Language Data Commons of Australia - Data Partnerships Project

ARDC AUSTRALIAN DATA PARTNERSHIPS PROGRAM PROJECT FINAL REPORT

Prepared by: Michael Haugh, Robert McLellan, Simon Musgrave, Peter Sefton, Ben Foley, Sam Hames, Sue Plunkett-Cole, Catherine Travis, Nick Thieberger

31/05/2023

CONTENTS

1 PROJECT INFORMATION

INVESTMENT ID	**DP768**
PROJECT START AND END DATES	March 2021 - June 2023
LEAD ORGANISATION	The University of Queensland
PARTNER ORGANISATIONS	ARC Centre of Excellence for the Dynamics of Language (CoEDL) via ANU Australian National University The University of Melbourne
PROJECT CONTACT PERSON	Michael Haugh - Project Director Robert McLellan - Program Manager Sue Plunkett-Cole - Project Coordinator

1.1 Background

Large collections of language data have been amassed in Australia but many remain under-utilised or at risk. The aim of establishing a Language Data Commons of Australia (LDaCA) was to federate these efforts into nationally integrated research infrastructure for collections of high strategic importance for the Australian research community, and for translational research related to the national interest.

LDaCA is enabling a sustainable long-term repository for ingesting and curating existing language data collections of national significance. These collections include intangible cultural heritage of the languages of some of the world's longest continuous cultures in one of the world's most linguistically diverse regions (Aboriginal and Torres Strait Islander languages and regional languages of the Pacific), and data which is important for cyber-security (AusTalk, Australian National Corpus, corpora of regional languages), for gauging popular opinions and sentiment (Australian Twitter Corpus), and for emergency communication (languages of the region and some Indigenous languages).

Some data is in well-established archives (Trove), but much is in institutional repositories subject to varying degrees of sustainability. As a portal to these language records, with associated metadata, LDaCA ensures long-lasting access for analysis and re-use of these invaluable data in a culturally, ethically and legally appropriate manner.

1.2 Achievement of project aims

The Language Data Commons of Australia (LDaCA) aims to make nationally significant language data available for academic and non-academic use and to provide a model for ensuring continued access with appropriate community control.

These broad aims are being implemented by:

1. creating research infrastructure for language collections to be used by the Australian research community and for translational research.
2. ensuring long-lasting access to these invaluable collections for analysis and reuse in a culturally, ethically and legally appropriate manner.

This project has:

- established a Governance Board and impact pathway for key stakeholders,
- developed a language data access policy framework,
- developed shared technical infrastructure and standards across institutions,
- built a portal for discovery and access of language data, and
- engaged researchers and stakeholder communities with this infrastructure.

This project work will continue in the extended LDaCA-RDC Project (2023-2024)

2 DESCRIPTION OF PROJECT OUTPUTS

2.1. Achievements against project work packages:

DELIVERABLE / WORK PACKAGE	DETAILS INCLUDING EXPLANATION OF ANY VARIATION	COMPLETION DATE
1.1: Formation of steering committee	A steering committee was formed for the combined LDaCA Data Partnership + LDaCA HASS Research Data Commons +	1/3/2022

DELIVERABLE / WORK PACKAGE	DETAILS INCLUDING EXPLANATION OF ANY VARIATION	COMPLETION DATE
	Australian Text Analytics Platform and has met quarterly.	
1.2: Establishment of working groups	An Indigenous languages working group was convened at the University of Queensland in December 2021. The project recruited an Indigenous Program Manager to lead this work on an ongoing basis. In consultation with the Steering Committee to further define structured governance and decision-making processes as well as amendments within the existing Terms of Reference with provisions for sub-committee representative groups (1. Science and 2. Communities).	12/2021
1.3: Formation of community stakeholder reference groups	As use case examples develop, we near opportunities to actively engage with various community groups on language data specific to those groups and through practical workshops such as skills workshops (e.g. Computational Workshop in Sydney, Aus Linguistic Society in Melbourne) as well as future conferences and symposiums (e.g. *'Breaking down the Jargon'* Puuliima Language and Tech, and Australian Languages Workshop)	1/7/2022
1.4: Report on stakeholder impact pathways	Impact pathways have been identified to the extent appropriate for the level of maturity of this project and this work will continue in the extended LDaCA Project (2023-2024).	1/3/2023
1.5: Formation of LDaCA Governance Board	Has been coordinated in tandem with the formation of governance structure for the continuing LDaCA-RDC project	31/05/2023
2.1: Report on existing rights restrictions on collections in LDaCA	Completed. Report: https://docs.google.com/document/d/1jqI5CVuVKXX_QklD6y5VOLV1GVrC8pdr/edit?usp=sharing&ouid=113061980073587528996&rtpof=true&sd=true	01/12/2021
2.2: Report on cultural, ethical and legal constraints relevant to LDaCA	Constraints & requirements have been reviewed, detailed at: Data and access policy: https://docs.google.com/document/d/1Ldm95xygIY3QRQlGhdHbIsfLIdmPJP9_sT23XAxDTcU/edit?usp=share_link Access Policy - Indigenous data https://docs.google.com/document/d/1MqQ7xjGXwOZQU79	30/06/2022

DELIVERABLE / WORK PACKAGE	DETAILS INCLUDING EXPLANATION OF ANY VARIATION	COMPLETION DATE
	4vq7plOn7cdZOqcb_/edit?usp=share_link&ouid=100825349241080722744&rtpof=true&sd=true	
2.3: Policy framework for appropriate access to language data	Documentation for the management of data, licencing, access consideration and onboarding to LDaCA has been developed, in the form of three documents: (1) guidelines for managing access to data (of general use to people working with data), (2) LDaCA policies and (3) LDaCA processes. These will be available in early June 2023.	30/06/2023
3.1: Consultation with data custodians	Completed. Documented here: https://docs.google.com/spreadsheets/d/1xMjQ2EPQHJbjOrQEiGv2_7eyxCn3wHWlNEyWTEGSBIs/edit?usp=sharing	31/12/2022
3.2: Migration of AusNC to common (meta)data structure	Completed. AusNC metadata has been migrated to LDaCA format. Data has been migrated from Griffith University to LDaCA storage and made accessible in an LDaCA portal. Some of the AusNC collections are in the portal, not all have been made public yet.	30/06/2022
3.3: Migration of CoEDL corpora	Completed for this project - work will continue in extended LDaCA project, with ongoing discussions with data custodians.	30/06/2022
3.4: Migration of selected collections	These collections have been migrated to the LDaCA format: Auslan signbank (not public yet); The University of Queensland Fryer Library Indigenous data (public at https://ild.ldaca.edu.au/search) Auslan signbank (not public yet)	31/05/2023
3.5: Migration of selected Australian English collections	These collections have been migrated to the LDaCA format: Collections in AusNC (as per 3.2); Sydney Speaks (collections we have received are public at https://ild.ldaca.edu.au/search) Farms to Freeways (public at https://ild.ldaca.edu.au/search)	31/05/2023
3.6: Migration of other selected additional collections	Underway and continuing: Batchelor College CALL collections; Iltyem-Iltyem Indigenous sign language data; Papunya manuscripts; Holmer manuscripts; AusTalk - an audio-visual corpus of Australian English; Auslan corpus data; Australian Accented English.	31/05/2023

DELIVERABLE / WORK PACKAGE	DETAILS INCLUDING EXPLANATION OF ANY VARIATION	COMPLETION DATE
4.1: Consultation with researcher communities about research needs	Consultation with CIs guided user research efforts and outreach to target user groups. A Usability Working Group was formed. Findings from user research sessions was incorporated in the user workflow model (cf. WP 2.3). A user group carried out several tasks in a pilot version of the portal and provided feedback which was used to guide further development.	30/06/2022
4.2: Pilot version of data portal (access and aggregation)	Completed. An initial version of the repository has been developed using the Oni software stack developed at UTS, uses CILogon for authentication and REMS for Authorization.	30/06/2022
4.3: Pilot version of data portal (harvesting workflow)	Harvesting of notebooks is implemented.	31/05/2023
4.4: Full deployment of LDaCA data portal	LDaCA data portal is available online at https://data.ldaca.edu.au	31/05/2023
5.1: Development of outreach and engagement strategy	Outreach and engagement strategy was developed and implemented and continues to evolve with the project. This includes the LDaCA website which was launched at https://www.ldaca.edu.au/	31/03/2022
5.2: Training package on working across collections in LDaCA	The initial version of this package is based on collections identified for early ingestion (e.g. collections in Alveo, Sydney Speaks). See also WP3.	31/12/2022
5.3: Training package on preparing collections for ingestion in LDaCA	Training material has been developed for data ingestion. See also WP2.3.	31/12/2022
5.4: Training package on using language data collections	Training material has been developed for working with language data, and workshops using material have been facilitated in collaboration with other organisations (e.g. Digital Observatory, Summer Institute for Computational Social Sciences, AARNet etc).	01/01/2023
6.1: Tracking data ingestion and access	On track in line with the maturity of the portal.	30/06/2023
6.2: Tracking users and user behaviours	On track in line with the maturity of the portal.	30/06/2023

DELIVERABLE / WORK PACKAGE	DETAILS INCLUDING EXPLANATION OF ANY VARIATION	COMPLETION DATE
6.3: Tracking data sightings in the wild	On track in line with the maturity of the portal.	30/06/2023
6.4: Advocacy with publishers, societies and funders	Project was represented at Digital Humanities Australasia (2021); VALA (2022); Australian Languages Workshop (2022); Australian Linguistics society (2022); eResearch Australasia (2021, 2022); Collections as Data (Canada, 2023), Open Repositories (South Africa, 2023). LDaCA hosted a day of discussion (Aug 2022) which brought together Aboriginal and Torres Strait Islander and non-Indigenous leaders from academia, galleries, libraries, archives, and museums (GLAM) institutions to discuss co-designing a Language Data Commons for Australia. https://ardc.edu.au/resource/bringing-data-to-life-report/	30/06/2023

2.2. Project outputs

Outputs of the Language Data Commons of Australian - Data Partnerships project are detailed below:

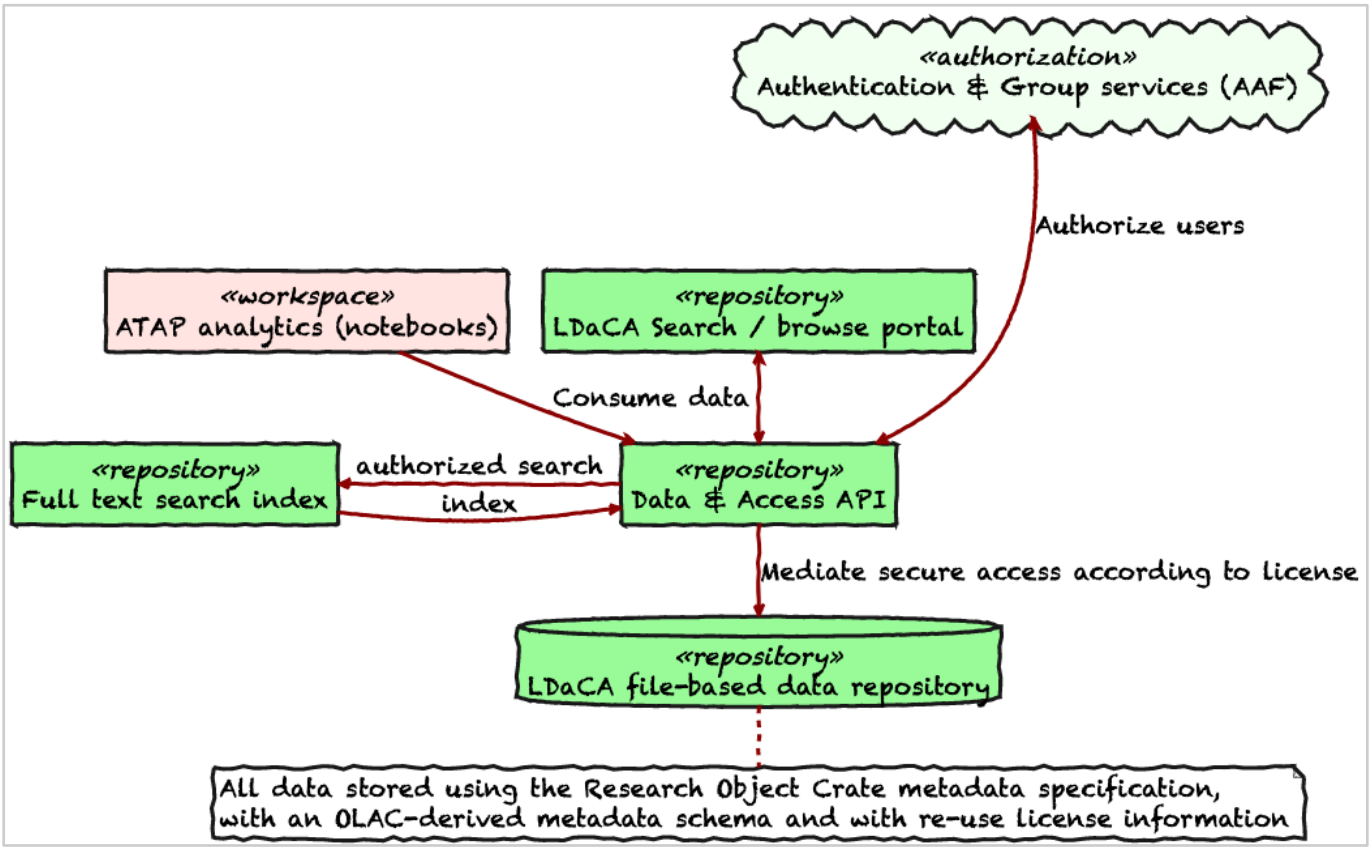
- OCFL - RO-Crate - Oni software stack customised for LDaCA purposes
- Data migration workflows:
    - Tools to input collections to LDaCA tech stack
    - Protocols for discussions with data stewards
    - Training package for data stewards
- Metadata vocabulary for language data
- Development of metadata editing tools commenced
- A CARE and FAIR-ready distributed access control system for language data (AAF/CILogon/REMS)
- Formulation of data access policy
- Documentation of development activities (e.g. Data Access Policy - see section 2.1 (WP2.2) above, Metadata documentation).
- Collections migrated to LDaCA - see section 2.1 (WP3) above.

LDaCA has initiated engagement with the GLAM sector as detailed in section 2.1 (WP6.4). Following on from that activity, work has begun on a catalogue of Oral Histories held in GLAM institutions.

Outputs available online:

- LDaCA data portal https://data.ldaca.edu.au
- Prototype portal which links to Australian Text Analytics Platform notebooks: https://data.atap.edu.au/main
- LDaCA website https://www.ldaca.edu.au

Technical architecture diagram of the LDaCA environment:



2.3 Outreach and training activities

ACTIVITY	ACTIVITY DESCRIPTION	LINK TO MATERIALS	NO. OF PARTICIPANTS	DATE OF ACTIVITY
Presentation to 4th Forum on Australian Englishes, LaTrobe University Accompanying blog post written by Simon Musgrave and published 19th October, 2021 on Sydney Corpus Lab website	Title: Advance Australia FAIR. Presenters: Musgrave, Haugh	Presentation video: https://sites.google.com/view/auseng/videos-2021 Blog post: https://sydneycorpuslab.com/what-are-the-fair-and-care-principles-and-why-should-corpus-linguists-know-about-them/		27/08/2021 19/10/2021
Presentation to RI-VIS Symposium on Europe-Australia Research Infrastructure Cooperation	Title: Language Data Commons of Australia and CLARIN Presenter: Musgrave	https://www.youtube.com/watch?v=0mr5HpRwZ4w		05/10/2021
Panel session as part of Digital Humanities Australasia 2021, University of Otago	Title: Commons and Communities - Initiatives in Research Infrastructure Organiser: Musgrave Participants: ARDC, NLA, ADA, LDaCA, IDN, AAH, AASS	bit.ly/33vrCaF		22/11/2021
Information meeting as part of Annual Conference of Australian Linguistic Society, LaTrobe University	Presenters: Haugh, Travis, Bednarek			09/12/2021

Presentation to workshop on Digital Approaches to Multimedia Text Analysis	Title: Infrastructure for multilingual text analysis Presenters: Musgrave, Sefton	https://ptsefton.com/2022/01/27/DAMTA_Slides_v1/index.html		27/01/2022
Presentation to Data61 Language and Social Computing Group	Title: Infrastructure for language based research Presenters: Foley, Kaiser, Musgrave, Sefton	Data61_18/02/2022 - Google Slides		18/02/2022
Lightning talk for UQ Digital Cultures and Societies Hub	Title: Infrastructure for Multilingual Text Analysis Presenter: Musgrave			01/04/2022
Presentation to the 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources	Integrating Auslan Resources into the Language Data Commons of Australia River Tae Smith, Louisa Willoughby and Trevor Johnston	Published version: https://www.sign-lang.uni-hamburg.de/lrec/pub/22017.html		25/06/2022
Presentation to Australian Languages Workshop	Title: Infrastructure for language based research Presenters: Foley, Musgrave	ALW2022		9/07/2022
Presentation to 3rd Workshop on Sociophonetic Variability in the English Varieties of Australia	Title: Maximising the value of digital infrastructure for sociophonetic research in Australia Presenters: Haugh, Travis, Musgrave	HaughMusgraveTravis_Sociophonetics workshop_12July2022.pptx	40	12/07/2022
Symposium/Roundtable with Indigenous representatives and GLAM institutions	Bringing Data to Life: Co-designing a Language Data Commons	https://ardc.edu.au/resource/bringing-data-to-life-report/		04/08/2022
Presentation to eResearch Australasia 2022	Title: Designing a metadata ecosystem for language research based	https://www.ldaca.edu.au/posts/ldaca-metadata-		18/10/2022

	on Research Object Crate (RO-Crate) Presenters: Sefton, Thieberger, La Rosa, Musgrave, Smith, Sacal Bonequi	ecosystem-eresearch-2022		
Presentation to eResearch Australasia 2022	Title: An interactive tour of the Language Data Commons of Australia and Australian Text Analytics platform Presenter: Sefton			19/10/2022
Presentation to eResearch Australasia 2022	Title: Designing technical platforms for Research Data Commons services Presenter: Sefton	BoF Session		19/10/2022
Presentation to eResearch Australasia 2022	Title: A CARE and FAIR-ready distributed access control system for human-created data using the Australian Access Federation and beyond Presenters: Sefton, Fewster, Sacal Bonequi, Johnstone, Travis, Smith, Carnuccio	https://www.ldaca.edu.au/posts/fair-care-eresearch-2022		19/10/2022
Presentation to 1st Monash Interdisciplinary Workshop on Language	Title: Monash and the Language Data Commons of Australia Presenter: Musgrave	Monash17Nov2022		17/11/2022
Presentation to 1st Monash Interdisciplinary Workshop on Language	Title: Auslan Signbank and Corpus (LDaCA) Presenter: Smith			17/11/2022
Workshop Preceding the conference of the Australian Linguistic	(Day of activities)	PublicDocuments - Google Drive		29/11/2022

Society				
Presentation to the Southeast Asia-Pacific Audiovisual Archive Association (SEAPAVAA) conference	Presenters: Thieberger, Harris	Archiving communities and new technologies in PARADISEC – improving access to recordings in Pacific languages		May 2023
Presentation to the Pacific Islands Universities Research Network Conference	Presenter: Thieberger	Digitisation for Pacific cultural materials – recent developments in the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)		July 2023

3 SUSTAINABILITY PLAN

This project work will be sustained as an element of the LDaCA Project (2023-24) which is a planned continuation of the HASS&I RDC co-investment with UQ and partners. The project plan details the extension and continuation of this project with expectations that, where approved, RIIP funding may be available for a further four years of coinvestment with UQ and partners in the LDaCA Project (2024-2028).

The LDaCA project uses standards-based metadata and data storage systems which are designed to ensure that data can be preserved for the long term, even in the absence of active data portals and other services. This means that when the LDaCA project ends planning for re-homing data is a straightforward process, and costs to reconstitute new services are vastly reduced compared to current practice where data and metadata are not standardised and significant effort is required to re-work data so it can be made accessible.

4 FAIR

4.1 Implementation of FAIR Data Guidelines

FAIR Data Actions

#	FAIR ACTION	DETAILS OF IMPLEMENTATION
1	All data outputs are assigned appropriate PIDs, preferably a DOI	Investigating the use of ARCP IDs for sub-parts of datasets. https://tools.ietf.org/id/draft-soilandreyes-arcp-03.html - these would potentially be linked from a DOI metadata page. A policy document has been produced. Roll out of PID will be in tandem with the release of collections on the LDaCA portal and is the responsibility of data stewards
2	All data outputs have metadata to enable discovery	The project is relying on the RO-Crate technology for data packaging. Inclusion of metadata is an essential aspect of this approach, and the project has developed an RO-Crate metadata profile tailored for language data (based on previous work by PARADISEC). See Metadata for Language Data - Metadata for Language Data (gitbook.io) The RO-Crate standard is well described. Data collections are prepared and converted into appropriate formats as they are published on the LDaCA portal.
3	All data outputs have a record in Research Data Australia	See: https://researchdata.edu.au/language-data-commons-australia/2761167 There are plans to establish an OAI-PMH end point to enable harvesting of metadata by RDA. The federated metadata version of the LDaCA portal will have an OAI-PMH endpoint (this work has been delayed)
4	All data outputs are registered with relevant discipline-specific discovery aggregators (**Recommended**, if they exist)	LDaCA persistent ID policy states that outputs can be pushed into an external data repository (e.g. Zenodo or an institutional repository) from where the user can obtain a persistent identifier and re-use the data.

5	The persistent identifier for the data output being described is included in the metadata	Persistent identifiers are provided by data custodians and released as part of the metadata record of each collection. Persistent identifiers are being rolled out with data collections released on the LDaCA portal when assigned by Data Stewards
6	All data outputs are made as openly available as possible; they are only closed where necessary	LDaCA is committed to making data available as freely as possible. However, this commitment is balanced against the application of the CARE Principles in all cases where the original source of the data has some moral or legal rights in the data. An access-control protocol has been implemented. (Described here – formal publications will follow)
7	All data outputs are made available through a repository	Preliminary version of the repository is currently available at https://data.ldaca.edu.au Data collections will be published through it as they become available. A number of target data sets have been identified and work is in progress for a roll out schedule.
8	All data outputs are available as a download and/or accessible through an open, documented API (where data is not closed)	All datasets published through the LDaCA portal will be available for download and through an API as per restrictions set by the data provider. API Access extends to notebooks and other analytical processes API documentation is available from: https://data.ldaca.edu.au/help
9	If the data outputs are not openly available, there is a clear description on the landing page on how to request access to the data outputs and the conditions that need to be met	Data sets contain statements on their accessibility – users an typically apply for a license (though in same cases it is by invitation only)
10	The persistent identifier for the data output points to a landing page about the data output, even if the data output is not public (open).	The onus on maintaining the persistent identifier will be on the data steward. The project provides advice and support to maintain best practice for persistent identifiers. Relevant documentations and workflows have been developed to assist data providers with persistent identifiers.

11	If the data output is not openly available there is an authorisation and authentication procedure to provide access to the data	See 6 – access control is implemented
12	The persistent identifier for the data output continues to point to a landing page, even if the data output is no longer available, and there is a policy to maintain these landing pages	See 10
13	Data outputs use community-agreed standard data formats (where such agreed formats exist)	Data are described using the Language Data Commons RO-Crate Metadata Profile: https://purl.archive.org/language-data-commons/profile Data formats being used with the LDaCA portal are designed to be easily transportable and adapted by other platforms to ensure maximum sustainability of the data and its independence from technical infrastructure.
14	Metadata for the data output uses community-agreed standards (where such agreed standards exist)	Same as 13
15	Data and metadata use community-agreed vocabularies, data models and ontologies (**Recommended**, preferably internationally agreed ones where they exist)	Same as 13.

16	Metadata contains persistent identifiers for research objects and entities (people, organisations) linked to the data outputs (including ORCIDs, grantIDs, RAIDs, DOIs, IGSNs)	Yes this is covered by the metadata profile in detail. See 13
17	All data outputs are assigned a machine readable licence (preferably CC-BY 4.0)	Yes this is covered by the metadata profile in detail. See 13 CC-BY is NOT preferred for human created data as data need to be able to be taken down at the request of participants. Documentation for data providers will provide advice on licences for data collections in case they don't already have one.
18	The licence information is available in a machine readable form on the landing page that the persistent identifier (for the data output) refers to	Licensing is determined by data custodians and released as part of the metadata record of each collection's RO-Crate.
19	There is a citation statement for the data output on the landing page that the persistent identifier refers to	Citation information is automatically produced as part of the metadata record.
20	Provenance information on the data output is attached alongside the data (**Recommended**)	Provenance information is in RO-Crates in LDaCA
21	Relevant discipline-specific metadata to enable reuse is captured and presented alongside the data output following research community best practice (**Recommended**)	Same as 13

5 PROJECT IMPACT

The ARDC and the Government wish to demonstrate the impact on researchers, industry and the general public of the NCRIS investment.

5.1 Impact stories

The EPIC project based at Monash University has an ongoing program recording medical consultations. This project has sought advice from LDaCA on good practice in collecting and managing data of that type.

The work of the sign language strand of the project (based at Monash University) has had international impact, being presented as part of the LREC -2023 workshop focussing on sign languages:

> Smith, R. T., Willoughby, L., & Johnston, T. (2022). Integrating Auslan Resources into the Language Data Commons of Australia. Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources, 181–186. https://aclanthology.org/2022.signlang-1.28

Closing the Gap through Supporting Australian Indigenous Languages

The role that the maintenance and reawakening of Indigenous languages can play in improving well-being in Australian Indigenous communities has been recognised in the recent National Agreement on Closing the Gap. The continuing LDaCA project will enable community researchers to more readily identify language data held across different archives. LDaCA will facilitate access to that data by creating resources for the GLAM sector to use, to code language content in existing collections, and indexing language resources in a portal. A key aspect of outreach activities will be to make community researchers aware of this capability and to provide training in using it effectively, ensuring timely access – as is critical if community researchers are to take advantage of momentum around support for Indigenous languages. An interdisciplinary team including Indigenous community language professionals, data scientists, linguists, librarians and information specialists has commenced work along these lines, linking institutions and communities and co-designing effective technological solutions.

https://ardc.edu.au/case-study/a-language-data-commons-for-australia/

The well-established PARADISEC research group (Pacific and Regional Archive for Digital Sources in Endangered Cultures) has been leading the charge for language researchers and archivists to work with LDaCA. In April 2023, they published a blog about their partnering with LDaCA, and advocating for the

Language Data Commons of Australia on their website, saying “this work allows us to plan more digitisation projects and to add textual transcripts and summaries to audio items in the collection”.

Blog post from April 2023: https://www.paradisec.org.au/blog/2023/04/paradisec-activity-update-april-2023/

5.2 Research Outcomes Planning

ACTIVITY	DETAILS
Establishment of a monitoring framework (for project outcomes and impacts)	The project management group monitors project outcomes (published notebooks, presentations, training activities and publications) continuously and project impact is being monitored by the Engagement team.
Inputs from research users in design of the infrastructure	The platform is continuously tested by various levels of users through the stages of development.
Partnerships with research translation specialists	Partnerships with the ATAP collaborators USyd (Sydney Corpus Lab), Sydney Informatics Hub and AARNet Engagement specialists continue with the LDaCA phase 2 work. Close liaison with ARDC Communications specialists has resulted in several publications detailing impact.
Establishment of policies, systems and workflows to track uptake (of project outcomes)	Outputs are tracked using Google Forms and Sheets with summary information flowing to reports to the project Steering Committee and to reports such as this one.

6 LESSONS LEARNED

6.1 What Went Well?

- Effective collaboration with partners has benefitted the project enormously.
- Recruiting of team members with excellent technical, training and engagement skills. Additionally, the project team provided access to their broader networks, locating the work in an international context.
- Building on existing technical components allowed rapid development of a stable architecture solution.
- Planned engagement across sectors can have unexpected beneficial outcomes. LDaCA's engagement with the GLAM sector has uncovered a wealth of untapped data in national archives and smaller museums/historical societies in the form of Oral Histories. This means that, for Australian English, there may be much more data available than we had previously thought.