# VESPA-Cloud

EOSC-hub Early Adopter Program 2d Call
https://www.eosc-hub.eu/eosc-early-adopter-programme-2nd-call

## Team

Principal investigator:

- Baptiste Cecconi, LESIA, Observatoire de Paris, CNRS, PSL, France baptiste.cecconi@obspm.fr

Collaborators:

- Pierre Le Sidaner, DIO, Observatoire de Paris, CNRS, PSL, France
- Stéphane Erard, LESIA, Observatoire de Paris, CNRS, PSL, France
- Angelo Pio Rossi, Jacobs Uni, Bremen, Germany
- Markus Demleitner, Heidelberg Uni, Germany
- Marco Molinaro, OATF-INAF, Trieste, Italy
- Albert Shih, DIO, Observatoire de Paris, CNRS, PSL, France
- Cyril Chauvin, DIO, Observatoire de Paris, CNRS, PSL, France
- Nicolas André, IRAP, Université de Toulouse, CNRS, France

## Project Description

VESPA (Virtual European Solar and Planetary Access) is a network of interoperable data services covering all fields of Solar System Sciences. It is a mature project, developed within EUROPLANET-FP7 and EUROPLANET-2020-RI. The latter ended in Aug. 2019. It will be further supported under the EUROPLANET-2024-RI project (starting in Feb. 2020).

The VESPA data providers are using a standard API (based on the *Table Access Protocol* of IVOA (International Virtual Observatory Alliance) and *EPNcore*, a common dictionary of metadata developed by the VESPA team). The VESPA services consist in searchable metadata tables, with links (URLs) to science data products (files, web-services...). The VESPA metadata includes relevant keywords for scientific data discovery, such as data coverage (temporal, spectral, spatial…), data content (physical parameters, processing level…), data origin (observatory, instrument, publisher…) or data access (format, URL, size…). VESPA hence provides a unified data discovery service for Solar System Sciences.

The architecture of the VESPA network is distributed (the metadata tables are hosted and maintained by the VESPA providers), but it is not redundant. The hosting and maintenance of VESPA provider's servers has proved to be a single point failure for small teams with little IT support. The VESPA-Cloud project with EOSC-Hub would greatly facilitate the sustainability of data sharing from small teams as well as teams, whose institutions have restrictive firewall policies (like labs hosted by space agencies, e.g., DLR in Germany). Each VESPA data provider is using the same server software, namely DaCHS (Data Centre Helper Suite), developed by the Heidelberg team included in the project.

VESPA-Cloud will provide a cloud-hosted facility to host VESPA compliant metadata tables in a controlled and maintained software environment. The VESPA providers will focus on the

science application configuration, whereas the VESPA core team will support them with the maintenance of the deployed instances. The development of the VESPA provider's data service will be done using a git versioning system (github or institute gitlab).

An instance of the VESPA query interface portal will also be implemented on an EOSC-hub provided virtual machine.

In the course of the VESPA-Cloud project, we will implement in the DaCHS framework cloud-storage API connectors (such as Amazon S3, iRODS, etc.) to reading data in the cloud and ingesting metadata. Since DacHS is used worldwide by many datacenters to share astronomical and solar system data collections, many teams will benefit from this development.

We also propose to setup a ***Europlanet*** Research Community, which will include VESPA-Cloud, and other *Europlanet-2024-RI* and *Europlanet-Society* related projects (such as SPIDER – *Sun Planet Interaction Digital Environment on Request*, or the previous services developed within PSWS – *Planetary Space Weather Services*).

Further development plans for VESPA-Cloud are listed below:

- **New VESPA portal architecture.**
  a new VESPA portal architecture based on Lucene-like technologies, will be developed in the frame of the upcoming EUROPLANET-2024-RI project. This would greatly enhance the portal search interface, especially for complex queries dealing with several services, where SQL-like queries are difficult set up and to generalize. This would also allow VESPA to be interoperable with NASA/PDS4 (Planetary Data Archive) Search Engine.
  As the development didn't started yet, we don't have quantitative elements for sizing our needs. The architecture of the search portal will be split into 3 elements:
  - a data ingestion server, which will harvest VESPA provider's servers (on VESPA-Cloud and on the classical VESPA network) regularly and update the Elastic database;
  - an Elastic nodes cluster (possibly using "Elastic Cloud Compute Cluster") with the VESPA network data
  - a front-end web query portal with the user interface querying the Elastic cluster.

  This new architecture is required with the growing number of services and data products served by the VESPA providers.
- **JupyterHub.**
  Access to VESPA data services through community based python scripts (astropy, pyvo...) with a JupyterHub facility (with "EGI Notebooks" applications). At the moment, we distribute jupyter notebook tutorials, which should be run locally by users on their own machine.
- **Run-on-demand.**
  On-demand computing services (models, cutouts, resampling...), using UWS (Universal Worker Service, an IVOA standard) as a job submission manager.
  The VESPA providers can serve data products as well as data services (like cutout or resampling services on both data or simulation runs), or even direct calls to

numerical codes through REST interfaces. The implementation of a UWS application, based on OPUS (Observatoire de Paris UWS System: https://uws-server.readthedocs.io/en/latest/) will enhance the overall interoperability between the VESPA network (and other IVOA based frameworks) with the EOSC resources.

- **Federated Authentication**
  User and Group management using federated authentication is not yet implemented in the VESPA network. Such capabilities will allow team to work with the VESPA infrastructure before their data is publicly released. This leads to a wider adoption of the VESPA network in the community as well as provision of better services, since the provider's will also be users of the data services.

## Description of the services and the technical environment that you have already in place

VESPA is a mature project, with 50 VESPA providers distributing open access datasets throughout the world (EU, Japan, USA). In October 2019, the current number of data products available within the VESPA network reaches 18.3 millions (among which 5 millions products from the ESA Planetary Science Archive).

Each VESPA provider is hosting and maintaining a server (physical or virtualized) with the same software distribution (DaCHS, Data Centre Helper Suite), which implements the interoperability layers (from IVOA and VESPA) and following FAIR principles. Each server hosts a table of standardized metadata with URLs to data files or data services. Data files can be hosted by the VESPA provider team, or in an external archive (e.g., ESA/PSA - Planetary Science Archive).

The VESPA query interface portal is developed and maintained at the Observatoire de Paris (Paris, France).

VESPA:

- http://www.europlanet-vespa.eu (project)
- http://vespa.obspm.fr (query portal)
- https://voparis-wiki.obspm.fr (wiki and documentation)

    - VESPA/EPNcore metadata dictionary: https://voparis-wiki.obspm.fr/display/VES/EPNcore+v2
    - Tutorials for implementing VESPA services: https://voparis-wiki.obspm.fr/display/VES/Implementing+a+VESPA+service

DaCHS:

- https://dachs-doc.readthedocs.io/

A preliminary prototype of a DaCHS instance on the EOSC-hub infrastructure has been tested earlier in 2019. This instance has been ordered through the EOSC-hub marketplace (https://marketplace.eosc-portal.eu), using *EGI Cloud container compute BETA*. This has been running for several weeks, with success. We could install and run the DaCHS framework, as well as serve a VESPA metadata table. This instance is now in *undeployed* status.

## Description of the services and resources that you need and expected benefits

The VESPA architecture relies on the assumption that data provider's servers are up and running continuously. The VESPA network is distributed but not redundant. For small teams with little or no IT support is available locally, the services are down regularly. We thus need a more stable and manageable platform for hosting those services. The EOSC-hub "cloud container compute" service would solve this problem.

We propose to use the EOSC infrastructure to host VESPA provider's servers (through a controlled deployment environment with git-managed containers).

The VESPA providers would be able to:

- order a VM with all the server framework installed,
- configure the server for their science application,
- co-administrate the server packages with the VESPA team,
- update the content and the tables.

In a second phase, we will implement an EOSC-hub hosted VESPA portal, using the web interface developed at Observatoire de Paris. The portal will be deployed from a git-based repository.

## Science Area

1.3 Physical sciences (Astronomy)

## Expected scientific impact of using services, data and other research outputs from EOSC-hub and its partners

VESPA is a distributed (although not redundant) data discovery and access framework. Hosting VESPA services in the cloud ensures their availability on the long term, and in turn the reliability of the full VESPA network. Starting in 2020, we propose to use the VESPA-Cloud infrastructure for the new VESPA providers selected after the yearly VESPA implementation workshop AO. This would add at 3 to 5 new services in 2020 from external teams. Science teams with the VESPA project will also use the VESPA-Cloud facility. We thus estimate to open about 10 services during year 2020.

The current VESPA network connects 50 data services, serving about 1.8M data products, with an average monthly visitor count of 100. With an enhanced visibility of the VESPA network through VESPA-Cloud, we expected to see a wider adoption.

VESPA-Cloud is a proof of concept, which will demonstrate the use and the efficiency of the EOSC-Hub infrastructure to the Solar System science community. Furthermore, in the course of the Europlanet-2024-RI project, the VESPA team will build strong interfaces with the planetary surfaces team building on GIS technologies (GMAP work package), as well as Space Weather (SPIDER work package) and Machine Learning (ML work package). This opens doors for reaching out similar communities focusing on Earth sciences. The VESPA-Cloud team has also existing contacts with the several ESCAPE work packages (e.g., on the interoperability, radio-astronomy, or solar physics topics).

## Contribution to Open Access and FAIR

The goal of VESPA is to make data Solar System Findable and Accessible through interoperable interfaces, and is recommending standard data and metadata formats, ensuring reusability.

All software developed for VESPA are open-source (mostly GPLv3, or Apache).

VESPA-Cloud enhances the accessibility, by providing a sustainable access for VESPA dataset

## Expected duration (from 6 to 12 months)

12 months for setting up services and finding a sustainable approach for further operations.

## Minimal Compute and Storage capacity needed for sustaining the Project

- 20 VM instances
- 2 CPU per VM,
- 4GB RAM per VM,
- 20 GB disk per VM,
- 1 fixed IP per VM
- 5 remote ssh-key access per VM
- deployment of containers from git-managed repository
- 10 TB of storage accessible from every VM

The minimal individual VESPA-Cloud provider's instance is: 2 CPU, 4GB RAM, 20 GB disk. We estimate to have about 20 instances to deploy in the first stage. Each instance must have a fixed and public IP address (customizable DNS names preferred). The instances are expected to be up and running all the time. Short unavailability of the services is acceptable, if the instance can be relaunched automatically.

Another need is a Storage Buffer for data ingestion. This is a global and temporary storage volume, which can be mounted on any VESPA-Cloud provider's instance for metadata extraction and ingestion. Data are pushed onto this volume for initial metadata extraction and ingestion and removed after this task is finished. This storage size should be 10TB. For providers needing a permanent storage capability, we will investigate with EUDAT, Zenodo or other EOSC partners.

For the current version of the VESPA portal, the compute and storage specifications are the same as for the individual VESPA-Cloud provider's instances.

For future prospects, as listed in the project description section, it is too early to propose a sizing of needs. We will work in parallel of the VESPA-Cloud project, with the Europlanet-2024-RI work program on this infrastructure sizing.

## Compute and Storage capacity to fully scale-up the Project after the completion of the pilot

With the new Europlanet-2024-RI program, we can expect at least 5 new providers per year, each with the same compute and storage needs.

Among the further developments foreseen after the completion of the pilot program:

- Next level data portal (lucent-like).
- JupyterHub access.
- On-demand computing services (models, cutouts, resampling...)

However, they are not yet not completely defined and sized, since this is part of the work to be accomplished in Europlanet-2024-RI.

## Minimal storage capacity for long-term archiving for sustaining the Project

Not fully applicable, see below.

## Long-term data management policies and long-term archiving capacity required by the Project

The VESPA-Cloud services are hosting metadata tables of data products hosted on independent facilities. Each metadata table can be rebuilt from a resource descriptor script maintained with a git versioning system. The archiving of the computed metadata tables is not planned yet, but any discussion on this specific point is welcome. The archiving of the resource descriptor script should also be discussed, but it may contain private data (such as logins and tokens), which shall not be openly available. The provider's teams are currently in charge of their own data preservation, outside of the VESPA or VESPA-Cloud projects.

However, the VESPA-Cloud provider’s instances are expected to run as long as the data providers are sharing their data. A long-term sustainability plan has to be prepared during the Pilot program, together with the new-born Europlanet Society (https://www.europlanet-society.org).

## Mention any classified and/or privacy-sensitive data

No sensitive data