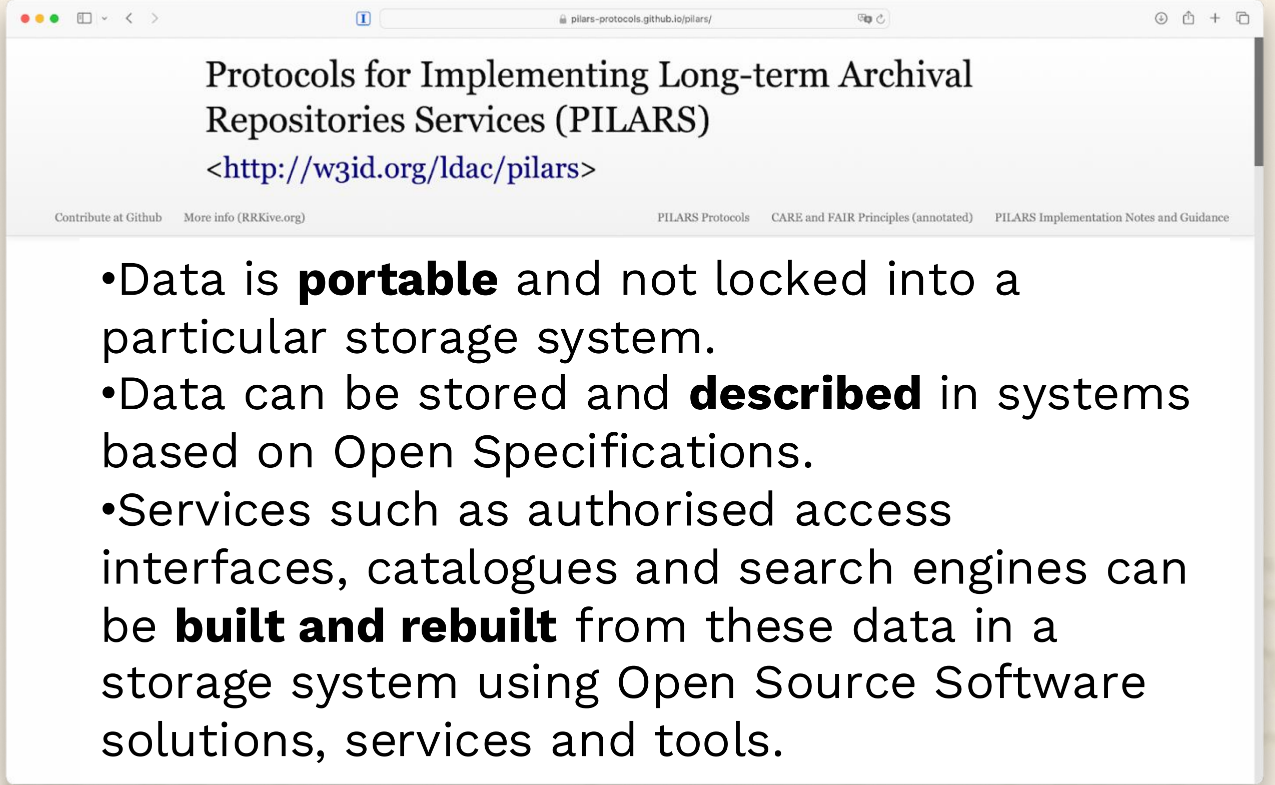
Implementing PILARS

Ensuring Digital Language and Cultural-Heritage Materials Remain Accessible, Usable, and Sustainably Managed Over Time

- Preserving digital language and cultural collections
- By adopting open standards and clear governance
- Sustainable stewardship protects past investments in research and infrastructure
- Addressing this problem isn’t just about technology

LDaCA Architecture

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



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

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

Contribute at Github More info (RRKive.org) PILARS Protocols CARE and FAIR Principles (annotated) PILARS Implementation Notes and Guidance

- 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.

PILARS

A framework of protocols to design sustainable archival systems.

Supports **FAIR** (Findable, Accessible, Interoperable, Reusable) and **CARE** (Collective Benefit, Authority to Control, Responsibility, Ethics) principles.

PILARS Goals

- Autonomy
- Sustainability
- Value

1. **Data Portability**
    1. Commodity Storage
    2. Storage Objects
    3. Store documentation within storage root

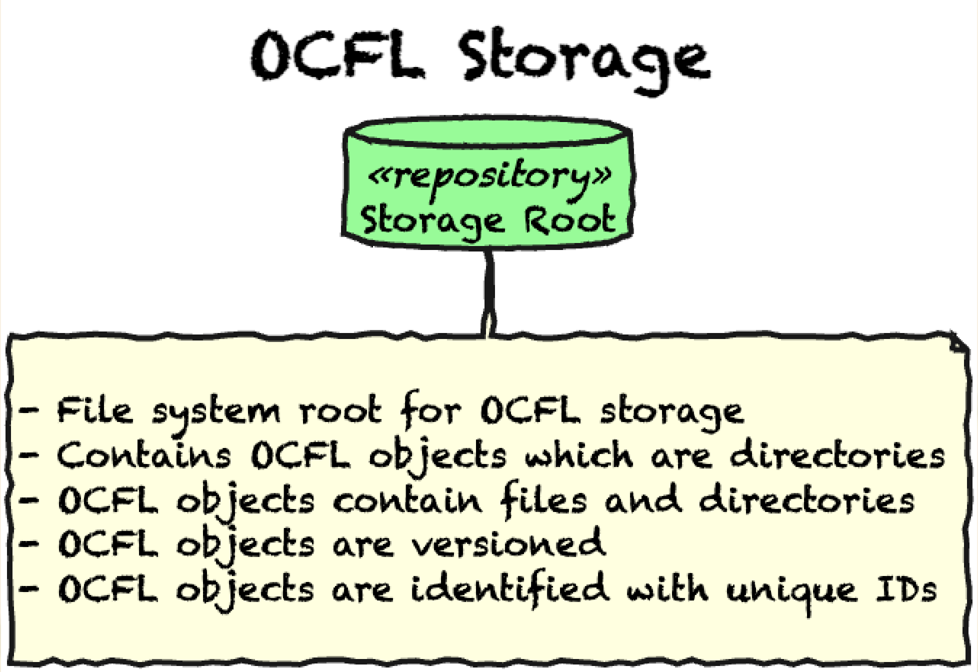
2. Metadata & Annotation

- Each object has descriptive metadata (usage rights, provenance)
- Use Linked Data, Represent high level structures

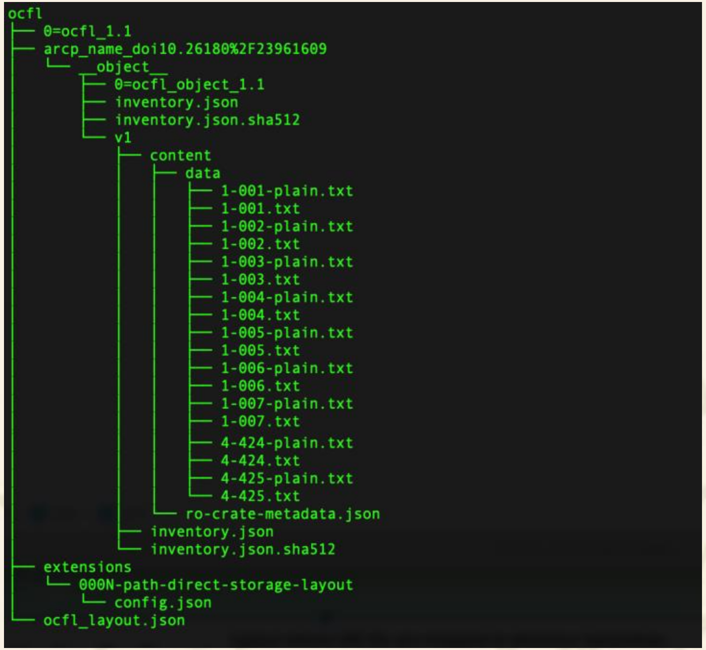
3. Governance

1 - Data is Portable

The **Oxford Common File Layout**



- File system root for OCFL storage
- Contains OCFL objects which are directories
- OCFL objects contain files and directories
- OCFL objects are versioned
- OCFL objects are identified with unique IDs



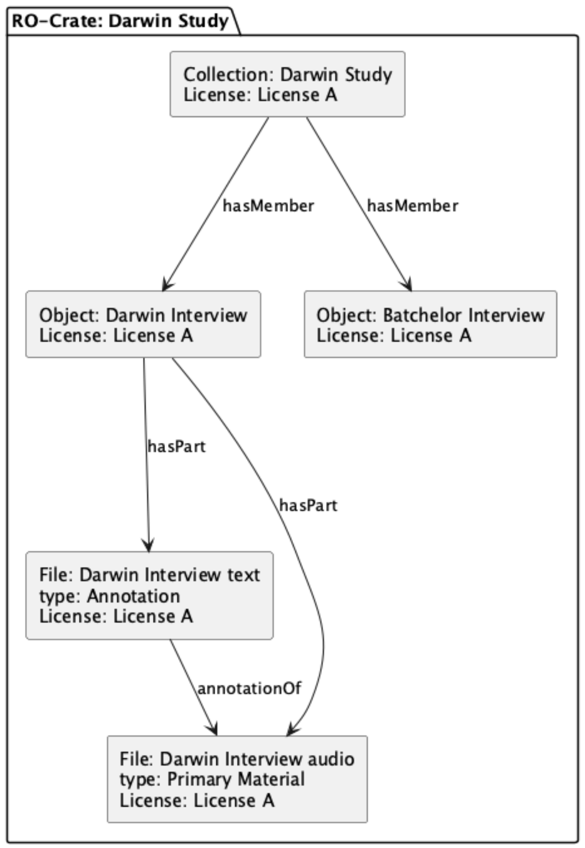
```
ocfl
├── 0=ocfl_1.1
├── arcp_name_doi10.26180%2F23961609
│   └── __object__
│       ├── 0=ocfl_object_1.1
│       ├── inventory.json
│       ├── inventory.json.sha512
│       └── v1
│           ├── content
│           │   ├── data
│           │   │   ├── 1-001-plain.txt
│           │   │   ├── 1-001.txt
│           │   │   ├── 1-002-plain.txt
│           │   │   ├── 1-002.txt
│           │   │   ├── 1-003-plain.txt
│           │   │   ├── 1-003.txt
│           │   │   ├── 1-004-plain.txt
│           │   │   ├── 1-004.txt
│           │   │   ├── 1-005-plain.txt
│           │   │   ├── 1-005.txt
│           │   │   ├── 1-006-plain.txt
│           │   │   ├── 1-006.txt
│           │   │   ├── 1-007-plain.txt
│           │   │   ├── 1-007.txt
│           │   │   ├── 4-424-plain.txt
│           │   │   ├── 4-424.txt
│           │   │   ├── 4-425-plain.txt
│           │   │   └── 4-425.txt
│           │   └── ro-crate-metadata.json
│           ├── inventory.json
│           └── inventory.json.sha512
├── extensions
│   └── 000N-path-direct-storage-layout
│       └── config.json
└── ocfl_layout.json
```

Storage

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

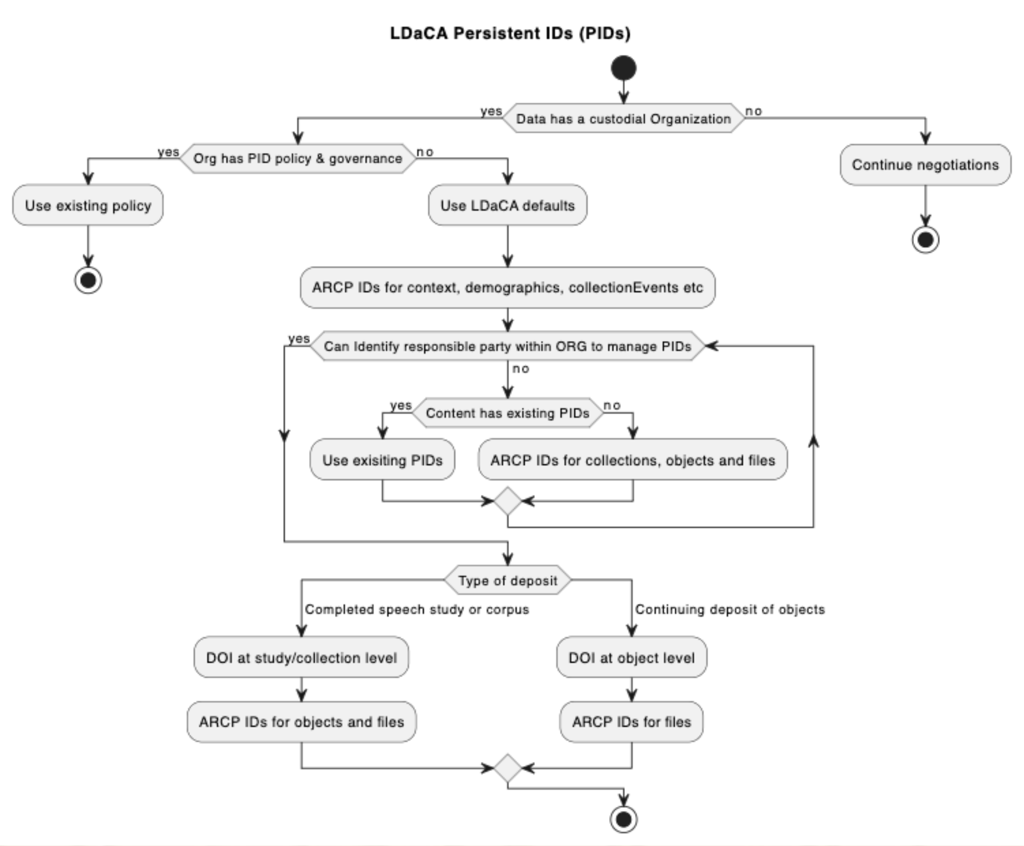
An RO-Crate is a Research Object (or RO) formed of a collection of data (a crate), a special **ro-crate-metadata.json** file which describes the collection and its license information.

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.

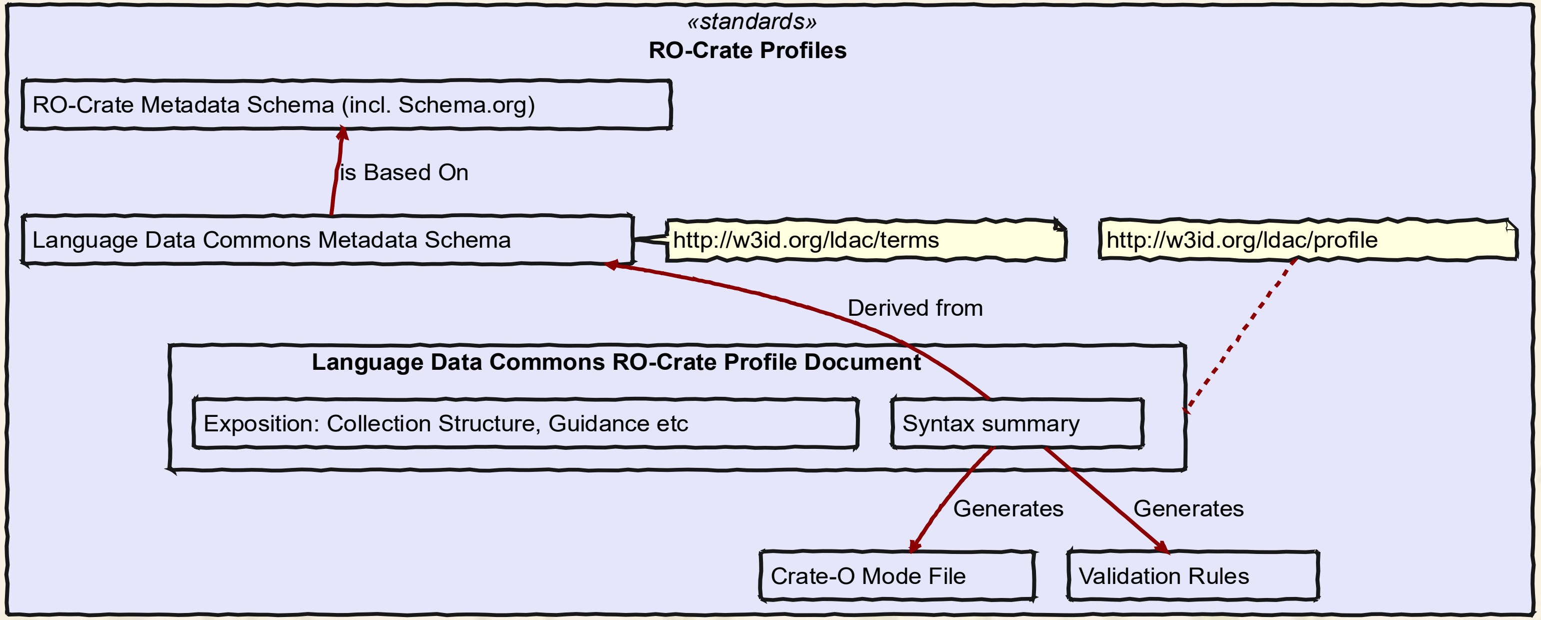


Persistant IDs

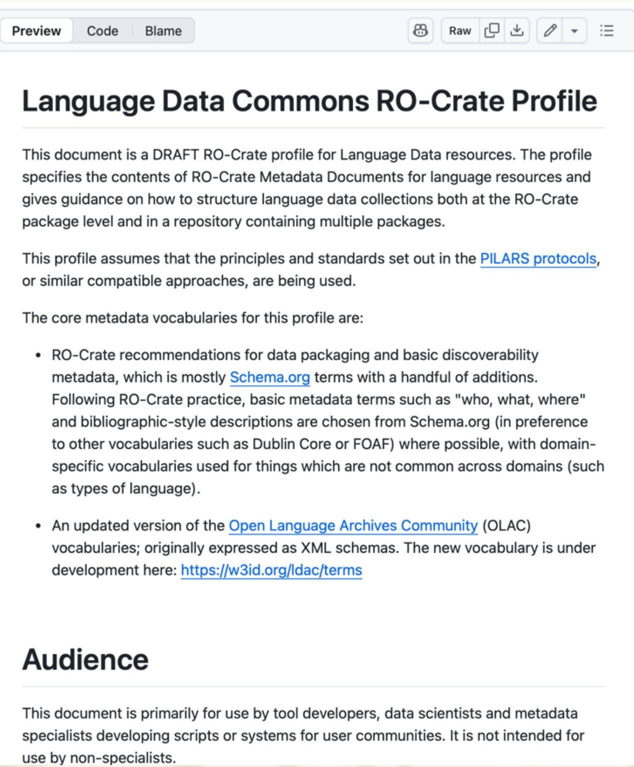
OCFL is laid out as URI IDs and mapped to directory hierarchies.


LDaCA Persistent IDs (PIDs)
Data has a custodial Organization
yes
no
Org has PID policy & governance
yes
no
Continue negotiations
Use existing policy
Use LDaCA defaults
ARCP IDs for context, demographics, collectionEvents etc
Can Identify responsible party within ORG to manage PIDs
yes
no
Content has existing PIDs
yes
no
Use exisiting PIDs
ARCP IDs for collections, objects and files
Type of deposit
Completed speech study or corpus
Continuing deposit of objects
DOI at study/collection level
DOI at object level
ARCP IDs for objects and files
ARCP IDs for files

2 Data is annotated


«standards»
RO-Crate Profiles
RO-Crate Metadata Schema (incl. Schema.org)
is Based On
Language Data Commons Metadata Schema
http://w3id.org/ldac/terms
http://w3id.org/ldac/profile
Derived from
Language Data Commons RO-Crate Profile Document
Exposition: Collection Structure, Guidance etc
Syntax summary
Generates
Generates
Crate-O Mode File
Validation Rules

Metadata Schemas



Language Data Commons RO-Crate Profile

This document is a DRAFT RO-Crate profile for Language Data resources. The profile specifies the contents of RO-Crate Metadata Documents for language resources and gives guidance on how to structure language data collections both at the RO-Crate package level and in a repository containing multiple packages.

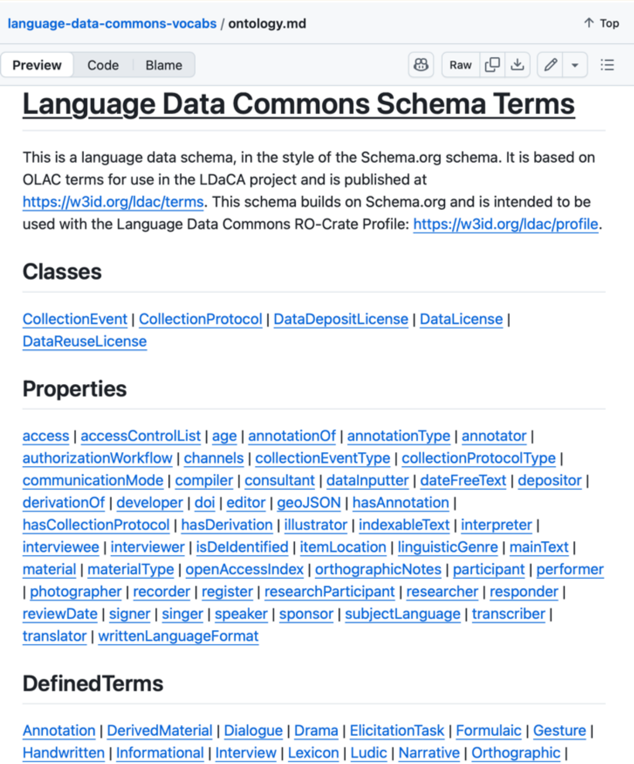
This profile assumes that the principles and standards set out in the PILARS protocols, or similar compatible approaches, are being used.

The core metadata vocabularies for this profile are:

- RO-Crate recommendations for data packaging and basic discoverability metadata, which is mostly Schema.org terms with a handful of additions. Following RO-Crate practice, basic metadata terms such as "who, what, where" and bibliographic-style descriptions are chosen from Schema.org (in preference to other vocabularies such as Dublin Core or FOAF) where possible, with domain-specific vocabularies used for things which are not common across domains (such as types of language).
- An updated version of the Open Language Archives Community (OLAC) vocabularies; originally expressed as XML schemas. The new vocabulary is under development here: https://w3id.org/ldac/terms

Audience

This document is primarily for use by tool developers, data scientists and metadata specialists developing scripts or systems for user communities. It is not intended for use by non-specialists.



Language Data Commons Schema Terms

This is a language data schema, in the style of the Schema.org schema. It is based on OLAC terms for use in the LDaCA project and is published at https://w3id.org/ldac/terms. This schema builds on Schema.org and is intended to be used with the Language Data Commons RO-Crate Profile: https://w3id.org/ldac/profile.

Classes

CollectionEvent | CollectionProtocol | DataDepositLicense | DataLicense | DataReuseLicense

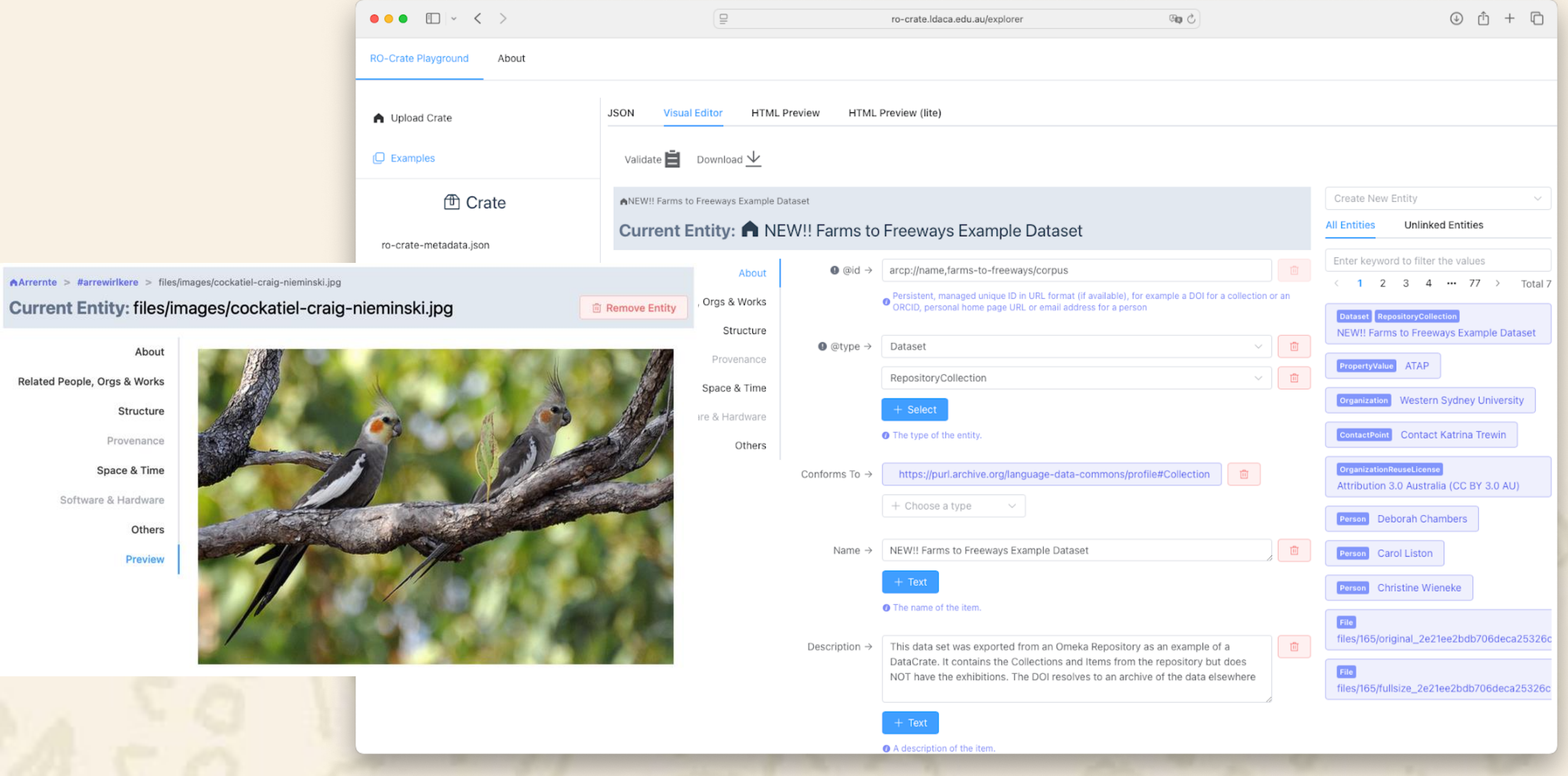
Properties

access | accessControlList | age | annotationOf | annotationType | annotator | authorizationWorkflow | channels | collectionEventType | collectionProtocolType | communicationMode | compiler | consultant | dataInputter | dateFreeText | depositor | derivationOf | developer | doi | editor | geoJSON | hasAnnotation | hasCollectionProtocol | hasDerivation | illustrator | indexableText | interpreter | interviewee | interviewer | isDeIdentified | itemLocation | linguisticGenre | mainText | material | materialType | openAccessIndex | orthographicNotes | participant | performer | photographer | recorder | register | researchParticipant | researcher | responder | reviewDate | signer | singer | speaker | sponsor | subjectLanguage | transcriber | translator | writtenLanguageFormat

DefinedTerms

Annotation | DerivedMaterial | Dialogue | Drama | ElicitationTask | Formulaic | Gesture | Handwritten | Informational | Interview | Lexicon | Ludic | Narrative | Orthographic |

Annotate



ro-crate.ldaca.edu.au/explorer

RO-Crate Playground | About

Upload Crate

Examples

Crate

ro-crate-metadata.json

JSON | Visual Editor | HTML Preview | HTML Preview (lite)

Validate | Download

NEW!! Farms to Freeways Example Dataset

Current Entity: NEW!! Farms to Freeways Example Dataset

About
, Orgs & Works
Structure
Provenance
Space & Time
ıre & Hardware
Others

@id → arcp://name,farms-to-freeways/corpus

Persistent, managed unique ID in URL format (if available), for example a DOI for a collection or an ORCID, personal home page URL or email address for a person

@type → Dataset

RepositoryCollection

+ Select

The type of the entity.

Conforms To → https://purl.archive.org/language-data-commons/profile#Collection

+ Choose a type

Name → NEW!! Farms to Freeways Example Dataset

+ Text

The name of the item.

Description → This data set was exported from an Omeka Repository as an example of a DataCrate. It contains the Collections and Items from the repository but does NOT have the exhibitions. The DOI resolves to an archive of the data elsewhere

+ Text

A description of the item.

Create New Entity

All Entities | Unlinked Entities

Enter keyword to filter the values

< 1 2 3 4 ••• 77 > Total 7

- Dataset RepositoryCollection NEW!! Farms to Freeways Example Dataset
- PropertyValue ATAP
- Organization Western Sydney University
- ContactPoint Contact Katrina Trewin
- OrganizationReuseLicense Attribution 3.0 Australia (CC BY 3.0 AU)
- Person Deborah Chambers
- Person Carol Liston
- Person Christine Wieneke
- File files/165/original_2e21ee2bdb706deca25326c
- File files/165/fullsize_2e21ee2bdb706deca25326c

Arrernte > #arrewirlkere > files/images/cockatiel-craig-nieminski.jpg

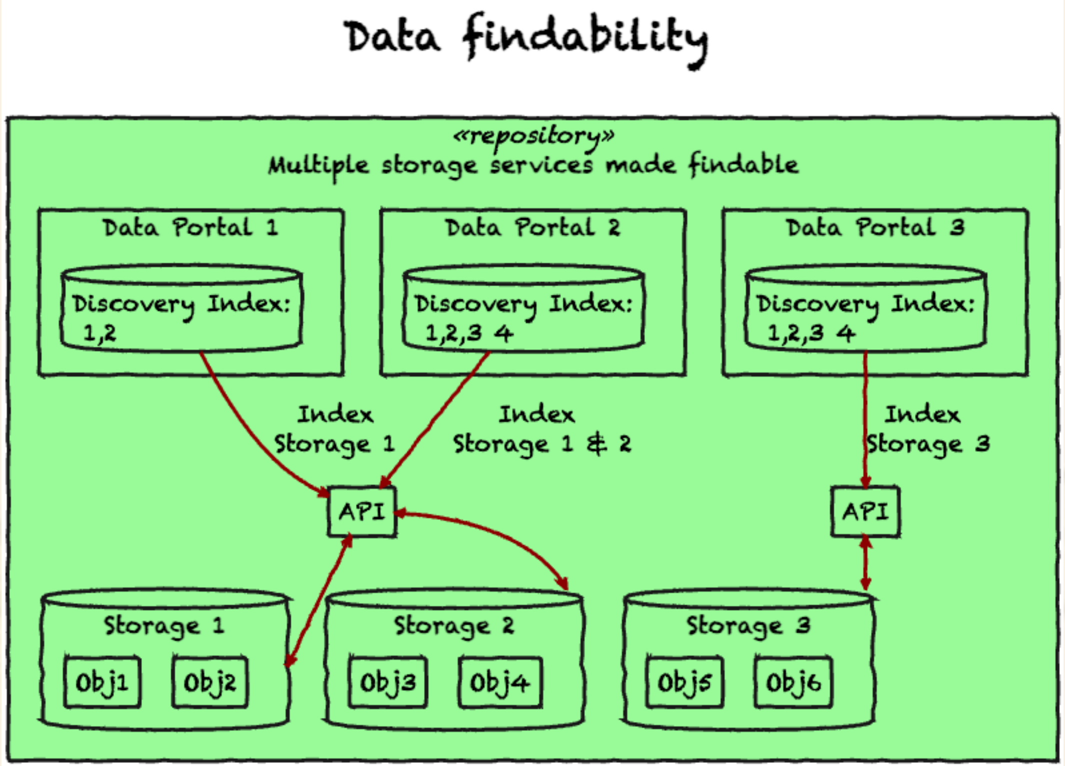
Current Entity: files/images/cockatiel-craig-nieminski.jpg

Remove Entity

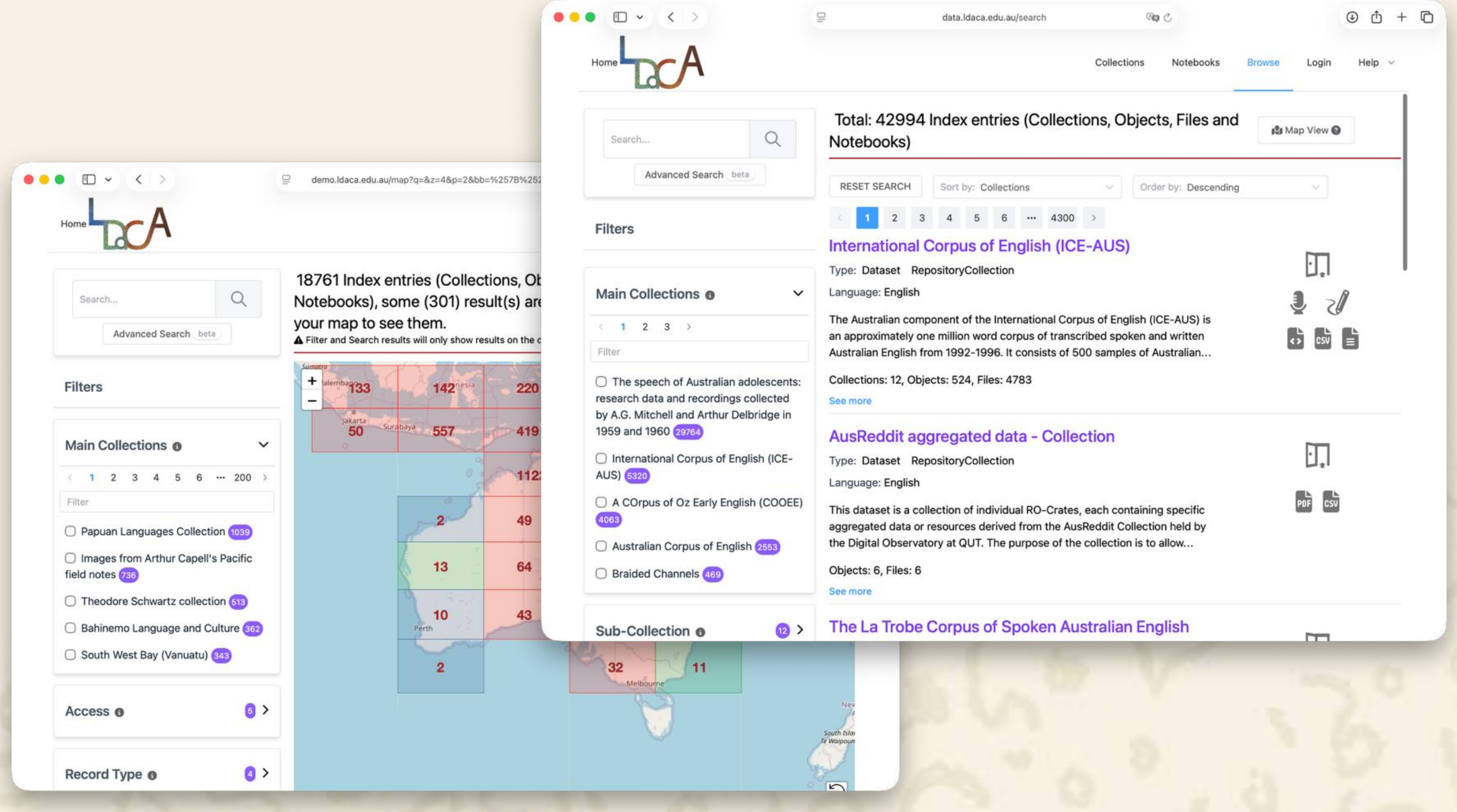
About
Related People, Orgs & Works
Structure
Provenance
Space & Time
Software & Hardware
Others
Preview

Index

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



Portal(s)



demo.ldaca.edu.au/map?q=&z=4&p=2&bb=%257B%252

Home

Search...

Advanced Search beta

18761 Index entries (Collections, Ob
Notebooks), some (301) result(s) are
your map to see them.

⚠ Filter and Search results will only show results on the c

Filters

Main Collections

‹ 1 2 3 4 5 6 ··· 200 ›

Filter

- ☐ Papuan Languages Collection 1039
- ☐ Images from Arthur Capell's Pacific field notes 736
- ☐ Theodore Schwartz collection 513
- ☐ Bahinemo Language and Culture 362
- ☐ South West Bay (Vanuatu) 343

Access 5 ›

Record Type 4 ›

data.ldaca.edu.au/search

Home

Collections Notebooks Browse Login Help

Search...

Advanced Search beta

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
- ☐ Braided Channels 469

Sub-Collection 12 ›

Total: 42994 Index entries (Collections, Objects, Files and Notebooks)

Map View

RESET SEARCH | Sort by: Collections | Order by: Descending

‹ 1 2 3 4 5 6 ··· 4300 ›

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 English from 1992-1996. It consists of 500 samples of Australian...

Collections: 12, Objects: 524, Files: 4783

See more

AusReddit aggregated data - Collection

Type: Dataset RepositoryCollection

Language: English

This dataset is a collection of individual RO-Crates, each containing specific aggregated data or resources derived from the AusReddit Collection held by the Digital Observatory at QUT. The purpose of the collection is to allow...

Objects: 6, Files: 6

See more

The La Trobe Corpus of Spoken Australian English

Access Control

A **distributed access control system** that leverages **federated authenication (AAF)** independently of **authorization services**.

Key features:

- License-based access control
- Enforcement points
- Interoperable protocols

Motivation

FAIR data principles require not just openness but **controlled access** in many contexts.

Traditional centralized access control solutions struggle with scalability, sustainability, cross-institutional trust, privacy, and fine-grained permissions.

Architecture & Workflow

1. User requests access
2. Enforcement point at repository
3. Repository polls authorization server if necessary
4. Decision point at authorization server
5. Audit & logging

Benefits	Challenges & Considerations
Scalability across organizations	Ensuring trust among domains
Fine-grained, dynamic access control	Performance overhead of distributed checks
Compliance with FAIR's "Accessible" principle	Handling license revocation, privacy, and interoperability

Access Control

Microsoft

AUSTRALIAN ACCESS FEDERATION

Google

eduGAIN

Email

CSC REMS

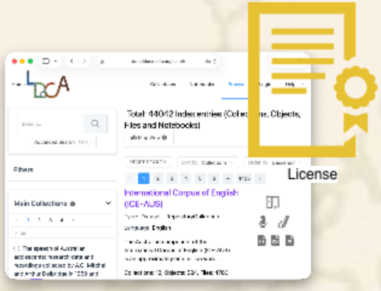
CILogon

Authentication: Who am I?

CADRE

Authorisation: What am I allowed to see?

AAI



LDaCA

Total: 44042 Index entries (Collections, Objects, Files and Notebooks)

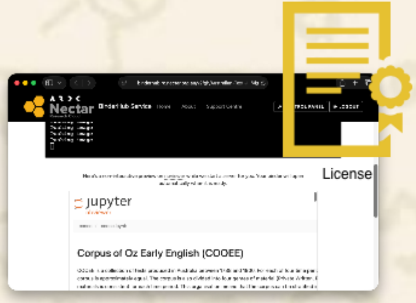
Filters

Main Collections

International Corpus of English (ICE-AUS)

Language: English

License



ARDC Nectar

BinderHub Service

jupyter

Corpus of Oz Early English (COOEE)

License

PORTALS

Key Learnings and Future Plans

Beyond project **websites; sustainable dashboards**

The focus is on **delivery**

- Decisions are made for **speed and appearance**,
- Code, data, and dependencies often become **conflated** .
- When the developer moves on, **knowledge and maintenance capacity disappear**.
- What began as a useful tool can become **a fragile, unmaintained system**

The focus shifts from quick delivery to **long-term value and maintainability**.

- Systems are built with **open standards,**
- Data and code are **portable and separate**
- Maintenance is part of the design
- The result is a system that **endures beyond individual projects and people**

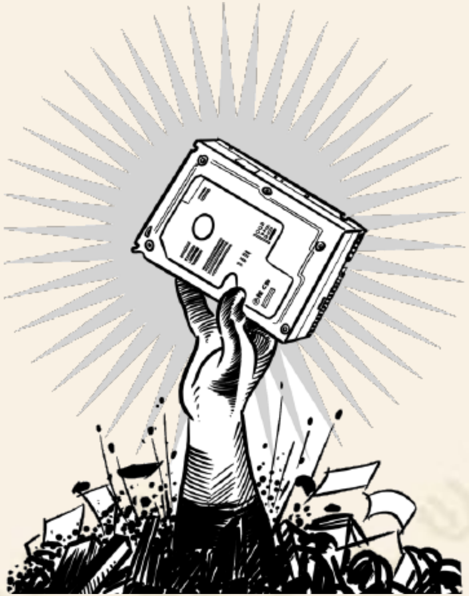
TODO

Fix bugs maintain our tools UX improvements

Design and implement complete Workflow for Interactive Deposits

Add more language data collections

Add more analytical notebooks and tools



https://ocfl.io/1.1.0/spec/

LDaCA

Language Data Commons of Australia

Implementing PILARS

Moises Sacal Bonequi

The University of Queensland Australia

Australian National University

aarnet

Batchelor Institute

First Languages Australia

QUT

The University of Melbourne

The University of Sydney

ARDC

Australian Government

NCRIS