

ESA Datalabs: Expanding Space Archives with Innovative Analysis and Collaboration Capabilities

V. Navarro¹, F. Marinic¹, S. del Rio², M. A. Diego³, M. Lopez-Caniego³, A. Anku¹, C. Arviset¹

¹ESA, Madrid, Spain (vicente.navarro@esa.int, filip.marinic@esa.int, anna.anku@esa.int, christophe.arviset@esa.int),

²RHEA for ESA (sara.del.rio@esa.int),

³AURORA for ESA (miguel.angel.diego@esa.int, marcos.lopez.caniego.alcarria@esa.int),

Introduction: Nowadays, the sheer amount of data collected from space-borne and ground-based sensors, is changing dramatically past approaches towards data processing and storage. In the era of Big Data, Machine Learning and seamless collaboration, data transfers are being replaced by algorithm transfers and computation close to the data.

Designed from the ground-up with a focus on end-user productivity, ESA Datalabs accelerates research, unleashing the full potential of ESA archives thanks to a software-as-a-service delivery model. Based on technologies such as, Docker containers, Kubernetes, and JupyterLab, ESA Datalabs provides a catalogue of computing capabilities co-located with ESAC archives.

The Space Archives: At the European Space Astronomy Centre (ESAC) near Madrid, the ESAC Science Data Centre (ESDC) hosts astronomy, planetary and heliophysics archives for ESA Space Science Missions. ESA Datalabs sits next to ESDC archives, expanding their capabilities with analysis and collaboration features connected to their data volumes for missions like Rosetta, BepiColombo ...



Figure 1: Data Volumes Manager

The Analysis: A Datalab is an application or software tool available for execution from the user's web browser through the ESA Datalabs platform. Datalabs can be purpose-built tools or existing external tools like Topcat, Aladin or JupyterLab. Users can search, star, comment, bookmark, and run any Datalab in their catalogue. Datalabs listed in the catalogue can be published by developers from ESA missions. Furthermore, external contributions can be added to the catalogue following a moderation process that includes automatic security scans. Once the user finds a Datalab of interest, a simple play icon allows to launch it immediately. As part of the launch wizard, the user can modify the default configuration of the Datalab set by its creator. At this point, it is possible to select a previous version of the Datalab, increase its computing resources, or adding data volumes from ESAC Archives. Upon

completion, the user is automatically redirected to start working on the selected Datalab. Following this first run, the user can go back to the catalogue to run up to five Datalabs in parallel (default profile configuration). Finally, the user can access the dashboard to switch to another Datalab or destroy it.



Figure 2: Datalab Lifecycle

Furthermore, leveraging on the powerful infrastructure developed for the execution of interactive data analysis applications, ESA Datalabs provides capabilities for the execution of Pipelines in the shape of complex batch data processing workflows. Pipelines represent an extension of the Datalab concept, defined as a set of an input area, a sequence of processing stages in between (Pipeline Steps), and an output area.

The Collaboration: The first collaboration level derives from the possibility to launch computing elements mounting ESA Archives, or even SPICE kernels data volumes. Along these lines, ESA Datalabs catalogue includes JupyterLabs, which feature notebooks shared with the community to show how to explore and analyse data in the archives. Planetary Science Archive users are expected to benefit soon from a set of JupyterLab based Datalabs for ESA planetary missions. More advanced collaboration levels in ESA Datalabs derive from the team data volumes, designed for groups of users sharing files and analysis tools.

Status and future work: At present, ESA Datalabs is available at datalabs.esa.int as a Private Beta. Implementation of Datalabs for multiple ESA missions such as Juice or JWST, has demonstrated its suitability to support cross-domain use cases. Researchers have fast access to pre-configured complex applications integrating ESA archives. ESA Datalabs next steps gear towards integration of parallel activities in the areas of GPU computing, machine learning pipelines and an integrated development environment for Datalabs.

References:

[1] V. Navarro et al (2021), “ESA Datalabs 2.0: Boosting Open Collaboration in ESA Space Science” ADASS 2021.

[2] C. Arviset et al (2021), “ESA Datalabs: Towards a collaborative e-science platform for ESA” BiDS 2019.

Acknowledgments: The authors would like to thank the ESA Datalabs User’s Group for their ideas and contributions. We thank all team members from the industrial consortium, in particular Edisoft and Uninova in Portugal, CGI in Estonia, ACRI-ST in France and AdwäisEO in Luxembourg. Finally, we also thank the Science and Operations technical IT Unit at ESAC, the members of ESA Space Science missions and the ESAC Science Data Centre who are building Datalabs to be deployed on the platform.