A lightweight approach to joining remote-sensiong earth-observation data and in-situ data in support of deep learning and reproducibility

Bernard Pruin, Nils Junike, Petabite GmbH, email: forename.surname@petabite.eu

Key points

A concept for the space-efficient combined storage of remote sensing and in-situ data is proposed, that has the key benefits

- Provides a useful separation of concerns when implementing algorithms
- Offers significant data input reduction for a large class of problems
- Allows to store economically and for long term the input data of process for future reference

Introduction

Earth Observation sets out to figure out the conditions of a point, area or volume on the earth by applying remote sensing techniques from space. In some cases relationships between the remotely received radiation and the situation and status on the observed location can be established through direct application of the laws of physics. In other cases data-analytical methods allow to create a verifiable connection between an in-situ value and its effect on remote sensing measurements. Validation, nowcasting and data-assimilation rely on these relations. A generic technical challenge for involved processes stem from the significant differences of the spatio-temporal characteristics of remote sensing and in-situ data. We summarize here our approach that lead us to the definition of joined remote sensing and in-situ data packages.

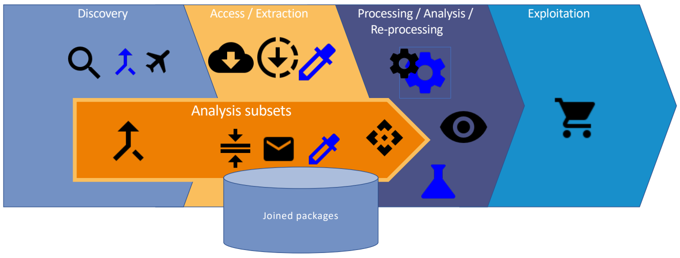


Figure: Realm and bundle product context

Remote-sensing vs. in-situ

Remote sensing data usually covers wider areas with a potentially high sampling rate due to the ever finer spatial resolution. Setting the emerging video data products aside, as a convention the EO data is provided in the form of data product of limited extent [3], where the granularity of the data products may be driven by various factors ranging from convention, existing reference systems, file size consideration, orbit characteristics or processing needs. In general remote sensing data products tend to be large and are getting larger due to technical advances.

In-situ data from localized sensors is by its nature spatially confined to points or individual areas and may be along trajectories. Depending on the acquisition rate, the data volume per sensor data stream may be small to negligible and there are generally a limited numbers of sensors. The sensor measurement data is often kept in data bases and data services tend to provide the data packaged per region, sensor or platform and for pre-set time intervals [2].

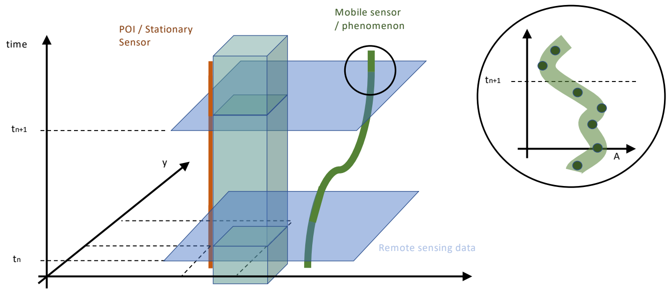


Figure: Remote sensing vs. in-situ data characteristics

Joined packages

Inherently there are two perspectives in defining the spatio-temporal area of interest for an application.

From the remote-sensing perspective sufficient in-situ data is needed to interpolate at the point of intersection of measurements and to take into account potential trends. The temporal extent is driven by inherent physical aspects as well as more technical requirements, e.g. the digitisation rate.

From the in-situ perspective there needs to be sufficient spatial data to allow for interpolation at the point of interest or to cover the observed phenomenon. In analogy to the first case, the size of the spatial extent is influenced by the spatial resolution of the remote sensing data and by inherent features of the phenomenon, e.g. when observing a ship, the ship‘s wake may be a feature of interest with a significant spatial extent.

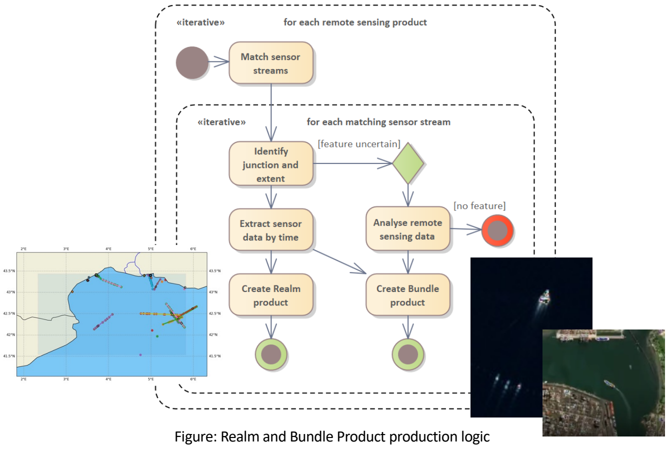


Figure: Realm and Bundle Product production logic

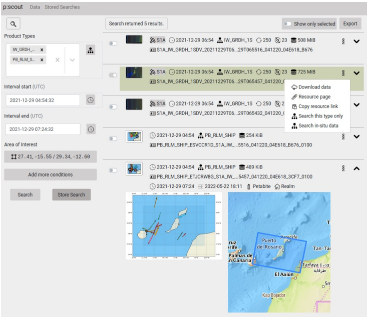
Storing the data in the given intermediate format provides a convenient mechanism to reduce the data to be stored and read by the algorithm and can be used to provide cost efficient long term reproducibility of the results. For results reproducibility it is not sufficient to reference the original data source, e.g. catalogue, as it’s content may change over time. Remote sensing products are mostly stored under a strict and published configuration management regime that may, as will be the case of Copernicus further secured through a traceability service. When looking at longer term reproducibility however, the actually used version of a product may eventually be replaced with a newer version, leading to an eventual discard of the old one. There may even be a storage policy shift that leads to an on-demand reproduction of a requested product from lower level original data.

In-situ data when taken directly from databases may change over time due to re-processing additional validation or data reduction process to save storage space. The change may even go unnoticed as there is usually no data-point level versioning information.

For both remote-sensing and in-situ data there is a strong case for storing the actual input into an analysis for future reference [4]. Using a storage concept that focusses on the junction of the observations can reduce the cost significantly, in particular when the processing function uses the intermediate data products from the outset.

Supporting service

The concept can benefit from a dedicated service that can provide the required packages up-front or on request and, as some of the packages may be quick to generate could also provide the required packages upon request and on-the-fly.



Search returned 5 results.

Show only selected

Product Types: IW_GRDH_... PB_RLM_S...

Interval start (UTC): 2021-12-29 04:54:32

Interval end (UTC): 2021-12-29 07:24:32

Area of Interest: 27.41, -15.55 / 29.34, -12.60

S1A 2021-12-29 06:54 IW_GRDH_1S 250 23 508 MiB
S1A_IW_GRDH_1SDV_20211229T06...29T065516_041220_04E618_B676

S1A 2021-12-29 06:54 IW_GRDH_1S 250 23 725 MiB
S1A_IW_GRDH_1SDV_20211229T06...29T065457_041220_0

S1A 2021-12-29 06:54 IW_GRDH_1S 250 2
S1A_IW_GRDH_1SDV_20211229T06...29T065432_041220_0

2021-12-29 04:54 PB_RLM_SHIP 254 KiB
PB_RLM_SHIP_ESVCCR1D_S1A_IW_...5516_041220_04E618_B676_0100

2021-12-29 04:54 PB_RLM_SHIP 489 KiB
PB_RLM_SHIP_ETJCRW8G_S1A_IW_...5457_041220_04E618_3CF7_0100
2021-12-29 07:24 2022-05-22 18:11 Petabite Realm

Currently available from Petabite

- Sentinel-1 GRDH ship sighting packages based on AIS data [5]
- Sentinel-2 MSIL2A ship sighting packages based on AIS data [5]
- Sentinel-1 and Sentinel-2 AIS data in product realm [5]

Internally used only

- Sentinel-5P UV aerosol index with community sensor PM10 data [1]
- Sentinel-2 MSIL2A and ASD-B data in product realm

Figure: Pre-packages in-situ data package in the Petabite data service

References

[1] Tesch, Eva: Estimating In-Situ Particulate Matter Concentration From Satellite Data Through Deep Learning. Master Thesis, Leuphana University of Lüneburg, 2022.
[2] Copernicus in situ TAC: Product User Manual for multiparameter Copernicus In Situ TAC (PUM), V1.12, 2021, DOI: 10.13155/43494.
[3] ESA: https://scihub.copernicus.eu/userguide/, Retrieved 05/2022.
[4] Liu, Z.,Wang, J.,Pan, S., and Meyer, D. (2019), Improving reproducibility in Earth science research, Eos, 100, https://doi.org/10.1029/2019EO136216. 10/2019.
[5] Petabite GmbH: Petabite Data Types, https://gitlab.com/petabite.eu/documentation/petabite-datatypes, Retrieved 05/2022