SciDataCon 2025

COMMITTEE ON DATA CODATA INTERNATIONAL SCIENCE COUNCIL

WORLD DATA SYSTEM

Contribution ID: **247** Type: **Poster**

Enabling transparent, open research processes using (not-always-open) RO-Crate data packages

Monday 13 October 2025 18:00 (1h 30m)

The RO-Crate (Research Object Crate) specification (Sefton et al. 2023) is a method for describing data sets with rich, interoperable Linked Data metadata. This presentation will show how we, the Language Data Commons of Australia project (LDaCA), use well described RO-Crate data packages (Soiland-Reyes et al. 2022) to enable CARE (Carroll et al. 2020) and FAIR (Wilkinson et al. 2016) compliant research with language data and also touch on some examples from other disciplines.

RO-Crates in the LDaCA environment are self contained packages that describe data resources, collections which aggregate the data at a variety of scales of granularity from a whole collection in one package to individual files in a set of packages, with linked-data metadata fields hasMember and memberOf that establish relationships between packages. The main reasons for the differences in granularity of packages are firstly, practical limitations on size; and secondly, licensing of data, where it is desirable to have a single licence apply to a package. We are dealing with human-created data which may be subject to a variety of participant rights, including copyright, university policies, privacy legislation and Indigenous Cultural and Intellectual Property Rights (ICIP) and these may apply in different ways to different parts of a collection.

The LDaCA team have developed data access systems, which are all available under open source licences. All mediate access to data packages held in data portals backed by Archival Repository systems which enforce licensing requirements; agents requesting data must hold an appropriate licence to use it. Sometimes licensing is automatic, as in the use of a Creative Commons license, but in other cases access may require permission from a researcher or a community.

With the licensing in place, the linked-data nature of RO-Crate allows seamless processing of collections of data, and the creation of virtual collections of data which aggregate distributed packages or reference metadata and data from multiple packages.

Researchers can openly publish code which accesses resources showing a transparent research workflow, while those wishing to re-run the code have to obtain licenses to the data, which may involve applying to a chief investigator on an academic study with an approved Ethics plan from their institution, or being vetted by a community authority.

We will illustrate the benefits of RO-Crate linked data packages with examples that show the possibility of text analytics such as topic modelling (Blei 2012) using a single tool on multiple datasets with vastly different structure and provenance. This interoperability is possible because the RO-Crate Linked Data metadata allows for declarative configuration files to map between different data sets.

One of the main advantages of RO-Crate is that it is discipline-agnostic and is now widely used in a variety of research contexts, mostly science based (e.g. Weiland et al. 2024). We will conclude with a ‘tour’of our tools that show how they can be used to create an RO-Crate environment introducing RO-Crate Metadata profiles that describe the method of storing data, data packaging and validation services, show how consistent RO-Crate metadata allows for data discovery by humans and enables machine access via an API and talk about how these are being applied in other disciples than the study of language.

References:

Blei, David M. 2012. Probabilistic topic models. Communications of the ACM 55(4). 77–84. https://doi.org/10.1145/2133806.2133826.

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.

Sefton, Peter, Ó Carragáin, Eoghan, Soiland-Reyes, Stian, Corcho, Oscar, Garijo, Daniel, Palma, Raul, Coppens, Frederik, et al. 2023. RO-Crate Metadata Specification 1.1.3. Zenodo. https://doi.org/10.5281/ZENODO.3406497.

Soiland-Reyes, Stian, Peter Sefton, Mercè Crosas, Leyla Jael Castro, Frederik Coppens, José M. Fernández, Daniel Garijo, et al. 2022. Packaging research artefacts with RO-Crate. Data Science. IOS Press 5(2). 97–138. https://doi.org/10.3233/DS-210053.

Weiland, Claus, Jonas Grieb, Daniel Bauer, Desalegn Chala, Erik Kusch, Carrie Andrew & Dag Endresen. 2024. Dataspace Integration for Agrobiodiversity Digital Twins with RO-Crate. Biodiversity Information Science and Standards 8. e134479. https://doi.org/10.3897/biss.8.134479.

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: SEFTON, Peter (Language Data Commons of Australia)

Presenter: SEFTON, Peter (Language Data Commons of Australia)

Session Classification: Poster Session

Track Classification: SciDataCon2025 Specific Themes: Open research through Interconnected, Interoperable, and Interdisciplinary Data