iCUPE data pilots, data and services (ver 2)

Steffen M. Noe (Estonian University of Life Sciences), Alexander Mahura (University of Helsinki), Roberto Salzano (Consiglio Nazionale delle Ricerche)

Tartu, Estonia, September 2019

WP5: Data provision, interoperability and facilitation of data and services

Task5.5: Facilitating iCUPE data pilots, data and services towards ERA-PLANET community, GEO and Copernicus / Deliverable 5.5.2 iCUPE data pilots, data and services (ver 2)

Version 2

Introduction

Following the FAIR data principle, the data pilot needs to provide

- an up-to-date Data Management Plan (DMP)
- data deposited into free accessible research data repositories
- ensure that third parties can freely access, mine, exploit, reproduce and disseminate the data
- provide the related information and identify (or provide) the tools needed to use the raw data to validate your research

The pilot should then apply to

- the data (and metadata) needed to validate results in scientific publications
- other curated and/or raw data (and metadata) that are specified in the DMP

The iCUPE Data Management Plan

The **iCUPE Data Management Plan (DMP)** has been delivered so far in version 1 in November 2017, with an appendix and in version 2 in November 2018. Version 3 has been delivered in NN 2020?

The data repositories

iCUPE datasets using different data repositories. Simple datasets, mostly those in a form of a tabular data are depending on the size of the data file available directly from the **iCUPE website** (https://www.atm.helsinki.fi/icupe/index.php/datasets/delivered-datasets). On this website, all datasets can be long-term freely/ openly accessed and by end of the project, those which are currently available on request will be made openly accessible as well.

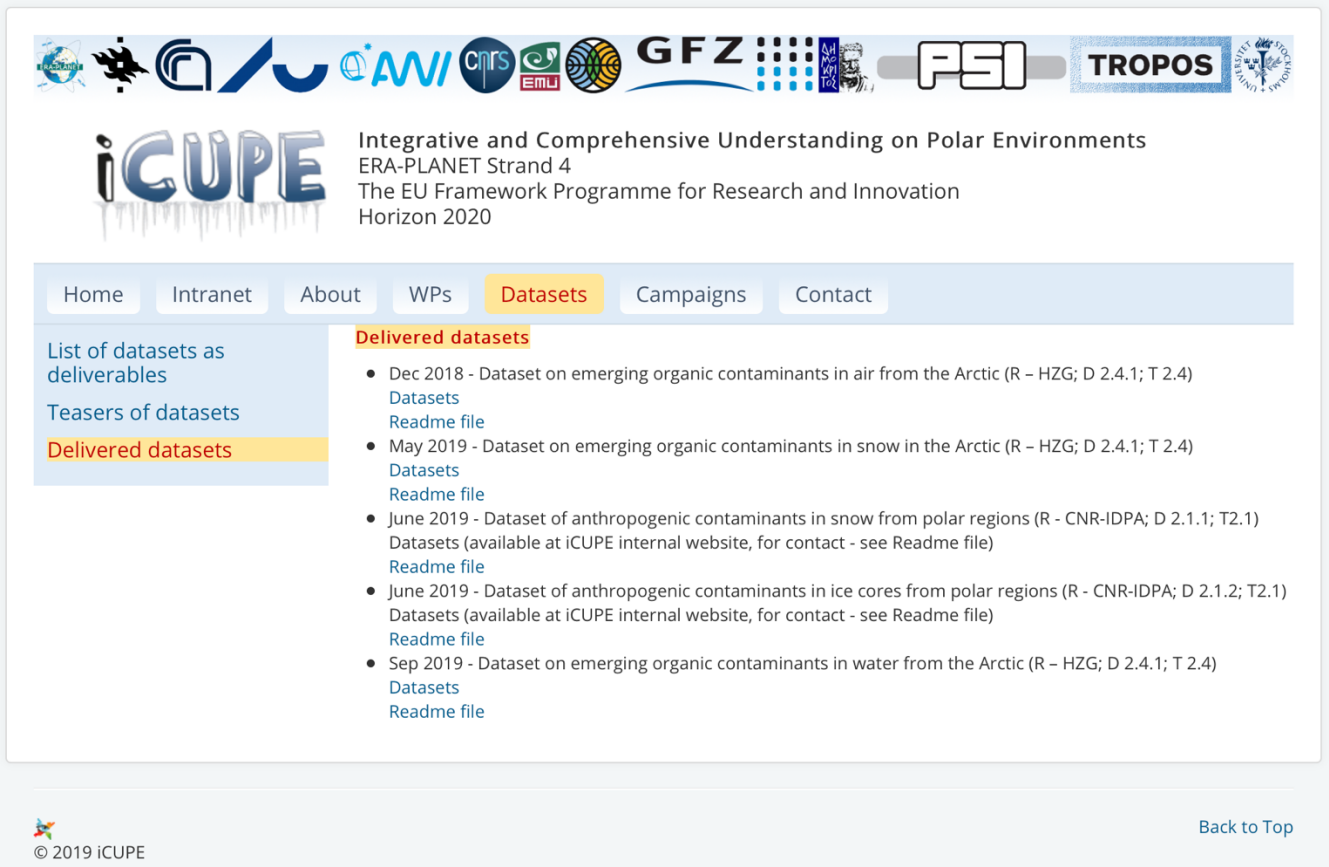

Figure 1. iCUPE data sets accessible via direct links and metadata information in linked readme files

Beside simple data sets, iCUPE data also includes data that are stored in other repositories. Examples of such repositories are the EBAS (http://ebas.nilu.no) database that is hosting data related to EMEP[1],

[1] European Monitoring and Evaluation Programme, https://www.emep.int

NILU[2], ACTRIS[3], WMO[4] activities and the LitDB (http://litdb.fmi.fi) database hosted by the Finnish Meteorological Institute (FMI) and includes ground-based in-situ data and satellite data products. Different datasets use as well the open data repository Zenodo (https://zenodo.org).

Access to iCUPE data

The access to iCUPE data is free. In terms of "simple" data, they can be directly downloaded from the iCUPE website via links provided there. We refer to simple data in the case the data have a simple tabular structure and their size is feasible to allow manual or automated download from a website or repository. All promised datasets have been delivered and made available on the project website. Each dataset on the website has added a "Readme" file that contains information on the dataset, i.e. data producer, address and contact details, data format, geographical location of measurements that contributed to the data set.

Data sets that have restrictions by the data producer are made available by personal contact to the data set producer/owner. This information is freely provided by the Readme files on the iCUPE website.

List of delivered datasets on the iCUPE website

- Dec 2018 - Dataset on emerging organic contaminants in air from the Arctic (R – HZG; D 2.4.1; T 2.4)
- May 2019 - Dataset on emerging organic contaminants in snow in the Arctic (R – HZG; D 2.4.1; T 2.4)
- June 2019 - Dataset of anthropogenic contaminants in snow from polar regions (R - CNR-IDPA; D 2.1.1; T2.1)
  Datasets (available at iCUPE internal website, for contact - see Readme file)
- June 2019 - Dataset of anthropogenic contaminants in ice cores from polar regions (R - CNR-IDPA; D 2.1.2; T2.1)
  Datasets (available at iCUPE internal website, for contact - see Readme file)
- Sep 2019 - Dataset on emerging organic contaminants in water from the Arctic (R – HZG; D 2.4.1; T 2.4)
- Sep 2019 - Pilot dataset on Near-Real Time aerosol absorption measurements from Zeppelin Station, Ny Ålesund, Svalbard (R – NSCR; D 1.2.1; T 1.2)

[2] Norwegian Institute for Air Research (Norsk institutt for luftforskning), https://www.nilu.no/en/
[3] European Research Infrastructure for the observation of Aerosol, Clouds and Trace Gases, https://www.actris.eu
[4] World Meteorological Organization, https://public.wmo.int/en

- December 2019 - Dataset on Arctic atmospheric Hg(II) observations (R – CNRS; from D 2.2.1; T 2.2)
  Datasets (available at iCUPE internal website, for contact - see Readme file)

- January 2020 - Dataset on Long-term monitoring of gaseous elementary mercury in background air at the polar station Amderma, Russian Arctic (R - INEP KSC RAS; from iCUPE collaborators)

- March 2020 - Dataset on classification of artificial light sources in the Yamal Peninsula, Western Siberia (R - GFZ; from D 3.2.2; T 3.2)

- March 2020 - Dataset on fractional snow cover area in selected sites of Svalbard islands (Norway) (R - CNR; from D 3.2.2; T 3.2)
  Datasets (available at iCUPE internal website, for contact - see Readme file)

- March 2020 - Dataset on small-scale vertical and horizontal variability of the atmospheric boundary layer aerosol using unmanned aerial systems (R - TROPOS; from D 1.3.2; T 1.3)
  Datasets (available at iCUPE internal website, for contact - see Readme file)

- April 2020 - Dataset on time series of lake size changes in Northeast Greenland (R – AWI; from D 3.3.2; T 3.3)

- May 2020 - Datasets of validated aerosol vertical profiles from ground-based and satellite observations above selected sites in Finland and Siberia (R – CNRS & UHEL; from D 4.3.1; T 4.3)
  Datasets for Russia (available from designated websites - registration, login, password are required; for contact - see Readme file)

- May 2020 - Dataset on Visible Near Infrared airborne and simulated EnMAP satellite hyperspectral imagery of Toolik Lake, Alaska (R - GFZ; from D 3.2.2; T 3.2)

- July 2020 - DS on aerosol physical and optical characteristics including equivalent black carbon at Ny-Alesund, Svalbard (R - CNR; from D1.1.2; T1.1)
  Datasets:
  - DS1 on aerosol ultrafine particle size distribution
  - DS2 on aerosol large particle size distribution
  - DS3 on scattering, absorption and equivalent black carbon

- July 2020 - DS on snow spectral reflectance measurements at Ny-Alesund, Svalbard (R - CNR; from D1.1.3; T1.1)

- July 2020 - DS on vertical profiles of equivalent black carbon in the Arctic boundary layer at Ny-Ålesund, Svalbard (R - CNR; from D1.3.2; T1.3)

- September 2020 - Dataset for ground-validation of precipitation measurements in high-latitudes (R – UHEL; from D 4.2.1; T 4.2)

- September 2020 - Dataset on Arctic atmospheric Hg(0) isotope observations (R – CNRS; from D 2.2.3; T 2.2)

- October 2020 - Dataset on organic aerosols in the Arctic (R – PSI; from D 1.4.2; T 1.4)

- December 2020 - A blueprint for novel proxy variables integrating in-situ and satellite remote sensing data with an exemplary dataset (R – UHEL; from D 4.1.1; T 4.1)
  Dataset on condensation sink
  Dataset on mixing layer height

Tools and services to access iCUPE data

All data sets have data teasers published on the iCUPE website[5]. These form together with the data set readme files as metadata information that is right available to the users.

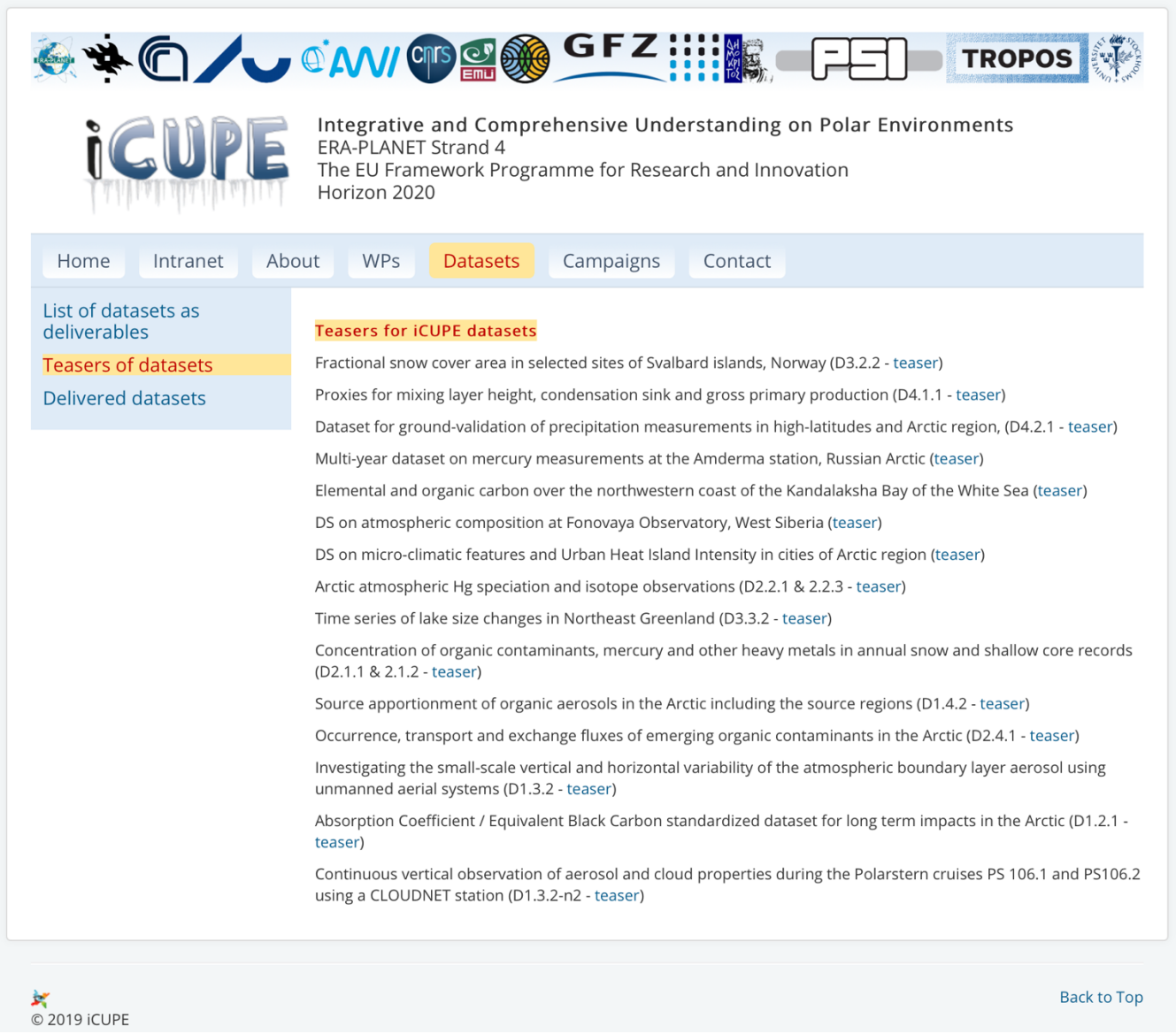

Figure 2. iCUPE data set teasers

[5] https://www.atm.helsinki.fi/icupe/index.php/datasets/submitted-datasets

iCUPE is further using VLAB[6], a virtual laboratory platform, that allows the generation of workflows to access data and to provide tools to that allow the use of the data to facilitate evidence-based decision-making. VLAB needs to get a set of models provided that operate on the data accessible in open repositories.

The dataset providing a snow cover assessment in Svalbard has a VLAB application utilizing satellite and in-situ data to produce visual results.

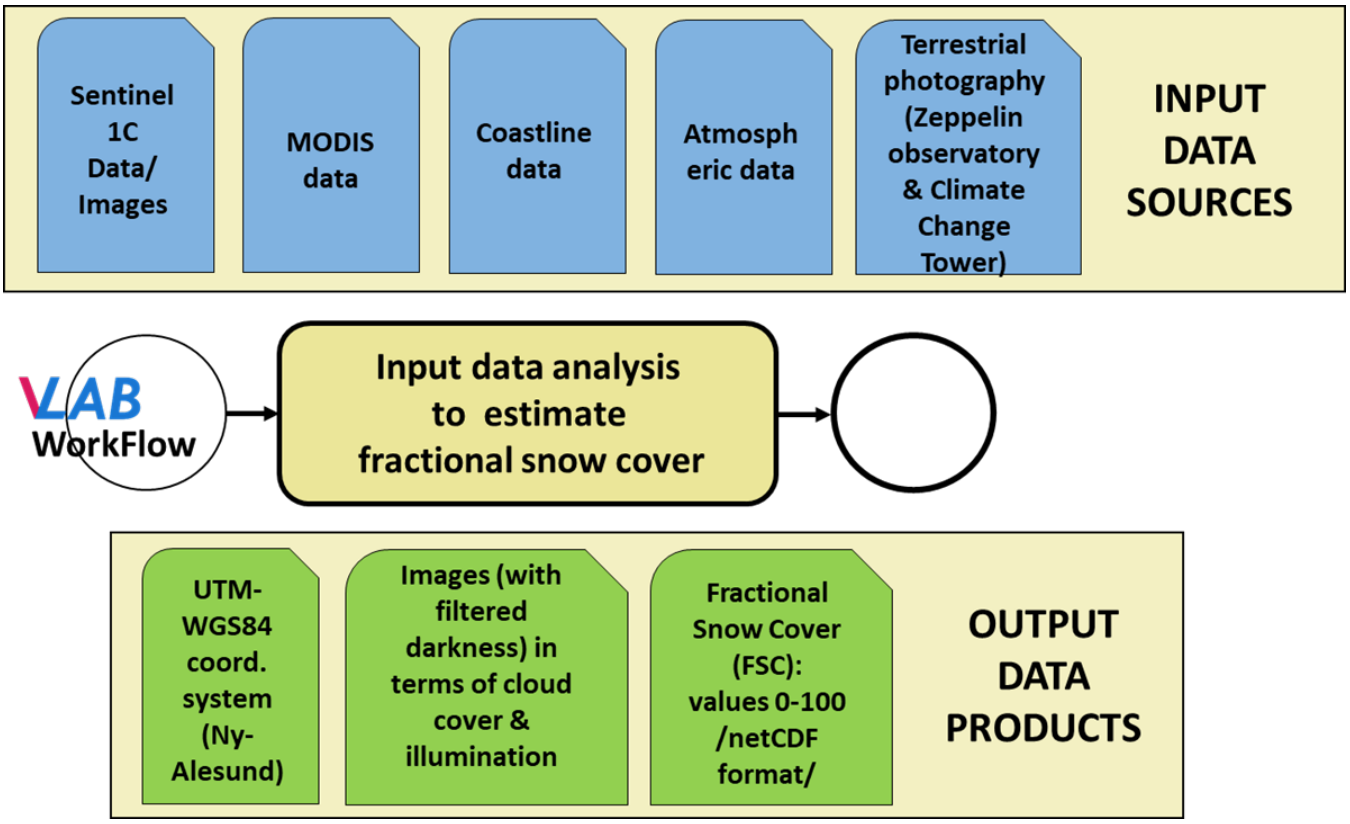

Figure 3. VLAB model workflow to generate a snow cover assessment utilizing iCUPE data and linking atmospheric parameters with satellite data products

[6] https://vlab.geodab.org

iCUPE pilot demonstrations are further available via GitHub[7] which ensures that the services are freely accessible and that they can be reused.

Interoperability

To demonstrate interoperability, we have implemented different examples where iCUPE data is processed in automatized way. Utilizing open cloud deployable services like VLAB and the docker platform as a service (PaaS) system together with version control and repository services like GitHub iCUPE data pilots offer an open approach to ensure interoperability. The paradigm used is that from existing data, a specialist in the field of interest creates workflows and interacts with the relevant repositories and PaaS systems. In VLAB, a workflow can be defined (Fig. 3) that includes different data streams and produces output. The Docker/Jupyter approach shows how to use another open system to provide similar services.

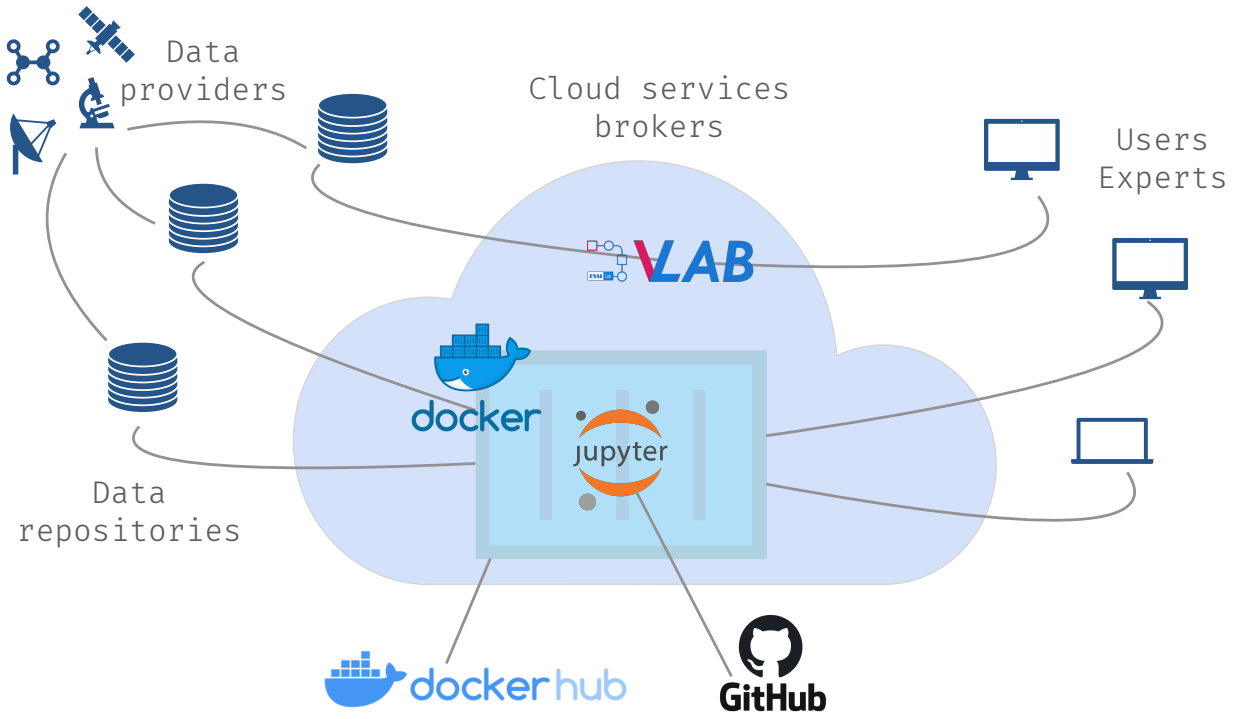

Figure 4. Structure of the iCUPE data pilot interoperability from data provision to the user.

[7] https://github.com

iCUPE acts via its structure as data provider, data repository and a data and knowledge broker. The latter is achieved by the iCUPE pilot services.

The Anderma mercury data example shows how the basic interoperability is achieved by a mixture of a data flow and code flow. The code of the workflow to create knowledge and information from data can be adapted to offer interfaces to other ERA-PLANET partners, GEO and Copernicus services in a machine-to-machine readable way.

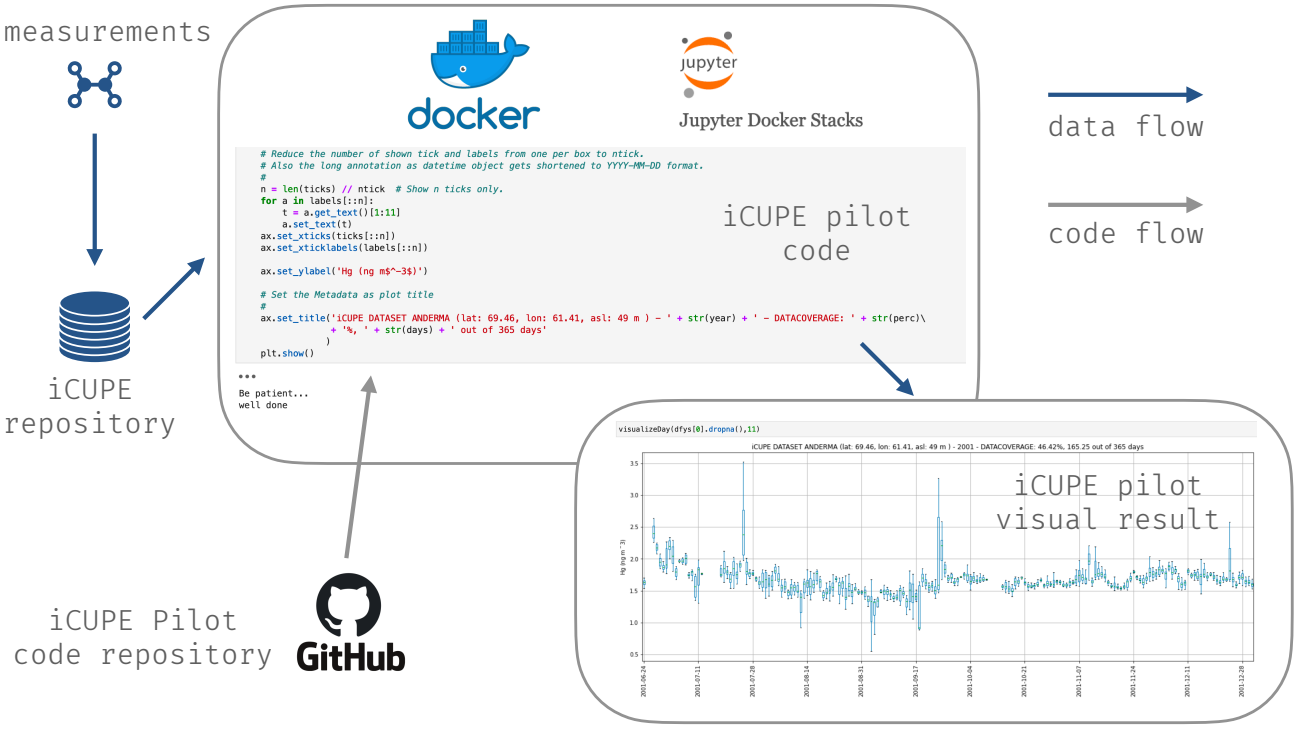

Figure 5. Anderma mercury data visualization pilot. The pilot is using the iCUPE dataset that is hosted on the project's website. The pilot's code is hosted on GitHub and is deployable into a docker container that is running the Jupyter Docker Stacks data science notebook. The data flow is indicated by blue arrows and the code flow by grey arrows.

.