A Multi-level Approach for Provenance Management and Exploration in Climate Workflows

F. Antonio[1], M. Rampazzo[2], L. Sacco[2], P. Nassisi[1], S. Fiore[2]

1. Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC), Lecce, Italy
2. University of Trento, Trento, Italy

Overview

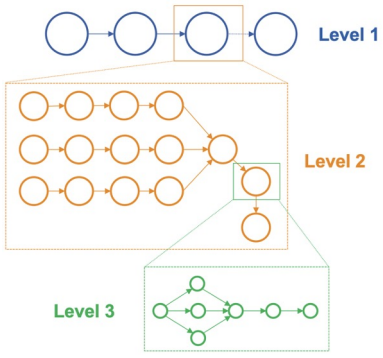
Provenance and reproducibility are key requirements for analytics workflows in Open Science contexts.

Climate research makes use of lots of data coming from the modelling and observational climate communities, so handling provenance at different levels of granularity and during the entire experiment lifecycle becomes key to properly and flexibly managing lineage information related to large-scale experiments as well as enabling reproducibility scenarios. Reproducibility, in turn, fosters re-usability, one of the FAIR guiding data principles.

Multi-level provenance

Climate experiments are rapidly evolving towards complex end-to-end workflows including simulation runs, data pre- and post-processing, data-driven and data intensive applications. This leads to different needs in terms of exploration of specific provenance information:

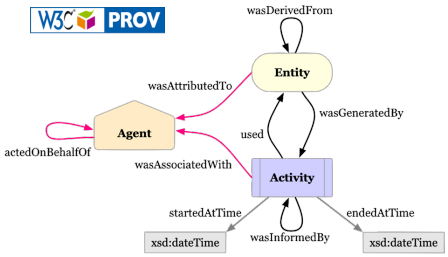
- Coarse-grain, regarding the overall set of tasks in a workflow (macro-view)
- Fine-grain, regarding in-depth details of a specific task (micro-view)



The proposed multi-level approach provides a way to manage provenance information in a structured way and navigate and explore the provenance space across multiple axes: the horizontal one related to the overall set of macro-tasks in the end-to-end workflow, and the vertical dimension in order to drill-down into a specific data-intensive task consisting of hundreds or thousands of data analytics kernels, thus exploring provenance at a finer level of detail.

W3C PROV Standard

The W3C PROV family of standards plays a crucial role in addressing interoperability. It defines a data model, serializations, and definitions to support the interchange of provenance information in heterogeneous environments such as the Web.

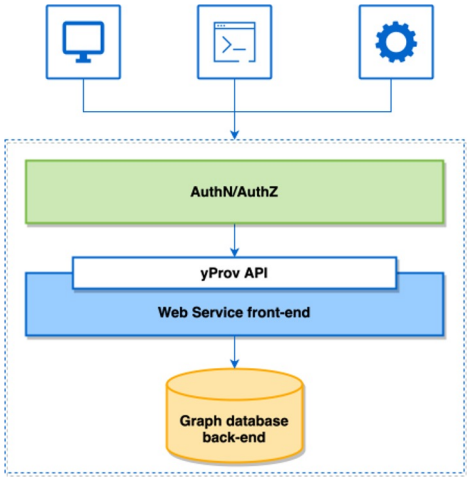


yProv service

yProv is a lightweight and interoperable service for multi-level provenance management within end-to-end scientific workflows. It consists of 3 main components:

- a Web Service front-end;
- a Graph database engine back-end;
- a Command Line Interface.

Authentication and authorization aspects are handled through the JSON Web Token standard. The service also features a RESTful API to perform the needed CRUD (Create, Read, Update, Delete) operations and easily manage provenance information.

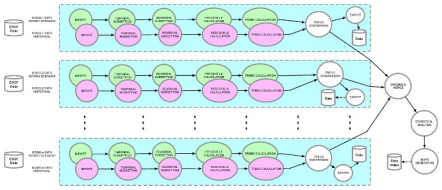


https://github.com/HPCI-Lab/yProv

Multi-model Precipitation Trend Analysis (PTA)

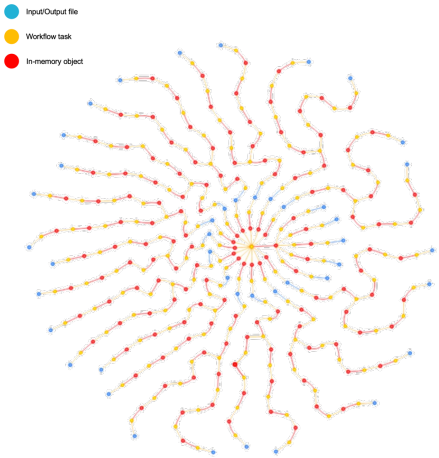
The multi-model Precipitation Trend Analysis is an example of climate models intercomparison data analysis consisting of hundreds of micro-tasks. The workflow includes two main stages:

- a single-model step, which evaluates and compares the precipitation trends for both the historical and future scenario data;
- a multi-model step, performing the ensemble analysis on the trend comparison data and the computation of five statistical indicators.



The PTA use case has been implemented by using the Ophidia High Performance Data Analytics framework, a CMCC research effort on big data analytics in the climate change domain. Inputs are precipitation data from the CMIP (Coupled Model Intercomparison Project) experiments, available through the ESGF (Earth System Grid Federation) data archive.

The PROV-compliant JSON document generated by Ophidia is sent to the yProv service, which adds nodes and edges to the associated provenance graph database running on the back-end.



This work was partially funded by the EU Climateurope2 project (Grant Agreement 101056933) and the EU InterTwin project (Grant Agreement 101058386). Moreover, this work was partially funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4 - Call for tender No. 1031 of 17/06/2022 of Italian Ministry for University and Research funded by the European Union – NextGenerationEU (proj. nr. CN_00000013).