

24/07/2025

Version 1.0



M23 Improving FAIRability of research with RO-Crates and Bioschemas

Author(s): Claus Weiland (SGN)

Contributor(s): Daniel Bauer (SGN), Caterina Bergami (CNR), Niels Billiet (APM), Chiara Bortoluzzi (SIB), Jonas Grieb (SGN) Ayco Holleman (Naturalis), Taimur Khan (UFZ), Julian Oeser (UFZ), Alessandro Oggioni (CNR-IREA), Christoph Wohner (EAA) and Rajapreethi Rajendran (SGN)



Co-funded by
the European Union

Project funded by



Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs,
Education and Research EAER
**State Secretariat for Education,
Research and Innovation SERI**

BMD (Biodiversity Meets Data) receives funding from the European Union's Horizon Europe Research and Innovation Programme and the Swiss State Secretariat for Education, Research and Innovation (SERI) (ID No 101181294). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union, the European Research Executive Agency (REA) or SERI. The EU, REA and SERI cannot be held responsible for them.

**Prepared under contract from the European Commission**

Grant agreement No. 101181294

EU Horizon Europe Research and Innovation Action

Project acronym:	BMD
Project full title:	Biodiversity Meets Data
Project duration:	01.03.2025 – 28.02.2029 (48 months)
Project coordinator:	Stichting Naturalis Biodiversity Center (Naturalis)
Call:	HORIZON-CL6-2024-BIODIV-01
Milestone title:	Workflow concept for FAIR workflow into VREs available and disseminated in in-person workshop <i>Improving FAIRability of research with RO-Crates and Bioschemas</i>
Milestone n°:	M23
Means of verification:	Report, Dataset, List of attendees (attached)
Work package:	WP5
Nature of the milestone:	Report
Contribution to deliverable n°:	D5.1
Licence of use:	This document is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license. You are free to share and adapt the material, provided appropriate credit is given.
Lead beneficiary:	SGN
Recommended citation:	Weiland, C., Bauer, D., Bergami, C., Billiet, N., Bortoluzzi, C., Grieb, J., Holleman, A., Khan, T., Oeser, J., Oggioni, A., Wohner, C. and Rajendran, R. et al. (2025). <i>Improving FAIRability of research with RO-Crates and Bioschema</i> . BMD project deliverable M23.
Due date of milestone:	M03
Actual submission date:	M04
Quality review:	Yes/No

Milestone status:

Version	Status	Date	Author(s)	Actions
0.1	Version 1.0	25 July 2025	Weiland, SGN + all participants	Sent for review
0.2				Reviewed





x.x	Finalised, with incorporation of feedback from reviewers
1.0	Submitted





Table of contents

Table of contents	5
Summary	6
List of abbreviations	6
1. Introduction	7
2. Hackathon Objectives	7
3. Hackathon Preparation	8
4. Hackathon Outcomes and Challenges	9
5. Future work	11
6. References	11
7. Resources	12
GitHub Repository with selected hackathon results:	12
8. Acknowledgements	13
9. References	14
10. Annex	17





Summary

Short (maximum 1 page) executive summary of the milestone.

List of abbreviations

BMD	Biodiversity Meets Data
eLTER	European Long-Term Ecosystem, Critical Zone and Socio-ecological Research Infrastructure
EU	European Union
CWL	Common Workflow Language
MLC	Machine Learning Commons
RDI	Research Data Infrastructure
RO-Crate	Research Object Crate
SDM	Species Distribution Model
WP	Work Package





1. Introduction

BMD's work package 5 provides through the implementation of thematic Virtual Research Environments the central tools for addressing needs and requirements of BMD's stakeholders regarding science- and particularly data-driven guidance concerning the management of conservation sites. In this respect, as presented in more detail in the emerging BMD WP5 Handbook (Oeser 2025), the Virtual Research Environments offer platforms that facilitate (i) assessments of biodiversity, (ii) analyses of biodiversity trends, and (ii) evaluations of the anticipated effects of climate and land cover changes on biodiversity.

To ensure scalability of computation, near-data processing close to the physical location of data and facilitated accessibility of the provided Species Distribution Models (SDMs), they will be effectively deployed and used as cloud services within the framework of BMD (Islam 2025).

The term 'thematic VRE' thus means that BMD provides these models in a coordinated combination of data, computational resources, and software including both domain-specific models as well as more generic tools to foster interoperability and re-usability. In BMD's toolkit, FAIR computational workflows (Wilkinson 2025) are essential building blocks for the emerging service ecosystem that will be provided within the planned Biodiversity Data Space (Islam 2025).

In this way, BMD will simplify the usage of SDMs for researchers and other stakeholders such as conservationists to model species distributions, investigate future projections on existing climate change scenarios and potentially use these projections to support adequate mitigation strategies.

2. Hackathon Objectives

The major aim of the hackathon was to advance the development of the fundamental models and methods to operate data-driven VREs leveraging on FAIR scientific workflows.

The chosen approach consisted of bringing together prototypical architectural components (T4.2 Cloud computing for VREs) and software specifications based on community standards, since data cannot be FAIR when an infrastructure does not implement policies, rules and procedures for FAIR (Lannom 2020).

To support the core objectives of BMD, particularly the provision of data from biodiversity monitoring harmonized across space-time-taxonomy (Raes 2025), a major focus of the hackathon has been placed on interoperable community formats involving RO-Crate (Soilant-Reyes 2022), the Common Workflow Language (CWL, Crusoe 2022), and [Schema.org](https://schema.org) and its extensions such as MLC Croissant.

RO-Crate provides a lightweight container for FAIR-compliant packaging of research data together with (i.a.) metadata, workflow descriptions in CWL and links to schema mappings to make data cross-domain interoperable. The Workflow Run RO-Crate is an enhancement of RO-Crate (Research Object Crate) and Schema.org, designed to document the provenance of computational workflow executions at various granularity levels and to consolidate all related outputs (Leo 2024).

CWL is an open standard widely adopted for describing and sharing computational workflows. It ensures that these workflows are portable and capable of running across various platforms and workflow engines, including Apache Airflow (Kotliar 2019), which is being considered as a workflow orchestration system for BMD. Primarily, CWL focuses on detailing the execution of command-line tools and their integration to form workflows, particularly in data-intensive scientific domains.





To foster machine-actionable data reuse of the products created in the workflows (Jacobsen 2019), the hackathon also put additional emphasis on AI-readiness involving the MLC Croissant specification, describing a high-level metadata format developed to open up the datasets for machine learning and autonomous processing. (Akthar 2024).

3. Hackathon Preparation

In preparation for the workshop, a workflow system developed in the context of the Biodiversity Digital Twin (BioDT) project was adapted for deployment in BMD's cloud (Figure 1, Weiland 2024).

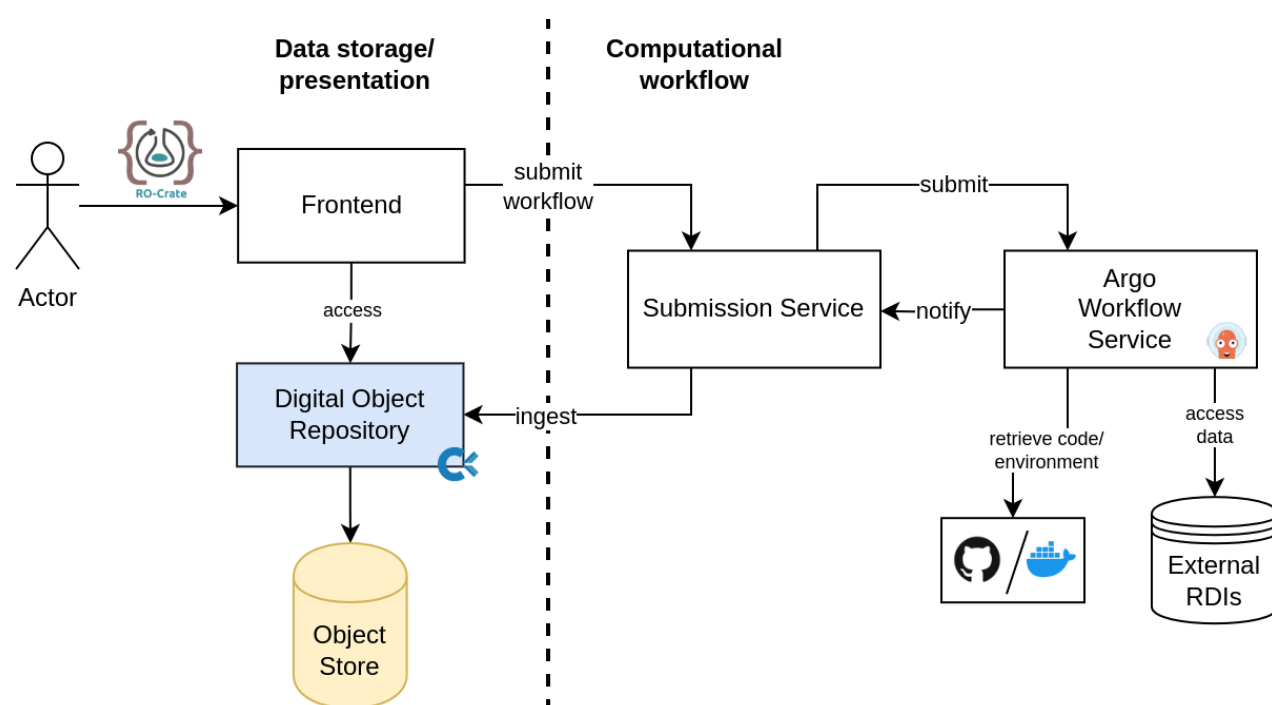


Figure 1: FAIR Workflow platform comprising the essential components Digital Object Repository, lightweight fronted, Workflow Submission Service and Workflow Service (Image: Bauer 2025).

The platform is built on a microservice architecture (as shown in Fig. 1), which allows for flexible extension and easy replacement of components during future integration into the Biodiversity Data Space (T4.1, T4.2). The design enables the effortless swapping of the underlying workflow engine with an alternative such as the aforementioned Apache Airflow framework.

The platform comprises four main components:

- **Workflow Submission Service (WSub):** This acts as a bridge between the workflow engine's API and other platform components. It abstracts the underlying workflow engine's API and





incorporates features necessary for managing workflow execution. The WSub queues workflows submitted to the Wserv and manages data ingestion into the DOR.

- **Workflow Service (Wserv):** This component is responsible for orchestrating and executing workflows. It retrieves workflows from the WSub, executes them, and temporarily stores the resulting artifacts.
- **Digital Object Repository (DOR):** This component utilizes an instance of CNRI's Cordra to store workflows, resulting data, and associated metadata as digital objects, each assigned a persistent identifier. These objects are structured using JSON-LD syntax, ensuring they remain meaningful and linkable beyond their use in constructing RO-Crates. Cordra interfaces with a storage backend and offers search functionality through a Lucene-based index. The DOR serves as the primary data storage for the application.
- **Frontend:** Developed as a Django application, the frontend interacts with both the WServ and DOR. Users can log in using their ORCID credentials to submit workflows for execution. The frontend also displays datasets stored in the DOR and dynamically constructs Workflow Run RO-Crates for download or linking. Additionally, the platform implements FAIR Signposting (not explored in depth during the hackathon) on each dataset's landing page to provide machine-interpretable metadata (Soilant-Reyes 2025).

4. Hackathon Outcomes and Challenges

The agenda was very clear given the time constraints (2 full workshop days lunch to lunch). The first day included a summary of the theoretical basics, particularly container formats such as RO-Crate and workflow-specific applications like Workflow Run RO-Crate. Additionally, the prepared workflow system was presented (see section 3). Following that, participants contributed presentations on their use cases. This included a longer discussion of a prototype for the user interfaces for the VRE building on an R-Shiny app - this was further elaborated into one of the hackathon projects. Following this discussion use cases were clustered and developed into the finally 4 hackathon projects:

- **Building RO-Crates for eLTER's measurement protocols,**
 - The aim was to evaluate the suitability of RO-Crates for documenting a 'typical' eLTER data flow—from initial data acquisition using a standard protocol, through data processing, to the final publication of the data in a dedicated repository, resulting in a citable dataset with accompanying metadata.
- **Torchgbif:** Facilitating access to GBIF data via Pytorch,
 - TorchGBIF is an open-source Python library offering PyTorch-compatible datasets and data loaders for GBIF biodiversity occurrence data, facilitating seamless integration of GBIF's vast data holdings into deep learning pipelines. It adheres to FAIR (Findable, Accessible, Interoperable, Reusable) data principles through features such as automated metadata generation, RO-Crate packaging, and provenance tracking for reproducible ecological modeling.
 - [Codebase: https://github.com/thisistaimur/torchgbif](https://github.com/thisistaimur/torchgbif)





- Deploying a bats species distribution modelling (SDM) workflow using Workflow Run RO-Crates in our prototypical workflow platform
 - We developed a Workflow RO-Crate for the bats SDM workflow, encapsulating the R-based script within an Argo workflow. The containerized environment for the SDM was adapted from a Shiny app image to support non-interactive execution of the SDM workflow within a workflow engine. The resulting Workflow RO-Crate is available on GitHub (see [resources](#)).
- Exploration of querying BMD's RO-Crates in a triple store
 - Building on RO-Crate's inherent interoperability with Linked Open Data principles by integrating the [schema.org](#) vocabulary as a metadata standard, we worked on Extract, Transform and Load (ETL) pipelines, with the aim to load the metadata of numerous RO-Crates into a single triple store. Subsequently, this would make aggregated queries using the SPARQL query language over all RO-Crates (containing e.g. the created model outputs or workflow definitions of the VREs in BMD) possible. In this project, Apache Jena Fuseki¹ was used as a lightweight and open source triple store software. In the scope of this hackathon, a first prototypical pipeline has been implemented. Future works will include the further expansion of this approach and, importantly, the discussion with other BMD stakeholders (e.g. WP4) regarding the integration of this feature in the overall BMD technical architecture.

In summary, the prototype for the visualization of the VREs was evaluated as a very good product for connecting and engaging end users. RO-Crates, on the other hand, are not a technology for the end users of the services but rather a part of the service architecture with which users interact only indirectly through user-friendly interfaces.

Building on the results from the hackathon, a new structural draft for the implementation of the VRE architecture was proposed (Figure 2). This structural draft describes the interaction between end user-friendly components (like the Single Access Point and the user interface(s) of the individual VRE) and the Workflow execution platform. The draft also highlights further components which will be required: A database for storing the results of the execution of an individual workflow and a task-queuing system. Computational tasks of the different VREs will possibly be quite extensive, so if several workflows are requested to run at the same time on a single workflow execution platform, this could overload the platform. Therefore it will be required to have several workflow execution platforms in place and a task-queuing system which acts as a load balancer to distribute incoming workflow execution tasks.

¹ <https://jena.apache.org/documentation/fuseki2/>



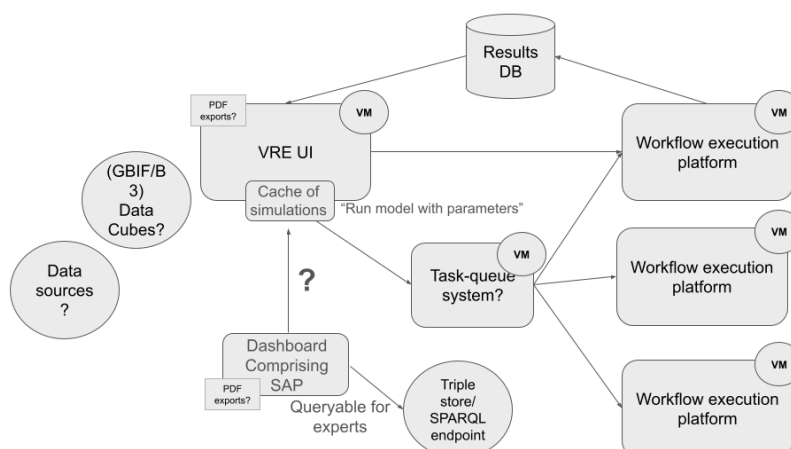


Figure 2: VRE architecture in BMD

5. Future work

During the hackathon, the application of RO-Crates was examined in relation to a series of use cases in the context of VRE. This led to a revised and expanded architectural concept that now needs to be refined and further developed in close coordination between the two work packages 4 and 5.

An essential focus of work that emerges from the progress made on the hackathon is the further development of the VRE-UI and the better functional integration of UI and Workflow Service.

Building on the results from the hackathon, we aim to develop the foundations for Deliverable 5.1 (FAIR Workflows into VREs) in the form of a second, extended hackathon (Q1 2026).

6. References

Mubashara, A., Benjelloun, O., Conforti, C. et al. (2024) *Croissant: A Metadata Format for ML-Ready Datasets*. In Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning (DEEM '24). Association for Computing Machinery, New York, NY, USA, 1–6. <https://doi.org/10.1145/3650203.3663326>

Bauer, D. (2025) *A FAIR workflow platform for biodiversity digital twins*. <https://github.com/Senckenberg-DCBiodivIT/fair-workflow-platform> [Retrieved 24. July 2025]

Crusoe, M.R., Abeln, S., Iosup, A., Amstutz, P. et al. (2022) *Methods Included: Standardizing Computational Reuse and Portability with the Common Workflow Language*. Com. ACM 65, 6 (June 2022), 54–63. <https://doi.org/10.1145/3486897>

Islam, S., (2025). *Initial data space design with implementation plan*. BMD project deliverable MS19, , unpublished manuscript





Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D. et al. *FAIR Principles: Interpretations and Implementation Considerations*. *Data Intelligence* 2020; 2 (1-2): 10–29. doi: https://doi.org/10.1162/dint_r_00024

Kotliar, M., Kartashov, A., Barski, A. (2019) *CWL-Airflow: a lightweight pipeline manager supporting Common Workflow Language*, *GigaScience*, Volume 8, Issue 7, <https://doi.org/10.1093/gigascience/giz084>

Lannom, L., Koureas, D., Hardisty, A.R. (2020) *FAIR Data and Services in Biodiversity Science and Geoscience*. *Data Intelligence*, 2020, 2 (1-2): 122-130. 2024-08-13. 2025-07-24. https://www.sciengine.com/doi/10.1162/dint_a_00034

Simone Leo, Michael R. Crusoe, Laura Rodríguez-Navas, Raúl Sirvent, Alexander Kanitz, Paul De Geest, Rudolf Wittner, Luca Pireddu, Daniel Garijo, José M. Fernández, Iacopo Colonnelli, Matej Gallo, Tazro Ohta, Hirotaka Suetake, Salvador Capella-Gutierrez, Renske de Wit, Bruno P. Kinoshita, Stian Soiland-Reyes (2024) Recording provenance of workflow runs with RO-Crate. *PLoS ONE* 19(9): e0309210. <https://doi.org/10.1371/journal.pone.0309210>

Oeser, J. et al. (2025). *BMD WP5 Handbook* <https://docs.google.com/document/d/1FcMswaoa4ed-muF6-amZN25KTzNfcVBSke46s8Rd1Tk/edit?usp=sharing> [Retrieved Jul 23. 2025]

Raes, Nils (2025), *Biodiversity Meets Data*, HORIZON-CL6-2024-BIODIV-01 proposal, unpublished manuscript

Stian Soiland-Reyes, Peter Sefton, Mercè Crosas, Leyla Jael Castro, Frederik Coppens, José M. Fernández, Daniel Garijo, Björn Grüning, Marco La Rosa, Simone Leo, Eoghan Ó Carragáin, Marc Portier, Ana Trisovic, RO-Crate Community, Paul Groth, Carole Goble (2022): Packaging research artefacts with RO-Crate. *Data Science* 5(2) <https://doi.org/10.3233/DS-210053>

Soiland-Reyes, S., Sefton, P., Leo, S., Castro, L. J., Weiland, C., & Van de Sompel, H. (2025). *Practical webby FDOs With RO-Crate and FAIR Signposting: Experiences and Lessons Learned*. Open Conference Proceedings, 5. <https://doi.org/10.52825/ocp.v5i.1273>

Weiland C, Grieb J, Bauer D, Chala D, Kusch E, Andrew C, Endresen D (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 2025 Wilkinson, S.R., Aloqalaa, M., Belhajjame, K. et al. *Applying the FAIR Principles to computational workflows*. *Sci Data* 12, 328 (2025). <https://doi.org/10.1038/s41597-025-04451-9>

7.Resources

GitHub Repository with selected hackathon results:





<https://github.com/Biodiversity-Meets-Data/2025-FAIR-Workflow-Hackathon>

FAIR workflow platform developed and adapted for the Hackathon by Daniel Bauer for the deployment of Workflow Run RO-Crates: <https://github.com/Senckenberg-DCBiodivIT/fair-workflow-platform>

TorchGBIF: <https://github.com/thisistaimur/torchgbif>

Slides with the theoretical content (day 1): <https://doi.org/10.5281/zenodo.16444312>

8. Acknowledgements

Niels Raes, Vânia Leonardo Feirrer, Michaela Kämmerer, Sybille Roller, Ingolf Kühn

9. References

Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Pieter Gijsbers, Joan Giner-Miguel, Nitisha Jain, Michael Kuchnik, Quentin Lhoest, Pierre Marcenac, Manil Maskey, Peter Mattson, Luis Oala, Pierre Ruyssen, Rajat Shinde, Elena Simperl, Goeffry Thomas, Slava Tykhonov, Joaquin Vanschoren, Jos van der Velde, Steffen Vogler, and Carole-Jean Wu. 2024. Croissant: A Metadata Format for ML-Ready Datasets. In Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning (DEEM '24). Association for Computing Machinery, New York, NY, USA, 1–6. <https://doi.org/10.1145/3650203.3663326>

Bauer, D. (2025) A FAIR workflow platform for biodiversity digital twins. <https://github.com/Senckenberg-DCBiodivIT/fair-workflow-platform> [Retrieved 24. July 2025]

Michael R. Crusoe, Sanne Abeln, Alexandru Iosup, Peter Amstutz, John Chilton, Nebojša Tijanić, Hervé Ménager, Stian Soiland-Reyes, Bogdan Gavrilović, Carole Goble, and The CWL Community. 2022. Methods Included: Standardizing Computational Reuse and Portability with the Common Workflow Language. Commun. ACM 65, 6 (June 2022), 54–63. <https://doi.org/10.1145/3486897>

Islam, S., (2025). Initial data space design with implementation plan. BMD project deliverable MS19, , unpublished manuscript

Annika Jacobsen, Ricardo de Miranda Azevedo, Nick Juty, Dominique Batista, Simon Coles, Ronald Cornet, Mélanie Courtot, Mercè Crosas, Michel Dumontier, Chris T. Evelo, Carole Goble, Giancarlo Guizzardi, Karsten Kryger Hansen, Ali Hasnain, Kristina Hettne, Jaap Heringa, Rob W.W. Hooft, Melanie Imming, Keith G. Jeffery, Rajaram Kaliyaperumal, Martijn G. Kersloot, Christine R. Kirkpatrick, Tobias Kuhn, Ignasi Labastida, Barbara Magagna, Peter McQuilton, Natalie Meyers, Annalisa Montesanti, Mirjam van Reisen, Philippe Rocca-Serra, Robert Pergl, Susanna-Assunta Sansone, Luiz Olavo Bonino da Silva Santos, Juliane Schneider, George Strawn, Mark Thompson, Andra Waagmeester, Tobias Weigel, Mark D. Wilkinson, Egon L. Willighagen, Peter Wittenburg, Marco Roos, Barend Mons, Erik Schultes; FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence 2020; 2 (1-2): 10–29. doi: https://doi.org/10.1162/dint_r_00024





Michael Kotliar, Andrey V Kartashov, Artem Barski, CWL-Airflow: a lightweight pipeline manager supporting Common Workflow Language, GigaScience, Volume 8, Issue 7, July 2019, giz084, <https://doi.org/10.1093/gigascience/giz084>

Larry Lannom, Dimitris Koureas, Alex R. Hardisty. FAIR Data and Services in Biodiversity Science and Geoscience (J/OL). Data Intelligence, 2020, 2 (1-2): 122-130. 2024-08-13. 2025-07-24. https://www.sciengine.com/doi/10.1162/dint_a_00034

Simone Leo, Michael R. Crusoe, Laura Rodríguez-Navas, Raül Sirvent, Alexander Kanitz, Paul De Geest, Rudolf Wittner, Luca Pireddu, Daniel Garijo, José M. Fernández, Iacopo Colonnelli, Matej Gallo, Tazro Ohta, Hirotaka Suetake, Salvador Capella-Gutierrez, Renske de Wit, Bruno P. Kinoshita, Stian Soiland-Reyes (2024) Recording provenance of workflow runs with RO-Crate. PLoS ONE 19(9): e0309210. <https://doi.org/10.1371/journal.pone.0309210>

Oeser, J. et al. (2025). BMD WP5 Handbook <https://docs.google.com/document/d/1FcMswaoa4ed-muF6-amZN25KTzNfcVBSke46s8Rd1Tk/edit?usp=sharing> [Retrieved Jul 23. 2025]

Raes, Nils (2025), Biodiversity Meets Data, HORIZON-CL6-2024-BIODIV-01 proposal, unpublished manuscript

Stian Soiland-Reyes, Peter Sefton, Mercè Crosas, Leyla Jael Castro, Frederik Coppens, José M. Fernández, Daniel Garijo, Björn Grüning, Marco La Rosa, Simone Leo, Eoghan Ó Carragáin, Marc Portier, Ana Trisovic, RO-Crate Community, Paul Groth, Carole Goble (2022): Packaging research artefacts with RO-Crate. Data Science 5(2) <https://doi.org/10.3233/DS-210053>

Soiland-Reyes, S., Sefton, P., Leo, S., Castro, L. J., Weiland, C., & Van de Sompel, H. (2025). Practical webby FDOs With RO-Crate and FAIR Signposting: Experiences and Lessons Learned. Open Conference Proceedings, 5. <https://doi.org/10.52825/ocp.v5i.1273>

Weiland C, Grieb J, Bauer D, Chala D, Kusch E, Andrew C, Endresen D (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 2025 Wilkinson, S.R., Alokala, M., Belhajjame, K. et al. Applying the FAIR Principles to computational workflows. Sci Data 12, 328 (2025). <https://doi.org/10.1038/s41597-025-04451-9>

10. Annex

