Australian Research Data Commons

Language Data Commons of Australia Project (2024-2028)

Draft Project Plan for Public Feedback

Michael Haugh, Ben Foley, Sam Hames, Robert McLellan, Simon Musgrave, Peter Sefton, Sue Plunkett-Cole

08/03/2024

NCRIS National Research Infrastructure for Australia An Australian Government Initiative

ARDC is enabled by NCRIS

CONTENTS

REVISION HISTORY

Version	Date	Editor	Summary of changes
1.0	5/1/2024	Sue Plunkett-Cole	Draft 1
2.0	23/02/2024	Sue Plunkett-Cole	Draft 2 - new template, update all sections
3.0	08/03/2024	Sue Plunkett-Cole	Submitted draft to ARDC for feedback

1. PROJECT INFORMATION

PROJECT TITLE	**Language Data Commons of Australia (HIR001)**
PROJECT START and END DATES	1 July 2024 - 30 June 2028
CONTRACTING ORGANISATION	The University of Queensland (UQ)
PROJECT LEAD CONTACT PERSON	Michael Haugh
PROJECT MANAGER	Robert McLellan (Program Manager) Sue Plunkett-Cole (Project Coordinator)
FOCUS AREA and ACTIVITY	Language Data Commons of Australia, HASS and Indigenous Research Data Commons, ARDC

1.1. Project aims and outcomes

Australia has many large collections of language data, but many remain under-utilised or at risk. This project leverages existing infrastructure to secure vulnerable and dispersed language collections of written, spoken, multimodal, and signed text, and to link these with improved analysis environments for new research outcomes. The Language Data Commons of Australia (LDaCA) was initiated in 2021 as a national infrastructure project that supports language work and language research as an ARDC co-investment project. It continues to establish sustainable long-term repositories for ingesting and curating language data collections of national significance:

- to democratise access where appropriate to Australia's rich linguistic heritage through enabling those collections to become more compliant with FAIR principles while upholding the same commitment to the CARE principles;
- to develop the computational capabilities, technical infrastructure, and support services to analyse language collections at scale;
- to increase the awareness and skills of researchers in applying digital methods; and
- to open up the social and economic possibilities of Australia's language data for impactful research with significant benefits to the nation.

The project:

- improves researchers' digital skills by providing tools and training which raise awareness of best practice in digital research;
- makes available valuable collections of national significance making them more findable, accessible, interoperable and reusable (FAIR) while adhering to CARE principles; and
- develops the integrated national technical infrastructure to analyse language collections at scale.

It supports researchers to deliver innovative research outcomes, and opens up the social and economic possibilities of Australia's language data for translational research in the national interest by:

- balancing research needs while respecting community rights for language and cultural collections;
- highlighting contributions that language research and HASS disciplines can make to STEM research and non-academic applications; and
- positioning Australia internationally as a leading contributor of language collections and digital infrastructure.

To deliver on the above mentioned aims, the project is executed within the following five key activity streams which provide the basis for monitoring project outcomes:

1. Develop the social and technical foundations for a national, distributed archival repository, including:
    (a) shared, collaborative data governance and standards framework;
    (b) shared data access, authentication and authorisation policies, procedures and processes;
    (c) shared technical infrastructure for curation and storage of language data;
    (d) shared technical infrastructure for collection and annotation of language data.
2. Continue securing vulnerable and nationally significant collections of Aboriginal and Torres Strait Islander languages, Indigenous languages in Australia's Pacific region, (varieties of) Australian English and migrant languages, and sign languages of Australia and its region.
3. Develop a national data portal for accessing and repurposing language data of significance to researchers and communities, both that is held in GLAM institutions, including libraries, archives and museums, as well as language collections held in the distributed archival repositories.

4. Establish an integrated analytics environment for researchers to create fully described, reproducible research on written, spoken, multimodal and signed text in accordance with Open Science principles, and aligned with community expectations for research of practical benefit.
5. Provide training and develop resources for researchers and communities to support best practice in accessing, analysing and archiving language data in line with FAIR and CARE principles.

1.2. Budget

Redacted for publication

1.3. Payment Schedule

Redacted for publication

1.4. Project Partners & Subcontractors

Please include a brief profile of the partner and their role in the Australian research sector, as well as a brief description of their role in the project

ORGANISATION	SUBCONTRACTOR (SELECT BOX IF YES)	SUMMARY
The University of Queensland (UQ) Project lead institution	☐	LDaCA ATAP Language Technology and Data Analysis Laboratory (LADAL) School of Languages and Cultures
Australia's Academic and Research Network (AARNet)	☑	Australia's Academic and Research Network Services for the research sector | AARNet
Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS)	☑	Australian Institute of Aboriginal and Torres Strait Islander Studies Indigenous Research Exchange | AIATSIS

Australian National University (ANU)	☑	Languages and Cultures Sydney Speaks Project | School of Literature, Languages and Linguistics Voices of Regional Australia project receives ARC Discovery Project funding | ANU College of Arts & Social Sciences
Batchelor Institute of Indigenous Tertiary Education (BI)	☑	Batchelor Institute » Research CALL | Centre for Australian Languages and Linguistics
First Languages Australia (FLA)	☑	First Languages Australia Projects — First Languages Australia
Queensland University of Technology's Digital Observatory (QUT-DO)	☑	QUT Digital Observatory (researchdata.edu.au)
The University of Melbourne (UoM)	☑	PARADISEC Nyingarn project — Research Unit for Indigenous Language | Faculty of Arts (unimelb.edu.au) Home — Research Unit for Indigenous Language | Faculty of Arts (unimelb.edu.au) Home — RUMACCC | Faculty of Arts (unimelb.edu.au)
The University of Sydney (USyd)	☑	School of Languages and Cultures - Faculty of Arts and Social Sciences (sydney.edu.au) Sydney Corpus Lab – Discover the Power of Computer-based Text Analysis
University of Western Australia (UWA)	☑	Conservatorium of Music : The University of Western Australia (uwa.edu.au) Mayakeniny Noongar performance resources Maatakitj music

1.5. Project team roles and responsibilities

Redacted for publication

1.6. Governance

The LDaCA Project is governed by a project steering committee which convenes within the scope of the following chartered document: see Appendix B

Key procedural functions of the project's ongoing governance are itemised below for reference:

- Steering Committee meetings - minimum six times a year (no upper limit)
- Key representation within ARDC Technical Advisory Group
- Participation in ARDC Tech Leads meetings
- Weekly meetings of the Project Management Group - including ARDC representation
- Weekly meetings with the Project Leads and (rotating) CIs for progress and monitoring

Communities Advisory Board (on scope):

An initial agenda item for 2024 is the development of a Communities Advisory Group (working title) in which the project may seek specialist advice on matters pertaining to the diverse community groups reflected within the project. As a function of the steering committee, this reform would enact a governance instrument to support the requirement for informed decision-making, incumbent upon all CIs and steering committee members. The terms of the Communities Advisory Group's convention will develop further if pursued by the steering committee and be subject to the terms of reference cited above.

1.7. Scope

Out of Scope

At this stage, we are open to solutions and activities which align to the already developed infrastructure outputs. As the project progresses, any out of scope activities will be defined, particularly during the half-way co-design period.

Naturally, as a research infrastructure project, research activities are out of scope. Unless, solely for all intents and purposes directly correlating with the development of research infrastructure.

In Scope

Activity Stream 1: Social, data and technical architecture of language data repositories

WP-1.1: Policy frameworks and implementation strategies

Enhance LDaCA's language data governance policy framework and implement it through a range of strategies. This work package will address Indigenous Data Governance principles and all language data users and custodians in a way that is consistent with the rights restrictions (moral, cultural and copyright) associated with language data within Australia's legal jurisdiction.

WP-1.2: Language data licences

Continue development of a suite of licensing policies; a language data licensing framework for the management of access, authentication and authorisation; and documentation of implementation processes.

WP-1.3: Tools and processes for authorised access to repositories and workspaces

LDaCA repositories and workspaces need to be trusted and secure, which depends on defining security requirements and monitoring and upgrading them as necessary. This work package further develops authentication and authorization processes and tools based on user identities and roles and conforming to data classification, licences and access conditions.

WP-1.4: Standards-based infrastructure

Development and documentation of sustainable standards-based language data archival repository infrastructure, using Oxford Common File Layout (OCFL) for storage and Research Object Crate (RO-Crate) for consistent linked-data description for FAIR digital objects, and testing of this technology stack at scale.

WP-1.5: Tools for preparing archival-quality language data collections

Language data is currently stored in a myriad of formats, in proprietary databases and following idiosyncratic metadata schemas. This work package aims to develop tools and processes to streamline the conversion of a range of language collections from legacy formats into archive-ready RO-Crate and OCFL standards-based formats.

WP-1.6: Tools for presenting and appropriately reusing language data

The work package will develop infrastructure for presenting language data that is stored in standards-based formats. The deliverables will range from creating simple individual webpages, through to complex websites showcasing vast collections of language material. The work will encompass a range of functions from static display through to potentially facilitating user-contributions to collections.

Activity Stream 2: Securing language data collections

WP-2.1: Aboriginal and Torres Strait Islander language data

This package includes migration of community-identified significant collections of Aboriginal and Torres Strait Islander language data (in all modes and varieties) into the Language Data Commons RO-Crate Profile & OCFL format. Where required, implementation of access frameworks designed by and for Indigenous researchers and communities and non-Indigenous researchers is included. The work package also includes the enhancement of language data workspaces.

WP-2.2: Indigenous language data in Australia’s Pacific region

Migration of community-identified significant collections of Indigenous language data from Australia's Pacific region into the Language Data Commons RO-Crate Profile & OCFL format, and implementation of an access framework that meets the needs of Indigenous researchers and communities and non-Indigenous researchers. Data to be stored on commercial or institutional platforms (including cloud) as a first step to long-term preservation (may involve planned hand-on to other parties).

WP-2.3: Australian English and migrant language data collections

Identification of significant collections of Australian English and migrant language data through consultation with relevant linguistic and cultural groups, and subsequent migration of data into RO-Crate/OCFL formats, with implementation of an access framework for researchers and communities.

Activity Stream 3: Increasing accessibility of language data for researchers and communities through LDaCA Data Portal

WP-3.1: Discovery and annotation of language materials in GLAM and other existing institutions

Develop support for community-driven identification and stand-off annotation of language materials in GLAM institutions by outreach to organisations, including describing project outputs and outcomes, and guiding institutional staff in the project's data management practices and policies. Pilot extension of LDaCA Data Portal to collections held in GLAM institutions by transforming metadata about language materials from existing formats in GLAM collections into RO-Crate and loading that data into an Oni portal for search and discovery.

WP-3.2: Tools for analysing social media language datasets

Extend and improve infrastructure for connecting researchers to social media datasets and text analytics notebooks, including API access, and support usingBinder Hub with the REMS licence manager (authentication framework).

WP-3.3: Language Data Portals

Use the flexibility of the Oni app (UI for the language data archival repository toolkit) to improve the usability and functionality of the main LDaCA portal (and thereby increase its attractiveness to the user community), to trial building discipline-specific discovery tools (potentially engaging with other digital infrastructure projects to enhance tools (Community Data Lab, Digital Observatory)), and to improve methods for taking datasets "on Country".

Activity Stream 4: Improving language and text data analysis environments for researchers and communities

WP-4.1: Consolidated research workflows across LDaCA workspaces and repositories

Demonstrate the existing functionalities and affordances of the LDaCA repositories and workspaces to create end-to-end workflows for users, especially researchers. Identify and close gaps and friction points that impede working between repositories and workspaces.

WP-4.2: Language and text analytics notebooks and tools for HASS researchers and communities

Work with HASS users from various disciplines and communities to develop test and document new text analytics notebooks and support them in their engagement with text analytics, large language models, corpus linguistics, data visualisation, and language collections, and develop notebooks to assist with corpus work, including integrating legacy tools and software.

WP-4.3: Community-controlled workspaces for researchers and communities

Develop flexible LDaCA workspaces (Community Interface Platforms) for analysis of language data (including multimodal data), on the basis of consultation with selected communities.

Activity Stream 5: Communication, engagement and training

WP-5.1: Communication outreach with Australian communities about LDaCA infrastructure and tools

Raise awareness of the LDaCA ecosystem in Australian communities and increase their engagement with the elements of LDaCA which are relevant to their needs.

WP-5.2: Training and support for Australian communities in the use of LDaCA infrastructure and tools

Undertake an active program providing training resources and support mechanisms tailored to the needs of various communities for language data management and offer training in using text analytics resources.

WP-5.3: Communication outreach to international research communities about LDaCA infrastructure and tools

Raise awareness of the LDaCA ecosystem amongst international researchers.

1.8. Implementation and Timing - Work packages and Deliverables

The following are the agreed work packages and deliverables for the project.

WORK PACKAGE / DELIVERABLE	RESPONSIBLE	START QTR	FINISH QTR	ARDC Resources
Activity Stream 0: Project Management				
WP-0.1: Project governance				
WP0.1.01: Steering Committee TOR and minutes of meetings are available	UQ	2024-Jul	2028-Jun	

WORK PACKAGE / DELIVERABLE		RESPONSIBLE	START QTR	FINISH QTR	ARDC Resources
	WP0.1.02: ARDC Technical advisory group minutes / meeting notes are available	UQ	2024-Jul	2028-Jun	Project Admin
	WP0.1.03: ARDC Tech Leads minutes / meeting notes are available	UQ	2024-Jul	2028-Jun	Project Admin
	WP0.1.04: Communities advisory group is created and TOR is available	UQ	2024-Jul	2028-Jun	
	WP0.1.05: Project Management Group meeting notes are available	UQ	2024-Jul	2028-Jun	
	WP0.1.06: Project Leads weekly CI meeting notes are available	UQ	2024-Jul	2028-Jun	
	WP0.1.07: 2024 Co-design input is reflected in project planning	UQ	2024-Jul	2024-Sep	Project Admin
	WP0.1.08: 2026 Co-design input is reflected in any project updates	UQ	2026-Apr	2026-Sep	Project Admin
WP-0.2: Project coordination					
	WP0.2.01: Maintained, up-to-date Gantt chart is available	UQ	2024-Jul	2028-Jun	
	WP0.2.02: Steering Committee meetings were held per agreed schedule	UQ	2024-Jul	2028-Jun	
	WP0.2.03: Accessible project documentation (as appropriate to project stakeholders)	UQ	2024-Jul	2028-Jun	
WP-0.3: Project budget, contracts and reporting					
	WP0.3.01: Executed Partner CRAs	UQ	2024-Jul	2024-Sep	
	WP0.3.02: Acquitted financial transactions to budget	UQ	2024-Jul	2028-Jun	
	WP0.3.03: ARDC reports are submitted and approved	UQ	2024-Jul	2028-Jun	
WP-0.4: Communications					
	WP0.4.01: LDaCA website remains accessible	UQ	2024-Jul	2028-Jun	
	WP0.4.02: Improvements applied to LDaCA website design	UQ	2024-Jul	2028-Jun	
	WP0.4.03: Reached content targets for each communication medium	UQ	2024-Jul	2028-Jun	
WP-0.5: Sustainability					

WORK PACKAGE / DELIVERABLE		RESPONSIBLE	START QTR	FINISH QTR	ARDC Resources
	WP0.5.01: Published website with archival repository principles agreed by LDaCA and other HASS & Indigenous RDC signatories	UQ	2024-Jul	2024-Sep	
	WP0.5.02: Produce written strategy for the adoption of sustainability principles	UQ, ARDC	2024-Oct	2025-Jun	
	WP0.5.03: Produce written strategy for the implementation of long-term repository(s)	UQ, ARDC	2025-Jul	2025-Dec	
	WP0.5.04: Produce written business model and case for long-term sustainability of repositories and workspaces beyond the life of the project	UQ, ARDC	2027-Jul	2028-Jun	
Activity Stream 1: Social, data and technical architecture of language data repositories					
WP-1.1: Policy frameworks and implementation strategies					
	WP1.1.01: Produce internal report, then a paper, about the LDaCA governance model	UQ	2024-Jul	2025-Jun	
	WP1.1.02: Produce internal report of identified consultation opportunities/ reviews with key peak bodies	UQ, FLA, AIATSIS	2026-Apr	2026-Dec	
	WP1.1.03: Produce updated written collection management strategy, based on existing FLA document	FLA & UQ	2025-Jul	2025-Dec	
WP-1.2: Language data licences					
	WP1.2.01: Produce written licensing guides for data custodians, specific to language data across the different stages of its lifecycle	UQ	2024-Oct	2025-Jun	
	WP1.2.02: Processes needed for licensing are available on LDaCA website	UQ	2025-Jul	2025-Dec	
	WP1.2.03: Licences and clauses are available on LDaCA website	UQ	2025-Jul	2028-Jun	
WP-1.3: Tools and processes for authorised access to repositories and workspaces					
	WP1.3.01: Users have working access to services using trusted identities	AARNet	2024-Jul	2028-Jun	
	WP1.3.02: Users are granted licences to access data and tools	AARNet	2024-Jul	2028-Jun	
	WP1.3.03: Services pass pen tests and meet cyber security requirements	AARNet	2024-Jul	2028-Jun	

WORK PACKAGE / DELIVERABLE		RESPONSIBLE	START QTR	FINISH QTR	ARDC Resources
	WP1.3.04: Data is appropriately accessible per collection via access framework and interfaces	UQ, Batchelor	2026-Jul	2027-Jun	
WP-1.4: Standards-based infrastructure					
	WP1.4.01: Crate-O is implemented in the main LDaCA portal, enabling GUI for uploading files and editing metadata	UQ	2024-Jul	2027-Sep	
	WP1.4.02: Test possible data loss and recovery; implement tools for disaster recovery of OCFL repositories	UQ	2024-Jul	2027-Mar	
	WP1.4.03: Create a data steward admin tool that can automate the creation of data portals for stand alone use, for data curation or sharing with small communities.	UQ, AARNet	2024-Jul	2028-Jun	
	WP1.4.04: Repositories and workspaces perform as expected, scale to demand and are continuously improved. Errors and bugs are resolved within service level agreement	AARNet, UQ, ARDC	2024-Jul	2028-Jun	
	WP1.4.05: Create and implement a process for LDaCA partners to request storage (rather than having to source it locally)	ARDC, AARNet	2024-Jul	2027-Sep	
WP-1.5: Tools for preparing archival-quality language data collections					
	WP1.5.01: Researchers without programming experience can create a web archive and export an RO-Crate with suitable metadata (with an explicit licence)	QUT-DO	2025-Jan	2025-Dec	
	WP1.5.02: Scripts/interfaces can be used to interface between specific Batchelor collections and the existing custom infrastructure (databases/ hosting/ archives).	Batchelor, UQ	2024-Jul	2025-Jun	
	WP1.5.03: Keeping Cultures data can be exported into RO-Crate format	Batchelor, UQ	2024-Jul	2025-Dec	
	WP1.5.04: Working RO-Crate export process is added to Lameta	UoM	2024-Jul	2025-Jun	
WP-1.6: Tools for presenting and appropriately reusing language data					
	WP1.6.01: Sites are developed to publish bird app data in new app/web interfaces, making legacy data discoverable	Batchelor, UQ	2024-Jul	2024-Dec	

WORK PACKAGE / DELIVERABLE		RESPONSIBLE	START QTR	FINISH QTR	ARDC Resources
	WP1.6.02: CALL Collection site functionality is replaced by custom Oni, making data discoverable and appropriately accessible	Batchelor, UQ	2025-Jan	2025-Dec	
	WP1.6.03: Deposit webpage or other appropriate community collection tool is created for CALL collection	Batchelor, UQ	2025-Oct	2026-Mar	
	WP1.6.04: Existing language resources and material are rehoused into person-centred collections, making data more discoverable and appropriately accessible	Batchelor, UQ	2026-Jan	2026-Jun	
Activity Stream 2: Securing language data collections					
WP-2.1: Aboriginal and Torres Strait Islander language data					
	WP2.1.01: Bird apps are converted to RO-Crate format	Batchelor, UQ	2025-Jan	2025-Mar	
	WP2.1.02: Batchelor Press, CALL Collection, person-centred collections, Akeyulerre data, bilingual language files and data are converted into RO-Crate format	Batchelor, UQ	2025-Jan	2025-Dec	
	WP2.1.03: A minimum of 4 workshops on working with Indigenous language data have been developed and run, and the related manuscripts/audio recordings have had their data/metadata enriched via the workshops	UoM	2024-Jul	2028-Jun	
	WP2.1.04: Dictionaries of Iltyem-iltyem, Adam Kendon files, person-centred data are built	UoM	2024-Jul	2028-Jun	
	WP2.1.05: Digitised files and associated metadata of at-risk audio recordings of Indigenous language data in WA language centres are converted into RO-Crate format	UWA	tba	tba	
	WP2.1.06: Developed and run a minimum of 3 workshops for community researchers, alongside individually tailored training, to increase content in the LDaCA data portal through use of Nyingarn	UWA	tba	tba	
WP-2.2: Indigenous language data in Australia's Pacific region					
	WP2.2.01: ANU Data Manager is employed and new materials are ingested to PARADISEC at ANU	UoM	2024-Jul	2028-Jun	

WORK PACKAGE / DELIVERABLE	RESPONSIBLE	START QTR	FINISH QTR	ARDC Resources
WP2.2.02: USyd Data Manager is employed and new materials are ingested to PARADISEC at USyd	UoM	2024-Jul	2028-Jun	
WP2.2.03: Digitise and ingest 200 analog tapes from the Pacific per year	UoM	2024-Jul	2028-Jun	
WP-2.3: Australian English and migrant language data collections				
WP2.3.01: Relevant collections of Australian English and community languages are identified, and a catalogue is created	ANU	2024-Jul	2028-Jun	
WP2.3.02: Relevant collections of Australian English and community languages are migrated into standard formats	ANU	2024-Jul	2028-Jun	
WP2.3.03: Relevant Australian English and community languages are received by state/national institution, or are onboarded to LDaCA, and the access framework for researchers and communities is implemented	ANU	2024-Jul	2028-Jun	
WP2.3.04: Project scheme has been run to train data stewards in preparing data for sharing	ANU	2024-Jul	2028-Jun	
WP2.3.05: Pathway has been tried with at least 1 oral history collection	ANU	2024-Oct	2025-Sep	
Activity Stream 3: Increasing accessibility of language data for researchers and communities through LDaCA Data Portal				
WP-3.1: Discovery and annotation of language materials in GLAM and other existing institutions				
WP3.1.01: Community annotations are reflected within the data portal	UQ	2025-Jan	2025-Dec	
WP3.1.02: Information sheets (infographics) on how to engage with project outreach on the LDaCA website	UQ, AARNet	2024-Jul	2025-Jun	
WP-3.2: Tools for analysing social media language datasets				
WP3.2.01: LDACA data portal contains searchable information about the Australian Twittersphere, and users are referred to the Digital Observatory to apply for full-text access	QUT-DO	2024-Jul	2025-Jun	
WP3.2.02: The LDACA notebook suite contains example material guiding people in the use of Australian Twittersphere data	QUT-DO	2025-Apr	2025-Dec	

WORK PACKAGE / DELIVERABLE		RESPONSIBLE	START QTR	FINISH QTR	ARDC Resources
	WP3.2.03: LDACA data portal contains searchable information about the NewsTalk, and users are referred to the Digital Observatory to apply for full-text access	QUT-DO	2026-Jan	2026-Dec	
	WP3.2.04: The LDACA notebook suite contains example material guiding people in the use of NewsTalk data	QUT-DO	2026-Oct	2027-Jun	
WP-3.3: Language Data Portals					
	WP3.3.01: Core search functionalities are identified and developed which satisfy multiple user groups of the data portal	UQ	2024-Jul	2028-Jun	
	WP3.3.02: Proof of concept developed using Sydney Speaks, and has similar functionality to Corpus of American History (COHA)	UQ, ANU, QUT-DO, USyd, CDL	2024-Jul	2028-Jun	
	WP3.3.03: Improved user friendliness and usability of LDaCA resources, documented in change-logs	UQ	2024-Jul	2028-Jun	
	WP3.3.04: At least one collection has been loaded onto a portable server and made available to a community and one made available in a centre	AARNet, UQ	2024-Jul	2028-Jun	
	WP3.3.05: LDaCA portal enables filtering, selecting and downloading metadata, collections (corpora), objects, code and files	UQ	2024-Jul	2028-Jun	
Activity Stream 4: Improving language and text data analysis environments for researchers and communities					
WP-4.1: Consolidated research workflows across LDaCA workspaces and repositories					
	WP4.1.01: LDaCA Data Portal provides clear and accessible mechanisms for (appropriate) bulk downloading of data in appropriate formats for analysis (including for LDaCA workspaces and discipline-standard tools)	UQ, USyd, ANU	2024-Jul	2025-Jun	
	WP4.1.02: LADAL web analytics show that new audiences have been reached for LDaCA data management fundamentals and workflows	UQ	2024-Jul	2028-Jun	
	WP4.1.03: A broad audience of different academic disciplines understand the value of the LDaCA style repositories as places to enhance the discoverability, accessibility and usability of their existing and future language data	UQ, USyd	2024-Jul	2028-Jun	

WORK PACKAGE / DELIVERABLE		RESPONSIBLE	START QTR	FINISH QTR	ARDC Resources
	WP4.1.04: Referenceable live notebook that articulates a vision for working with access-controlled collections	UQ	2025-Jan	2025-Dec	
WP-4.2: Language and text analytics notebooks and tools for HASS researchers and communities					
	WP4.2.01: Documented use of notebooks, tools, associated resources by users from a range of disciplines or backgrounds	USyd, UQ	2024-Jul	2028-Jun	
	WP4.2.02: Availability of redesigned notebooks, tools, associated resources via LDaCA and ATAP	UQ, USyd	2024-Jul	2028-Jun	
	WP4.2.03: Documented use of notebooks, and tools to integrate data from LDaCA and elsewhere with standard disciplinary applications	ANU, UoM, USyd	2024-Jul	2026-Jun	
	WP4.2.04: An ephemeral instance of common corpus tools can be launched on demand and easily connected with the desired data, and data products made FAIR	UQ, ANU, USyd, UoM	2025-Jan	2026-Jun	
WP-4.3: Community-controlled workspaces for researchers and communities					
	WP4.3.01: Produced internal report on scoped study investigating the usefulness, practicality, prospective uptake and costs of an audio visual workspace	UQ, UoM	2025-Jul	2026-Jun	
	WP4.3.02: Produced internal report on scoped study for an online workspace which enables curation and annotation of data collections based on Individuals	UQ	2026-Jul	2027-Jun	
	WP4.3.03: Online workspace created which allows upload, OCR and annotation for historical documents	UQ	2024-Jul	2025-Jun	
	WP4.3.04: Enhanced features are in place: search mechanism, fuzzy search, synonym search and others according to user feedback	UoM	2024-Jul	2028-Jun	
	WP4.3.05: Nyingarn is hosted and available publicly	AIATSIS	2024-Jul	2028-Jun	
Activity Stream 5: Communication, engagement and training					
WP-5.1: Communication outreach with Australian communities about LDaCA infrastructure and tools					

WORK PACKAGE / DELIVERABLE		RESPONSIBLE	START QTR	FINISH QTR	ARDC Resources
	WP5.1.01: Deliver at least one language-focused event (e.g. ALS), at least one Indigenous focused event (e.g., Puliima), at least one other event (e.g. AHA) through the project period	UQ	2024-Jul	2028-Jun	
	WP5.1.02: Outputs are tracked and reported to ARDC	UQ, All partners	2024-Jul	2028-Jun	
	WP5.1.03: Deliver four events per year, two at universities and two in other forums	UQ	2024-Jul	2028-Jun	
	WP5.1.04: Deliver or contribute to one event per year at a different location each year, plus one event per year in association with AIATSIS	UQ	2024-Jul	2028-Jun	
WP-5.2: Training and support for Australian communities in the use of LDaCA infrastructure and tools					
	WP5.2.01: Batchelor course material includes up to date info about archiving technology (example of the guides in use)	Batchelor, UQ	2024-Jul	2024-Dec	
	WP5.2.02: Online training of LDaCA text analytics notebooks and tools is available; and regular workshops delivered on training in text analytics and related fields such as corpus linguistics	UQ, USyd	2024-Jul	2028-Jun	
	WP5.2.03: Online training of LDaCA data management practice is available; and regular workshops delivered on managing language data	UQ	2024-Jul	2028-Jun	
	WP5.2.04: Track outputs of providing advice to researchers on language data management practice and report outputs to ARDC	UQ	2024-Jul	2028-Jun	
	WP5.2.05: Workshops delivered to Language Centres about collection management (developing local strategies & implementation)	UQ, FLA	2024-Jul	2027-Mar	
	WP5.2.06: Workshops delivered to Language Centres about rights and licensing	UQ, FLA	2025-Jan	2025-Sep	
	WP5.2.07: Track outputs which raise awareness of contemporary concepts and methods for the use of language data in computational HASS and report outputs to ARDC	USyd, UQ	2024-Jul	2028-Jun	
WP-5.3: Communication outreach to international research communities about LDaCA infrastructure and tools					

WORK PACKAGE / DELIVERABLE		RESPONSIBLE	START QTR	FINISH QTR	ARDC Resources
	WP5.3.01: Publications, workshops, and presentations with international audiences using or showcasing LDaCA resources are tracked and reported to ARDC	UQ, ANU, USyd	2024-Jul	2028-Jun	
	WP5.3.02: Engagement with international researchers regarding LDaCA resources are tracked and reported to ARDC	UQ, ANU, USyd	2024-Jul	2028-Jun	
	WP5.3.03: Engagement of international researchers with LDaCA resources is tracked and reported to ARDC	UQ, ANU, USyd	2024-Jul	2028-Jun	
	WP5.3.04: A register is retained of individuals or institutions with shared interests or goals with whom closer engagement might be explored	UQ, ANU, USyd	2024-Jul	2028-Jun	
End of table					

1.9. Visual Summary Overview

Please refer to Appendix C for a visual summary of the project.

1.10. Assumptions

The following assumptions are made in order to deliver successful project outcomes.

ITEM #	CATEGORY (Scope/cost/quality)	DESCRIPTION
1	Quality	Data custodians are willing to work with the project to allow users to access collections while preserving and enforcing the requirements of data owners and communities.
2	Scope, quality	Community groups as reflected within the project are willing to engage with the project in good-will as both contributors to and beneficiaries of the outcomes.
3	Scope, quality	Machine-access to national data collections will be made possible by custodians whenever permissible.

4	Cost, scope, quality	Transformation and migration of existing collections will be facilitated by custodians whenever permissible.

1.11. (Inter)dependencies

DEPENDENCY	RELATIONSHIP TO / IMPACT ON PROJECT	HOW AND WHO WILL MANAGE THE DEPENDENCY
ARDC cloud computer and storage resources	As required	Project Director, Program Manager and Project Coordinator will meet frequently with the HASS&I RDC Director
ARDC services	As required	Project Director, Program Manager and Project Coordinator will meet frequently with the HASS&I RDC Director
Access to skilled staff to support the infrastructure	Skilled staff will continue to be employed for LDaCA and ATAP projects, and new staff will be recruited as needs arise	
Access to particular ARDC expertise	Participation of HASS&I RDC Director	Project Director, Program Manager and Project Coordinator will meet frequently with the HASS&I RDC Director
Outcomes of the 2021-24 LDaCA-RDC project	This project ends in mid-2024, and if co-investment continues, the LDaCA project (2024-2028) will continue to build on these outputs	The project management group are continuing on to manage the LDaCA project and outputs are expected to be delivered successfully

1.12. Risks

In the Controls/Mitigation Strategy section include what preventative actions you plan to take and/or actions you might take should the preventative actions fail to control the risk (i.e. what's your plan B?). These might include applying other in-kind resources, reviewing the plan and reducing scope etc.

Risk Rating Key

		Consequence				
		Insignificant (1)	Minor (2)	Moderate (3)	Major (4)	Significant (5)
Likelihood	Almost certain (5)	5	10	15	20	25
	Likely (4)	4	8	12	16	20
	Possible (3)	3	6	9	12	15
	Unlikely (2)	2	4	6	8	10
	Rare (1)	1	2	3	4	5

RISK	**IMPACT TYPE AND HOW WILL IMPACT PROJECT** (Scope/cost/quality/ schedule)	**CONTROLS/MITIGATION STRATEGY**	**RESIDUAL RISK RATING** (after controls are in place)	**RISK OWNER**
Difficulty recruiting specialised staff in the data/language community	Quality, cost, schedule	Leverage the collaborative partners, community board, and networks of the previous LDaCA and ATAP projects. All critical roles are already filled with skilled staff.	Moderate	UQ & partner organisations
Access to some language collections may be restricted	Scope, quality	Continue to seek from broad sources to secure access to language collections	Moderate	UQ & partner organisations
Language infrastructure may be lacking in some language fields	Quality, schedule, cost	Identify early the language fields requiring investment, and assess when and how these will play a role in the project	Moderate	UQ & partner organisations

**Only include risks that have a rating of greater than 14 to the project.*

1.13. Outputs and Outcomes Monitoring and Evaluation Plan

The indicators below specify what will be measured in the project M&E process in order to assess whether, and to what extent, the project's key intended outputs outcomes have been achieved.

End of Project Outputs are the deliverables achieved as part of the project.

OUTPUT	INDICATOR/S	MEASURE	DATA SOURCE/S	TIMELINE FOR DATA COLLECTION	RESPONSIBILITY	BASELINE
[Output delivered as part of the project]	**[An indicator that will demonstrate the success of the outcome]**	**[A measure of the indicator, that shows the change]**	**[Where will you source the data to support the indicator? How will you collect the indicator data? Is it quantitative / qualitative e.g. surveys, interviews]**	**[when will the data be collected and how often]**	**[Who will be collecting this information]**	**[Where available, what is the baseline Indicator for the outcome]**
Delivery of open source platform	Demand for use, number of citations of the platform	% subscribed/oversubscribed, # of citations by publications by authors using the platform	Platform data, Data from DOI cites.	6 monthly	Project manager	No baseline

OUTPUT	INDICATOR/S	MEASURE	DATA SOURCE/S	TIMELINE FOR DATA COLLECTION	RESPONSIBILITY	BASELINE
EOP Output 1 - Delivery of multiple portals for discovery of and access to language collections	Portals publicly available	Usage statistics, # of citations by publications by authors using the platform Users of the data portal are able to: - download search results - do different kinds of searches - see key collection information in the results	Data from portals listing numbers of collections / objects / languages covered / number of different licences - size	Six monthly	Program Manager	1 Language Data Portal publicly available
EOP Output 2 - Secured language collections	Collections secured	Amount of data stored 16 new collections secured Report on compliance with FAIR	Data from portals listing numbers of collections / objects / languages covered / number of different licences	Six monthly	Program Manager	39KB index entries stored 80GB data stored 8 collections

OUTPUT	INDICATOR/S	MEASURE	DATA SOURCE/S	TIMELINE FOR DATA COLLECTION	RESPONSIBILITY	BASELINE
		CARE principles in progress reporting Appropriate governance is documented				
EOP Output 3 - Resources for language analysis	Additional tools developed Tools and data linked in usable workflows	Number of tools available (number to be determined at a later date) Publication of documented workflows	ATAP and LADAL platforms	Six monthly	Project manager	Tools: 6 Workflows: 0

End of Project Outcomes are the direct changes that occur from the outputs of the project that can be achieved within the timeframe of the investment.

OUTCOME	INDICATOR/S	MEASURE	DATA SOURCE/S	TIMELINE FOR DATA COLLECTION	RESPONSIBILITY	BASELINE
e.g. Families are eating a more healthy diet	Household budget spend for fruit and vegetables increases	Percentage of total household spend on fruit and vegetables increases by 20 percent	Household grocery receipts Household survey	Receipts collected as part of a survey at the beginning of the project. Receipts collected as part of a survey mid way and at the end of the project.	Project Manager	Of the study group, less than 5 percent of the household grocery budget was spent on fruit and vegetables.

EOP Outcome 1 *Enable effective archival infrastructure*	Technology stack and platform enables systems-neutral migration of language (meta)data (as per Activity Stream 1)	Technology stack functions at scale	Technical Advisory Group in collaboration with ARDC	Across project period, stack functioning at scale across multiple repositories	UQ and partners	Language data held existing archive infrastructure is tied into that infrastructure (i.e. it cannot be easily extracted from it)
EOP Outcome 2 *Secure relevant data*	A select number of language collections will be secured through migrating	Selected language collections are held in a secure digital format	Project Manager in collaboration with ARDC	Across project period	UQ and partners	Targeted collections are currently at risk of being lost

	into RO-Crate/OCFL format (as per Activity Stream 2)					
EOP Outcome 3 *Increase accessibility of language data for researchers and communities through LDaCA Data Portal*	A select number of language collections will be made available through the LDaCA Data Portal (as per Activity Stream 3)	Selected languages are made more FAIR (as per Activity Stream 3)	Project Manager in collaboration with ARDC	Across project period	UQ and partners	Targeted collections are currently difficult for researchers to access
EOP Outcome 4 *Improving language and text data analysis environments for researchers and communities*	A select number of analytics notebooks will be made available to researchers (as per Activity Stream 4)	Record uptake: number of users [quantitative] Impact case studies [qualitative]	Project Manager in collaboration with ARDC	Across project period	UQ and partners	Open source (i.e. non-proprietary) analytics tools are not easily implemented by HASS researchers
EOP Outcome 5 *Enable capable people and effective institutions*	A select number of workshops and research showcases will be developed and run	Survey participants of workshops [qualitative] Impact case studies [qualitative]	Project Manager in collaboration with ARDC	Across project period	UQ and partners	A relatively small number of HASS researchers are using analytics in their research

1.14. Communications & engagement

The LDaCA project 2024-2028 will expand upon the existing engagement strategy of the Language Data Commons of Australia. This will be done through:

- Conducting stakeholder surveys and user community consultations
- Organising webinars
- Delivering education and training workshops building on our already established networks
- Identifying additional user groups and expanding our network to include them
- Organising workshops targeting communities
- Participating in researcher development and mentoring programs
- Extending our communication strategy to reach a larger audience
- Developing strategic partnerships with selected academic and non-academic institutions to develop discipline-specific applications.

The stakeholders with which this plan is concerned include the following groups. Most of these groups can take different roles, such as data contributors or data users, depending on context:

- Research-related:
    - Research institutions, consortiums and programs
    - Research communities, e.g. researchers, research support staff, research librarians
    - Independent researchers
    - Citizen researchers
- Institutions and service providers:
    - Higher education providers
    - Other education providers
    - GLAM (galleries, libraries, archives, and museums) institutions
    - Other cultural institutions
    - Government organisations and bodies
    - Aboriginal and Torres Strait Islander-controlled organisations
    - National broadcasters
- Community-related:
    - Community organisations and programs
    - Aboriginal and Torres Strait Islander communities
    - Migrant communities

- Sign language communities
- Other communities, e.g. Australian English speech community

- General public

There are particular groups within these stakeholders that are a greater focus for engagement and who will form the core audiences for our activities, namely key data contributors (such as GLAM institutions) and key data users such (as researchers).

Another group of stakeholders in the LDaCA project consists of organisations and individuals involved in project administration, such as the Australian Research Data Commons (ARDC), other HASS & Indigenous Research Data Commons streams and project partners.

2. GLOSSARY OF TERMS

TERM	DESCRIPTION
ATAP	Australian Text Analytics Platform
Bird app / bird app data	The bird apps were made many years ago as a way to test low-cost language resource app development. The apps featured licensed photos and bird-call audio for bird species across the continent. Each community involved in the project selected a collection of birds, and published a localised version of the app with audio and text of the name in language of each bird, an example sentence and translation.
CDL	Community Data Lab (an ARDC integration project)
FAIR and CARE principles	The CARE principles complement the existing FAIR principles, which require data to be **findable, accessible, interoperable and reusable**. While the FAIR principles are about making it easier to share and reuse data, the CARE principles ensure that data is used ethically. CARE stands for: **Collective benefits; Authority to control; Responsibility; Ethics** CARE Principles | ARDC
HASS & Indigenous RDC	Humanities, Arts and Social Sciences (HASS) and Indigenous Research Data Commons
IDN	Indigenous Data Network
IRISS	Integrated Research Infrastructure for the Social Sciences
LDaCA	Language Data Commons of Australia
OCFL format	Oxford Common File Layout format
PARADISEC	Pacific and Regional Archive for Digital Sources in Endangered Cultures
RO-Crate	Research Object-Crate: a community effort to establish a lightweight approach to packaging research data with their metadata.
Partner organisations’ acronyms	
AARNet	AARNet
AIATSIS	Australian Institute of Aboriginal and Torres Strait Islander Studies
ANU	Australian National University

BI	Batchelor Institute of Indigenous Tertiary Education
FLA	First Languages Australia
QUT-DO	Queensland University of Technology’s Digital Observatory
UoM	The University of Melbourne
UQ	The University of Queensland
USyd	The University of Sydney
UWA	University of Western Australia

3. CHANGE CONTROL (for ARDC information only)

Approval of this Project Plan will comprise the baseline for the project. Changes to any of the following are considered variances:

- Project details
- Project outcomes & aims
- Budget
- Project partners
- Project team roles and responsibilities
- Governance
- Milestones and deliverables
- (Inter)dependencies

Variances to the Project Plan require endorsement by the Steering Committee and then ARDC approval. If approved, the Project Plan will be revised and project reports from that point forward will report project progress against the revised Project Plan, not the original.

To request a variance:

1. The Steering Committee submits a request to ARDC for variance to the approved project plan.
2. ARDC reviews the changes and advises the project manager or project lead of the outcome.

4. APPENDICES

Redacted for publication