





FAIR Digital Objects in Autosubmit Workflows

Bruno de Paula Kinoshita^{1,*} , Edgar Garriga Nogales¹ ,
Manuel Giménez de Castro Marciani¹ , and Miguel Castrillo Melguizo¹ 

¹ Barcelona Supercomputing Center, ES

*Correspondence: Bruno de Paula Kinoshita, bruno.depaulakinoshita@bsc.es

Abstract. Climate and weather models are, respectively, complex computer programs used to better understand and predict the climate and to forecast weather. Running these programs involves a series of tasks that may require the use of other computer programs and scripts to prepare the input data, replace values in configuration files (e.g. Fortran namelists), prepare parameters for a batch server used in a high-performance computing facility (e.g. Slurm user configuration, project details, job request, etc.), among other possible variables. There are multiple challenges to make these programs and their outputs FAIR (Findable, Accessible, Interoperable, and Reusable) digital objects. The use of computational workflows executed with workflow managers eases the configuration and execution of climate and weather models, encapsulating some of its complexity, and contributes to a more uniform collection of provenance information about the execution of these models. In this article we focus on Autosubmit, an experiment and workflow manager used for running weather, air quality, and climate experiments that implements the Workflow Run RO-Crate profile to archive provenance information. We also highlight points that are not covered by RO-Crate and explain how we plan to address these shortcomings through recommendations found in FDO specifications.

Keywords: Climate Models, Weather Models, Workflows, Workflow Management, Autosubmit, RO-Crate, FDO

FAIR Digital Objects in Autosubmit Workflows

Climate models are computer programs that are able to simulate interactions of different climate components (atmosphere, ocean, land surface, ice) to create projections for future climate, normally run over a long period, e.g. 30+ years. Weather models are used to predict future weather, normally run for forecasts of days. These models may require from hours to days to execute a simulation depending on variables such as input parameters, resolution of the data produced, and size of the grid used. Earth modelling workflows are structured processes for carrying out these simulations [1]. Workflow software automates this process in a flexible and repeatable way and can encapsulate all the complexity that these simulations involve. This is particularly important for ensemble climate simulations, where each member of the ensemble is run with slightly different initial conditions to cover different scenarios [2]. Provenance information is vital in the execution of these programs and workflows both to provide information on what happened during execution, but also from a scientific point concerning reproducibility and replicability.

Over the years workflow managers such as Autosubmit [3], ecFlow [4], and Cylc [5] have been extensively used to run climate and weather experiments. For example, Autosubmit is used at the Barcelona Supercomputing Center (BSC) to run the climate (e.g. EC-Earth model

[6]) and atmospheric compositions experiments (e.g. MONARCH model [7]), ecFlow is used at the European Centre for Medium-Range Forecast (ECMWF) to run the Integrated Forecasting System (IFS) model [8], and Cylc is used at the United Kingdom Met Office and at the New Zealand National Institute of Water and Atmospheric Research (NIWA) to run the Unified Model [9][10]. Provenance has normally been handled in a custom or unstructured way by these workflow managers. Autosubmit is an open-source Python workflow manager and meta-scheduler. It was created in 2011 for use in climate research, to configure and run scientific experiments on high-performance computing (HPC) environments. It supports scheduling jobs via SSH to run on remote batch servers, such as Slurm, PBS, PJM, and others.

Autosubmit has had a feature to archive experiments since its 3.1.0 release, in 2015. This option produces a compressed file with the complete experiment data, logs, traces, and configuration. This archive includes important provenance information (prospective and retrospective [11]) about climate simulations, but it still requires consumers to know how to navigate through tens, hundreds, or thousands of files in order to understand the experiment and workflow configuration and traces and data produced during the workflow execution.

The release 4.0.100 of Autosubmit in November 2023 added support to RO-Crate. This version produces archives that comply with the Workflow Run RO-Crate profile [12]. Consequently, users can upload their workflows to WorkflowHub and Zenodo using RO-Crate, which gives consumers of these digital objects a more structured way to parse and consume the provenance information, enabling not only these objects to be machine-readable, but also machine-interpretable and machine-actionable [13]. Figure 1 has an example of an Autosubmit experiment published in WorkflowHub.

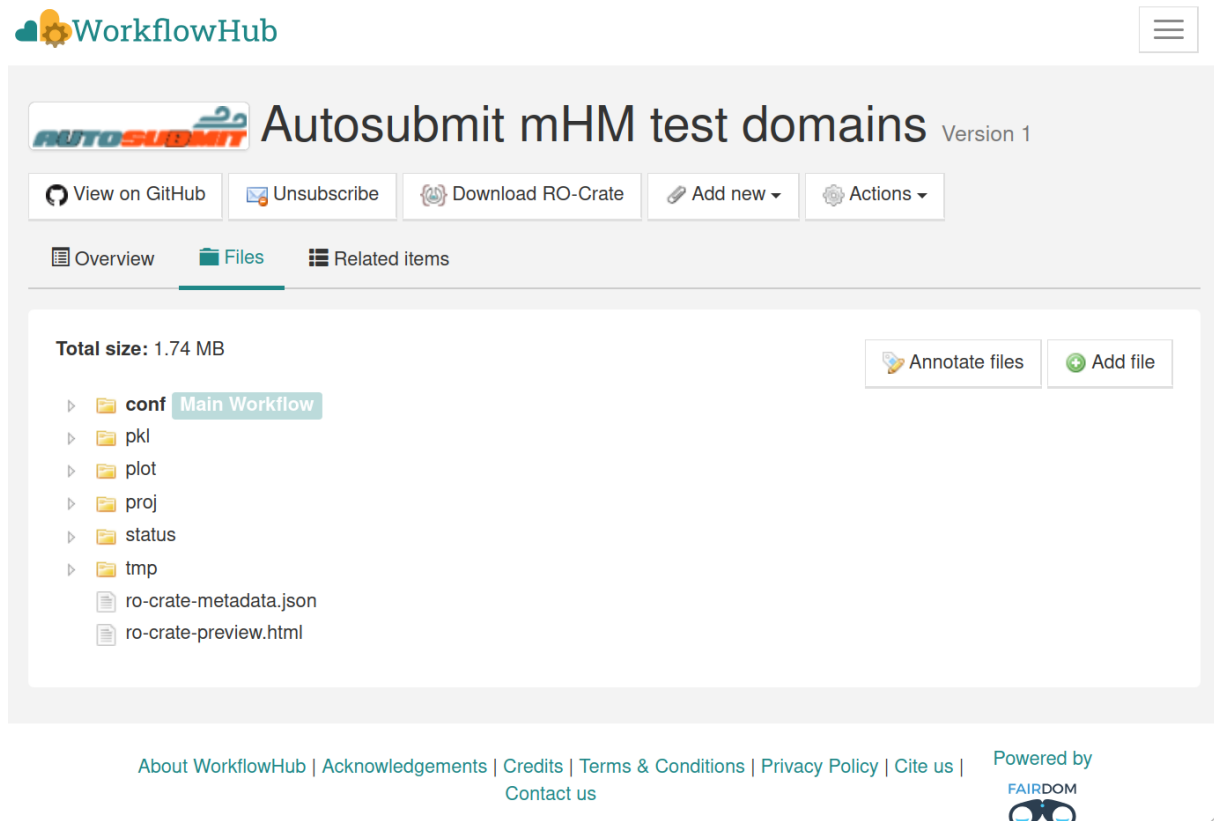


Figure 1. An Autosubmit experiment uploaded to WorkflowHub, showing the RO-Crate archive. The *ro-crate-metadata.json* is the main file in the RO-Crate archive, that describes its contents and metadata.

The `ro-crate-metadata.json` file uses JSON-LD (JSON for Linked Data [14]), a format that encodes linked data - which uses ontologies to structure data. Figure 2 shows a part of a JSON-LD file with information about the workflow structured according to the RO-Crate specification (e.g. license, publisher, workflow configuration, and parts that comprise the workflow). RO-Crate profiles act as a set of conventions of types and properties required or expected to be present in the `ro-crate-metadata.json` file, and the data types and relationships in the RO-Crate provenance graph are modelled through the use of ontologies.

```
    },
    {
      "@id": "proj/git_project/docs/plot.gif"
    }
  ],
  "license": "Apache-2.0",
  "mainEntity": {
    "@id": "conf/metadata/experiment_data.yml"
  },
  "mentions": {
    "@id": "#create-action"
  },
  "publisher": {
    "@id": "https://ror.org/05sd8tv96"
  }
},
{
  "@id": "conf/metadata/experiment_data.yml",
  "@type": [
    "File",
    "SoftwareSourceCode",
    "ComputationalWorkflow"
  ],
  "hasPart": [
    {
      "@id": "https://github.com/kinow/auto-mhm-test-domains.git"
    }
  ],
  "input": [
    {
```

Figure 2. Part of the contents of `ro-crate-metadata.json` of an Autosubmit, showing the workflow license, the main configuration file (“mainEntity”), the publisher, and parts of the workflow such as the Git repository used.

A modeling climate experiment may be composed of one or more models running individually or in coupled mode and may also include steps to handle initial data, prepare a model, or run alongside a model. They often also include post-processing tools operating on the model output or other applications consuming that output to generate a different kind of information. The current RO-Crate provenance information collected by Autosubmit only documents up to the point where the model and the other tools are executed without going into details about what happened inside each step of the experiment. This is due to the fact that climate models and other tools used for the simulation do not provide enough metadata in a structured way, or standardized provenance traces. Without this information, it is difficult to port, adapt, and re-use these simulations as one must first try to retrieve this information from multiple configuration files, model logs, and other traces that may or may not be produced by models and tools.

We are currently investigating the inclusion of other PROV-based [15] provenance records into the RO-Crate archives, to create a provenance graph of a workflow run that starts at the configuration and execution of models, going all the way into what is happening inside the models and tools that are part of the whole experiment, improving the FAIRness of the experiment and of the data it produced. A recent successful step in this direction was the use of METACLIP [16] with SUNSET (SUBseasonal to decadal climate forecast post-processing and asSEssmentT suite) [17]. METACLIP is an ontology for provenance in climate products

based on PROV, which was recently integrated into SUNSET, an R post-processing utility created within the BSC-CES group and used by different groups. In Figure 3, we have a plot generated running SUNSET with the retrospective provenance information utilizing META-CLIP.

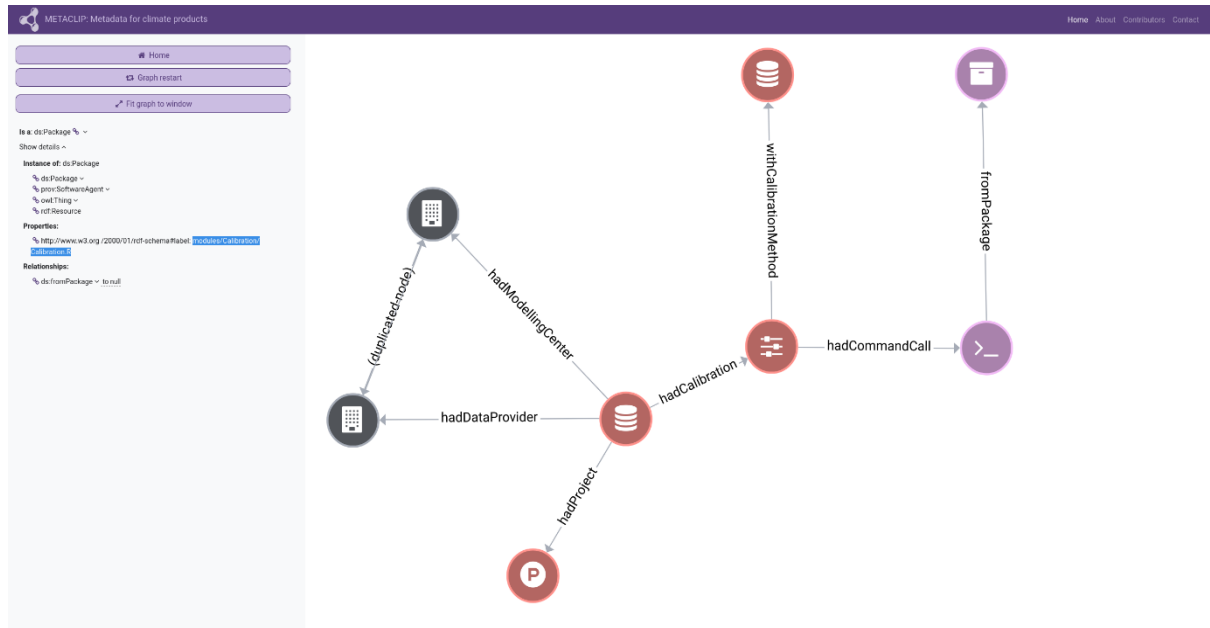


Figure 3. Screenshot of METACLIP Web Interpreter with a graph showing what happened during the execution of SUNSET. The left sidebar is displaying information about the top right icon (as the user clicked on that node) which represents a “package” (*ds:Package*) executed by a command call (“*ds:Command*”). The package represents a script (modules/Calibration/Calibration.R) called by SUNSET. “*ds:*” is a prefix for Dataset, part of the METACLIP ontology. On the left sidebar it is possible to see “*prov:*” too, from PROV.

Following this initial adoption and implementation of open standards related to provenance and FAIR in Autosubmit, the next step is to improve the FAIRness of digital objects from Autosubmit experiments, following the recommendations of the FDO specifications [18]. The goal is to start by addressing issues present in the digital objects produced by Autosubmit, or in Autosubmit itself, using the FDO specifications as a guideline to improve iteratively.

For example, **1)** the PIDs produced by current versions of Autosubmit are not globally unique. They are automatically generated following a consistent format (“a000”, “a001”, “a002”), but PIDs are repeated across different installations of Autosubmit. This means that if two organizations install Autosubmit, both might have experiments with the exact same PID (“a000”, “a001”, etc.) without a formal way to reference and distinguish these PIDs. The PID System from Anders et al. [19] is one of the resources being used as a reference to design a solution for this issue that should implement some of the features described in that proposed note document, like being globally unique, persistent, resolvable, scalable, and robust.

Another example is **2)** the set of attributes associated with an Autosubmit PID. Currently, it includes attributes that are part of the Workflow Run RO-Crate profile (e.g. license, creator, date published) and other information that may be included by using or extending Schema.org. However, given a PID that identifies a climate simulation workflow, it would be useful to know at least what models are involved and information about grid resolution and input data used. To provide better machine-readability and -actionability, we are studying the combination of Signposting [20] with Kernel Attributes and PID Profiles [21][22] to define schemas and attributes that will be for certain types of experiments commonly used in Autosubmit.

The use of standards like RO-Crate, Signposting, and FDO specifications in Autosubmit workflows improves FAIRness in climate and weather experiments and also addresses the requirements and regulations of European projects where Autosubmit is used [23]. Future work includes the two examples above – **1)** and **2)** - and also other points for the digital objects in Autosubmit, such as security, auditing, licensing, versioning, and mutability. The work with METACLIP, described earlier in this text, will be finalized and used to study how that improves the provenance in climate and weather workflows, if that gives operators and developers better insight into what happened during the execution of these programs, and also to archive and reuse these workflows. We hope this work opens the path for other workflow managers to learn from our experience with climate and weather workflows using these standards and specifications, thus improving the interoperability and FAIRness of digital objects produced by this type of experiment.

Data availability statement

The version of Autosubmit that supports RO-Crate, 4.0.100, can be found on Zenodo: <https://zenodo.org/records/10199020>

The WorkflowHub workflow used for Figure 1 and for Figure 2. can be found at this address: <https://workflowhub.eu/workflows/640?version=1>

The METACLIP JSON used for Figure 3 can be found on Zenodo: <https://zenodo.org/records/14042298>

The rest of the software cited in the paper contains references that allow readers to learn more about them, and access the source code or contact their authors.

Author contributions

Bruno de Paula Kinoshita: Conceptualization, Investigation, Software, Writing – original draft

Edgar Garriga Nogales: Writing - review & editing

Manuel Giménez de Castro Marciani: Writing - review & editing

Miguel Castrillo Melguizo: Writing - review & editing, Supervision, Project administration

Competing interests

The authors declare that they have no competing interests.

References

- [1] I. Anders, K. P. Gehlen, and H. Thiemann, "Canonical Workflows in Simulation-based Climate Sciences," *Data Intelligence*, vol. 4, no. 2, pp. 212–225, Apr. 2022, doi: 10.1162/dint_a_00127.
- [2] C. Tebaldi, "The use of the multi-model ensemble in probabilistic climate projections," *Philosophical transactions of the Royal Society of London*, vol. 365, no. 1857, Jun. 2007, doi: 10.1098/rsta.2007.2076.
- [3] D. Manubens-Gil, J. Vegas-Regidor, C. Prodhomme, O. Mula-Valls, and F. J. Doblas-Reyes, "Seamless management of ensemble climate prediction experiments on HPC platforms," in *2016 International Conference on High Performance Computing & Simulation (HPCS)*, 2016, pp. 895–900. doi: 10.1109/HPCSim.2016.7568429.

- [4] A. Bahra, Managing work flows with ecFlow. ECMWF, 2011, pp. 30–32. doi: 10.21957/nr843dob.
- [5] H. Oliver et al., "Workflow Automation for Cycling Systems," *Computing in Science & Engineering*, vol. 21, no. 4, pp. 7–21, 2019, doi: 10.1109/MCSE.2019.2906593.
- [6] F. Massonnet, M. Ménégos, M. Acosta, X. Yepes-Arbós, E. Exarchou, and F. J. Doblas-Reyes, "Replicability of the EC-Earth3 Earth system model under a change in computing environment," *Geoscientific Model Development*, vol. 13, no. 3, pp. 1165–1178, 2020, doi: 10.5194/gmd-13-1165-2020.
- [7] M. Klose et al., "Mineral dust cycle in the Multiscale Online Nonhydrostatic Atmosphere Chemistry model (MONARCH) Version 2.0," *Geoscientific Model Development*, vol. 14, no. 10, pp. 6403–6444, 2021, doi: 10.5194/gmd-14-6403-2021.
- [8] European Centre for Medium-Range Weather Forecasts - ECMWF, "IFS DOCUMENTATION - Cy48r1 Operational implementation 27 June 2023. PART VI: TECHNICAL AND COMPUTATIONAL PROCEDURES.," pp. 40–48, Jun. 2023, Accessed: Jul. 02, 2024. [Online]. Available: <https://www.ecmwf.int/sites/default/files/elibrary/2023/81372-ifs-documentation-cy48r1-part-vi-technical-and-computational-procedures.pdf>
- [9] M. Bush et al., "The second Met Office Unified Model–JULES Regional Atmosphere and Land configuration, RAL2," *Geoscientific Model Development*, vol. 16, no. 6, pp. 1713–1734, 2023, doi: 10.5194/gmd-16-1713-2023.
- [10] R. Santana et al., "Wave forecast investigations on downscaling, source terms, and tides for Aotearoa New Zealand," *Geoscientific Model Development Discussions*, vol. 2024, pp. 1–35, 2024, doi: 10.5194/gmd-2024-110.
- [11] J. Freire, D. Koop, E. Santos, and C. T. Silva, "Provenance for Computational Tasks: A Survey," *Computing in Science & Engineering*, vol. 10, no. 3, pp. 11–21, 2008, doi: 10.1109/MCSE.2008.79.
- [12] S. Leo et al., "Recording provenance of workflow runs with RO-Crate," *PLOS ONE*, vol. 19, no. 9, p. e0309210, Sep. 2024, doi: 10.1371/journal.pone.0309210.
- [13] W. Claus, I. Sharif, B. Daan, A. Ivonne and W. Peter, "FDO Machine Actionability", Nov. 2022, doi: 10.5281/zenodo.7825650.
- [14] M. Sporny, D. Longley, G. Kellogg, M. Lanthaler, P.-A. Champin, and N. Lindström, "JSON-LD 1.1. A JSON-based Serialization for Linked Data.," W3C, Jul. 2020. Available: <https://www.w3.org/TR/json-ld11/>
- [15] L. Moreau et al., PROV-DM: The PROV Data Model. W3C. 2013. Accessed: Jul. 4th, 2024. Available: <https://www.w3.org/TR/2013/REC-prov-dm-20130430/>
- [16] J. Bedia, D. San-Martín, M. Iturbide, S. Herrera, R. Manzananas, and J. M. Gutiérrez, "The METACLIP semantic provenance framework for climate products", *Environmental Modelling & Software*, vol. 119, pp. 445–457, 2019, doi: 10.1016/j.envsoft.2019.07.005.
- [17] N. Pérez-Zanón et al., "SUNSET: SUBseasonal to decadal climate forecast post-processing and assessment suite", presented at EMS Annual Meeting 2024, Barcelona, Spain, Sep. 1–6, 2024, abstract EMS2024-361, doi: 10.5194/ems2024-361.
- [18] FAIR Digital Objects Forum. "Specifications". Accessed: Jul. 2, 2024. Available: <https://fairdo.org/specifications/>
- [19] A. Ivonne et al., "FAIR Digital Object Technical Overview", Apr. 2023, doi: 10.5281/zenodo.7824714.
- [20] H. Van de Sompel, FAIR Digital Objects and FAIR Signposting. Zenodo, 2023. doi: 10.5281/zenodo.7977333.
- [21] A. Ivonne et al., "FDO PID Profiles & Attributes", Oct. 2022, doi: 10.5281/zenodo.7825630.
- [22] B. Christophe, H. Maggie, L. Larry, P. Andreas, S. Ulrich and W. Peter, "Implementation of Attributes, Types, Profiles and Registries", Mar. 2023, doi: 10.5281/zenodo.7825573.
- [23] European Centre for Medium-Range Weather Forecasts - ECMWF. "The Fast Development of DestinE's Climate Change Adaptation Digital Twin." *Destination Earth*, 19 Jul. 2024, <https://destine.ecmwf.int/news/the-fast-development-of-destines-climate-change-adaptation-digital-twin/>. Accessed 3 Jul. 2024.