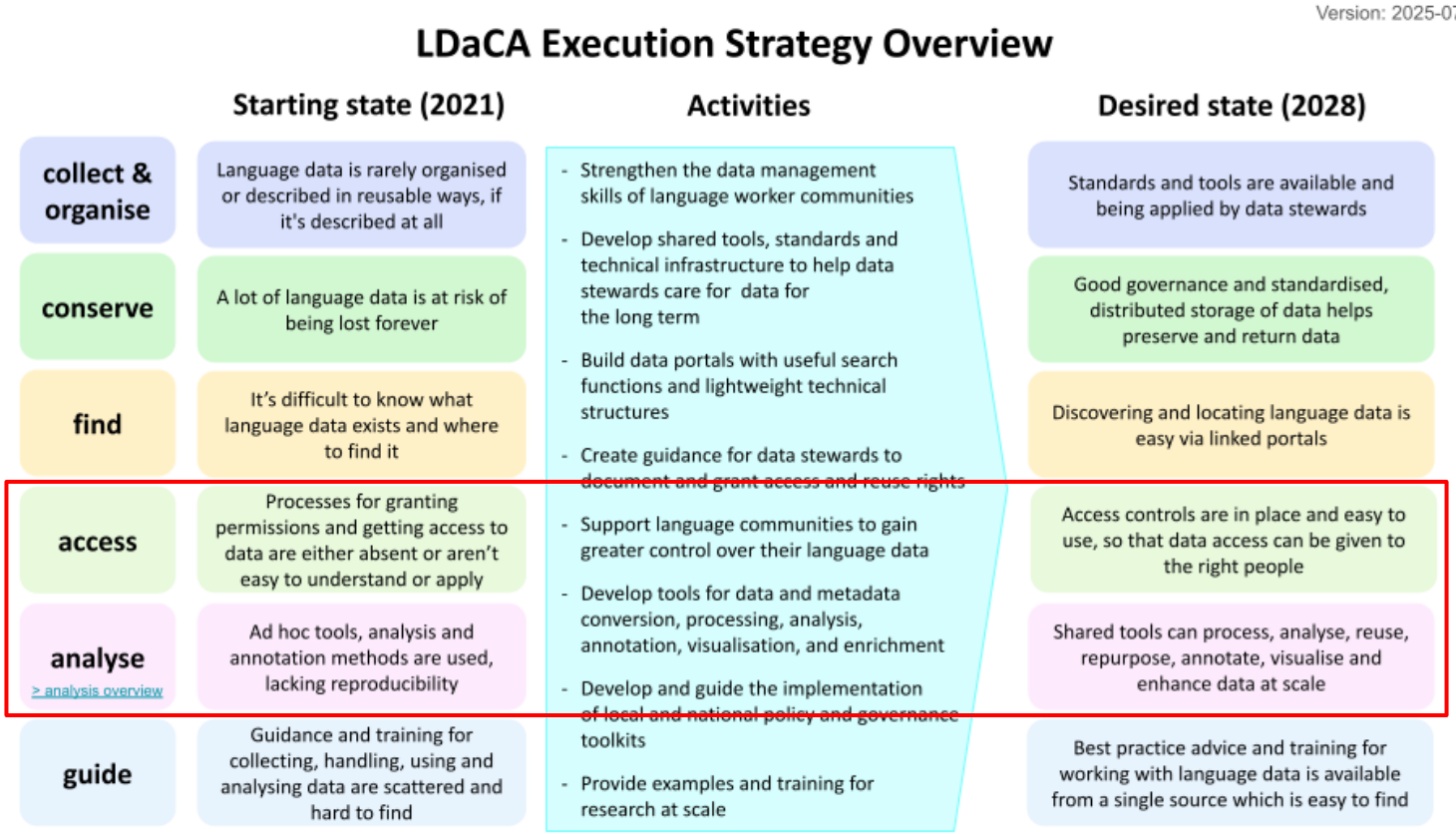
Low friction FAIR interoperability using RO-Crate metadata in text analytics pipelines

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).



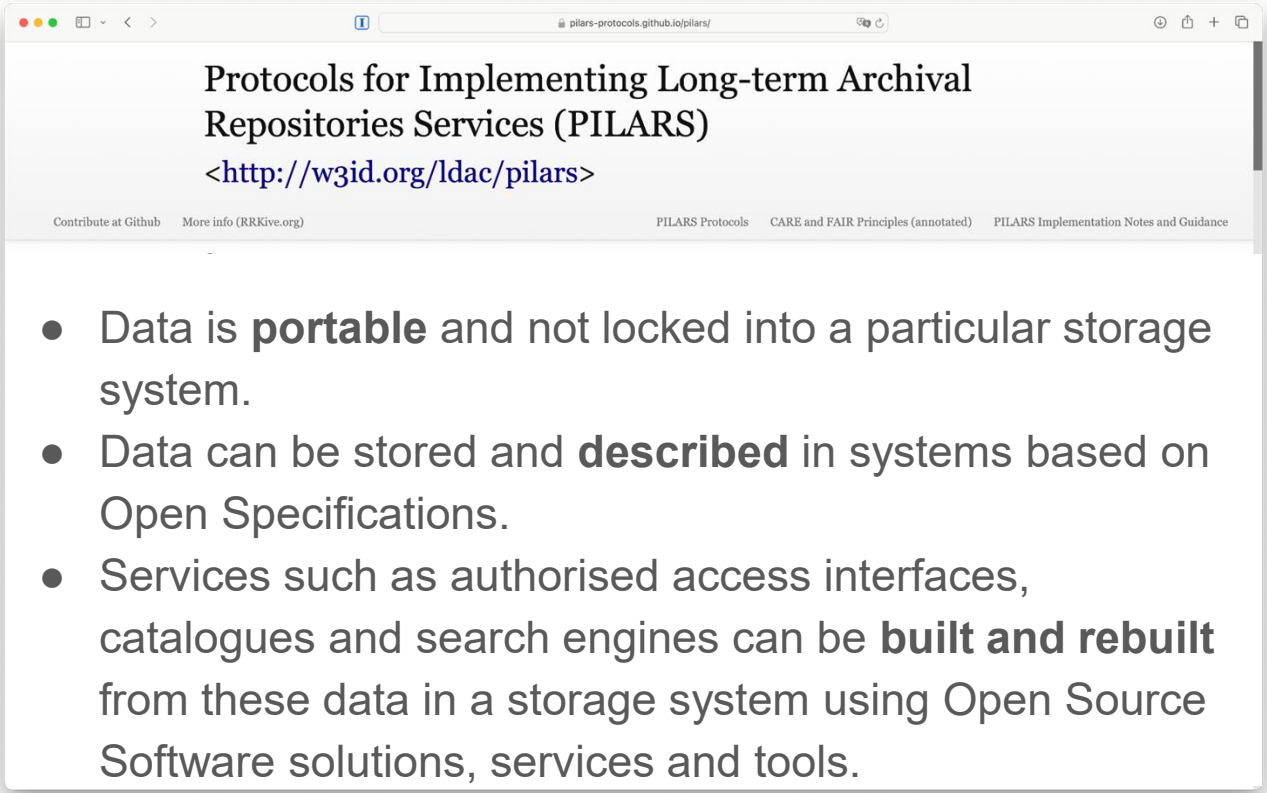
Version: 2025-07-31

LDaCA 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 - Train researchers in data management, standardised data formats, preparation, transformation and wrangling of data for analysis - Document methods and develop toolkits to transform BYO (meta)data to standard formats without compromising data integrity - 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 - Identify promising methods, practices and workflows, including emerging methods (AI) and adapt them across research contexts ethically and appropriately - Develop data connections that link Language Data Commons compliant data - in portal and BYO - to analytical tools - Develop and demonstrate end-to-end best-practice workflows to connect researchers, data and computational tools	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		Documentation and training programs available to help researchers adopt appropriate standards
findable	Appropriate implementations of analytical methods are hard to find		LDaCA infrastructure is interconnected, with suitable interfaces, data formats and guidance on appropriate usage
adaptable	Methods are specialized to particular studies or research cohorts		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		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		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)



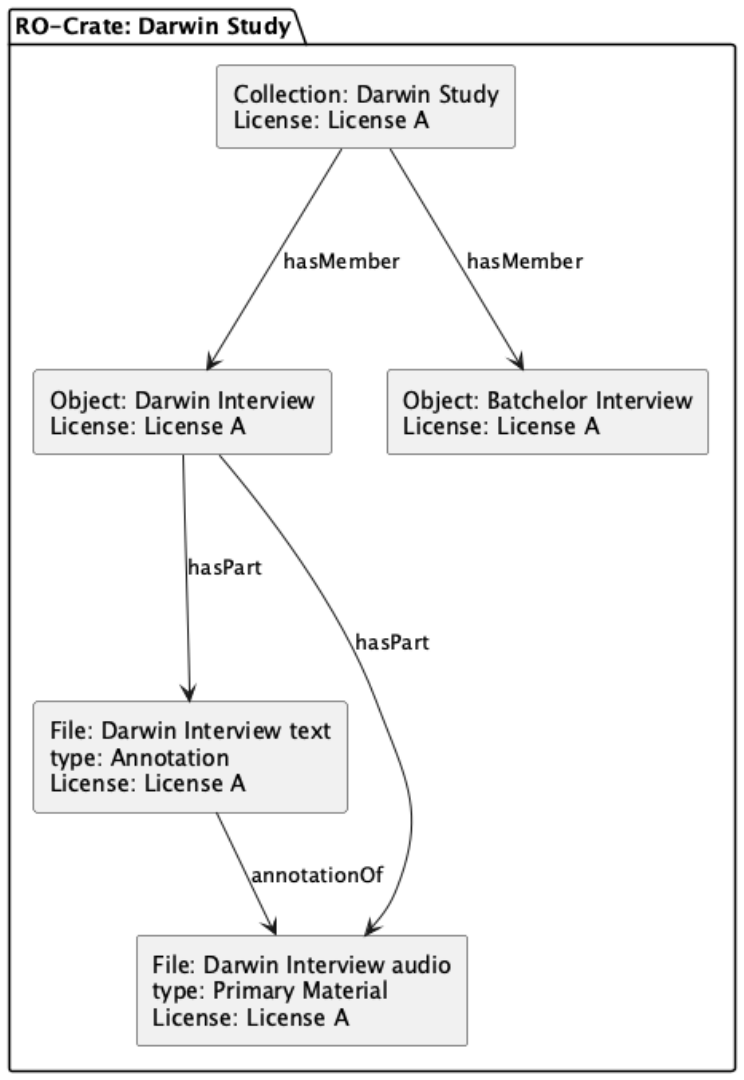
https://w3id.org/ldac/pilars

Storage

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

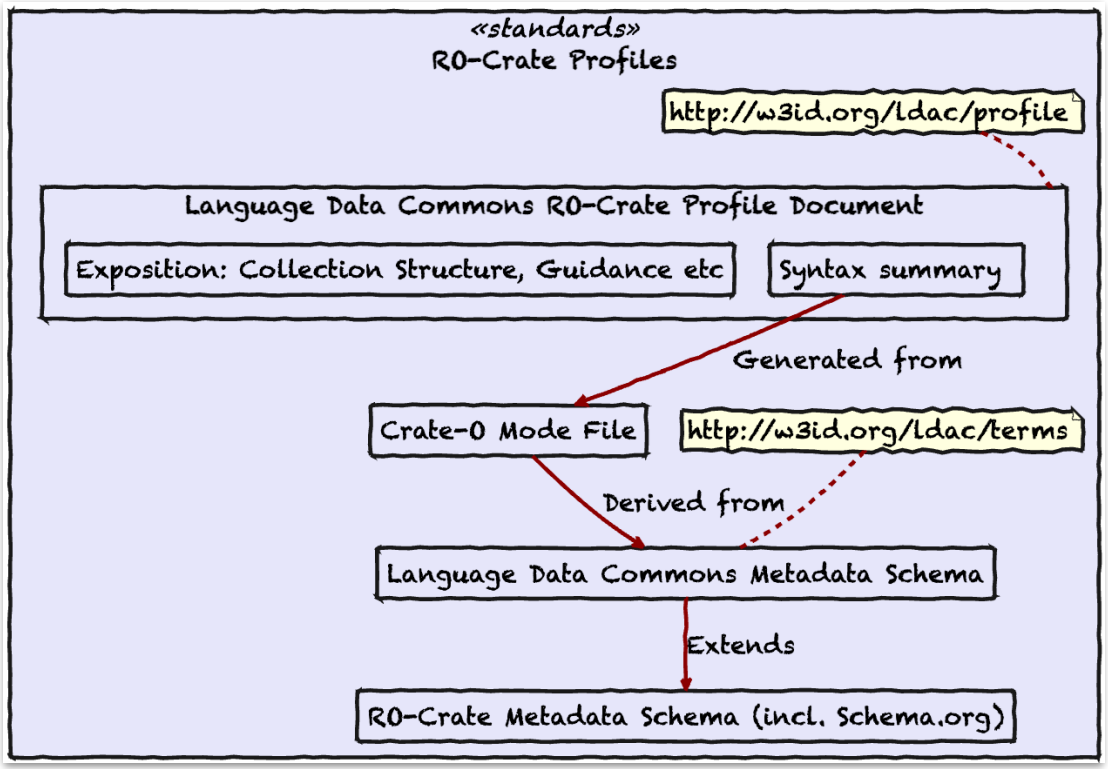
An RO-Crate is a Research Object (or RO) formed of a collection of data (a crate), a special `ro-crate-metadata.json` file which describes the collection and its license information.

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.



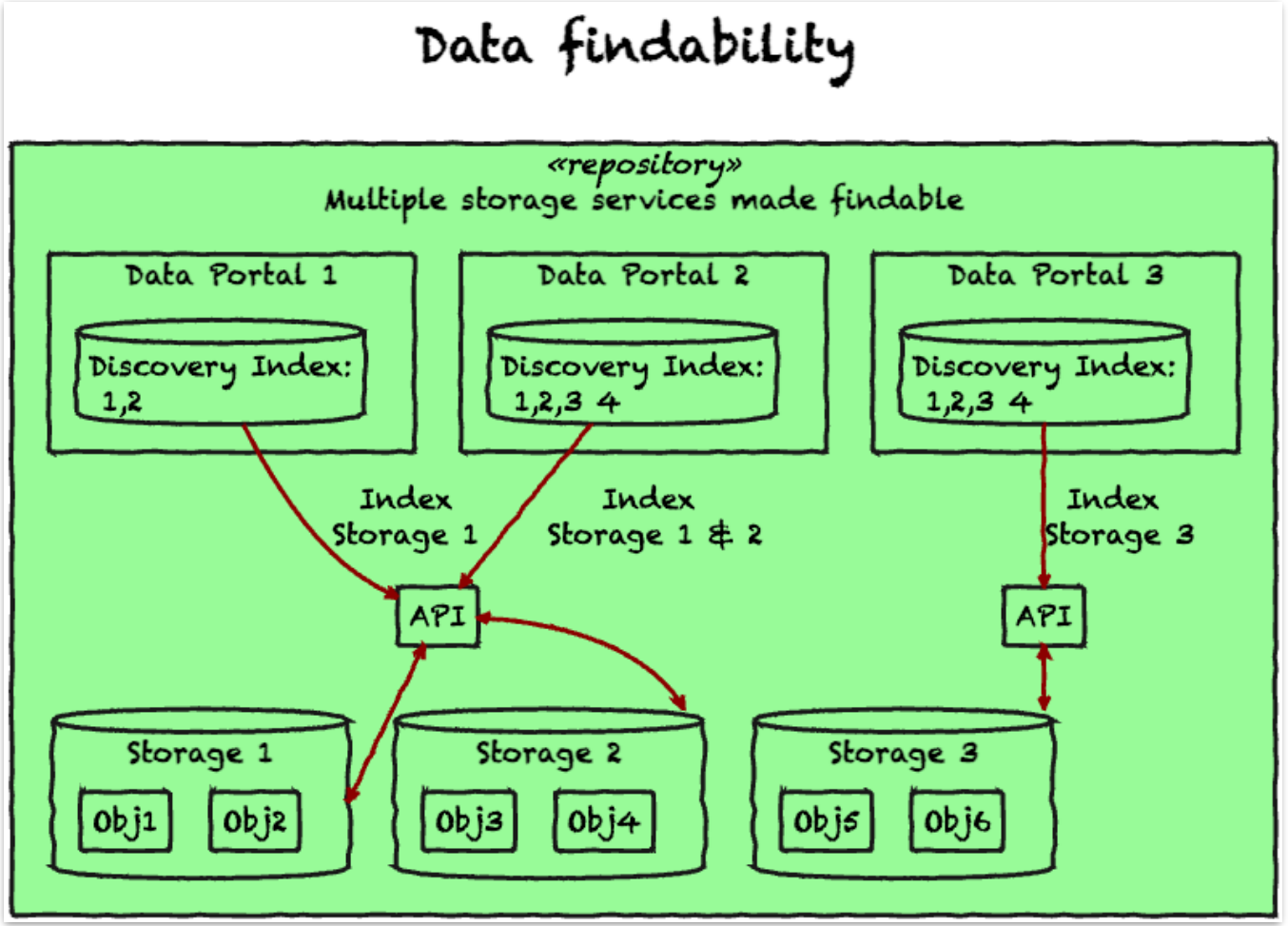
Data Annotation

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

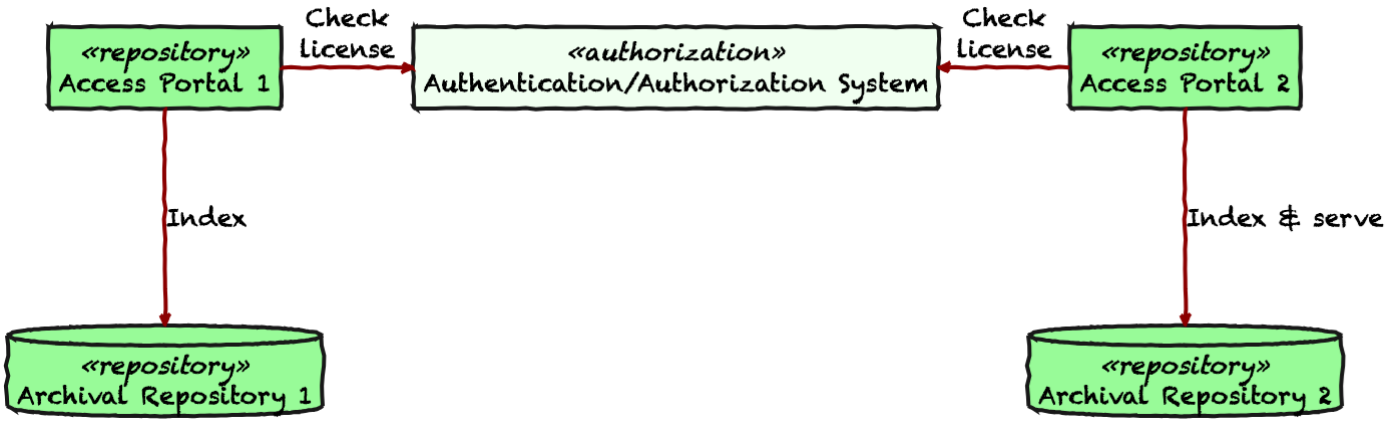
Portals can be indexed from the storage to make them findable.



Distributed Access Control

Motivation

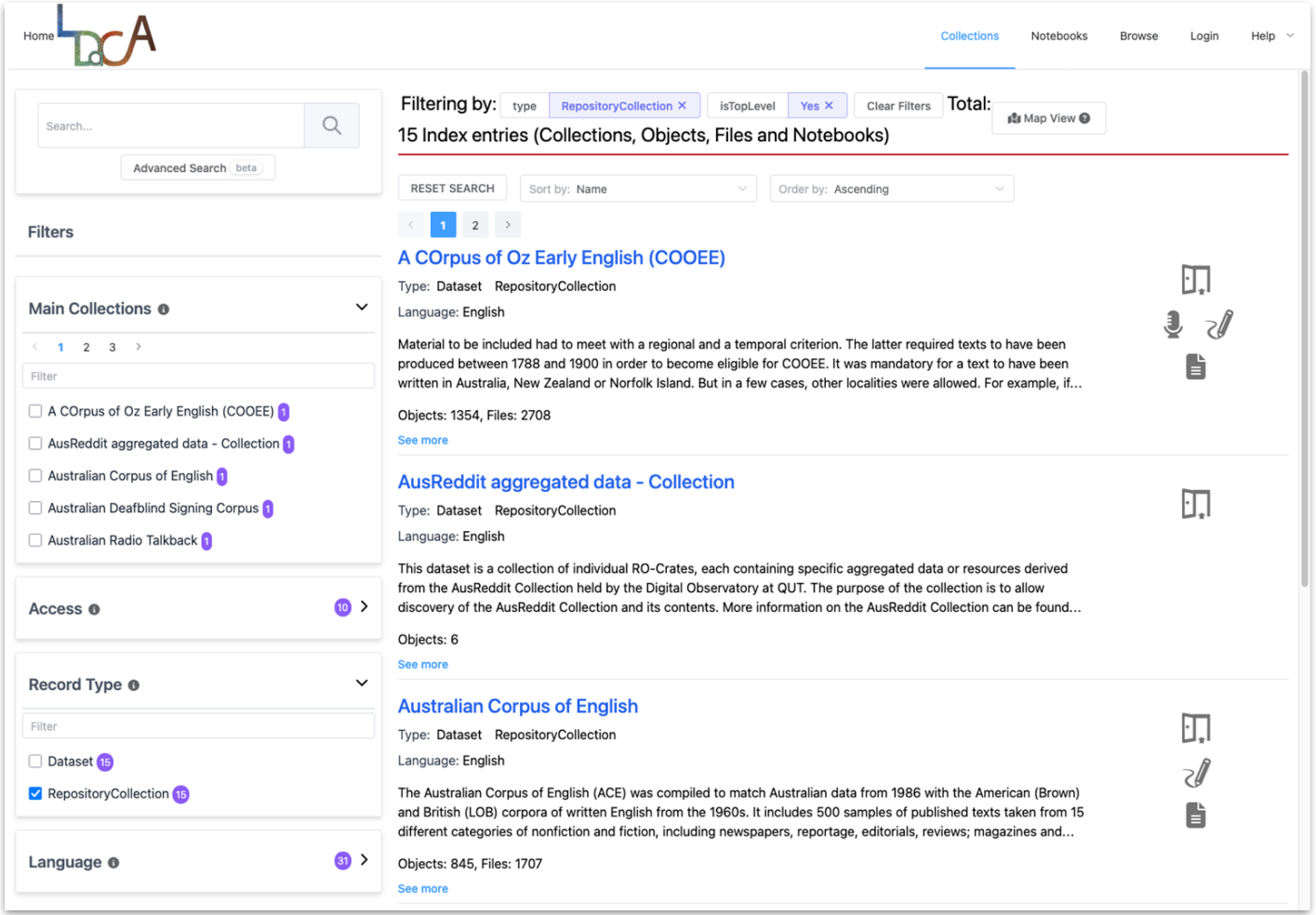
- FAIR (Findable, Accessible, Interoperable, Reusable) data principles require not just openness but **controlled access** in many contexts.
- Traditional centralized access control solutions struggle with scalability, cross-institutional trust, privacy, and fine-grained permissions.



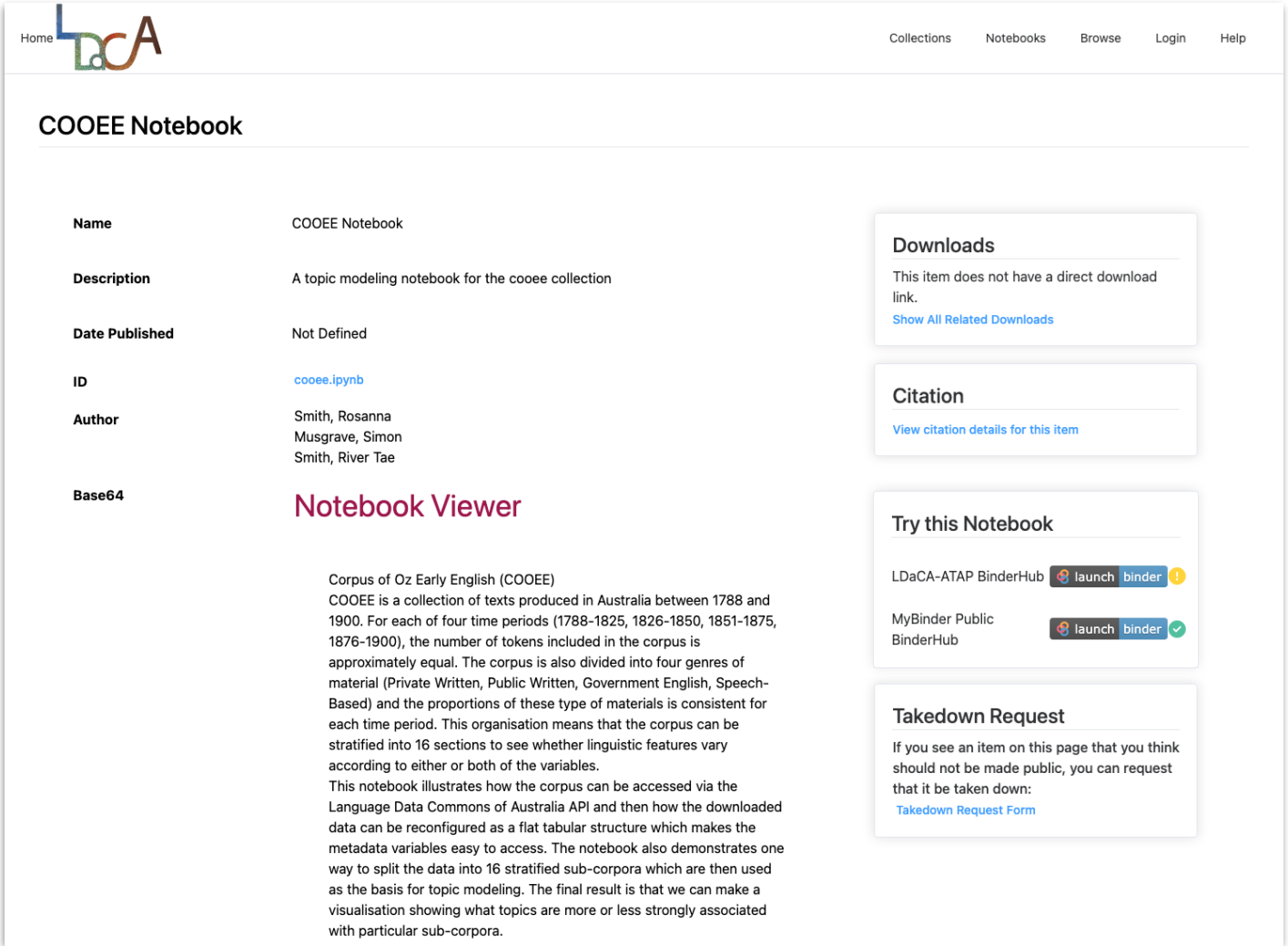
Portal(s)

Main portal: data.ldaca.edu.au

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



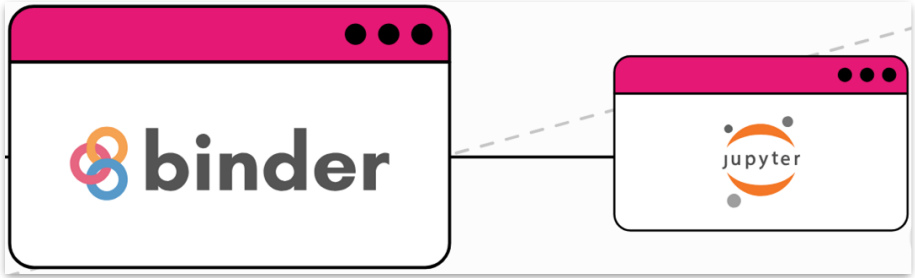
Analysis


Home
Collections
Notebooks
Browse
Login
Help
COOEE Notebook
Name
COOEE Notebook
Description
A topic modeling notebook for the cooee collection
Date Published
Not Defined
ID
cooee.ipynb
Author
Smith, Rosanna
Musgrave, Simon
Smith, River Tae
Base64
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
launch binder
MyBinder Public BinderHub
launch binder
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)

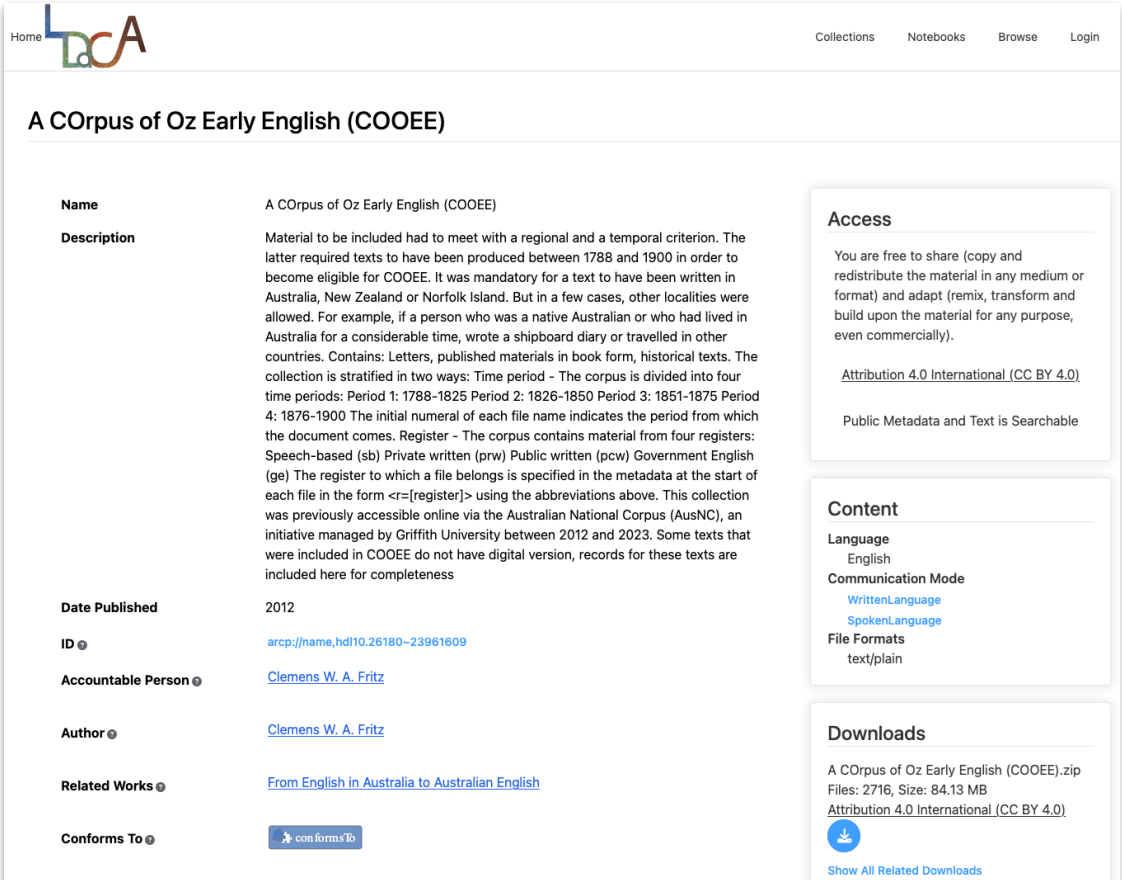
A collection of texts written in Australia between 1788 and 1900.

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)



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 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 `ldac: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

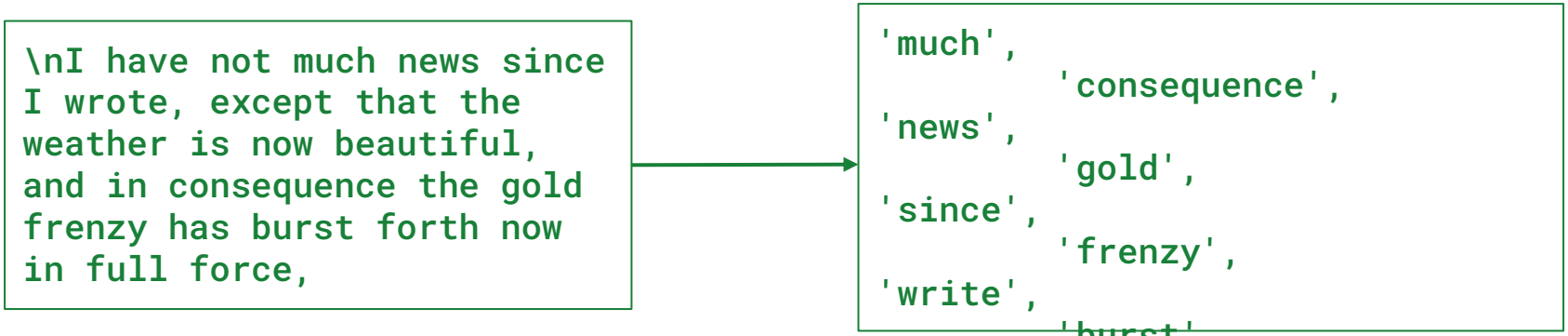

h-croser Hamish Croser

Languages

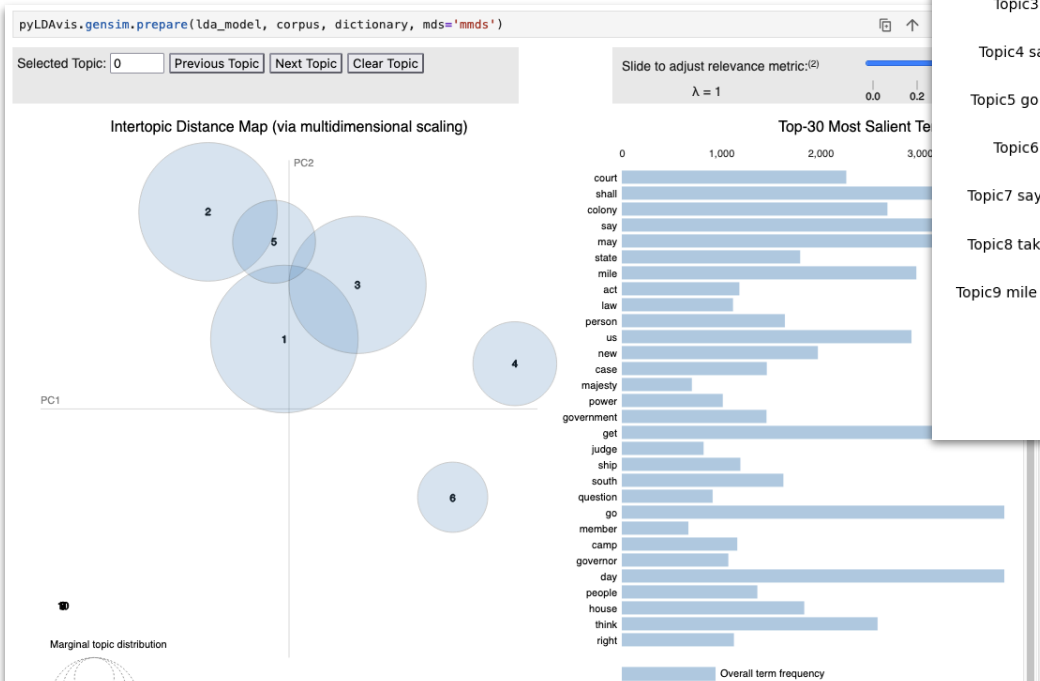
HTML 74.0% **Python** 26.0%

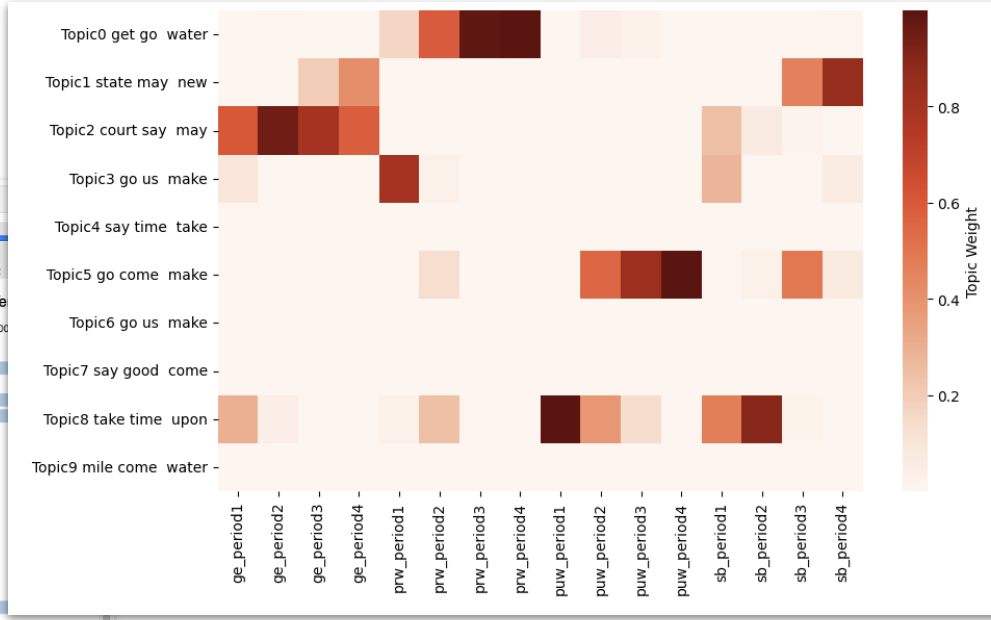
COOEE Notebook - Tokenization

- 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.)



COOEE Notebook - Visualising the data



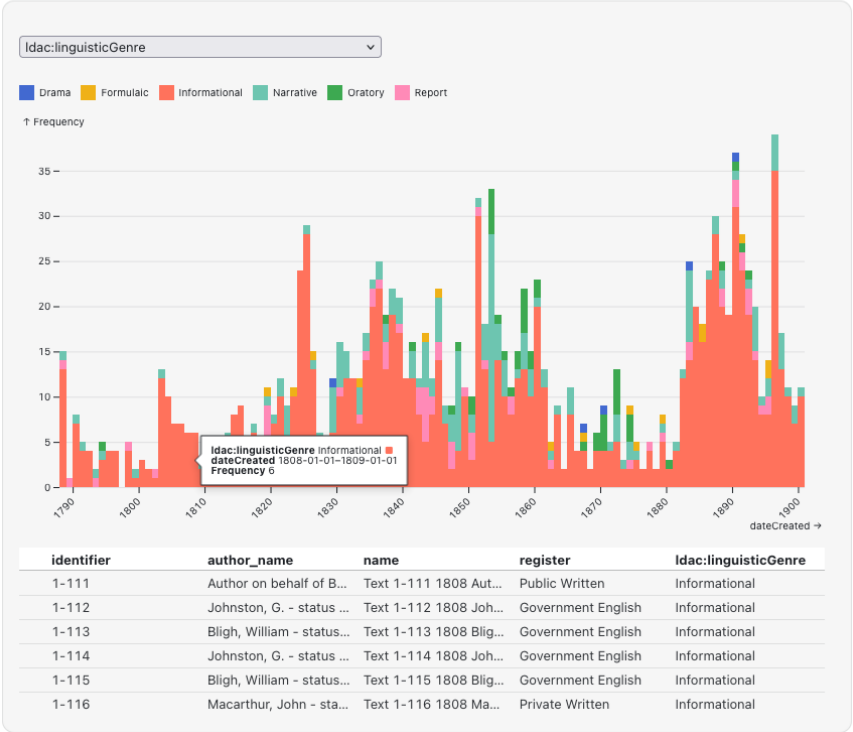


Tabulator

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

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

COOEE



identifier	author_name	name	register	ldac:linguisticGenre
1-111	Author on behalf of B...	Text 1-111 1808 Aut...	Public Written	Informational
1-112	Johnston, G. - status ...	Text 1-112 1808 Joh...	Government English	Informational
1-113	Bligh, William - status...	Text 1-113 1808 Blig...	Government English	Informational
1-114	Johnston, G. - status ...	Text 1-114 1808 Joh...	Government English	Informational
1-115	Bligh, William - status...	Text 1-115 1808 Blig...	Government English	Informational
1-116	Macarthur, John - sta...	Text 1-116 1808 Ma...	Private Written	Informational

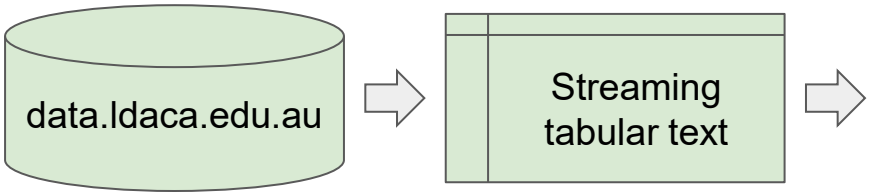
Annotated CSV exports

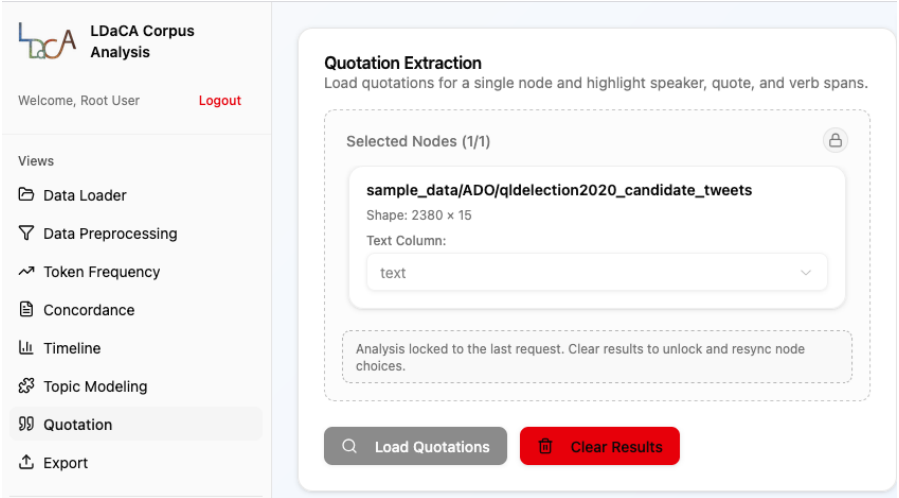
	A	B	C	D	E	F	G	H	I
1	entity_id	@type	name	birthDate	local:birthDateEstimateStart	local:birthDateEstimateEnd	birthPlace	birthPlace_id	gender
2	https://www.pe	Person	Clemens W. A. Fritz						
3	arcp://name,hd	Person	Philip, Arthur - status 1788 text #1-001	1738	1738	1738	Great Britain	#place_GB	m
4	arcp://name,hd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


```
{
      "name": "author_local:birthDateEstimateStart",
      "label": "local:birthDateEstimateStart",
      "propertyUrl": "arcp://name,hdl10.26180~23961609/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





Version: 2025-07-31

LDaCA Execution Strategy Overview

	Starting state (2021)	Activities	Desired state (2028)
collect & organise	Language data is rarely organised or described in reusable ways, if it's described at all	- 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	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
analyse > analysis overview	Ad hoc tools, analysis and annotation methods are used, lacking reproducibility		
guide	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)
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 - Document methods and develop toolkits to transform BYO (meta)data to standard formats without compromising data integrity	Documentation and training programs available to help researchers adopt appropriate standards
findable	Appropriate implementations of analytical methods are hard to find	- Train researchers in computational methods application and development	LDaCA infrastructure is interconnected, with suitable interfaces, data formats and guidance on appropriate usage
adaptable	Methods are specialized to particular studies or research cohorts	- Develop guidance and train researchers on how to choose appropriate analytical approaches eg ethical and appropriate use of AI - raise awareness of computational methods - Identify promising methods, practices and workflows, including emerging methods (AI) and adapt them across research contexts ethically and appropriately	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	- Develop data connections that link Language Data Commons compliant data - in portal and BYO - to analytical tools	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