Enhancing RDM in Galaxy by integrating RO-Crate

Paul De Geest[‡], Frederik Coppens[‡], Stian Soiland-Reyes[§,|], Ignacio Eguinoa[‡], Simone Leo[¶]

‡ Center for Plant Systems Biology (PSB), The Flemish Institute for Biotechnology (VIB), Gent, Belgium
§ Department of Computer Science, The University of Manchester, Manchester, United Kingdom
| Informatics Institute, University of Amsterdam, Amsterdam, Netherlands
¶ Center for Advanced Studies, Research, and Development in Sardinia (CRS4), Pula (CA), Italy

Corresponding author: Paul De Geest (paul.degeest@psb.ugent.be), Frederik Coppens (frcop@psb.vib-ugent.be), Stian Soiland-Reyes (soiland-reyes@manchester.ac.uk), Ignacio Eguinoa (ignacio.eguinoa@psb.vib-ugent.be), Simone Leo (simone.leo@crs4.it)

Abstract

We introduce how the Galaxy research environment (Jalili et al. 2020) integrates with RO-Crate as an implementation of Findable Accessible Interoperable Reproducible Digital Objects (FAIR Digital Objects / FDO) (Wilkinson et al. 2016, Schultes and Wittenburg 2018) and how using RO-Crate as an exchange mechanism of workflows and their execution history helps integrate Galaxy with the wider ecosystem of ELIXIR (Harrow et al. 2021) and the European Open Science Cloud (EOSC-Life) to enable FAIR and reproducible data analysis.

RO-Crate (Soiland-Reyes et al. 2022) is a generic packaging format containing datasets and their description using standards for FAIR Linked Data. The format is based on schema.org (Guha et al. 2016) annotations in JSON-LD, which allows for rich metadata representation. The RO-Crate effort aims to make best-practice in formal metadata description accessible and practical for use in a wider variety of situations, from an individual researcher working with a folder of data, to large data-intensive computational research environments.

The RO-Crate community brings together practitioners from very different backgrounds, and with different motivations and use cases. Among the core target users are:

- researchers engaged with computation and data-intensive, workflow-driven analysis;
- digital repository managers and infrastructure providers;
- individual researchers looking for a straightforward tool or how-to guide to “FAIRify” their data;
- data stewards supporting research projects in creating and curating datasets.

Given the wide applicability of RO-Crate and the lack of practical implementations of FDOs, ELIXIR (Harrow et al. 2021) co-opted this initiative as the project to define a common format for research data exchange and repository entries. Thus, during the last

year it's been implemented in a wide range of services, such as: WorkflowHub (Goble et al. 2021) (a registry for describing, sharing and publishing scientific computational workflows) uses RO-Crates as an exchange format to improve reproducibility of computational workflows that follow the Workflow RO-Crate profile (Bacall et al. 2022); LifeMonitor (Leo et al. 2022) (a service to support the sustainability of computational workflows being developed as part of the EOSC-Life project) uses RO-Crate as an exchange format for describing test suites associated with workflows.

Tools have been developed towards aiding the previously mentioned use cases and increasing the general usability of RO-Crates by providing a user-friendly (programmatic) interface for consumption and production of RO-Crates through programmatic libraries for consuming/producing RO-Crates (ro-crate-py De Geest et al. 2022, ro-crate-ruby Bacall and Whitwell 2022, ro-crate-js Lynch et al. 2021).

The Galaxy project provides a research environment with data analysis and data management functionalities as a multi user platform, aiming to make computational biology accessible to research scientists that do not have computer programming or systems administration experience. As such, it stores not just analysis related data but also the complete analytical workflow, including its metadata. The internal data model involves the history entity, including all steps performed in a specific analysis, and the workflow entity, defining the structure of an analytical pipeline. From the start, Galaxy aims to enable reproducible analyses by providing capabilities to export (and import) all the analysis history details and workflow data and metadata in a FAIR way. As such it helps its users with the daily research data management. The Galaxy community is continuously improving and adding features, the integration of the FAIR Digital Object principles is a natural next step in this.

To be able to support these FDOs, Galaxy leverages the RO-Crate Python client library (De Geest et al. 2022) and provides multiple entry points to import and export different research data objects representing its internal entities and associated metadata. These objects include:

1. a workflow definition, which is used to share/publish the details of an analysis pipeline, including the graph of tools that need to be executed, and metadata about the data types required
2. individual data files or a collection of datasets related to an analysis history
3. a compressed archive of the entire analysis history including the metadata associated with it such as the tools used, their versions, the parameters chosen, workflow invocation related metadata, inputs, outputs, license, author, CWLProv description (Khan et al. 2019) of the workflow, contextual references in the form of Digital Object Identifiers (DOIs), 'EMBRACE Data And Methods' ontology (EDAM) terms (Ison et al. 2013), etc.

The adoption of RO-crate by Galaxy allows a standardised exchange of FDOs with other platforms in the ELIXIR Tools ecosystem, such as WorkflowHub and LifeMonitor. Integrating RO-Crate deeply into Galaxy and offering import and export options of various

Galaxy objects such as Research Objects allows for increased standardisation, improved Research Data Management (RDM) functionalities, smoother user experience (UX) as well as improved interoperability with other systems. The integration in a platform used by biologists to do data intensive analysis, facilitates the publication of workflows and workflow invocations for all skill levels and democratises the ability to perform Open Science.

Keywords

FAIR digital object, workflow, exchange format, packaging format, metadata description, FAIR, reproducible data analysis

Presenting author

Paul De Geest

Presented at

First International Conference on FAIR Digital Objects, poster

Conflicts of interest

References

- Bacall F, Whitwell M (2022) ResearchObject/ro-crate-ruby: v0.4.17. Zenodo https://doi.org/10.5281/zenodo.6810687
- 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.
- De Geest P, Droesbeke B, Eguinoa I, Gaignard A, Huber S, Leo S, Pireddu L, Rodríguez-Navas L, Sirvent R, Soiland-Reyes S (2022) ro-crate-py. Zenodo https://doi.org/10.5281/zenodo.3956493
- Goble C, Soiland-Reyes S, Bacall F, Owen S, Williams A, Eguinoa I, Droesbeke B, Leo S, Pireddu L, Rodríguez-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
- Harrow J, Drysdale R, Smith A, Repo S, Lanfear J, Blomberg N (2021) ELIXIR: providing a sustainable infrastructure for life science data at European scale. Bioinformatics 37 (16): 2506-2511. https://doi.org/10.1093/bioinformatics/btab481

- Ison J, Kalas M, Jonassen I, Bolser D, Uludag M, McWilliam H, Malone J, Lopez R, Pettifer S, Rice P (2013) EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. Bioinformatics (Oxford, England) 29 (10): 1325-32. https://doi.org/10.1093/bioinformatics/btt113
- Jalili V, Afgan E, Gu Q, Clements D, Blankenberg D, Goecks J, Taylor J, Nekrutenko A (2020) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. Nucleic Acids Research 48 https://doi.org/10.1093/nar/gkaa434
- Khan FZ, Soiland-Reyes S, Sinnott RO, Lonie A, Goble C, Crusoe MR (2019) Sharing interoperable workflow provenance: A review of best practices and their practical application in CWLProv. GigaScience 8 (11). https://doi.org/10.1093/gigascience/giz095
- Leo S, Piras ME, Pireddu L (2022) Welcome to Life-Monitor. https://about.lifemonitor.eu/ . Accessed on: 2022-7-10.
- Lynch M, Sefton P, Soiland-Reyes S (2021) UTS-eResearch/ro-crate-js. v2.0.7. GitHub. Release date: 2021-12-07. URL: https://github.com/UTS-eResearch/ro-crate-js
- Schultes E, Wittenburg P (2018) FAIR Principles and Digital Objects: Accelerating Convergence on a Data Infrastructure Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2018. Communications in Computer and Information Science, 1003. In: Manolopoulos Y, Stupnikov S (Eds) Communications in Computer and Information Science. Springer https://doi.org/10.1007/978-3-030-23584-0
- Soiland-Reyes S, Sefton P, Crosas M, Castro LJ, Coppens F, Fernández J, Garijo D, Grüning B, La Rosa M, Leo S, Ó Carragáin E, Portier M, Trisovic A, RO-Crate Community, Groth P, Goble C (2022) Packaging research artefacts with RO-Crate. Data Science1-42. https://doi.org/10.3233/ds-210053
- Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific data 3: 160018. https://doi.org/10.1038/sdata.2016.18