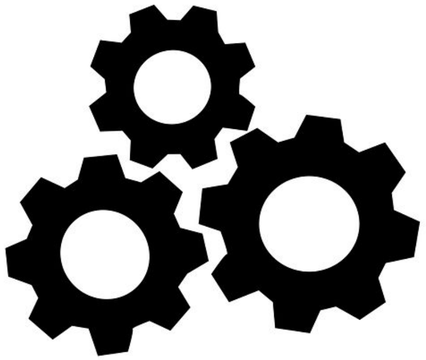
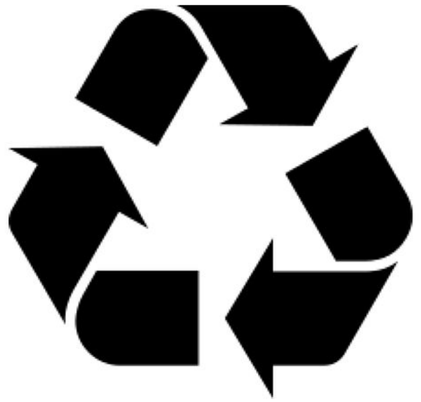
Big Data, organisation and analysis

FAIR data principles

Steffen M. Noe, Spring 2023

Findable Accessible Interoperable Reusable

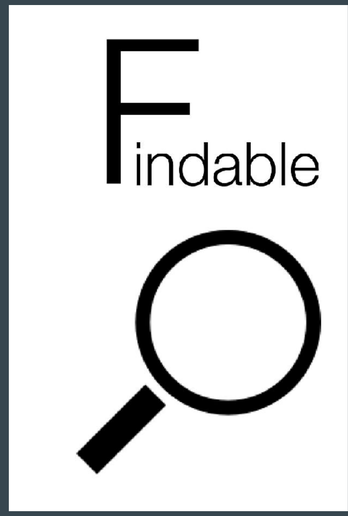




https://www.go-fair.org/fair-principles/

Findable

First step in using or re-using data is to find them!



- Metadata and data should be easy to find for humans and computers
- Machine readable metadata for automatic discovery

- F1: (Meta)data are assigned a globally unique and persistent identifier
- F2: Data described by metadata
- F3: Metadata contain clearly and explicit the identifier of the data they describe
- F4: (Meta)data are registered or indexed in a searchable repository

Accessible

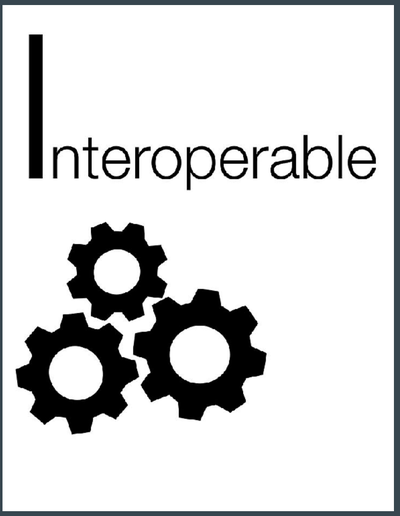
Creating access for both, humans and machines

- Users need to know how to access the data, once these are found
- Need of standardised communication protocols

- A1: (Meta)data can be retrieved by their identifier using a standard protocol
  - A1.1: Free, open, and universally implementable protocol
  - A1.2: The protocol allows for authentication and authorisation where necessary
- A2: Metadata are accessible even when data are not anymore available

Interoperable

Data integration and data pipelines



- Data need to be integrated with other data
- Need of interoperate with applications and workflows for analysis, storage, and processing

- I1: (Meta)data use a formal, accessible, shared and broadly applicable language
- I2: (Meta)data should use a vocabulary that follows FAIR principles
- I3: (Meta)data include qualified references to other (meta)data

Reusable

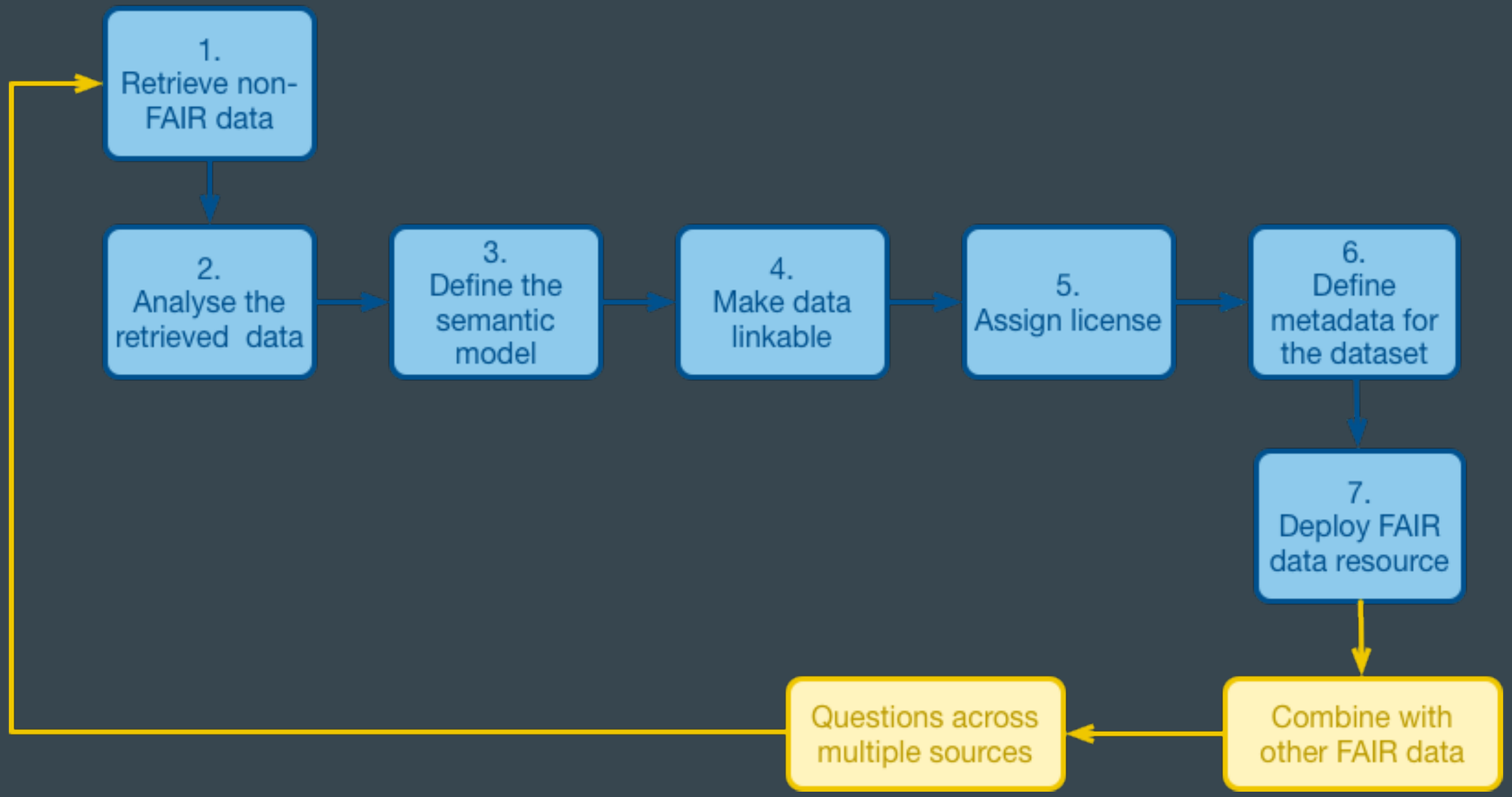
Reuse via replication and combination

- Ultimate goal of FAIR principles
- (Meta)data should be well-described to enable replication and combination

- R1: (Meta)data are richly described with a plurality of accurate and relevant attributes
  - R1.1: (Meta)data are released with a clear and accessible data usage license
  - R1.2: (Meta)data are associated with a detailed provenance
  - R1.3: (Meta)data meet domain-relevant community standards

The “FAIRification” process

This is as well working for any data!



Metadata creation

Needed to make the data findable - often time needed to provide good metadata is underestimated

- There are numerous schemata out (Dublin Core, Schema.org, ...)
- Usually these are standardised sets of JSON, XMLS, RDS descriptive languanges
- These are readable by both, humans and machines
- There are helper tools available to create metadata for your own datasets

Example: Poster on Arctic data sets of the ERA PLANET project iCUPE

iCUPE datasets applicability for evaluating impacts on social-economical activities in the Arctic

Arctic datasets
UN SDG's
Science policy integration

This research is mainly carried out by **Steffen Noe** et al. and collaborators from INAR, the iCUPE team and the Estonian University of Life Sciences.

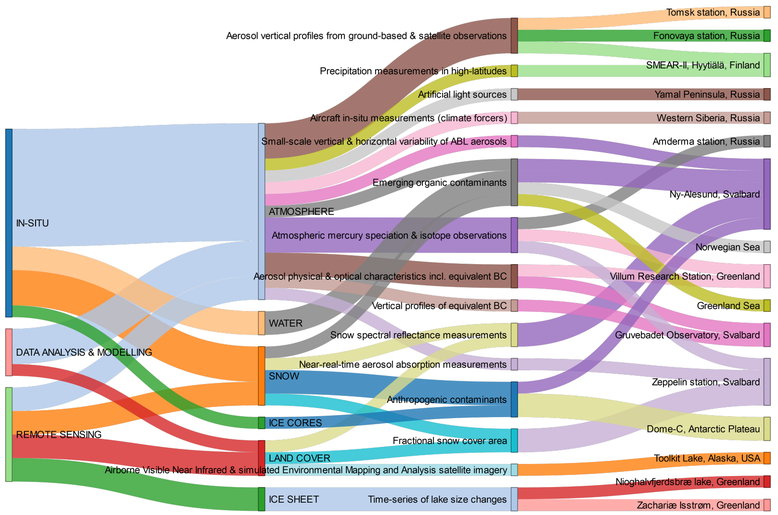
Why this study?

- Integrative and Comprehensive Understanding on Polar Environments (iCUPE) project developed datasets (DS) utilizing in-situ observational capacities within the Arctic or remote sensing observations from ground or from space.
- The DS covered atmospheric, cryospheric and terrestrial domains.
- Connecting the iCUPE datasets to United Nations' Sustainable Development Goals (SDGs).
- Facilitate scientific data driven decisions and enable data – policy feedbacks.
- Showcase selected DSs and their next steps facilitating future use of the iCUPE datasets.

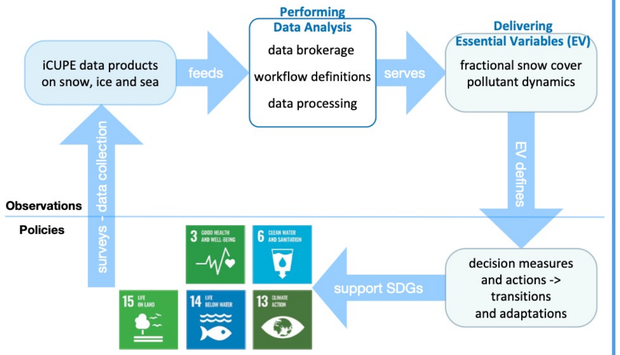
iCUPE data sets

Are available at:
www.atm.helsinki.fi/icupe/index.php/datasets/submitted-datasets

- Included into science and education, (e.g., hackthearctic.com)
- Enable science-policy feedbacks and inclusion of indigenous knowledge
- Accessible via cloud services and open repositories
- Interoperable towards GEOSS, Copernicus DIAS



Results



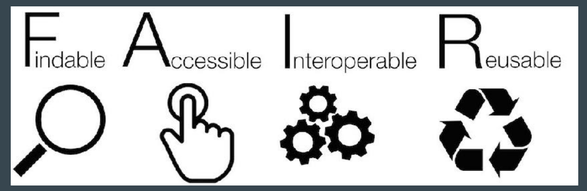
- DSs on aerosols, black carbon, physico-chemical properties and spatio-temporal variability - atm. pollution and climate change - impact on environment and population, health for Artic states, long-range transport of pollution to remote populated regions - climate adaptation and changing social life-style and economic activities in Arctic regions
- DSs on atmospheric mercury observations - atmospheric pollution and deposition on underlying surfaces – contamination of seas/lands – impact on fishery and reindeer economical activities – impact on environment and population health through food chains
- DSs emerging organic contaminants in water - contamination of seas - impact on fishery industry - impact on population health through food chains
- DSs on emerging organic and anthropogenic contaminants in snow - contamination of food supply for reindeers – impact on economic activities and style of the life of indigenous people & impact on population health through food chains
- DS time series of lake size changes in Northeast Greenland - changes in water resources availability - influence on hydropower plans of the Greenlandic government to foster economic development in Greenland

Contact information: steffen.noe@emu.ee

iCUPE ERA-PLANET SMEARstonia www.emu.ee Eesti Maaülikool Estonian University of Life Sciences EMU INAR Institute for Atmospheric and Earth System Research

Data sets are publicly available

We use Dataset Teasers and ReadMe files. These contain metadata on the dataset.
iCUPE data follows the FAIR principle.



DS Teasers
https://www.atm.helsinki.fi/icupe/index.php/datasets/submitted-datasets
Delivered DS
https://www.atm.helsinki.fi/icupe/index.php/datasets/delivered-datasets

Example:


DS Teaser

iCUPE Integrative and Comprehensive Understanding on Polar Environments
ERA-PLANET strand 4

Monitoring, modeling and assessment of potential sources, dynamics and atmospheric transport for low and elevated mercury concentrations in Arctic regions

Fidel Pankratov, Institute of Northern Environmental Problem, Kola Science Centre of the Russian Academy of Sciences (INEP KSC RAS)
fidel_ru@mail.ru
Moscow, 23.04.2018

iCUPE Collaborators Datasets
DS on atmospheric mercury measurements at Amderma station
Document version number: 1

The development of a model for the dynamics of mercury (Hg) in the surface layer of the atmosphere is logical extension of the long-term monitoring of Hg in Russian Arctic. The Hg input from the southern and middle latitudes to the Arctic will be assessed using the long-term high-resolution data (concentrations of elemental Hg in the atmosphere with a resolution of 1 hour from 2001 to the present, as well as the meteorological parameters (temperature, wind direction, humidity) with a resolution of 3 hours). Model of the global mercury transport in atmosphere of the northern hemisphere and especially in the Arctic atmosphere will also be tested. These data will be used to calculate the deposition rates of mercury to the underlying tundra surface, and uptake of the organic forms of mercury through biological chains will be assessed subsequently.

At the polar station Amderma the phenomenon of the atmospheric mercury depletion events (AMDEs) was confirmed using the long-term monitoring data. The unique experiment when the atmospheric mercury collection point during the long-term monitoring was consequently moved from the mainland to the coast line of the Kara Sea fixed the increasing number of the AMDEs. The results obtained will be helpful in better understanding of the mercury behavior in the Arctic region.

The obtained long-term monitoring data at the Amderma station are compared with the results of measurements made at other international Polar Stations. High convergence of the results is shown for all polar stations.

The volcanic eruptions in Iceland are identified as the cause of the unusually high atmospheric mercury concentrations in the background layer at the Amderma Station. These data can be used to identify and evaluate local anthropogenic and natural sources that affect Arctic pollution.

References

Pankratov F.F., Mahura A., Karpusova I.V., Katz O.V. Dynamics of atmospheric mercury in the Russian Arctic depending on the measurement position versus coastline. // Extended Abstract and Poster, Atmospheric Sciences, The 1st Pan-Eurasian Experiment (PEEX) Science Conference & The 5th PEEX Meeting Helsinki, Finland 10-13 February 2015; http://www.atm.helsinki.fi/FAAR/reportseries/rr-163.pdf, (pp. 324 - 329).
Pankratov F., Mahura A., Popov V., Katz O. Long-term continuous monitoring of mercury in the Russian arctic: winter increase of atmospheric mercury depletion events. // Synopsis and Poster, Atmospheric Sciences, Pan Eurasian Experiment (PEEX). European Geosciences Union, General Assembly 2014. April 27 – 02 May, 2014, Vienna, Austria: Abstract. http://meetingorganizer.copernicus.org/EGU2014/posters/15039
Pankratov F. Dynamic of atmospheric mercury in the Russian Arctic. Thesis, November 2015, DOI: 10.13140/RG.2.1.4155.1767.1

April 24, 2018 1


DS Metadata / ReadMe

Integrative and Comprehensive Understanding on Polar Environments
ERA-PLANET strand 4
iCUPE

Dataset name: iCUPE Dataset (DS) from iCUPE Collaborators

Long-term monitoring of gaseous elementary mercury in background air at the polar station Amderma (Russian Arctic)

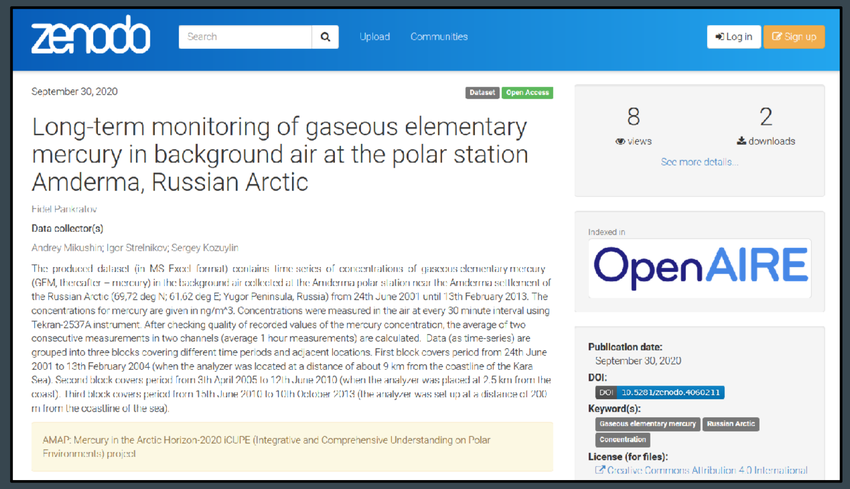
Author(s) and affiliations: Fidel Pankratov
Institute of Northern Environmental Problems, Kola Science Center, Russian Academy of Sciences (INEP KSC RAS)

Place and date: Apatity, Russia, 28 January 2020

Dr. Fidel Pankratov
Institute of Northern Environmental Problems
Kola Science Center, Russian Academy of Sciences
14A Fersman Str., Apatity
Murmansk region, 184200, Russia
E-mail: fidel_ru@mail.ru

The produced dataset (in MS Excel format) contains time-series of concentrations of gaseous elementary mercury (GEM, thereafter – mercury) in the background air collected at the Amderma polar station near the Amderma settlement of the Russian Arctic (69,720N; 61,620E; Yugor Peninsula, Russia) from 24th June 2001 until 13th February 2013. The concentrations for mercury are given in ng/m3. Concentrations were measured in the air at every 30 minute interval using Tekran-2537A instrument. After checking quality of recorded values of the mercury concentration, the average of two consecutive measurements in two channels [average 1 hour measurements] are calculated.
Data [as time-series] are grouped into three blocks covering different time periods and adjacent locations (Figure 1). First block covers period from 24th June 2001 to 13th February 2004 (point №1, when the analyzer was located at a distance of about 9 km from the coastline of the Kara Sea). Second block covers period from 3th April 2005 to 12th June 2010 (point №2, when the analyzer was placed at 2.5 km from the coast). Third block covers period from 15th June 2010 to 10th October 2013 (point №3, the analyzer was set up at a distance of 200 m from the coastline of the sea).

28 January 2020 1


zenodo

Search | Upload | Communities | Log in | Sign up

September 30, 2020 Dataset Open Access

Long-term monitoring of gaseous elementary mercury in background air at the polar station Amderma, Russian Arctic

Fidel Pankratov

Data collector(s)

Andrey Mikushin; Igor Strelnikov; Sergey Kozuylin

The produced dataset (in MS Excel format) contains time series of concentrations of gaseous elementary mercury (GEM, thereafter – mercury) in the background air collected at the Amderma polar station near the Amderma settlement of the Russian Arctic (69,72 deg N; 61,62 deg E; Yugor Peninsula, Russia) from 24th June 2001 until 13th February 2013. The concentrations for mercury are given in ng/m^3. Concentrations were measured in the air at every 30 minute interval using Tekran-2537A instrument. After checking quality of recorded values of the mercury concentration, the average of two consecutive measurements in two channels (average 1 hour measurements) are calculated. Data (as time-series) are grouped into three blocks covering different time periods and adjacent locations. First block covers period from 24th June 2001 to 13th February 2004 (when the analyzer was located at a distance of about 9 km from the coastline of the Kara Sea). Second block covers period from 3th April 2005 to 12th June 2010 (when the analyzer was placed at 2.5 km from the coast). Third block covers period from 15th June 2010 to 10th October 2013 (the analyzer was set up at a distance of 200 m from the coastline of the sea).

AMAP, Mercury in the Arctic Horizon-2020 iCUPE (Integrative and Comprehensive Understanding on Polar Environments) project

8 views 2 downloads
See more details...

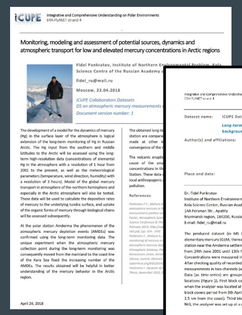
Indexed in OpenAIRE

Publication date: September 30, 2020
DOI: 10.5281/zenodo.4060211
Keyword(s): Gaseous elementary mercury | Russian Arctic | Concentration
License (for files): Creative Commons Attribution 4.0 International

iCUPE project repository weblink:
https://www.atm.helsinki.fi/icupe/images/Datasets/DS_Hg-Amderma_20200125.zip
ZENODO repository weblink:
https://doi.org/10.5281/zenodo.4060211

Data set processing towards users

DS Teaser



DS Metadata / ReadMe

access

Lexical analysis
Machine learning

Data analysis, visualisation
Machine learning

User accessible document

Location
Data description
…

Data identifier
Visual representation
Data access
…

Data

Data transfer

iCUPE website / repositories

Cloud computing infrastructures

Data set processing towards users

Given the data processing chain shown before:

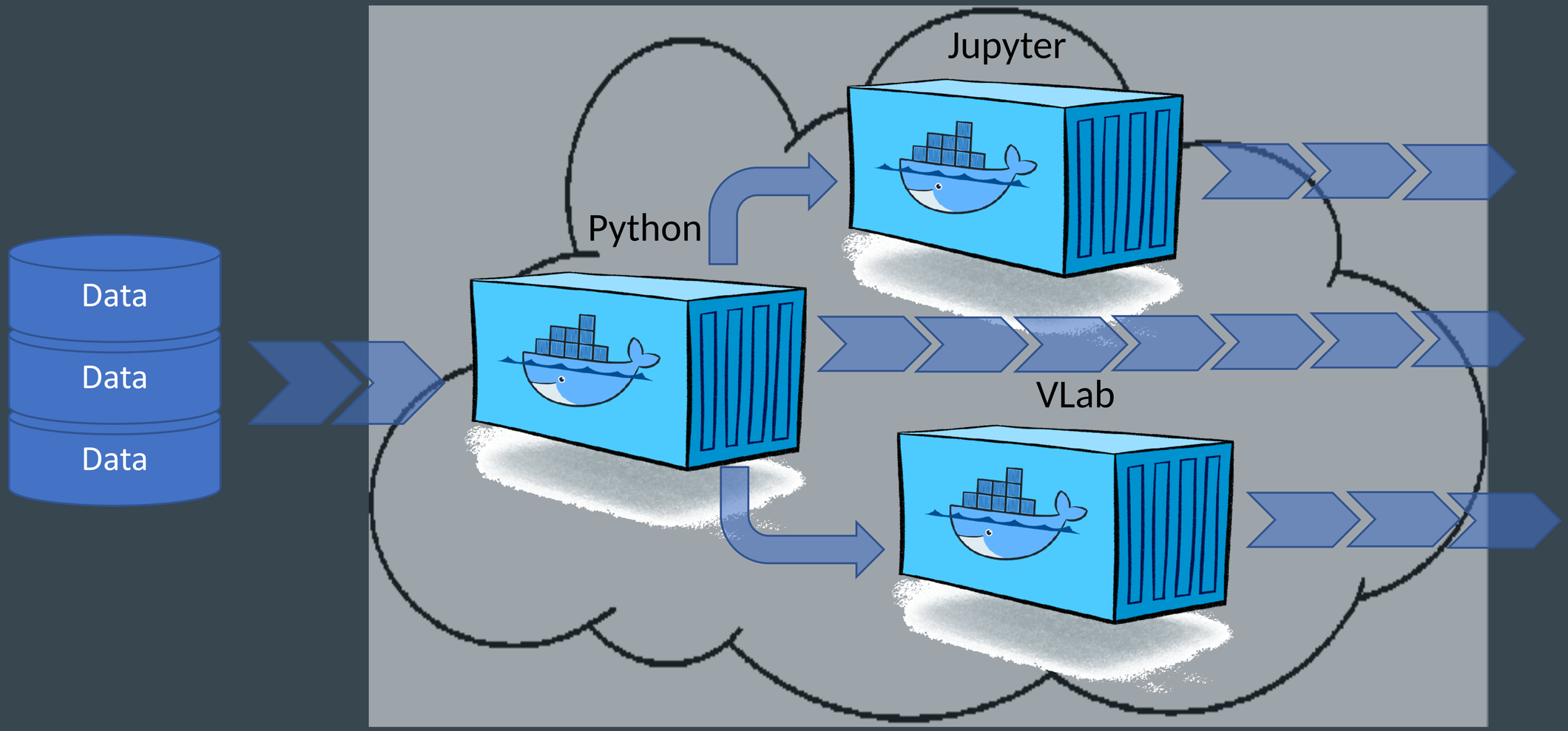
- The constant updating of data need to be taken into account for metadata creation
- The data might need to be sliced into appropriate time slots. (yearly, monthly, …)
- If continuous data are linked to static data a check on matching time is needed

Solutions:

Automatise metadata creation as much as possible

Need a growing number of microservices with growing number of data

Interoperability through cloud based microservices



Another example: Mercury pollution pilot workflow

measurements

iCUPE repository

Jupyter Docker Stacks

```
    # Reduce the number of shown tick and labels from one per box to ntick.
    # Also the long annotation as datetime object gets shortened to YYYY-MM-DD format.
    #
    n = len(ticks) // ntick  # Show n ticks only.
    for a in labels[::n]:
        t = a.get_text()[1:11]
        a.set_text(t)
    ax.set_xticks(ticks[::n])
    ax.set_xticklabels(labels[::n])

    ax.set_ylabel('Hg (ng m$^-3$)')

    # Set the Metadata as plot title
    #
    ax.set_title('iCUPE DATASET ANDERMA (lat: 69.46, lon: 61.41, asl: 49 m ) - ' + str(year) + ' - DATACOVERAGE: ' + str(perc)\
                 + '%, ' + str(days) + ' out of 365 days'
                )
    plt.show()
```

iCUPE pilot code

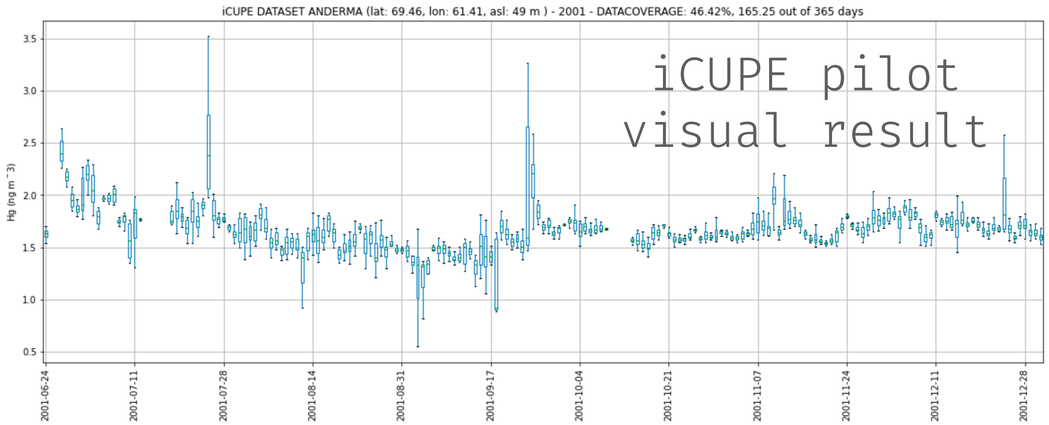
```
...
Be patient...
well done
```

data flow

code flow

iCUPE Pilot code repository


```
visualizeDay(dfys[0].dropna(),11)
```



iCUPE pilot visual result

Open Science and FAIR data structures at EMÜ

Maaülikool web: https://library.emu.ee/en/research/

https://library.emu.ee/en/research/open-access/

https://library.emu.ee/en/research/datacite-estonia/

https://library.emu.ee/en/research/open-science/

https://library.emu.ee/en/research/research-data-management/

https://library.emu.ee/en/research/rights-and-licences/