SWIRRL API for provenance-aware and reproducible workspaces. The EPOS and IS-ENES approach.

Alessandro Spinuso, Friedrich Striewski, Ian van der Neut, Mats Veldhuizen, Tor Langeland, Christian Page, and Daniele Bailo

1KNMI, R&D Observation and Data Technology, Utrecht, Netherlands
2NORCE, Norwegian Research Centre, Bergen, Norway
3CERFACS, Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique, Tulouse, France
4INGV, Istituto Nazionale Geofisica e Vulcanologica, Rome, Italy

Modern interactive tools for data analysis and visualisation are designed to expose their functionalities as a service through the web. We present an open source web API (SWIRRL) that allows Science Gateways to easily integrate such tools in their websites and re-purpose them to their users. The API, developed in the context of the ENVRIFair and IS-ENES3 EU projects, deals on behalf of the clients with the underlying complexity of allocating and managing resources within a target container orchestration platform on the cloud. By combining storage and third parties' tools, such as JupyterLab and the Enlighten visualisation software, the API creates dedicated working sessions on-demand. Thanks to the API's staging workflows, SWIRRL sessions can be populated with data of interest collected from external data providers. The system is designed to offer customisation and reproducibility thanks to the recording of provenance, which is performed for each method of the API's affecting the session. This is implemented by combining a PROV-Templates catalogue and a graph database, which are deployed as independent microservices. Notebooks can be customised with new or updated libraries, and the provenance of such changes is then exposed to users via the SWIRRL interactive JupyterLab extension. Here, users can control different types of reproducibility actions. For instance, they can restore the libraries and data used within the notebook in the past, as well as creating snapshots of the running environment. This allows users to share and rebuild full Jupyter workspaces, including raw data and user generated methods. Snapshots are stored to Git as Binder repositories, thereby compatible with mybinder.org. Finally, we will discuss how SWIRRL is and will be adopted by existing portals for Climate analysis (Climate4Impact) and for Solid Earth Science (EPOS), where advanced data discovery capabilities are combined with customisable, recoverable and reproducible workspaces.