Processing in ACTRIS

Context of processing in ACTRIS

Complete ACTRIS report on Processing available at: https://envriplus.manageprojects.com/projects/requirements/notebooks/470/pages/36/comments/389/attachments/611/download

Summary of ACTRIS requirements for processing

Detailed requirements

Background

The ACTRIS Data Centre consists of three topical data repositories archiving the measurement data, which are all linked through the ACTRIS data portal to provide a single access point to all data. Hence, the ACTRIS Data Centre is founded on 3 topical data repositories:

- Near-surface in-situ and trace gas data are reported to **EBAS ---> data per year**
- Aerosol profile data are reported to the **EARLINET** Data base --> data per month
    - **LIDAR** data (acronym for Light Detection And Ranging), which is a remote sensing technology that measures distance by illuminating a target with a laser and analyzing the reflected light.
- Cloud profile data are reported to the **CLOUDNET** data base ---> data per day

Inputs

Mostly ACTRIS input data are tabular, but they could be also numners, SDF files and Matrix. The size of the data depends of the component of they belong. For example **Near-surface** in situ are on order of a few 100 Gb per year. In contrast, **Cloud profile** data have several TB per year because they collect data daily. For addressing the velocity of the data, once again it depends on the the component that we are referring to. **Cloud profile** collects several GB per instrument and day, while the other components nodes collects several GB per instrument and year or month.

The data can be very heregoeneous or not. Low-data-rate (e.g. **Near surface**) are very heterogeneous and they have different documentation needs. However, high data rate (e.g. **Lidar and Cloud radar**) rather homogeneous

The following information corresponds to **Lidar** node:

- The input data are numbers , which are stored in NetCDF files as matrix. The input data is around 100KB (per station and per component and per file). Currently, the data rate is defined as:
    - *30 stations, 3 measures, 100 KB per file, per week à 30 * 100KB * 3 * per week.*

Data are made available for the analytics phase using http and/or https protocols [examples in the active collab]. And the plan is to use OGC services

Analytics

It is quite difficult to quantify at the moment since (NRT) processing is being expanded right now. Largest processing needs clearly needed for cloud radar data.

The following information belongs to **EARLINET** component:

- Data processing, will need 1050 processing per month, divided in two steps (which can be done in parallel):
    - Pre-processing (preparing the data for the real process), which takes below 20 seconds.
    - Processing, which takes 10 seconds.
- Both steps can be done in parallel

The programming languages that ACTRIS uses are Python, C and 3Pascal, and it uses 3 Linux servers as a platform as Hardware. Each Linux server has with 4 core processor, and 16 GB RAM. The software requirements are:

- **Lidar** node: Linux, Open Source software, 3 Pascal compiler, and many other libraries (e.g. NetCDF libraries).
- **CLOUDNET** uses decentralist processing.
- **Near surface** in-situ uses dedicated database server (SYBASE) and VM server (Linux) for processing.
- Web applications also use .net.

ACTRIS suuports OAI-PMH, OGC WCS standards.

ACTRIS plans to have all software that they provided with an open source license. In such a way, that everyone can use them and contribute to the processes/algorithm. But, a coordinator will be needed to review contributions perform by users.

ACTRIS recognised that will be a good idea to have a sandbox as stable process/algorithm, and use it to compare others.

ACTRIS uses batch and interactive processing

ACTRIS mostly uses interactive-processing mode, and developers could use a monitor console.

ACTRIS uses a proprietary workflow.

Output

Data output answers are quite similar to the input data. Mostly ACTRIS output data are tabular, but they could be also numners, SDF files and Matrix. The voulume of output data can be clasiffied by component:

- **Near surface**: The output data size is ~ input data size. The velocity is ~ year.
- **EARLINET**: The output data size ~ input data size. Velocity ~ year.
- **CLOUDNET**: The output data size is smaller than the input data size.

The following information corresponds to **Lidar** node:

- ACTRIS has 5 different topologies (each one for capturing different aerosol optical properties); each one produces datasets around 100KB. Besides, it also produces images. The data rate is: *30 stations * 100 KB * 5 topologies * per week*

For making available the analytics outcomes, ACTRIS expect to use the website and protocols like http and https.

Statistical

CLOUDNET and **EBAS** components have automatic collection of data (continues monitoring). However, **EARLINET** component has scheduler for measurements (no automatic collection), which allows configuring the collection of data with different hypothesis in mind, which can be refined later (for **EARLINET**).

To analysis the responses ACTRIS uses all data available (multivariate analysis), which can be continuous for **EBAS** and **CLOUDNET** components and discrete for **EARLINET** component. However, ACTRIS is trying to have continues data for **EARLINET** too. The data is bounded on regional region (just European regions). For profiling data, ACTRIS is limited into vertical: upper atmosphere and stratosphere. Currently, the EARLINET database has 505 files. But, for EARLINET it is expected to grow the data to 12GB/year.

ACTRIS is involved with the GAIA-CLIM project [1], which its aim is to improve the ability to use ground-based and sub-orbital observations to characterise satellite observations for a number of atmospheric Essential Climate Variables (ECVs). ACTRIS is working there in the measurements errors with data mining paradigm.

ACTRIS is planning for the next year to set up some data quality check (e.g. to check if there is any anomalies or strange values in the data or anomalies into the atmosphere) for the 3 components.

ACTRIS wants to work with different approaches and understand the different between them approaches. Therefore, depending of the situations, ACTRIS wants to use the most suitable approach.

Formalities (who & when)

Go-between	Rosa Filgueira
RI representative	Lucia Mona and Markus Fiebig
Period of requirements collection	July to November 2015
Status	Finished