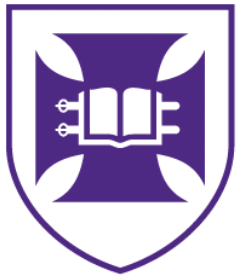
A report on two years work on a standards-based architecture for a Data Commons

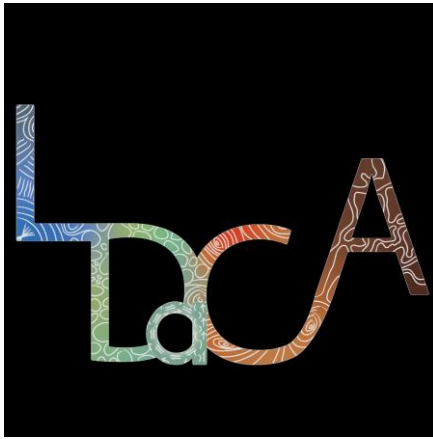
Peter Sefton, Simon Musgrave University Of Queensland, St Lucia, Queensland, Australia



THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

CREATE CHANGE

With thanks for their contribution:



The Language Data Commons of Australia (LDaCA) and Australian Text Analytics Platform (ATAP) projects received investment (https://doi.org/10.47486/DP768 and https://doi.org/10.47486/PL074) from the Australian Research Data Commons (ARDC). The ARDC is funded by the National Collaborative Research Infrastructure Strategy (NCRIS).

What is a data commons?

Grossman *et al*. (2016) give the following description:
"Data commons collocate data, storage, and computing infrastructure with core services and commonly used tools and applications for managing, analyzing, and sharing data to create an interoperable resource for the research community"

(Grossman, Robert L., Allison Heath, Mark Murphy, Maria Patterson & Walt Wells. 2016. A Case for Data Commons: Toward Data Science as a Service. *Computing in Science & Engineering* 18(5). 10–20. https://doi.org/10.1109/MCSE.2016.92.)

Australia’s linguistic diversity

- Speakers of more than 300 languages live in Australia
- In 2021, 5.8 million people (22.8%) reported using a language other than English at home
- Top five:
  - Mandarin 685,274
  - Arabic 367,159
  - Vietnamese 320,758
  - Cantonese 295,281
  - Punjabi 239,033

Nationally significant language data

- Existing language collections to capitalise on prior investments
  - Australian National Corpus
  - PARADISEC
  - Austalk
  - Signbank
  - …
- Australia's Indigenous language holdings of global importance
  - Indigenous cultural heritage
  - Language research
  - Importance for communities
- New and emerging types of language data
- To illustrate language behaviour in Australia across time

Australian Research Data Commons

Academic and non-academic use

- Language data is not only for linguists
- Used by other disciplines including:
  - (Social) History
  - Literary studies
  - Cultural studies
  - Language teaching
  - ….
- Language data is important to communities which are the source of data
  - Language and identity are closely linked
  - Indigenous communities are of special importance in this regard

(Language data science methods are relevant across many other disciplines also – the Australian Text Analytics Platform is part of LDaCA)

Continued access

- Sustainability is a key objective
- Data is sustainable, access systems less so
- LDaCA relies on standards-based data storage and packaging models
  - Our portal solutions will become obsolete
  - New solutions can be built on top of the data and description
- Sustainability also means finding long-term data storage solutions
  - This is an ongoing problem…….

Appropriate control

- The A of FAIR does not mean everyone can access everything
- The CARE principles are a guide to responsible data governance
  - Applied to any data in which people have rights
- Explicit licences are required in our data packaging solution
  - Who can access data and for what uses
- Decisions about access are the responsibility of Data Stewards

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 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.

Whatever else – we need long term data management

AARNet is decommissioning CloudStor on Friday 15 December 2023. FileSender will continue as a standalone service. See FAQs for more information.

Articles in this section

Is there a replacement for CloudStor?

3 months ago · Updated

AARNet does not currently offer a replacement for CloudStor. However, we remain committed to supporting the sector by identifying and filling gaps in the market.

Please contact your institution's IT team to discuss alternative services a

Was this article helpful?

Yes No

1 out of 7 found this helpful

Was this article helpful?

Yes No

1 out of 7 found this helpful

Follow

What Are The Benefits Of Using CloudStor Collections?

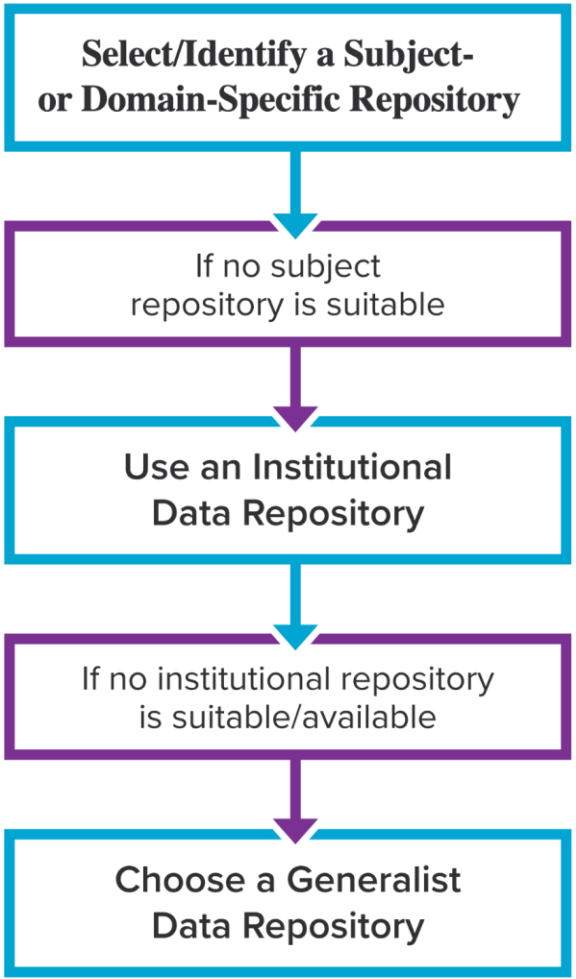
- Streamline data packaging by using the same application used to manage and

- Store valuable research data store in a safe and secure location.

- Store valuable research data store in a safe and secure location.
- Simplify data management planning by using a purposebuilt application for archiving and packaging.
- Have confidence that Collections you send and receive can be validated to determine authenticity and integrity.

RELATED ARTICLES

How does CloudStor Collections Work?
What is CloudStor Rocket?
What are the Cloudstor Collections


Select/Identify a Subject-
or Domain-Specific Repository
If no subject
repository is suitable
Use an Institutional
Data Repository
If no institutional repository
is suitable/available
Choose a Generalist
Data Repository

Annotated Guide to Choosing a Data Repository

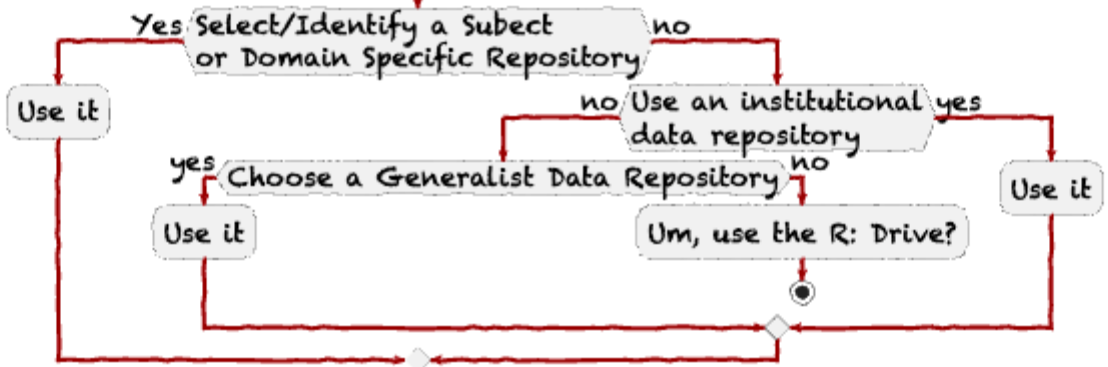
https://ardc.edu.au/resource/guide-to-choosing-a-data-repository/

When deciding what data repository to publish in key considerations include:
- security obligations
- publisher requirements
- community conventions
- institutional policy.

Consider

Also consider:
- the type of data you are publishing
- its format and file size
- its potential reuse and reuse conditions in the form of, say, a licence.

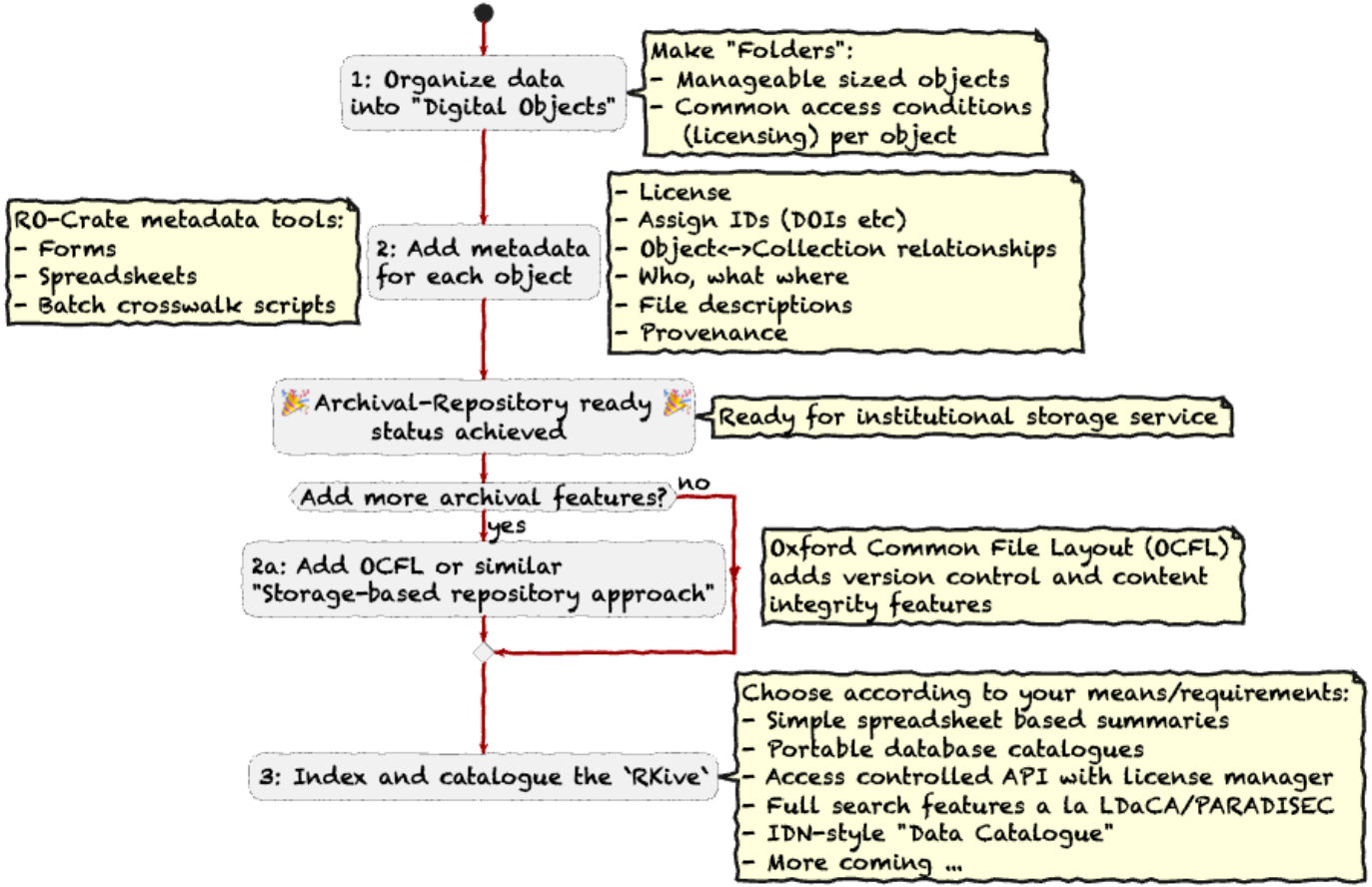
Other considerations can include:
- versioning of the data
- storage location of the data
- sensitivity of the data
- access conditions.

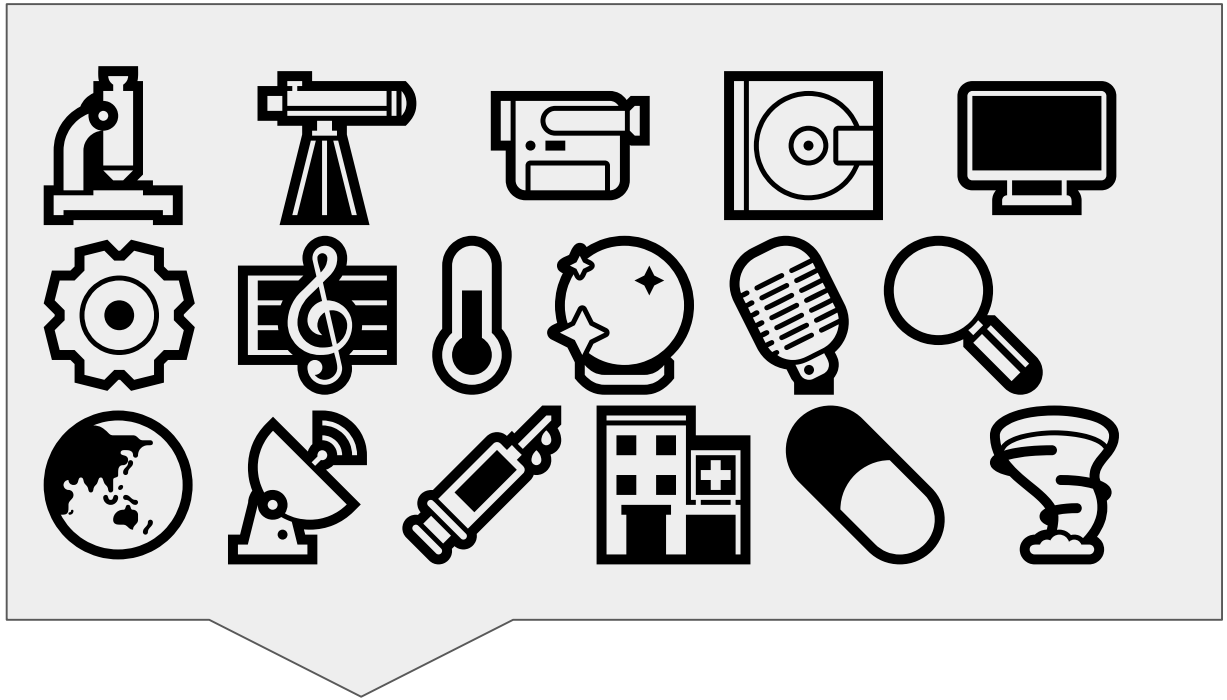


done

If you made it here, you found a disipline, institutional or generalist repo that can address all of the above considerations, sensitivity, access control, data volume

3 Steps to Archival-Repositories (Turn that R: drive into an 'RKive')





The structural elements of a Language Data Commons RO-Crate are:

- A Collection/Object hierarchy to allow language data to be grouped - for example a corpus with sub-corpora, or collections of items (objects) from a particular region.
- Dataset and File entities (as per RO-Crate). Files may be referenced locally or via URI - eg from an API. If an RO-Crate contains files they MUST be linked to the root dataset using `hasPart` relationships as per the RO-Crate specification.

NOTE: The terms Collection and Object are encoded in RO-Crate metadata using RepositoryCollection and RepositoryObject types respectively. These in turn are re-named versions of the Portland Common Data Model types, pcdm:Collection and pcdm:Object.

A conformant RO-Crate:

- MUST have a `@type` attribute that that includes in its values `Dataset` and either `RepositoryCollection` or `RepositoryObject`

Recordings in South Efate

Download all the metadata for Recordings in South Efate in JSON-LD format

Check this crate

Browse files Recordings in South Efate

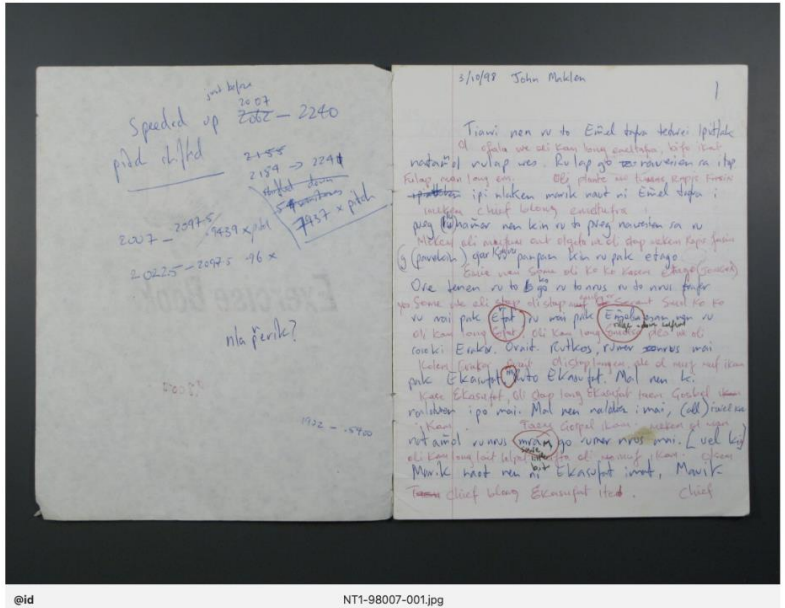
@id	/
name [?]	Recordings in South Efate
@type	• Dataset • RepositoryObject
description [?]	NT1-98007. Text #047 (speaker is John Maklen. Text title: History of villages before Erakor); T Erromango); Text #049. Text title: Asaraf (speaker is John Maklen);Text #050. Text title: Mumu Erakor—the spirit who lives at Erakor (speaker is John Maklen);Text #038. Text title: The need There are time-aligned transcripts of this item and handwritten transcripts by Manuel Wayane
memberOf [?]	https://catalog.paradisec.org.au/collections/NT1
additionalType [?]	item
collector	Nick Thieberger
contentLanguages	• Bislama • South Efate
countries	Vanuatu
dateCreated [?]	2012-09-27T10:08:01.000Z
dateModified [?]	2018-05-17T04:13:04.000Z
depositor	Nick Thieberger
digitisedOn	Mon Jan 01 2001 00:00:00 GMT+1100 (Australian Eastern Daylight Time)
hasPart [?]	• NT1-98007-001.jpg • NT1-98007-002.jpg • NT1-98007-003.jpg • NT1-98007-004.jpg • NT1-98007-005.jpg • NT1-98007-006.jpg • NT1-98007-007.jpg • NT1-98007-008.jpg • NT1-98007-009.jpg • NT1-98007-010.jpg • NT1-98007-011.jpg • NT1-98007-012.jpg • NT1-98007-013.jpg • NT1-98007-014.jpg

Recordings in South Efate

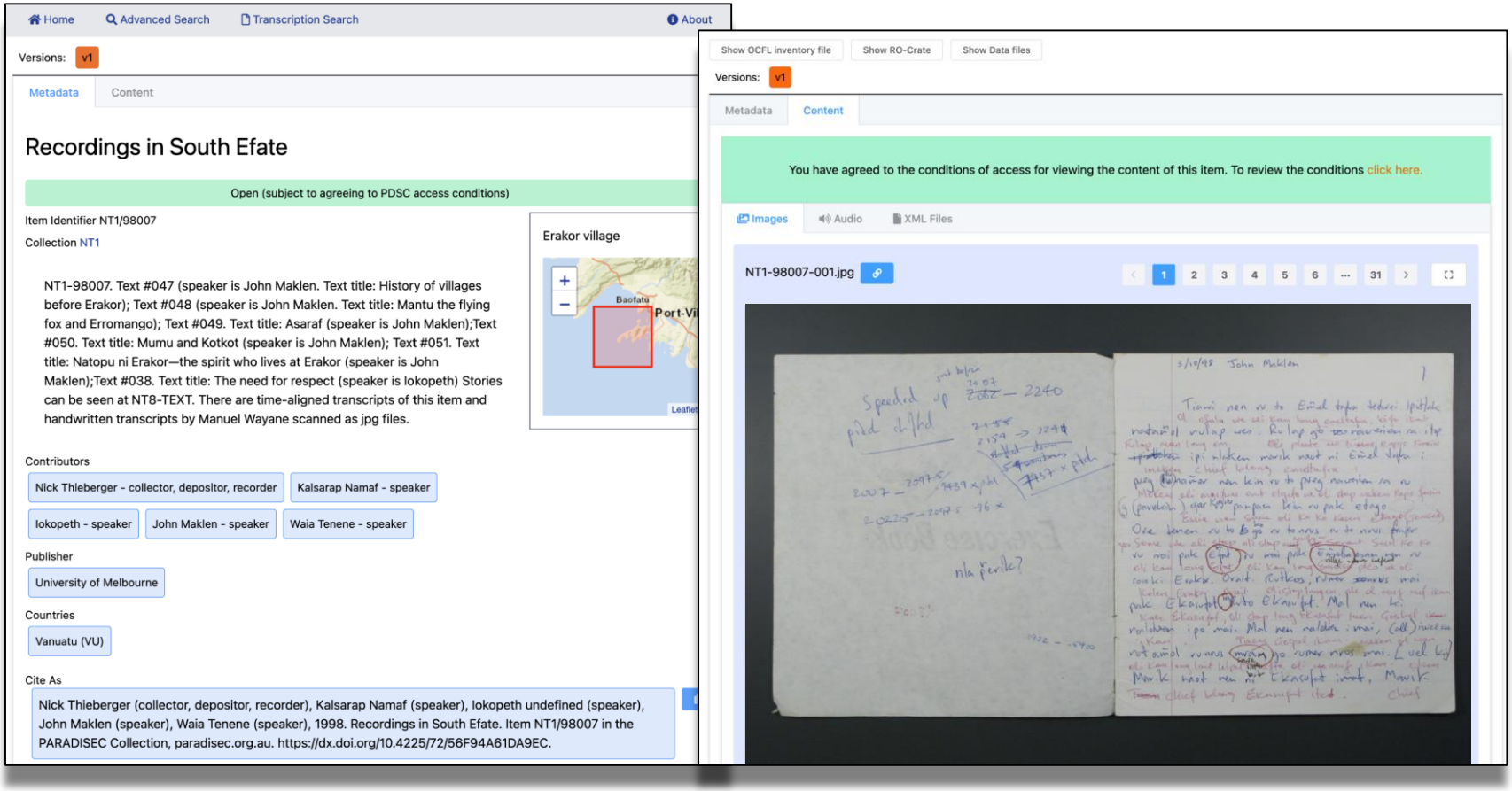
Download all the metadata for Recordings in South Efate in JSON-LD format

Check this crate

Download: NT1-98007-001.jpg



@id	NT1-98007-001.jpg
name [?]	NT1-98007-001.jpg
@type	File
encodingFormat [?]	image/jpeg
contentSize [?]	1658368
dateCreated [?]	2012-09-27T10:08:01.000Z



https://mod.paradisec.org.au

show fields included in search

Collection

1 2 3 4

Filter

- A COrpus of Oz Early English (COOEE) 4071
- Australian Corpus of English 3400
- Braided Channels 395
- AustLit 268
- ICE: S1A: Conversations 202

Access

Filter

- Attribution 4.0 International (CC BY 4.0) 8720
- Attribution-NoDerivs 3.0 Australia (CC BY-ND 3.0 AU) 396
- Data License for AustLit 269

Record Type 12

Language 1

Filtering by: license | Data License for AustLit ×

Total: 269 Index entries (Collections, Objects, Files and Notebooks)

RESTART SEARCH | Sort by: Relevance | Order by: Descending

1 2 3 4 5 6 ... 27

Convict Once

Type: RepositoryObject

Language: English

Member of: AustLit

Search Score: 1

See more

Login Required

convict-once-origin

Type: File DerivedMaterial

Member of: AustLit

XML derivative of the original

Search Score: 1

See more

Login Required

convict-once-raw

Type: File DerivedMaterial

Member of: AustLit

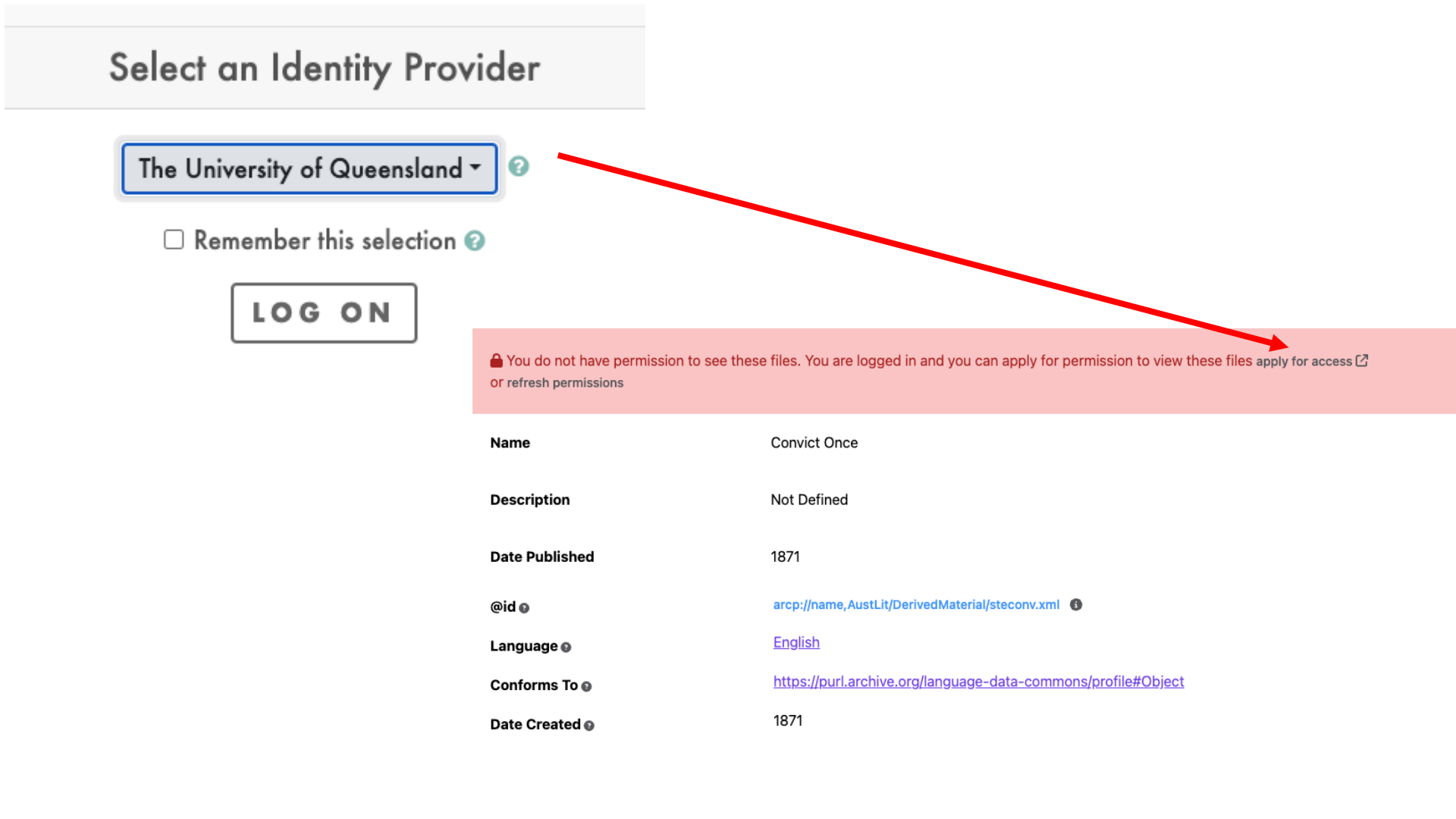
TXT derivative of the original work, contains TEI markup

Search Score: 1

See more

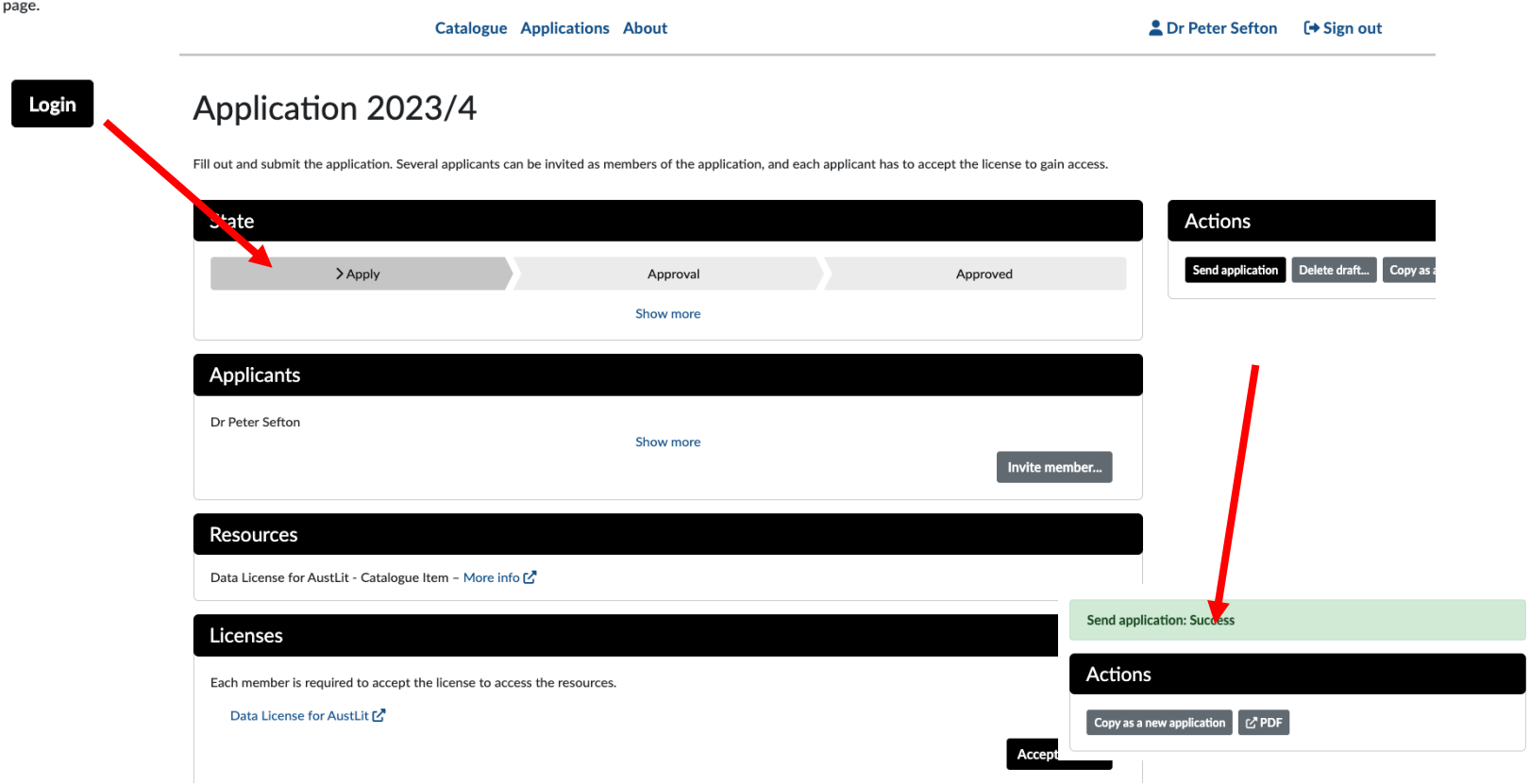
Login Required

Portal shows that accessing this data requires authorization. First step: Log in.


Select an Identity Provider
The University of Queensland
Remember this selection
LOG ON
You do not have permission to see these files. You are logged in and you can apply for permission to view these files apply for access
or refresh permissions
Name
Convict Once
Description
Not Defined
Date Published
1871
@id
arcp://name,AustLit/DerivedMaterial/steconv.xml
Language
English
Conforms To
https://purl.archive.org/language-data-commons/profile#Object
Date Created
1871

Welcome to LDaCA REMS

This is the Resource Entitlement Management System for LDaCA Program. More information is available on About page.
Please, login to access REMS.



State

✓ Apply | ✓ Approval | ✓ Approved

Show more

Access to **Data License for AustLit** granted to Dr Peter Sefton

Name	Convict Once
Description	Not Defined
Date Published	1871
@id	arcp://name,AustLit/DerivedMaterial/steconv.xml
Language	English
Conforms To	https://purl.archive.org/language-data-commons/profile#Object
Date Created	1871
Creator	Stephens, J. Brunton (James Brunton) (1835-1902)
Publisher	University of Sydney Library
Citation	Convict Once