Access control framework for language collections

Ben Foley, Peter Sefton, Simon Musgrave, Moises Sacal Bonequi
The University of Queensland
Australia
{b.foley, p.sefton, s.musgrave, m.sacalbonequi}@uq.edu.au

Abstract

This paper introduces the licence-based access control framework developed by the Language Data Commons of Australia (LDaCA) for a range of language collections, with examples given of implementation for significant Indigenous and Australian English collections. Language collections may be curated for many reasons, such as documentation for language revival, for research, security or commercial purposes. Some language collections are created with the intention of being "Open Access"; publicly available with no restriction. Other collections require that access be limited to individuals or groups of people, either at the collection level or at the level of individual items, such as a recording. To facilitate access, while respecting the intended access conditions for a collection, or collection items, some form of user identification and authorisation process is typically required. The access control framework described in this paper is based upon descriptions of access conditions in easy-to-read licences which are stored alongside data files in the collections; and is implemented using identity-based authentication and authorisation systems where required. The framework accommodates accessibility needs from unrestricted to extremely limited access, is dynamic, and able to be modified in response to changes in access needs. Storing licences with the data is a significant development in separating language data and access requirements from access infrastructure.
Keywords: access, archiving, collection management, metadata, research object

1. Introduction

The Language Data Commons of Australia (LDaCA) project aims to ensure that significant language collections are accessible for future use, with required community controls where appropriate. The first tranche of collections which LDaCA is working with includes Indigenous languages, languages of the Pacific region, collections of Australian English, and multilingual translated government publications. These datasets are multimodal—comprising spoken, written and signed language materials—and cover varied access conditions. Some collections are "open" and publicly available, intended to be accessed without restriction. Other collections are restricted to individual researchers, community groups or some other groups of users.

LDaCA is developing policies and technical infrastructure to house and provide appropriate access to language collections. LDaCA's approach to language material management is based on the guiding principles of FAIR and CARE, which facilitate appropriate ongoing use of collections. These principles inform the development of LDaCA's policy frameworks and technical infrastructure, which are then implemented according to the specific needs of those who are responsible for particular language collections.

1.1. Community contexts

A challenge for people and organisations responsible for managing access aligned with community needs is to provide appropriate access which satisfies community needs—whether for academic or non-academic communities—while also being convenient to users. Approaches to digital access management have historically ranged from extremely complex efforts to replicate community relationships in infrastructure design: such as those presented by the Ajamurnda project, understanding and responding to community values and practices around knowledge ownership and sensitivities, and modelling access modes on these information systems (Nathan, 2018); through to projects which attempt to avoid the need for access control by requesting all data to be open. Projects such as CADRE[1] are currently attempting to design attribute-based authorisation systems, based on attributes that are common to groups of users, for example, whether a user is a member of a professional group. Previous access control systems tend to have been limited by the specific collections used to inform the design processes. A major benefit of LDaCA's approach is the range of language collections that have informed policy development and infrastructure design—ranging from Indigenous community language collections, research institution collections, government collections, even collections developed over a lifetime by individual researchers.

[1] https://cadre5safes.org.au

LREC-COLING 2024, pages 113–121

1.2. Principles for data management

Language data management in research contexts is trending towards data-driven and reproducible ways of working (L. Berez-Kroeker et al., 2022). The FAIR (Wilkinson et al., 2016) and CARE (Carroll et al., 2020) principles have been developed as a foundation for good, scholarly ways of finding and using data, providing key guideposts (Findable, Accessible, Interoperable, Reusable) for transparent, reproducible, and reusable language data. Further key principles (Collective benefit, Authority, Responsibility, Ethics), for advancing Indigenous innovation and self-determination, guide ethical work with Indigenous language collections. In non-research community contexts such as those in which community-based Language Centres operate, moral and ethical principles tend to be paramount as a virtue of projects being embedded in community-led work. The FAIR and CARE principles arose from the research community, yet can be a useful framework for reviewing data management decisions in non-research contexts.

Further guideposts used to assist ethical work with language material include Traditional Knowledge (TK) labels. TK labelling is a strategy for expressing Indigenous community practices around appropriate use of digital items. Labels can describe specific conditions of use for language materials that are relevant to community rules, governance and protocols. Conditions of use are typically expressed as icons or labels published with data as an indicator of how the data is intended to be accessed (see Figure 1 for a set of labels developed by the Local Contexts organisation). For example, an item intended to be used only by specific women would be tagged with the Local Contexts TK Women Restricted protocol label. The titles and descriptions of Local Contexts Labels can be translated and localised by individual communities to suit their own languages and requirements[2].

1.3. Data licences

Licences[3] are legal instruments which detail the extent to which data can be used or shared. They are agreements between copyright holders and users. Licensing provides a simple mechanism to satisfy a complex range of access requirements, such as the use of material for individual researchers to perform linguistic analysis, for a commercial entity to train an AI system, or for someone to use data to support their language learning. Data licences may allow completely open use of language data, set some minimal conditions (such as attribution of ownership when a user shares the data), or may specify geographic, demographic, time-based or some other limitation of use.

Figure 1: Local Contexts Traditional Knowledge (TK) Labels, showing a selection of icons representing community protocols for data use.

Data licences are based upon the rights which data owners hold, and vary from one jurisdiction to another. Australian law currently protects material rights to language data, with inadequate protection for communal or community rights. The focus on material rights results in situations where language data is typically owned by those who recorded or published the material, rather than by the language communities themselves. Movements to better protect Indigenous Cultural and Intellectual Property (ICIP) rights in Australia are ongoing, in an effort to provide legal pathways to uphold ICIP rights for Indigenous peoples to "maintain and control their cultural knowledge and expressions" (Janke, 2021).

2. Language collection access

Recognising the need for appropriate community control, LDaCA designed an adaptable, revocable, implementable, tiered approach to language data access, and using data licences as part of an access framework. The access framework consists of licence-based access conditions, along with user identification and authorisation systems (Sefton et al., 2023a).

Access conditions are designated by assigning a licence to a collection, or by assigning different licences to individual items in a collection. A licence is determined by considering whether the material can be reused, shared, or modified, along with other restrictions of use. Licences may be cho-

[2] https://localcontexts.org/labels/traditional-knowledge-labels

[3] Throughout we follow the convention of UK English, distinguishing *licence* as a noun from *license* as a verb. An exception to this convention occurs in Figure 4, where *license* as a noun appears in a screenshot of (part of) a JSON metadata file. In this case, the usage aligns with the schema.org property *license*.

Table 1: Access types

Access type	Description	Authentication or authorisation required?
Open access	Data is licensed for use without interaction by the user. This form of licence is commonly known as Open Access. May use a Creative Commons licence or similar.	No
Indication of assent	Access requires acceptance of licence, demonstrated by indication of assent to licence conditions via a minimal form of interaction (eg a click-through licence).	Optional. In some cases, identity authentication may be required. Authorisation by a data steward[a] is not required.
Access by application	Specific licence conditions apply. Users apply for access. A user's application is reviewed by the data steward to confirm the user is authorised to access data. Approval may depend on the community or academic affiliation of the applicant, the intended use of the data, payment of a fee, or other conditions. Requires ongoing engagement from the data steward in order to manage access lists and approve/decline access requests.	Users must authenticate their identity and authorisation must be granted by the data steward.
Access by invitation	Specific licence conditions apply. Users are invited by the data steward. Requires ongoing engagement from the data steward in order to manage access lists and initiate access invitations.	Users must authenticate their identity and authorisation must be granted by the data steward.

[a] A data steward is the copyright owner of the collection, or their representative.

sen from existing sources such as Creative Commons[4], derived from licence templates, or written from scratch.

LDaCA uses the RO-Crate (Soiland-Reyes et al., 2022) standard for organising collection items. RO-Crate is a collection management approach which groups files with metadata about those files in "crates", which are simply folders on a disk. For example, a video recording of a language user would be grouped with material derived from the recording (transcription and translation files, images from the recording session), along with the licence for using the material, and a metadata file describing all the files in the crate. In order to make access control systems as simple and robust as possible, best practice is to have all items in a crate covered by a single licence, so far as this is possible. This may mean an item is composed of multiple crates to adequately cover the access requirements of parts of the collection.

At the most open end of the access spectrum, a collection may be viewed by the user reading a licence, without requiring user authentication. Other collections may require acceptance of terms and conditions, or by applying to access the data (see Table 1). Access to restricted datasets requires identification of the user, and verification of their authority to use the material, according to the authorisation level required.

2.1. Technical notes

A distributed access control system comprising user identification and authorisation services automates user access to collections. The LDaCA system uses a data portal/repository of collections, along with off-the-shelf software for authentication and authorisation, namely CILogon[5] (an identity management service operated by the national research identify provider AAF[6]) and REMS[7] (a licence management tool).

[4] https://creativecommons.org

[5] https://www.cilogon.org

[6] https://ror.org/033jftm25

[7] https://github.com/CSCfi/rems

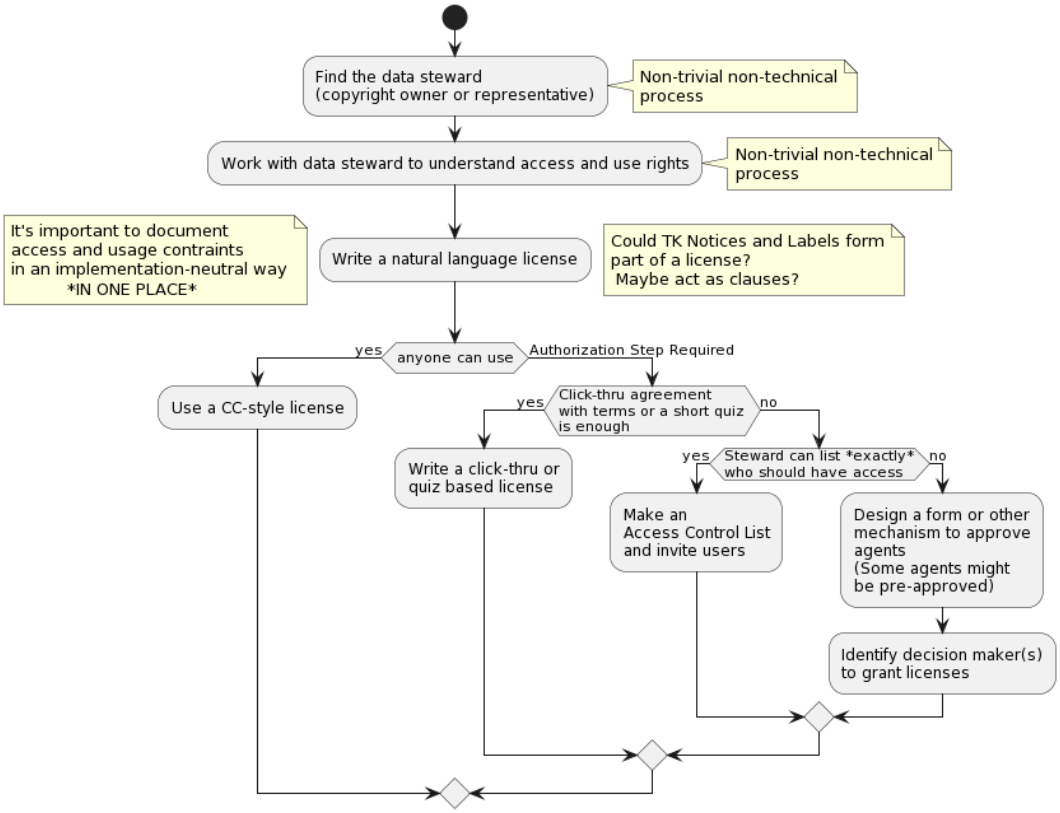

Figure 2: Data licensing process

2.2. Data licensing

The selection of an appropriate licence to best represent the access conditions required for a given collection is made by the collection's data steward. To assist data stewards to make informed decisions, LDaCA provides tools and resources to facilitate the process of choosing or writing licences. In the licensing process (see Figure 2), the first step is to identify the data steward responsible for the collection's data governance. This person may be the copyright owner or their representative. This data steward will determine the access conditions and usage rights for the collection. An easy-to-read licence is then written or selected to satisfy the required conditions.

Depending on the conditions, the licensing process diverges. For some collections, an existing licence such as a Creative Commons licence may be suitable, or a template may be used to assist in writing a custom licence for the collection. For other situations, a click-through agreement describing terms, or a short quiz, may be sufficient to obtain a user's assent.

For collections with restricted access conditions, and when users' identities are known, an Access Control List of approved people's email addresses can be created to provide the mechanism for authorisation. People included on the list can be invited to access the collection. In restrictive cases where the users are not yet known, a form or some other mechanism may be created for users to complete when requesting access. In these cases, form responses would be reviewed by the data steward or their representatives, to determine whether access is granted or denied.

In some situations, items within a collection that are created by or involve a particular participant may need to be withdrawn. Withdrawal may be temporary, for example, when showing cultural respect for a contributor's work by adhering to Sorry Business[8] protocols for short term removal of material. Long-term removal of material from a collection may be required if a participant withdraws consent. To accommodate revocation of access, a licence may stipulate that re-use for the purpose granted is allowed, but data must not be further shared. Or, data stewards may decide to restrict use entirely for a period.

Licences may limit sharing of files by users, but allow sharing by referral to the persistent identifier

[8] Sorry Business is a period in which cultural practices such as ceremonies take place to commemorate death. Passing-away protocols of bereavement may include: not using the name, or broadcasting the voice or image of the person who passed away, for some period of time.

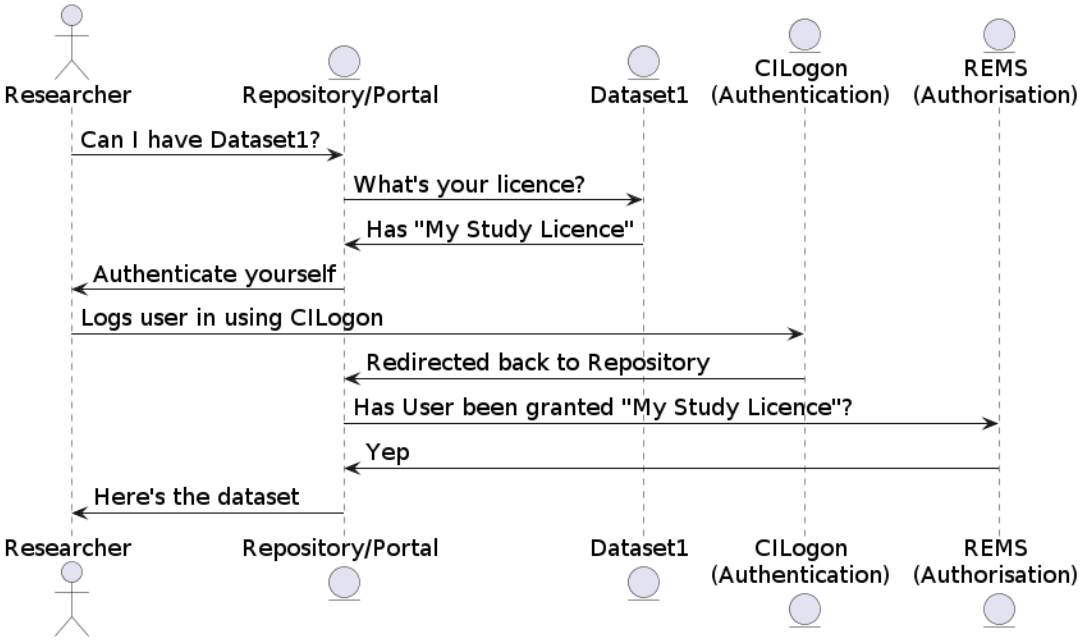

Figure 3: Access-control dance, showing the process for a user, who has been granted a licence in REMS, obtaining access to a dataset.

of the collection, so subsequent users can apply for a licence themselves. This approach is recommended for situations where participants may request permanent removal of their data from collections. Removing participant data from copies of collections that have been shared adhoc is nigh-impossible; whereas if collections are shared by link referral, the collection can be modified as required.

In addition to an easy-to-read licences, the metadata vocabulary which is being developed by LDaCA includes terms to record the type of access allowed and the authorisation workflow required for data objects . These terms do not replace specific licences but can be used, for example, for grouping material by broad access level for display purposes.

2.3. Access process

Depending on a collection's conditions, a user may access restricted data from a data portal by following a process of simply accepting terms and conditions, or by requesting permission. To access a restricted collection which requires authentication, firstly a user authenticates their identity using CILogon. Many common identification methods are supported, including Gmail, university logins, and ORCID[9]. Once identified, the REMS tool checks whether the user has been granted a licence for the data. If the user has an authority to access the data from prior invitation or approval, they are given access to the collection in the portal (see Figure 3). If the user is not yet approved, access approval is requested via a form integrated in REMS, or an external form.

3. System benefits

Benefits of a licence-based access system include:

1. **Access conditions are encapsulated in a licence document which can be included in the collection.** Documenting access conditions in a licence, which is stored in the file-system alongside the collection data, ensures that access information persists with the collection, independent of the mechanisms for identifying and authenticating users. This approach reflects the FAIR principle of reusability—ensuring the collection is reusable with intended access conditions, independent of the identification and authorisation infrastructure.

2. **Authority to grant access can be delegated to the people who hold rights in the data.** Delegation of authorisation supports the CARE principle of "Authority to control", in that the authority to control access can be handled by community members or those who hold rights in the dataset.

3. **Access is identity-based rather than attribute-based.** Although efforts are underway through other projects to develop authentication systems based on attributes of users (e.g., whether a user is a researcher, or a teacher etc), these systems are not

[9] https://orcid.org

Figure 4: File-based storage, showing an item's files alongside a licence file and metadata JSON file. The licence is also referenced in the metadata JSON file.

yet available as a service. The projects developing attribute-based authentication have a narrow focus on government-provided data which will consequently use a very limited range of attributes to authenticate. It is unlikely that the attribute-based authentication systems being developed by such programs will have the scope to handle the range of collections encountered by LDaCA.

4. **Licences can be easily changed over time.** Licensing provides a mechanism to modify collection availability over time as access requirements change. An example of this adaptation is in provision of a passing-away protocol in a licence, to limit use during periods of respect following the death of a collection contributor.

4. Case studies

4.1. The CALL Collection, Batchelor Institute Library

The Centre for Australian Languages and Linguistics (CALL) Collection is an important archive of physical and digital works in and about Indigenous languages, containing thousands of items and representing hundreds of languages. The archive includes texts, audio and video resources that have been contributed over the past 50 or so years by students and staff of the Batchelor Institute of Indigenous Tertiary Education, as well as other teachers, linguists and language workers. Protocols for the Collection and website, along with a suite of consent forms, licences, and website terms and conditions, have been developed by the project officer Karen Manton in close collaboration with Terri Janke, a leading activist for ICIP rights. The access conditions of items in the collection range from items that are openly viewable online with no restriction, to some with access limitations according to gender, and to some items being highly restricted.

The CALL Collection includes physical and digital items which are described in an online catalogue containing metadata about collection items. Some of the physical items in the collection have digital versions—scanned images, digitised and transcribed tapes etc. The rights of items in the collection are complex, being subject to Intellectual Property laws and ICIP rights and protocols. Some of the items have cultural restrictions and are excluded from public view. Other items are able to be viewed online, according to the collection's terms of use. Items either have one of three CALL Collection End User Licences, or a Creative Commons licence. The licence and relevant TK Labels or TK Notices for a particular item are shown in the current user interface in the item record view.

For all licences, users are not permitted to use the material to make money. Users are required to name and respect the people who made the materials, and their work. No derivatives are allowed, except as specified under the cultural maintenance licence.

The collection is covered by a blanket Terms for Use, which points out that:

- If you see a Creative Commons licence and icons displayed with a work, you can use the material under that Creative Commons licence.
- If you see the CALL Collection End User Licences and icons, you must follow the licence and protocols that are right for who you are and why you want to use the material.

All items on the site are covered by the CALL Collection Website Terms of Use, which detail the range of permissible uses. Permissible uses are based

upon three types of users: General Public, who can only make personal use of material; Education users, who are allowed teaching, learning and research use; and Culture and Language users, who may use items for cultural maintenance, teaching and learning in the user's own language.

The availability of an item is covered by one of the following four categories: Public access, Waiting for approval, Restricted, External Link. Public access material is typically in a digital format and can be viewed or downloaded. Items that are tagged as Waiting for approval may be waiting for permissions to be granted for publishing the item. For these works, users who have the authority to approve access are encouraged to contact the Collection staff to assist in the approval process. Some of the items in the collection are restricted, perhaps due to cultural protocols. Items for which publication permission has not been gained, yet are published online elsewhere, are designated as External Links.

LDaCA is working with Batchelor Institute to backup and reformat item record information, copying metadata from an ageing database into RO-Crate metadata format (Soiland-Reyes et al., 2022), and embedding cultural protocols to ensure long-term and appropriate access to this significant collection. The reformatting process of copying data from a database into a file-based format ensures long-term availability to the materials. During the reformatting process, item licence information and access guides in the form of TK Labels or Notices (where these are available) are written to a directory alongside the data item, and a reference to the licence is included in the item-specific metadata file. Locating the licence alongside the item in a metadata file provides access conditions in a future-proof way that is human- and machine-readable into the (near) future (see Figure 4).

4.2. Sydney Speaks

The Sydney Speaks collection (Travis, 2014) is a dataset made up of recordings of people in Sydney, a major Australian city, telling stories about their lives and experiences. The research aims of this collection are to investigate language change over 100 years, and analyse the impact of ethnic diversity on the way Australian English is spoken. Access to the collection is managed on three broad levels —

- Accessible, where participants have provided full consent for sharing and reuse, and audio recordings and transcripts have been de-identified and time aligned.
- No access, where participants have provided full consent for sharing and reuse, and audio recordings and transcripts have not been de-identified and are not time aligned.
- Data is unavailable, where participants have not provided consent or have provided partial consent, or audio recordings have not been transcribed.

The collection consists of multiple sub-collections, which were created under different conditions. It includes the NSW Bicentennial Oral History Project, recordings made in the 1980s which are freely accessible through the National Library of Australia; the Sydney Social Dialect Survey, recordings made in the late 1970s to early 1980s by a researcher at the University of Sydney without the kind of formal ethics approval process that is required today; and recordings from 2016 onwards that are restricted according to agreements with the participants about this data collection. As such, different data access licences have been developed to meet the various access restrictions across the collection.

Users intending to access the data are required to complete a form, providing their email address and an account of their intended use, as usage of the collection is limited to research that will advance the scientific study of language or of Australian society, as per the aims of Sydney Speaks, and that does not duplicate ongoing research within the team (Travis and Johnston, 2023). Administration of applications is managed by lead researcher, Catherine Travis.

5. Conclusion

The licence-based access system developed by LDaCA supports a range of access types, and is suitable for managing access to research and non-research language collections alike. The system has been demonstrated to be effective in providing open, unrestricted access to data, and for restricting access to entire collections or to individual items in a collection. Licence information is stored within a collection's metadata, as human and machine-readable information about the collection's access conditions. A suite of off-the shelf services (CILogon and REMS) along with a modern metadata standard (RO-Crate) are used to administer and implement access conditions for identity-based authentication and authorisation to collections.

6. Acknowledgements

We thank Catherine Travis and Cale Johnstone for providing the information on the Sydney Speaks project, and Karen Manton for providing information on the CALL Collection project.

7. Bibliographical References

Oliver Adams, Benjamin Galliot, Guillaume Wisniewski, Nicholas Lambourne, Ben Foley, Rahasya Sanders-Dwyer, Janet Wiles, Alexis Michaud, Séverine Guillaume, Laurent Besacier, Christopher Cox, Katya Aplonova, Guillaume Jacques, and Nathan Hill. 2021. User-friendly automatic transcription of low-resource languages: Plugging ESPnet into Elpis.

Alëna Aksënova, Daan van Esch, James Flynn, and Pavel Golik. 2021. How Might We Create Better Benchmarks for Speech Recognition? In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 22–34, Online. Association for Computational Linguistics.

Australian Research Data Commons (ARDC). Research data management framework for institutions.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. ArXiv:2006.11477 [cs, eess].

Jeffrey Beall. 2010. Metadata for Name Disambiguation and Collocation. *Future Internet*, 2(1):1–15. Number: 1 Publisher: Molecular Diversity Preservation International.

Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. 2020. The CARE Principles for Indigenous Data Governance. *Data Science Journal*, 19:43.

Ben Foley, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T Mark Ellison, Daan van Esch, Scott Heath, František Kratochvíl, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nicholas Thieberger, and Janet Wiles. 2018. Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System.

Yashesh Gaur and Walter S Lasecki. 2016. The Effects of Automatic Speech Recognition Quality on Human Transcription Latency. page 7.

Terri Janke. 2018. Indigenous Knowledge: Issues for protection and management. Discussion Paper, IP Australia.

Terri Janke. 2019. *True Tracks: Indigenous Cultural and Intellectual Property Principles for putting Self-Determination into practice*. Ph.D. thesis, Australian National University.

Terri Janke. 2021. *True Tracks: Respecting Indigenous knowledge and culture*. NewSouth Publishing. Google-Books-ID: GTQ4EAAAQBAJ.

Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, and Lauren B. Collister, editors. 2022. *The Open Handbook of Linguistic Data Management*. The MIT Press.

Alexis Michaud. 2018. Integrating Automatic Transcription into the Language Documentation Workflow: Experiments with Na Data and the Persephone Toolkit. *Language Documentation*, 12.

David Nathan. 2018. Safety before sanctions, sanctions before barriers: Digital access protocol for Anindilyakwa people of Groote Eylandt.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic Speech Recognition for Supporting Endangered Language Documentation. *Language Documentation*, 15.

Nay San, Martijn Bartelds, Blaine Billings, Ella de Falco, Hendi Feriza, Johan Safri, Wawan Sahrozi, Ben Foley, Bradley McDonnell, and Dan Jurafsky. 2023. Leveraging supplementary text data to kick-start automatic speech recognition system development with limited transcriptions.

Peter Sefton. 2023. Towards a Generic Research Data Commons: A highly scalable standard-based repository framework for Language and other Humanities data. Section: posts.

Peter Sefton, Moises Sacal Bonequi, Simon Musgrave, and Jenny Fewster. 2023a. A CARE- and FAIR-Ready Distributed Access Control System for Human-Created Data. *Proceedings of the 2nd International Workshop on Digital Language Archives: LangArc 2023*, pages 23–27. Publisher: University of North Texas.

Peter Sefton, Eoghan Ó Carragáin, Stian Soiland-Reyes, Oscar Corcho, Daniel Garijo, Raul Palma, Frederik Coppens, Carole Goble, José M. Fernández, Kyle Chard, Jose Manuel Gomez-Perez, Michael R. Crusoe, Ignacio Eguinoa, Nick Juty, Kristi Holmes, Jason A. Clark, Salvador Capella-Gutierrez, Alasdair J. G. Gray, Stuart Owen, Alan R. Williams, Giacomo Tartari, Finn Bacall, Thomas Thelen, Hervé Ménager, Laura Rodríguez-Navas, Paul Walk, brandon whitehead, Mark Wilkinson, Paul Groth, Erich Bremer, Leyla Jael Castro, Karl Sebby, Alexander Kanitz, Ana Trisovic, Gavin Kennedy, Mark Graves, Jasper Koehorst, Simone Leo, Marc Portier, Paul Brack, Milan Ojsteršek, Bert Droesbeke, Chenxu Niu, Kosuke Tanabe, Tomasz Miksa, Marco La Rosa, Cedric Decruw, Andreas Czerniak, Jeremy Jay, Sergio Serra, Ronald Siebes, Shaun de Witt, Shady El Damaty, Douglas Lowe, Xuanqi Li, Sveinung Gundersen, and Muhammad Radifar. 2023b. RO-Crate Metadata Specification 1.1.3. Publisher: Zenodo Version Number: 1.1.3.

River Tae Smith, Louisa Willoughby, and Trevor Johnston. 2022. Integrating Auslan Resources into the Language Data Commons of Australia. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 181–186, Marseille, France. European Language Resources Association.

Stian Soiland-Reyes, Peter Sefton, Mercè Crosas, Leyla Jael Castro, Frederik Coppens, José M. Fernández, Daniel Garijo, Björn Grüning, Marco La Rosa, Simone Leo, Eoghan Ó Carragáin, Marc Portier, Ana Trisovic, RO-Crate Community, Paul Groth, and Carole Goble. 2022. Packaging research artefacts with RO-Crate. *Data Science*, 5(2):97–138. Publisher: IOS Press.

Huessein Suleman. 2023a. Designing Repositories in Poor Countries.

Huessein Suleman. 2023b. Designing Repositories in Poor Countries.

Catherine E. Travis. 2014. Sydney Speaks: Variation and Change in Australian English. Australian Research Council Centre of Excellence for the Dynamics of Language, Australian National University. Australian Research Council Centre of Excellence for the Dynamics of Language, Australian National University.

Catherine E. Travis. 2023. Sydney Speaks Corpora - Google Form.

Catherine E. Travis and Cale Johnston. 2023. Introducing the Sydney Speaks project: Compiling legacy and contemporary data collections.

Daan van Esch, Ben Foley, and Nay San. 2018. Future Directions in Technological Support for Language Documentation.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino Da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 'T Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene Van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan Van Der Lei, Erik Van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018.

Greg Wilson, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K. Teal. 2017. Good enough practices in scientific computing. *PLOS Computational Biology*, 13(6):e1005510. Publisher: Public Library of Science.

8. Language Resource References

Travis, Catherine E. 2014. *Sydney Speaks: Variation and Change in Australian English*. Australian Research Council Centre of Excellence for the Dynamics of Language, Australian National University. Australian Research Council Centre of Excellence for the Dynamics of Language, Australian National University.