Creating lightweight FAIR Digital Objects with RO-Crate

Stian Soiland-Reyes[‡,§], Peter Sefton[|], Leyla Jael Castro[¶], Frederik Coppens[#], Daniel Garijo[¤], Simone Leo[«], Marc Portier[»], Paul Groth[§]

‡ The University of Manchester, Manchester, United Kingdom
§ Informatics Institute, Faculty of Science, University of Amsterdam, Amsterdam, Netherlands
| The University of Queensland School of Languages and Cultures, The University of Queensland, Brisbane, Queensland, Australia
¶ Informationszentrum Lebenswissenschaften (ZB Med), Cologne, Germany
# Vlaams Instituut voor Biotechnologie & Universiteit Ghent (VIB-UGent) Center for Plant Systems Biology, Ghent, Belgium
¤ Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain
« Center for Advanced Studies, Research, and Development in Sardinia (CRS4), Pula (CA), Italy
» Vlaams Instituut voor de Zee (VLIZ), Oostende, Belgium

Corresponding author: Stian Soiland-Reyes (soiland-reyes@manchester.ac.uk)

Abstract

RO-Crate (Soiland-Reyes et al. 2022) is a lightweight method to package research outputs along with their metadata, based on Linked Data principles (Bizer et al. 2009) and W3C standards. RO-Crate provides a flexible mechanism for researchers archiving and publishing rich data packages (or any other research outcome) by capturing their dependencies and context. However, additional measures should be taken to ensure that a crate is also following the FAIR principles (Wilkinson 2016), including consistent use of persistent identifiers, provenance, community standards, clear machine/human-readable licensing for metadata and data, and Web publication of RO-Crates.

The FAIR Digital Object (FDO) approach (De Smedt et al. 2020) gives a set of recommendations that aims to improve findability, accessibility, interoperability and reproducibility for any digital object, allowing implementation through different protocols or standards.

Here we present how we have followed the FDO recommendations and turned research outcomes into FDOs by publishing RO-Crates on the Web using HTTP, following best practices for Linked Data. We highlight challenges and advantages of the FDO approach, and reflect on what is required for an FDO profile to achieve FAIR RO-Crates.

The implementation allows for a broad range of use cases, across scientific domains. A minimal RO-Crate may be represented as a persistent URI resolving to a summary website describing the outputs in a scientific investigation (e.g. https://w3id.org/dgarijo/ro/sepln2022 with links to the used datasets along with software).

One of the advantages of RO-Crates is flexibility, particularly regarding the metadata accompanying the actual research outcome. RO-Crate extends schema.org, a popular vocabulary for describing resources on the Web (Guha et al. 2016). A generic RO-Crate is not required to be typed beyond *Dataset*[1]. In practice, RO-Crates declare conformance to particular profiles, allowing processing based on the specific needs and assumptions of a community or usage scenario. This, effectively, makes RO-Crates typed and thus machine-actionable. RO-Crate profiles serve as metadata templates, making it easier for communities to agree and build upon their own metadata needs.

RO-Crates have been combined with *machine-actionable Data Management Plans* (maDMPs) to automate and facilitate management of research data (Miksa et al. 2020). This mapping allows RO-Crates to be generated out of maDMPs and vice versa. The ELIXIR Software Management Plans (Alves et al. 2021) is planning to move their questionnaire to a machine-actionable format with RO-Crate. ELIXIR Biohackathon 2022 will explore integration of RO-Crate and the Data Stewardship Wizard (Pergl et al. 2019) with Galaxy, which can automate FDO creation that also follows data management plans.

A tailored RO-Crate profile has been defined to represent Electronic Lab Notebooks (ELN) protocols bundled together with metadata and related datasets. Schröder et al. (2022) uses RO-Crates to encode provenance information at different levels, including researchers, manufacturers, biological and chemical resources, activities, measurements, and resulting research data. The use of RO-Crates makes it easier to programmatically question-answer information related to the protocols, for instance activities, resources and equipment used to create data.

Another example is WorkflowHub (Goble et al. 2021) which defines the Workflow RO-Crate profile (Bacall et al. 2022), imposing additional constraints such as the presence of a main workflow and a license. It also specifies which entity types and properties must be used to provide such information, implicitly defining a set of operations (e.g., get the main workflow and its language) that are valid on all complying crates. The workflow system Galaxy (The Galaxy Community 2022) retrieves such Workflow Crates using GA4GH TRS API.

The workflow profile has been further extended (with OOP-like inheritance) in Workflow Testing RO-Crate, adding formal workflow testing components: this adds operations such as getting remote test instances and test definitions, used by the LifeMonitor service to keep track of the health status of multiple published workflows.

While RO-Crates use Web technologies, they are also *self-contained*, moving data along with their metadata. This is a powerful construct for interoperability across FAIR repositories, but this raises some challenges with regards to mutability and persistence of crates.

To illustrate how such challenges can be handled, we detail how the WorkflowHub repository follows several FDO principles:

1. Workflow entries must be *frozen* for editing and have complete kernel metadata (title, authors, license, description) [FDOF4] before they can be assigned a

persistent identifier, e.g. https://doi.org/10.48546/workflowhub.workflow.255.1 [FDOF1]

2. Computational workflows can be composed of multiple files used as a whole, e.g. CWL files in a GitHub repository. These are snapshotted as a single RO-Crate ZIP, indicating the main workflow. [FDOF11]

3. PID resolution can content-negotiate to Datacite's PID metadata [FDOF2] or use FAIR Signposting to find an RO-Crate containing the workflow [FDOF3] and richer JSON-LD metadata resources [FDOF5,FDOF8], see Fig. 1

4. Metadata uses schema.org [FDOF7] following the community-developed Bioschemas ComputationalWorkflow profile [FDOF10].

5. Workflows are discovered using the GA4GH TRS API [FDOF5,FDOF6,FDOF11] and created/modified using CRUD operations [FDOF6]

6. The RO-Crate profile, effectively the FDO Type [FDOF7], is declared as https://w3id.org/workflowhub/workflow-ro-crate/1.0; the workflow language (e.g. https://w3id.org/workflowhub/workflow-ro-crate#galaxy) is defined in metadata of the main workflow.

Further work on RO-Crate profiles include to formalise links to the API operations and repositories [FDOF5,FDOF7], to include PIDs of profiles and types in the FAIR Signposting, and HTTP navigation to individual resources within the RO-Crate.

RO-Crate has shown a broad adoption by communities across many scientific disciplines, providing a lightweight, and therefore easy to adopt, approach to generating FAIR Digital Objects. It is rapidly becoming an integral part of the interoperability fabric between the different components as demonstrated here for WorkflowHub, contributing to building the European Open Science Cloud.

Keywords

FAIR, research object, linked data, RO-Crate, JSON-LD, FDO, WorkflowHub

Presenting author

Stian Soiland-Reyes

Presented at

First International Conference on FAIR Digital Objects, poster

Acknowledgements

We would like to acknowledge the RO-Crate community and the WorkflowHub Club.

Funding program

Stian Soiland-Reyes is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement numbers 823830 (BioExcel-2), 824087 (EOSC-Life) and the Horizon Europe programme under grant agreement 101046203 (BY-COVID).

Daniel Garijo is supported by the Madrid Government (Comunidad de Madrid-Spain) under the Multiannual Agreement with Universidad Politécnica de Madrid in the line Support for R&D projects for Beatriz Galindo researchers, in the context of the V PRICIT (Regional Programme of Research and Technological Innovation).

Leyla Jael Garcia is supported by German Research Foundation DFG grant for NFDI4DataScience.

Frederik Coppens is supported by Research Foundation - Flanders (FWO) for ELIXIR Belgium (I002819N).

Author contributions

Author contributions to this article according to the Contributor Roles Taxonomy CASRAI CrEDiT:

- **Stian Soiland-Reyes**: Conceptualization, Funding acquisition, Project administration, Software, Writing – original draft, Writing – review & editing
- **Peter Sefton**: Funding acquisition, Project administration, Software
- **Leyla Jael Castro**: Writing – original draft, Writing – review & editing
- **Frederik Coppens**: Funding acquisition, Supervision, Writing – review & editing
- **Daniel Garijo**: Software, Writing - review and editing
- **Simone Leo**: Conceptualization, Project administration, Software, Writing – original draft
- **Marc Portier**: Writing – review & editing
- **Paul Groth**: Supervision

Conflicts of interest

References

- Alves R, Bampalikis D, Castro LJ, Fernández JM, Harrow J, Kuzak M, Martin E, Psomopoulos F, Via A (2021) ELIXIR Software Management Plan for Life Sciences. BioHackrXiv https://doi.org/10.37044/osf.io/k8znb
- Bacall F, Williams AR, Owen S, Soiland-Reyes S (2022) Workflow RO-Crate profile 1.0. https://w3id.org/workflowhub/workflow-ro-crate/1.0. Accessed on: 2022-7-10.
- Bayarri G, Hospital A (2022) Automatic Ligand parameterization for GROMACS. WorkflowHub https://doi.org/10.48546/workflowhub.workflow.255.1
- Bizer C, Heath T, Berners-Lee T (2009) Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems 5 (3): 1-22. https://doi.org/10.4018/jswis.2009081901
- De Smedt K, Koureas D, Wittenburg P (2020) FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. Publications 8 (2). https://doi.org/10.3390/publications8020021
- Goble C, Soiland-Reyes S, Bacall F, Owen S, Williams A, Eguinoa I, Droesbeke B, Leo S, Pireddu L, Rodriguez-Navas L, Fernández JM, Capella-Gutierrez S, Ménager H, Grüning B, Serrano-Solano B, Ewels P, Coppens F (2021) Implementing FAIR Digital Objects in the EOSC-Life Workflow Collaboratory. Zenodo https://doi.org/10.5281/zenodo.4605654
- Guha RV, Brickley D, Macbeth S (2016) Schema.org. Communications of the ACM 59 (2): 44-51. https://doi.org/10.1145/2844544
- Miksa T, Jaoua M, Arfaoui G (2020) Research Object Crates and Machine-actionable Data Management Plans. *First Workshop on Data and Research Objects Management for Linked Open Science (DaMaLOS)* https://doi.org/10.4126/frl01-006423291
- Pergl R, Hooft R, Suchánek M, Knaisl V, Slifka J (2019) “Data Stewardship Wizard”: A Tool Bringing Together Researchers, Data Stewards, and Data Experts around Data Management Planning. Data Science Journal 18 (1). https://doi.org/10.5334/dsj-2019-059
- Schröder M, Staehlke S, Groth P, Nebe JB, Spors S, Krüger F (2022) Structure-based knowledge acquisition from electronic lab notebooks for research data provenance documentation. Journal of Biomedical Semantics 13 https://doi.org/10.1186/s13326-021-00257-x
- Soiland-Reyes S (2022) stain/signposting: signposting v0.7.0. Zenodo https://doi.org/10.5281/zenodo.6815412
- Soiland-Reyes S, Sefton P, Crosas M, Castro LJ, Coppens F, Fernández J, Garijo D, Grüning B, Rosa ML, Leo S, Ó Carragáin E, Portier M, Trisovic A, RO-Crate Community, Groth P, Goble C (2022) Packaging research artefacts with RO-Crate. *Data Science* 5 (2). https://doi.org/10.3233/DS-210053

- The Galaxy Community (2022) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. Nucleic Acids Research 50 https://doi.org/10.1093/nar/gkac247
- Wilkinson M, et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18

Endnotes

[*1] Resources described by an RO-Crate are also typed, e.g. *Person*, *Organization*, *ScholarlyArticle*, *ImageObject*

```
(a2a) stain@xena:~$ signposting https://doi.org/10.48546/workflowhub.workflow.255.1
Signposting for https://workflowhub.eu/workflows/255?version=1
CiteAs: <https://doi.org/10.48546/workflowhub.workflow.255.1>
DescribedBy: <https://workflowhub.eu/workflows/255?version=1> application/vnd.datacite.datacite+xml
             <https://workflowhub.eu/workflows/255?version=1> application/ld+json
Item: <https://workflowhub.eu/workflows/255/ro_crate?version=1> application/zip
```

Figure 1.

FAIR Signposting on a workflow PID (Bayarri and Hospital 2022) discovered from HTTP *Link:* headers using the Signposting tool (Soiland-Reyes 2022) shows machine-actionable navigation to content-negotiate for the metadata FDOs, as well as download bit sequence [FDOF3] as an RO-Crate zip. JSON-LD from workflowhub.eu follows the BioSchemas ComputationalWorkflow profile to give workflow details not included in DataCite's general JSON-LD.