Five ways RO-Crate data packages are important for repositories

OR2024 Presentation Submission

The 19th International Conference on Open Repositories, June 3-6th 2024, Göteborg, Sweden

Five ways RO-Crate data packages are important for repositories

Peter Sefton, University of Queensland, p.sefton@uq.edu.au
Stian Soiland-Reyes, The University of Manchester, soiland-reyes@manchester.ac.uk

Abstract

Research Object Crate is a linked data metadata packaging standard which has been widely adopted in research contexts. In this presentation we will briefly explain what RO-Crate is, how it is being adopted worldwide, then go on to list ways that RO-Crate is growing in importance in the repository world:

1. Uploading of complex multi-file objects means RO-Crate is compatible with any general purpose repository that can accept a zip file (with some coding, repository services can do more with RO-Crates)
2. Download for well-described data objects complete with metadata from a repository rather than just a zip or file with no metadata
3. Using RO-Crate metadata reduces the amount of customisation that is required in repository software, as ALL the metadata is described using the same simple, self-documenting linked-data structures, so generic display templates
4. Sufficiently well-described RO-Crates can be used to make data FAIR compliant, aiding in Findability, Accessibility, Interoperability and Reusability thanks to standardised metadata and mature tooling
5. And if you're looking for a sustainable repository solution, there are tools which can run a repository from a set of static files on a storage service, in line with the ideas put forward by Suleman in the closing keynote for OR2023

Metadata Standards; RO-Crate; Repository Tools; Open Source

Audience

This presentation should be of interest to the general Open Repositories audience, but will have particular relevance for metadata specialists, software developers, and those choosing new repository solutions.

Uploading of complex multi-file objects

RO-Crate [1], [2] is a data packaging format and can be used to put multiple data files together with their metadata into a package such as a zip, tar or disk image file. This means that as long as your repository can handle a zip file it can take RO-Crates.

Beyond simply allowing upload of opaque RO-Crates there are opportunities for repository software to recognize metadata in an uploaded package and to pre-populate built-in metadata forms and/or datastores. This is not a pattern the authors have seen widely implemented in comprehensive institutionally focussed repositories, although at the time of writing it is being explored in Dataverse and InvenioRDM/Zenodo. We would encourage repository developers to explore this further,

particularly those working with research data. RO-Crate support is increasing in research-domain repositories; eg RO-Crate upload with metadata extract is supported by WorkflowHub and ROHub),

RO-Crate is a packaging format suitable for downloads

One of the perennial problems with downloads is that once a user has the data, it often does not come with metadata as shown on the landing page, or if present, metadata provided is in an ad hoc format. RO-Crate solves this by specifying an extensible way to put linked-data metadata with data assets and to provide an HTML page or small website with the data to explain it.

RO-Crate download is already available in many data repositories, examples include:

- WorkflowHub: A registry for describing, sharing and publishing scientific computational workflows
- ROHub: A repository of Earth Science datasets and computational methods
- TLCMap: The Time Layered Cultural map is a set of tools that work together for mapping Australian history and culture
- The Language Data Commons of Australia data portal and its API is entirely built on RO-Crates
- Dataverse: at time of writing, RO-Crate downloads are in development

We will encourage developers from other repository platforms to follow the Dataverse project's lead and add RO-Crate support.

Less user interface customisation will be needed for different types of metadata

One of the key benefits of linked-data metadata over previous 'legacy' approaches, is that multiple vocabularies can be combined into a single metadata document in a way that is not possible with, say MARC, or MODS XML and that all these vocabularies can use the same syntax, and approach to describing data. This means that a simple generic RO-Crate viewer can be used to visualise any metadata whether it is basic "Who, What, Where" metadata (a la Dublin Core) or domain-specific metadata like the RO-Crate metadata profile (https://w3id.org/ldac/profile) used by the Language Data Commons of Australia. This can be displayed alongside the core RO-Crate metadata without any expensive configuration or coding. If the recommendations are followed, the RO-Crate metadata terms are self-documenting, e.g. all the Language Data Commons terms which use a Schema.org Style approach, are defined here: https://w3id.org/ldac/terms.

The availability of RO-Crate editing tools opens the way for repository software to focus on access and discoverability

We argue that the core functionality of a repository is keeping data safe and making it available with appropriate access controls (remember, not all data can be made Open Access - the A for accessibility in FAIR is about giving the **right people** (or other agents) access to the **right data**).

RO-Crates require clear licensing statements to travel with data, and we will demonstrate how these have been integrated into access-control systems.

There is an opportunity, if RO-Crate is adopted as an interchange format, for the metadata editing functions of a repository to be decoupled from it so the editor components for a particular metadata profile can be shared between repository instances, or handled in a more distributed architecture than in typical current repositories.

With a repository to keep data safe and serve it using persistent Identifiers, RO-Crates help make data FAIR

RO-Crate is increasingly being used to describe the provenance [3] of derived data in such a way that the workflows/computation that produced it can be re-run automatically to validate it, or as a basis for new research. This might be a button on a repository to run a bioinformatics workflow, or re-run a Jupyter notebook that produces a set of plots – we will demonstrate a selection of these.

There are tools which can run a repository from a set of static files on a storage service, in line with the ideas put forward by

The team at the Language Data Commons of Australia, with partner institutions and colleagues, has been working to produce a set of tools for building Archival Repository software stacks that is based on a principled approach to keeping data safe, based on the principles presented in the Arkisto website[4] and more recently at RRKive.org; the core idea is that a collection of RO-Crates in a storage service can be the basis of a repository – either using a simple on-disk directory layout or something more complicated such as an Oxford Common File Layout (OCFL) specification.

References

[1] Sefton, Peter *et al.*, "RO-Crate Metadata Specification 1.1.3," Apr. 2023, https://doi.org/10.5281/zenodo.7867028

[2] S. Soiland-Reyes *et al.*, "Packaging research artefacts with RO-Crate," *Data Science*, vol. 5, no. 2, pp. 97–138, Jan. 2022 https://doi.org/10.3233/DS-210053

[3] S. Leo *et al.*, "Recording provenance of workflow runs with RO-Crate." arXiv, Dec. 12, 2023. https://doi.org/10.48550/arXiv.2312.07852

[4] P. Sefton, M. La Rosa, and Mi. Lynch, "Arkisto: a repository based platform for managing all kinds of research data," in *ptsefton.com*, Jun. 2021. Accessed: Jan. 31, 2022. http://ptsefton.com/2021/06/11/or-2021-arkisto/