Language Data Commons of Australia (LDaCA)

Australia is a massively multilingual country, in one of the world's most linguistically diverse regions. More than a quarter of the world's languages are spoken in Australia and its region. However, while Australia's rich linguistic heritage is well documented, much of this data remains inaccessible or under-utilised due to barriers to accessing collections, as well as a lack of access to tools and skills for analysing that data at scale. The goal of the **Language Data Commons of Australia** (LDaCA) is to work collaboratively with researchers, communities and institutions to develop an integrated national infrastructure for analysing spoken, written, signed and multimodal text collections at scale in order to open up the social and economic possibilities of Australia's rich linguistic and cultural heritage for impactful research with significant benefits to the nation. LDaCA is making available valuable collections of national significance more findable, accessible, interoperable and reusable (FAIR) while adhering to CARE principles; developing the computational infrastructure and tools required to analyse language collections at scale; and increasing the awareness and capabilities of researchers in applying digital methods to language and text data.

LDaCA was initiated in 2021 as a national infrastructure project that supports language work and language research through co-investment from the Australian Research Data Commons (ARDC). LDaCA is led by the University of Queensland (UQ) in partnership with AARNet, ANU, Batchelor Institute of Indigenous Tertiary Education, First Languages Australia, Queensland University of Technology, University of Melbourne, University of Sydney, and University of Western Australia. As a key focus is on Indigenous languages, LDaCA adheres to an Indigenous Data Governance framework developed in collaboration with ARDC and IDN.

The first two phases of LDaCA (2021-2024: HIR001; 2024-2028: HIR024) have focused on:

1. Developing the social and technical foundations for a national, distributed archival repository for language and text data, 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; and(d) shared technical infrastructure for collection and annotation of language data.
2. Continuing to secure 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. Developing a national data portal for accessing and repurposing language and text data of significance to researchers and communities, both that is held in GLAM institutions, including libraries, archives and museums, as well as language and text collections held in other distributed archival repositories.
4. Establishing an integrated analytics environment for researchers to create fully described, reproducible research on written, spoken, signed and multimodal text in accordance with Open Science principles, and aligned with community expectations for research of practical benefit.
5. Providing training and develop resources for researchers and communities to support best practice in accessing, analysing and archiving language and text data in line with FAIR and CARE principles.

Key existing components of the Language Data Commons of Australia (LDaCA) include: the PILARS protocols for sustainable research infrastructure, RO-Crate as an implementation-neutral approach to describing data, and the ONI data portal for making data available to human and machine agents with appropriate security controls for data capture and access, and LDaCA Analytics for language and text analysis, including the Australian Text Analytics Platform (ATAP) and the Language Technology and Data Analysis Lab (LADAL). These components of the LDaCA infrastructure have been

purposefully designed to be maximally adaptable to a wide range of research disciplines across HASS and beyond.

The goal of LDaCA in the next NCRIS Roadmap is to provide the technical architecture and collaborative blueprint for an integrated HASS NCRIS capability, and to contribute to the technical foundations of an Indigenous NCRIS capability in line with Indigenous Data Governance principles and community expectations. The LDaCA social and technical framework provides for the provision of key cross-cutting services for research with unstructured text data for HASS and Indigenous researchers and communities, including:

1. National Research Data Archival Repository
2. Unstructured Data Transformation and Repurposing
3. AI-Enabled Text Data Capture and Research Workflows
4. Text Analytics: Tools and Workbenches
2. Digital Methods Training and Research Support
3. Data and Infrastructure Governance

Given Australia is in one of the most linguistically diverse regions in the world, and the world is shifting to language based technologies, there are huge economic opportunities for Australia in developing world-leading infrastructure for language and text data. The establishment of a National Research Data Archival Repository for not only language and text data, but HASS research data more broadly, is central to the success of this endeavour.

From 2028-2032, LDaCA anticipates working in collaboration with other key partners to develop an integrated HASS NCRIS capability, as well as contributing to the establishment of an Indigenous NCRIS capability as appropriate. This will involve working strategically with other focus areas of the ARDC-supported HASS and Indigenous Research Data Commons and other relevant existing NCRIS capabilities in order to leverage existing and emerging relationships with key partners, including GLAM (e.g. AIATSIS, NSLA), community organisations (e.g. Language Centres), private industry (e.g. Amazon, Google), international research infrastructures (e.g. CLARIN, DARIAH), as well as relevant Government and NGO stakeholders. This will require a significant new investment that capitalises on investments to date in the development of LDaCA and other focus areas of the HASS and Indigenous RDC, as well as the development of research infrastructure for HASS and Indigenous researchers and communities more broadly.