A CARE- and FAIR-Ready Distributed Access Control System for Human-Created Data

Peter Sefton
Language Data Commons of Australia
The University of Queensland
Brisbane QLD Australia
p.sefton@uq.edu.au

Moises Sacal Bonequi
Language Data Commons of Australia
The University of Queensland
Brisbane QLD Australia
m.sacalbonequi@uq.edu.au

Simon Musgrave
Language Data Commons of Australia
The University of Queensland
Brisbane QLD Australia
s.musgrave@uq.edu.au

Jenny Fewster
Australian Research Data Commons
Monash University
Melbourne VIC Australia
jenny.fewster@ardc.edu.au

ABSTRACT

The Language Data Commons of Australia (LDaCA) makes nationally significant language data available for academic and non-academic use, managing the data in a culturally, ethically, and legally appropriate manner guided by FAIR and CARE principles. Here, we describe the approach which we are taking to access control and a design for a distributed access control system which can look after the A-is-for-accessible in FAIR data while respecting the CARE principles. We also describe and demonstrate a pilot system based on that design, showing how data licenses that allow access by identified groups of people can be used by adding functionality, CILogon for non-institutional identification and REMS for managing access to resources, to the existing Australian Access Federation infrastructure.

CCS CONCEPTS

• Applied computing → Computers in other domains → Digital libraries and archives • Applied computing → Arts and humanities • Security and privacy → Human and societal aspects of security and privacy

KEYWORDS

Language archive, FAIR, CARE, access control

ACM Reference format:

Peter Sefton, Moises Sacal Bonequi, Simon Musgrave, and Jenny Fewster. 2023. A CARE- and FAIR-Ready Distributed Access Control System for Human-Created Data. In Proceedings of the 2nd International Workshop on Digital Language Archives (LangArc-2023), ACM/IEEE Joint Conference on Digital Libraries. USA, 5 pages. https://doi.org/10.12794/langarc2114304

1 INTRODUCTION

The Language Data Commons of Australia (LDaCA) focuses on preservation and discovery of distributed multi-modal language data collections under a variety of governance frameworks. This will include access control that reflects ethical constraints and intellectual property rights, including those of Aboriginal and Torres Strait Islander, migrant and Pacific communities. Regarding rights, our project is informed by the CARE principle (https://www.gida-global.org/care) for Indigenous data which also describe the level of respect which should be given to any data collected from individuals or communities.

Language archiving has received considerable attention in the last 20 years because of the importance of the practice in the documentary linguistics tradition originating with Himmelmann [4]. Discussions of access to language archives [1,3] concentrate on the need for access control, who should be involved in making decisions and how those decisions can be documented. Perhaps understandably, the details and implementation of processes at a technical level have received less attention. Two exceptions to this generalisation must be mentioned. Broeder et al. [2] present a technical architecture for access control in a federated repository system, and Nathan [6] discusses a system based on the roles which can be taken by those interacting with the archive, an approach which emphasises that technical solutions must be based on human behaviour. In this paper, we present a design for a distributed access control system which could look after the A-is-for-accessible in FAIR data while respecting the CARE principles; and describe and demonstrate a pilot system based on that design, showing how data licenses that allow access by identified groups of people can be used with an Australian Access Federation (AAF) pilot system (CILogon) to give the right people access to data resources. We suggest that our approach combines desirable features of the designs described by [2] and by [6].

2. BASIC PRINCIPLES

Our system must be able to implement data access policies with real-world complexity and one of our challenges has been developing a data access policy that works across a range of different collections of language data. Accessibility, the A of FAIR data [8], means that data is accessible to the right people and who is included in 'right people' varies from collection to collection and even within a single collection. Another challenge is to make sure

that the information about access is sustainable; that is, the information is not locked in a specific software solution and can be easily reused when delivery systems change.

The key idea is to separate safe storage of data from its delivery. Each item in a repository is stored with licensing information in natural language and the repository defers access decisions to an Authorization system, where data custodians can design whatever process they like for granting license access.

3. LICENSES

A license in this context is *a natural language document* in which a copyright holder sets out the terms and conditions of use for data. Licenses *may* have metadata that describes them, e.g., a property to say that this is an open license and such metadata *about* a license can be used to automate decision making. If it is labelled as being an open license, then a repository can serve data and include that data, if it is labelled as "closed" or more aptly, "authorization-required" then repository software can perform an authorization step, which we cover in detail later.

In the world of research data generated by or about human participants, licenses can't always allow unauthenticated access and data redistribution, and they may permit distribution only to certain people, or classes of person. So, a license is a document that expresses conditions such as "Data can be used by other researchers", but unfortunately we don't have systems in the research-data ecosystem which can automatically identify a user as "a researcher" (see also [2]).

The access control system we have been prototyping is based on licenses. For any data object, we store a license with it, and we give the license an ID which is a URL we can use to identify it uniquely (see Figure 1). Figure 2 shows how a license is explicitly linked to the data using a metadata description standard known as "Research Object Crate" (RO-Crate) [7]. Each object in the repository is a crate, with a metadata file that describes the object and (optionally) its component files, including the data license. Every item in a repository has a license, which may be an open one like CC Share Alike or a custom license derived from the ethics and participant agreements for a study in the context of local laws and institutional policy.

Using this license, distributed access portals in our architecture can check against an authorization system for each request for data. The portals may host data with the same licensing but do not need to maintain access control lists.

4. AUTHENTICATION

When we first developed access controls for LDaCA in 2021 it was a requirement that data licensing and access control decisions be decoupled from each other, and from particular repository software. We could not find an available open-source system for managing license-based access to data, so our starting approach used groups as a proxy for granting licenses on the basis that all common user-directory services such as LDAP include the concept of user groups.

A proof-of-concept Github based system demonstrated that authorization can be delegated from a data repository service to an external service. For each of the licenses there was a Github group (organization). The data_repository, when requested to serve data would get the user to login using the Github Authentication services (no Github repositories were used), then check if the user was in the correct license group. Although this worked, there were no workflow options and it supported only a single logon service, which is not widely used in academia or by community groups.

The AAF were already working with other research groups on a service called CILogon (https://www.cilogon.org/). Like Github, this service has groups but also allows users to log in with a variety of Authentication providers, including research institutions, via the AAF as well as social logins such as Google and Microsoft (and our old friend Github).

Again, this worked, but the current version of CILogon does not have particularly easy-to-use ways for a license-holder to create groups. The AAF team made us aware of the Resource Entitlement Management System, (REM: https://github.com/CSCfi/REMS), which is an open source application out of Finland which has been used previously in at least one language data repository [5]. This software is the missing link for LDaCA in that it allows a data custodian to grant licenses to users. And it works with CILogon as an Authentication layer so we can let users log in using a variety of services.

At the core of REMS is a set of licenses which can be associated with Resources - in our design this is (almost always) a one-to-one correspondence, for example we would have a license "Sydney Speaks Data Researcher Access license" corresponding to a resource that represents ALL data with that license. These Resources can then be made available through a catalogue, and workflows can be set up for pre-authorization processes ranging from single-click authorizations where a user just accepts a license and a bot approves it, to complex forms where users upload credentials, and one or more data custodians approve their request and grant them the license (see Figure 3). Once a user has been granted a license then a repository can authorize access to a resource by checking with REMS to see if a given user holds the license. Users do not have to find REMS on their own - they will be directed to it from data and computing services when they need to apply for authorization. Figure 4 shows the interactions involved in accessing data once a user has been granted the license in REMS via a data portal which gives access to data in a repository or archive.

5. DISCUSSION

Access Control Lists (ACLs) are a popular approach to the problem we are addressing but we suggest that the more modular approach which we advocate has several advantages over ACLs. Firstly, ACLs need maintenance over time - people's identities change, they retire and die, so storing a list of identifiers such as email addresses alongside content is not a viable long-term preservation strategy. Rather, we will encourage data custodians to describe in words what are permitted uses for the data, and by whom, in a license, then allow whoever is the current data custodian to manage that access in a separate administrative system.

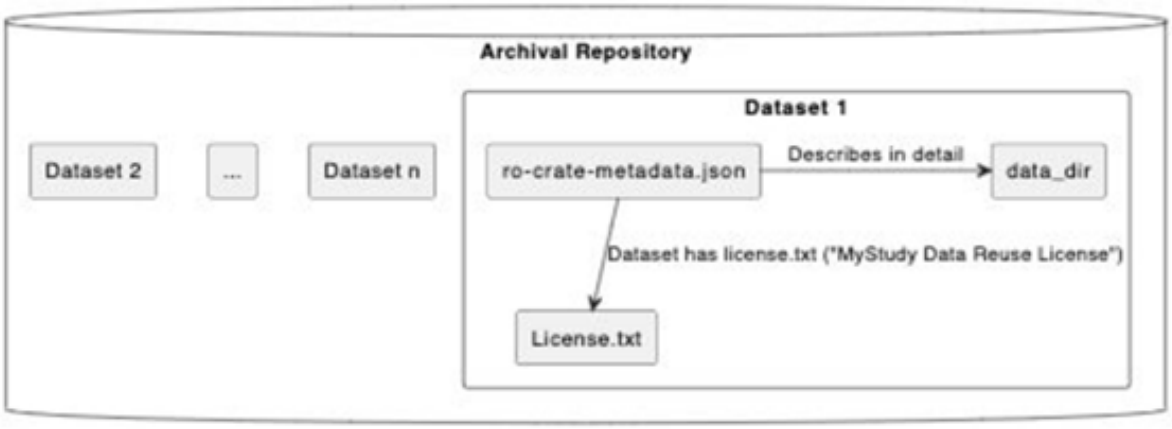

Figure 1: Data packaging architecture

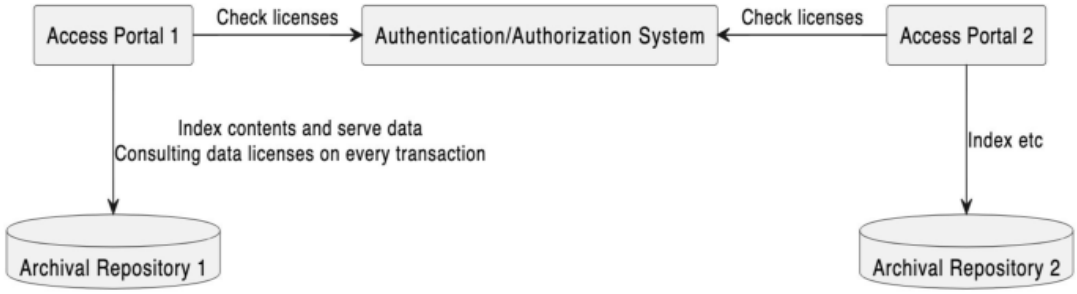

Figure 2: Relationships between repositories, portals, and the Authentication System

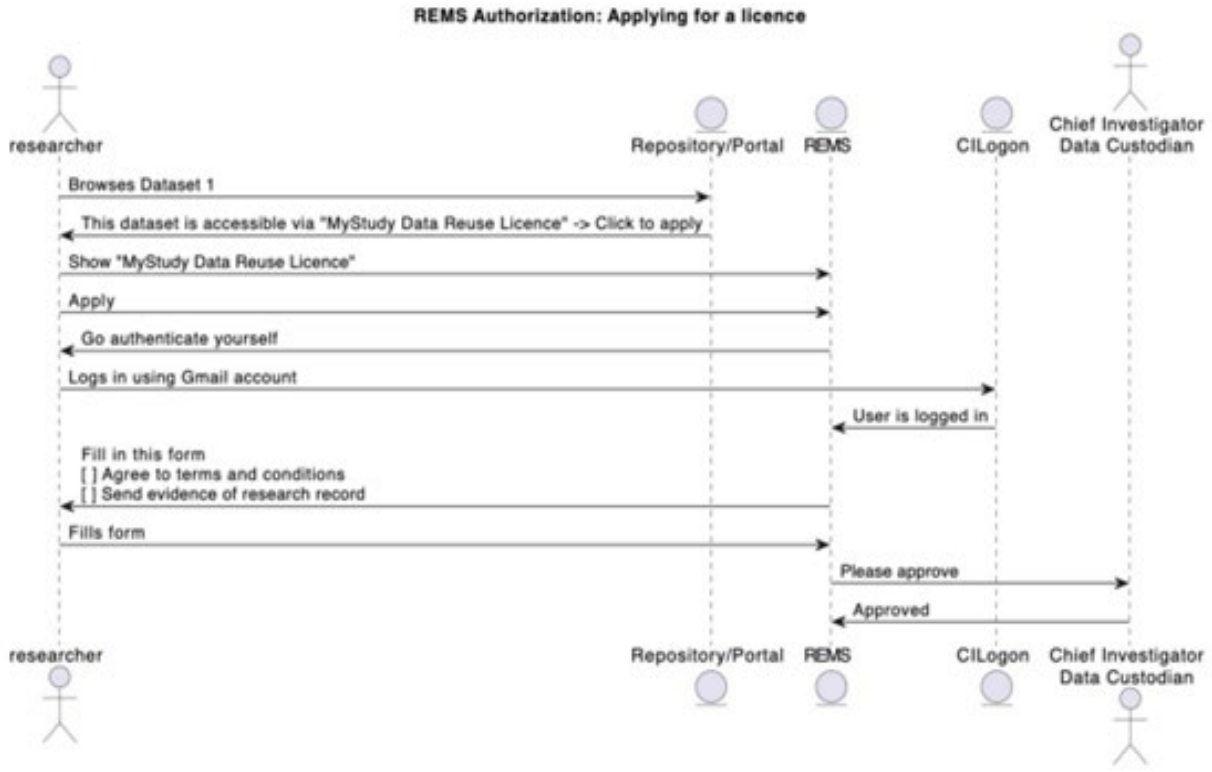

Figure 3: Interaction diagram showing the flow involved in a user applying for a data license via REMS.

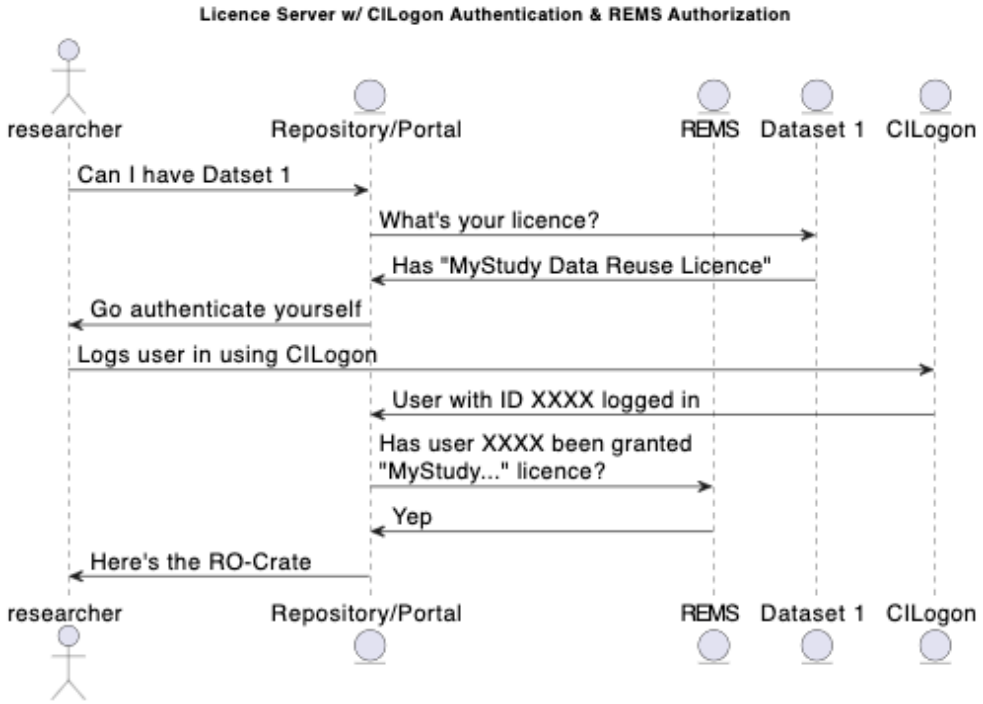

Figure 4: The "access-control dance" for a user who has been granted a license in REMS

Secondly, LDaCA data will be stored in a variety of places with separate portal applications serving data for specific purposes; if these systems all have in-built authorization schemes, even if they are the same, then we have the problem of synchronizing access control lists around a network of services. Thirdly, accessing data that requires some sort of authorization process is not a language or humanities specific problem, so working with an existing application that can handle pre-authorization workflows and access-control authorization decisions is an attractive choice and should allow LDaCA to take advantage of centrally managed services with relevant functionality. Fourthly, if complex access controls are implemented inside a system, then there is a risk that data becomes stranded inside that system and cannot be reused without completely re-implementing the access control. For example, imagine an archive of cultural material with complex access controls encoded into the business logic such as "this item is accessible only to male initiates". Applications like this need to store user accounts with attributes on both data and user records that can be used to authorize access. There is a high risk of data being stranded in a system such as this if it is no longer supported.

Our approach may seem to involve more work than an ACL based system. We believe that our emphasis on licenses as the basis for access control has advantages which outweigh the possibility of additional work (although we are not convinced that extra work will be needed in the long term). Reuse of data (the R in FAIR) means that users, including researchers and community members, should be able to download data for certain authorised purposes and activities. The license is the way that data custodians communicate to data users (and future administrators) what those purposes and activities are. A license, which is always packaged with data will allow:

- A user to inspect a five-year-old dataset in their downloads folder and work out what they are allowed to do with it.
- An IT professional to clean up laptop that has been handed in by (or seized from – it happens) a departing faculty member.
- A developer to re-create an access control replacing a decommissioned system.

We expect that the overhead of writing licenses will diminish greatly over time and standard clauses and complete licenses will be established.

It might seem that using REMS to administer access control means that we are locked into a specific software solution. This is not really the case; REMS is an app for establishing relationships between resources (licenses) and users. Both these components can be exported and used in another system for other purposes (e.g., auditing). In other words, if there is lock-in, it is temporary. But, because our process requires a governance step *first* in writing a license, then there is a statement of intent for re-building those processes later if needed - a step which is very likely to be missing in a system with built-in access control.

ACKNOWLEDGMENTS

The Language Data Commons of Australia (LDaCA) project received investment from the Australian Research Data Commons (ARDC). The ARDC is funded by the National Collaborative Research Infrastructure Strategy (NCRIS).

REFERENCES

[1] Andrea L. Berez-Kroeker and Ryan Henke. 2018. Language Archiving. In *The Oxford Handbook of Endangered Languages*, Kenneth L. Rehg and Lyle Campbell (eds.). Oxford University Press, 346–369. https://doi.org/10.1093/oxfordhb/9780190610029.013.18

[2] Daan Broeder, Freddy Offenga, Peter Wittenburg, Peter Van de Kamp, David Nathan, and Sven Strömqvist. 2006. Technologies for a federation of language resource archive. In *5th international conference on language resources and evaluation (LREC 2006)*, 2291–2294.

[3] Lisa Conathan. 2011. Archiving and language documentation. In *The Cambridge Handbook of Endangered Languages* (1st ed.), Peter K. Austin and Julia Sallabank (eds.). Cambridge University Press, 235–254. https://doi.org/10.1017/CBO9780511975981.012

[4] Nikolaus Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics* 36, 1 (1998), 161–196.

[5] Martin Matthiesen. 2015. REMS – Access Management at The Language Bank of Finland. In *DEIC Conference*. Middelfart, Denmark. Retrieved May 19, 2023 from https://gl.deic.dk/sites/default/files/uploads/PDF/Martin_Matthiesen_REMS_at_the_Language_Bank_of_Finland.pdf

[6] David Nathan. 2014. Access and accessibility at ELAR, an archive for endangered languages documentation. In *Language Documentation and Description, vol 12: Special Issue on Language Documentation and Archiving*, David Nathan and Peter K. Austin (eds.). SOAS, London, 187–208.

[7] Sefton, Peter, Ó Carragáin, Eoghan, Soiland-Reyes, Stian, Corcho, Oscar, Garijo, Daniel, Palma, Raul, Coppens, Frederik, Goble, Carole, Fernández, José M., Chard, Kyle, Gomez-Perez, Jose Manuel, Crusoe, Michael R., Eguinoa, Ignacio, Juty, Nick, Holmes, Kristi, Clark, Jason A., Capella-Gutierrez, Salvador, Gray, Alasdair J. G., Owen, Stuart, Williams, Alan R., Tartari, Giacomo, Bacall, Finn, Thelen, Thomas, Ménager, Hervé, Rodríguez-Navas, Laura, Walk, Paul, whitehead, brandon, Wilkinson, Mark, Groth, Paul, Bremer, Erich, Castro, Leyla Jael, Sebby, Karl, Kanitz, Alexander, Trisovic, Ana, Kennedy, Gavin, Graves, Mark, Koehorst, Jasper, Leo, Simone, Portier, Marc, Brack, Paul, Ojsteršek, Milan, Droesbeke, Bert, Niu, Chenxu, Tanabe, Kosuke, Miksa, Tomasz, La Rosa, Marco, Decruw, Cedric, Czerniak, Andreas, Jay, Jeremy, Serra, Sergio, Siebes, Ronald, de Witt, Shaun, El Damaty, Shady, Lowe, Douglas, Li, Xuanqi, Gundersen, Sveinung, and Radifar, Muhammad. 2023. RO-Crate Metadata Specification 1.1.3. (April 2023). https://doi.org/10.5281/ZENODO.3406497

[8] 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. *Sci. Data* 3, 1 (March 2016), 160018. https://doi.org/10.1038/sdata.2016.18