SYLVA Training Package 2: Accessing the SYLVA Data Portal and IT infrastructure

Contents

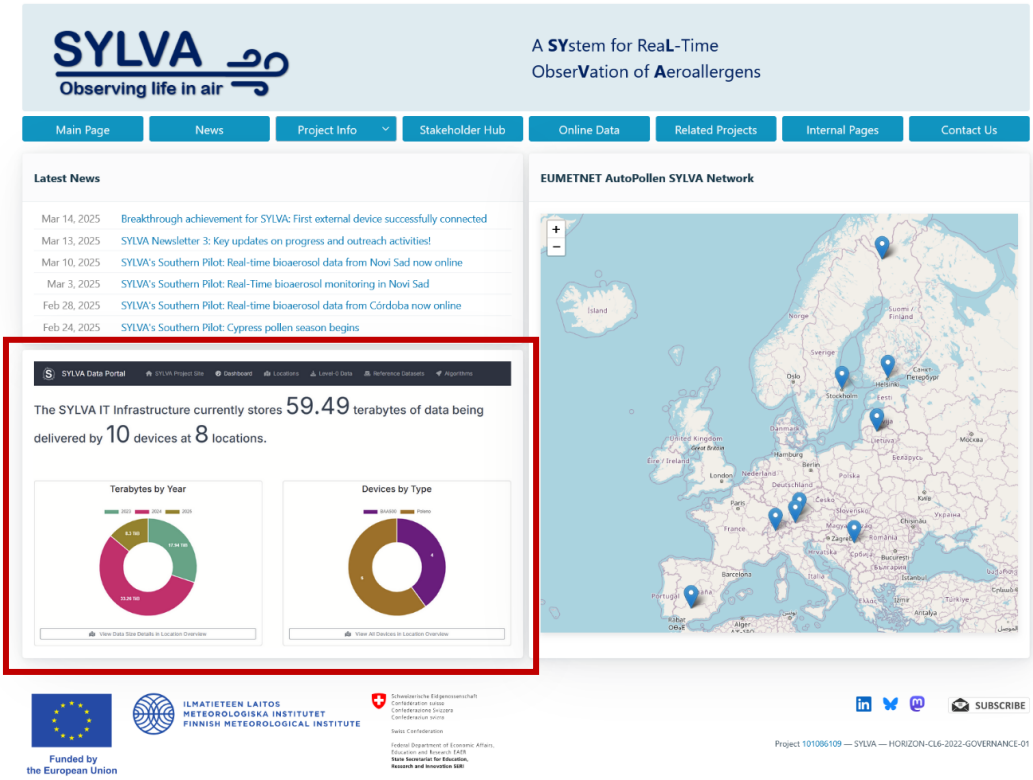
Introduction

The second stakeholder training package focuses on the SYLVA data portal and the data it contains. Meant for technical users of SYLVA technologies and infrastructure, the training package explains in detail how to access the various datasets, how they can be used, as well as how end users can make use of the SYLVA infrastructure themselves. The aim of the training package is to provide as much useful information as possible in a concise and clear way. To facilitate this, images have been used to supplement the text descriptions provided.

To note, level-0 or raw data refer to the measurements directly made by the instrument. These data need to be analysed by a particle classification algorithm to obtain airborne concentrations (level-1 data).

The SYLVA Data Portal

Go to the SYLVA website https://sylva.bioaerosol.eu and from there to get to the SYLVA data portal, click on the link of the widget of the SYLVA data portal – highlighted by the red box in the screenshot below.



The entry page, or Dashboard, of the data portal shows you how much level-0 data are available per year (on the left) and for how many devices of different types there are data (on the right).

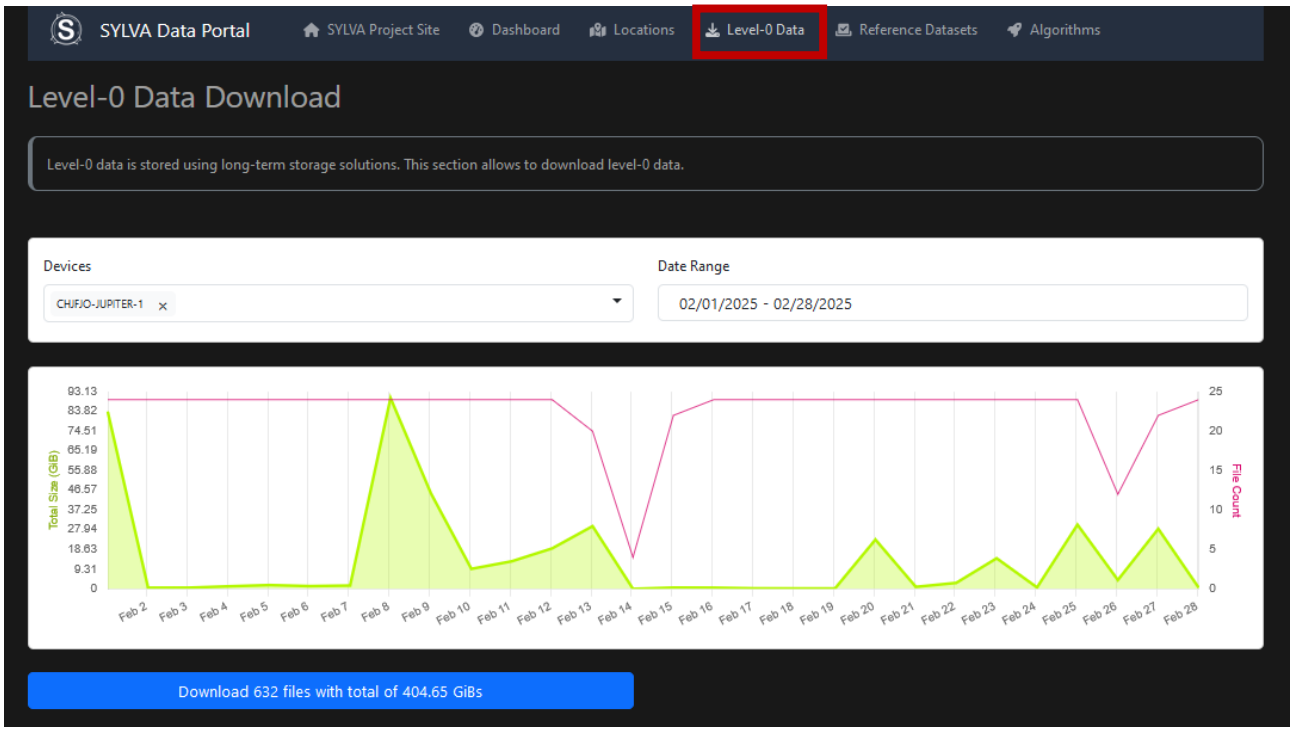
This brings you directly to the “locations” page of the SYLVA data portal, which provides information about each of the sites reporting data to the SYLVA infrastructure. For each site, the following information is provided:

Level-1 concentration data

To access the level-1 data (airborne concentrations of different pollen and fungal spores) for each site click on the link “Open level-1 data in EBAS Database” (see red box above). This brings you to the EBAS website where one needs to search for SYLVA_NRT in the “Framework” column to obtain the list SYLVA stations reporting data. From there, one can obtain the level-1 data by clicking on the stations individually.

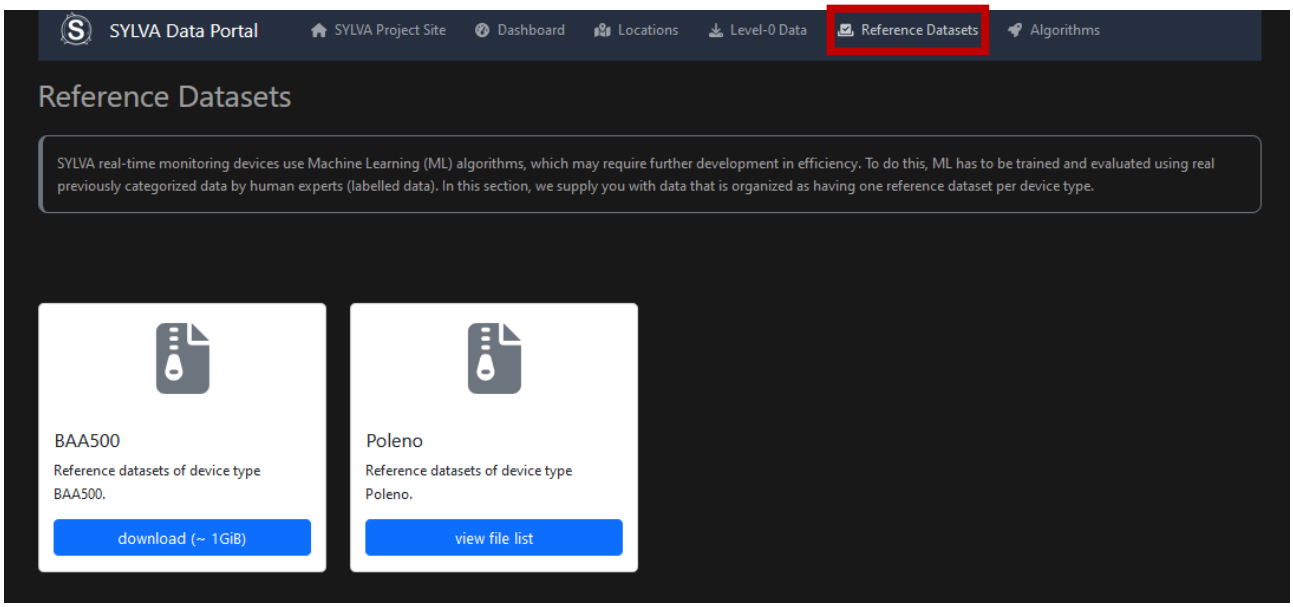
Level-0 raw data

The level-0 raw data can be obtained by clicking on the “Level-0 Data” tab (see red box in the figure below). From this page one can select the device from the drop-down menu on the left (the example below shows the Swisens Poleno Jupiter at the Swiss Jungfraujoch site – CHJFO-Jupiter-1) as well as the period for which one would like data from the drop-down calendar on the right (in this example from 1-28 February 2025). To download the data, click on the blue button at the bottom of the page once you have selected a device and time period from the drop-down calendar for which you would like data. The number of files as well as the accumulative size is indicated on the blue button. Note that the raw data are in the format provided directly by each instrument and therefore differ from one manufacturer to another. In the case of the SYLVA data portal, data are available from the Swisens Poleno device (Jupiter model) and the Hund Wetzlar BAA500 device.



Training datasets

Datasets for developing classification algorithms are available from both the Hund Wetzlar BAA500 as well as the Swisens Poleno Jupiter. These training datasets can be found under the tab “Reference datasets” (see the red block in the image below).

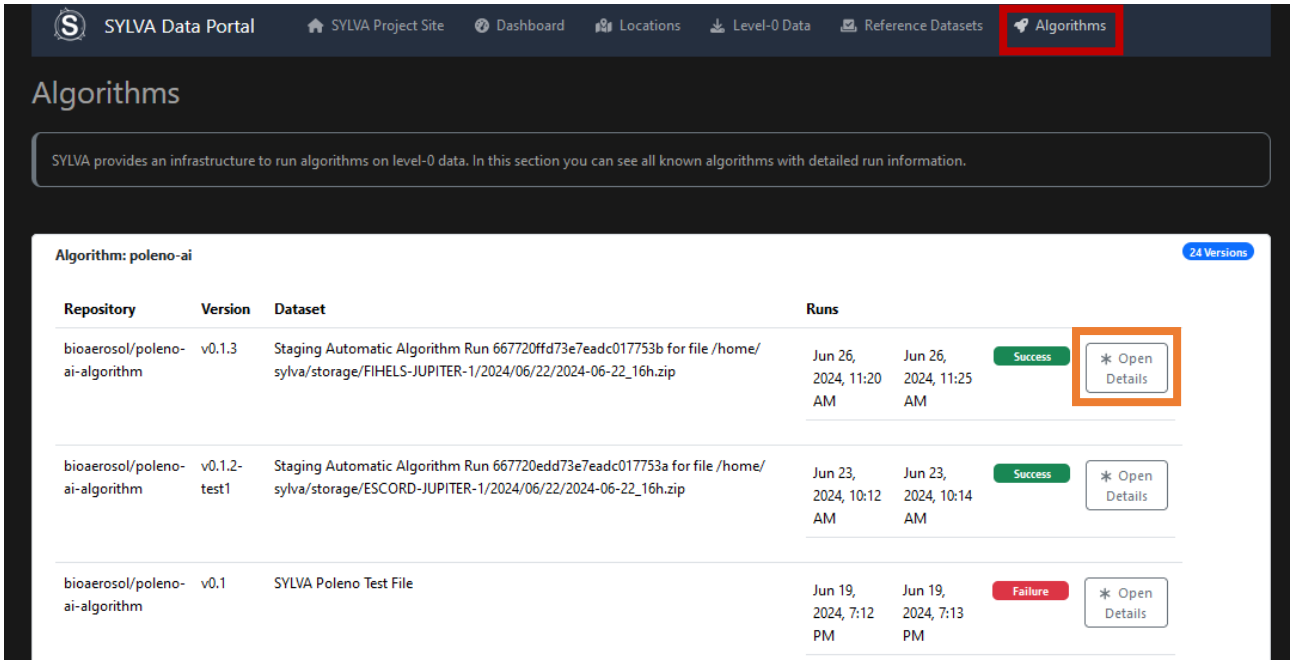


By clicking on the white block for the BAA500 the download of all BAA500 training datasets start directly. The downloaded zip file includes a meta data which describes the content of dataset. Clicking on the white block for the Poleno beings one to a separate page listing all the datasets available for this device. Note that one training dataset in .zip format is available for each plant/fungal spore taxa. At the bottom of the page an excel file (take the one with the most recent

date) provides extensive metadata about which file corresponds with which taxa, where the pollen/fungal spores were collected, etc.

The SYLVA algorithm test space

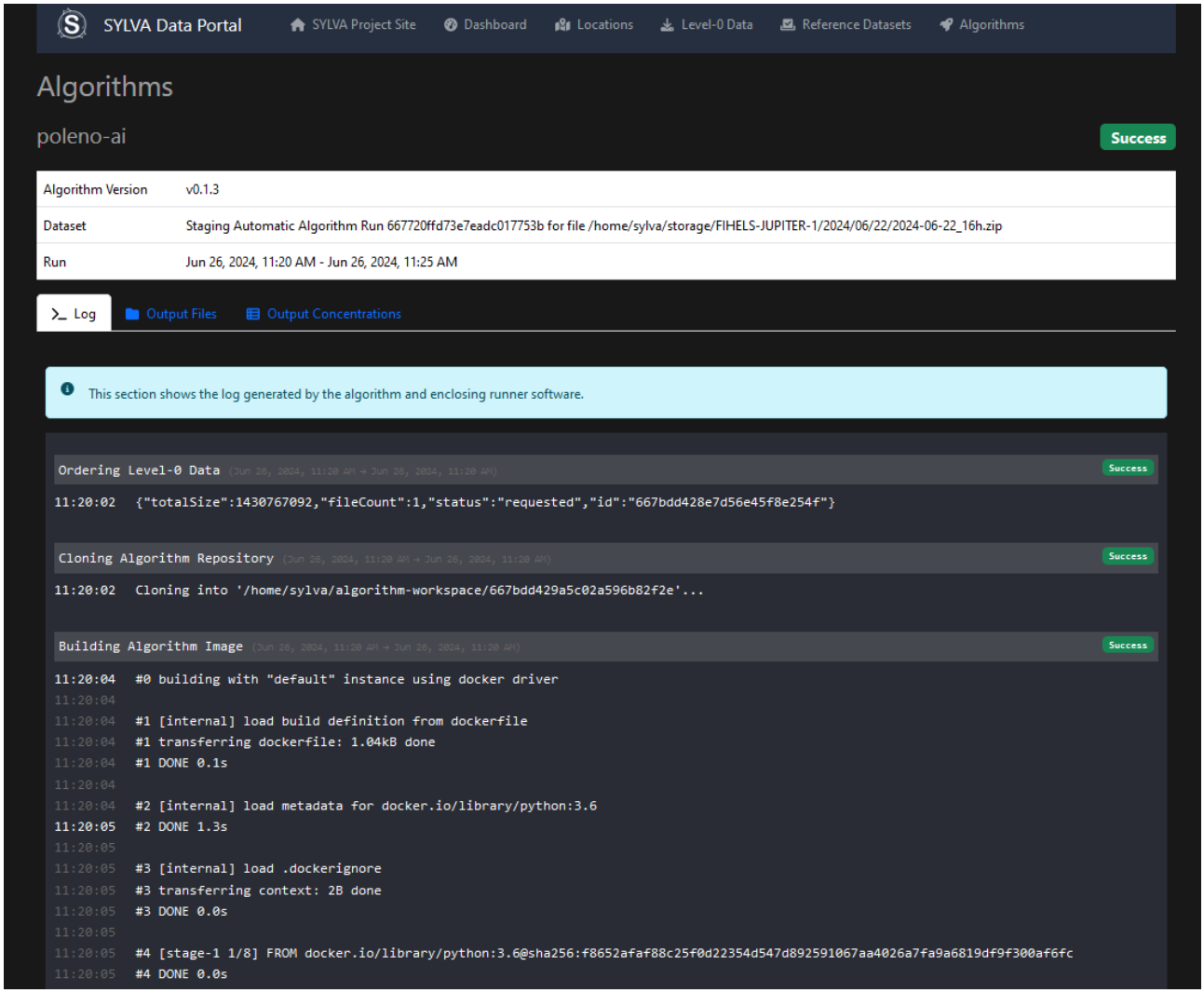
The final “Algorithms” tab on the data portal (see red box in figure below) brings you to a page where the results of different algorithm tests are shown. This is a unique aspect of the SYLVA data portal and will be extremely useful to algorithm developers aiming to evaluate their algorithm against a reference dataset to find out if algorithm performance is improved.



Algorithms need to be trained locally, that is, they cannot be trained using the SYLVA infrastructure directly. Once an algorithm has been developed, it can be tested against the SYLVA reference dataset which has been created particularly for the purposes of testing pollen and fungal spore classification algorithms from across Europe.

The testing process is carried out through the SYLVA algorithm runner which can be found on GitHub (https://github.com/bioaerosol/sylva-algorithm-runner). It entails uploading your model to a GitHub repository, uploading the algorithm and the environment required (which needs a docker and a container) to the SYLVA ftp, and then running the algorithm on the SYLVA reference dataset. Thereafter, the results from the test are directly made visible on the SYLVA Data Portal algorithm tab. By clicking on the “Open Details” button (orange block in the figure above) one can see the output of the algorithm test to check where there might have been problems or if the test was a success (see an example below).

It is important to note that algorithm developers need to contact the SYLVA IT infrastructure team (Samer Alashhab (s.alashhab@tum.de) or Robert Gebauer (robert.gebauer@gmx.de)) to create a Github account that is required to upload the developed algorithm to the SYLVA testing space. Either of the experts can assist in the process of testing the algorithm, if needed.



Conclusions

The second SYLVA training package focuses on how to use the SYLVA data portal. This provides a useful tool to all end users who would like to access any of the data or services available through the portal. This includes raw (level-0) data from 8 devices at 6 different sites, training datasets for developing particle classification algorithms, as well as the possibility to test algorithms against a reference dataset once developed.