Contribution ID: **245** Type: **Presentation**

# The Language Data Commons of Australia: Supporting research for diverse communities

*Monday 13 October 2025 15:36 (11 minutes)*

Australia is a massively multilingual country, in one of the world's most linguistically diverse regions. Significant collections of this intangible cultural heritage have been amassed, including collections of Aboriginal and Torres Strait Islander languages, Australian Englishes, and regional languages of the Pacific, as well as collections important for cyber-security and for emergency communication. The Language Data Commons of Australia (LDaCA) is integrating this existing work into a national research infrastructure while also securing at-risk collections and improving access to under-utilised collections. LDaCA is thus ensuring that these invaluable resources will be available for analysis and reuse in the future, and that they will be managed in a culturally, ethically and legally appropriate manner guided by FAIR (Wilkinson et al. 2016) and CARE (Carroll et al. 2020) principles. The project aims to make nationally significant language data available for academic and non-academic use while providing a model for ensuring continued access with appropriate community control.

To deliver on the above-mentioned aims, the LDaCA is being developed through five key activity streams:

1. Developing the social and technical foundations for a national, distributed archival repositories ecosystem.
2. 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. Developing a national portal for accessing and repurposing language data of significance to researchers and communities.
4. Establishing 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. Providing training and resources for researchers and communities to support best practice in accessing, analysing and archiving language data in line with FAIR and CARE principles.

LDaCA is based at The University of Queensland (Brisbane, Australia). The project is part of the Australian Research Data Commons (ARDC) HASS and Indigenous Research Data Commons (HASS&I RDC) currently with co-investment from nine institutions and organisations. Given its aims, the establishment of LDaCA over the past four years has been dependent on developing collaborative partnerships across institutional boundaries and closely engaging with a range of different communities. This has involved, in turn, the need to foster new connections and raise awareness and capacity amongst a diverse range of stakeholders. The success of LDaCA depends not only on leveraging existing and previous language infrastructures in Australia in ways that support their respective aims (Musgrave & Haugh 2020), but also in ways that ensure researchers and communities have a real voice in the development of Australia's languages infrastructure. It has become increasingly evident that the success of LDaCA not only involves meeting a diverse range of needs, but also ensuring that researchers and communities can see themselves in that infrastructure.

We illustrate these points with two case studies from opposite ends of the continuum between local and global. Locating, securing and improving access for materials from Indigenous languages is a key activity for LDaCA. In many cases, the group for whom such material is relevant (even crucial) is a small community with close internal connections. Access control for the material may be important to such a community and LDaCA works to implement such control under community guidance, even if this may restrict the possibilities for academic research (Foley et al. 2024). At the opposite end of the scale, LDaCA has begun working with public interest documents, such as Federal Hansard, the record of the Commonwealth parliament. This material is openly accessible, but there is nevertheless a role for LDaCA in making it easier to work with for researchers from

a wide range of disciplines, and also in collaborating with the ParlaMint project associated with the CLARIN network in Europe (Erjavec et al. 2024) to make Australian data available and useable for an international research community.

References:

Carroll, Stephanie Russo, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, et al. 2020. The CARE Principles for Indigenous Data Governance. Data Science Journal 19. 43. https://doi.org/10.5334/dsj-2020-043.

Erjavec, Tomaž, Matyáš Kopp, Nikola Ljubešić, Taja Kuzman, Paul Rayson, Petya Osenova, Maciej Ogrodniczuk, et al. 2024. ParlaMint II: advancing comparable parliamentary corpora across Europe. Language Resources and Evaluation. https://doi.org/10.1007/s10579-024-09798-w.

Foley, Ben, Peter Sefton, Simon Musgrave & Moises Sacal Bonequi. 2024. Access control framework for language collections. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti & Nianwen Xue (eds.), Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024), 113–121. Torino, Italia: ELRA and ICCL. https://aclanthology.org/2024.lrec-main.10.

Musgrave, Simon & Michael Haugh. 2020. The Australian National Corpus (and beyond). In Louisa Willoughby & Howard Manns (eds.), Australian English Reimagined. Abingdon: Routledge.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3(1). 160018. https://doi.org/10.1038/sdata.2016.18.

**Primary author:** Prof. HAUGH, Michael (The University of Queensland)

**Co-authors:** Dr FOLEY, Ben (Language Data Commons of Australia); Dr HAMES, Sam (Language Data Commons of Australia); MUSGRAVE, Simon

**Presenter:** Prof. HAUGH, Michael (The University of Queensland)

**Session Classification:** Presentations Session 1: CAREful Indigenous Data Governance

**Track Classification:** SciDataCon2025 Specific Themes: Infrastructures to Support Data-Intensive Research - Local to Global