Identification and citation in ACTRIS

Context of identification and citation in ACTRIS

Complete ACTRIS report on Identification and Citation available at: https://envriplus.manageprojects.com/projects/requirements/notebooks/470/pages/36/comments/389/attachments/609/download

Summary of ACTRIS requirements for identification and citation

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 aerosol and trace gas data are reported to **EBAS**
- Aerosol profile data are reported to the **EARLINET** Data base
    - **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

Identification

There seems to be general agreement to use DOIs for data identification. Different approaches exist on granularity. For example, **LIDAR** community uses DOIs on pre-defined data collections, by using a semi-manual method. Surface in situ community looks at auto-DOI, one for each manually annual submission per individual instrument at a station. Finally, Cloud community (**CLOUDNET**), may have challenge with issuing DOI before data is collected. Typically the data are separated by parameters, so each data product contains information about 1 or 2 parameters, which means that they have a lot of files. However, they are planning to have combined data products with all parameters integrated. For example, within **CLOUDNET** they have some products data are specific for clouds, but now they are working in some new products combining aerosol and clouds. And this is the same for three components. Therefore, they are working on different level of products, some of them related with the instrument information, some others related with the outcomes information of other infrastructures inside ACTRIS.

The temporally resolution of data, depends of the data product/component:

- Data in near real-time (Only for **CLOUDNET** component, planned for the other 2 components)
- Data delivered each hour
- Daily products
- But currently they are working on seasonal and yearly data.

The data products are spatially by station, and ACTRIS is also planning to have them by region.

Once more, depending on the data center associated with each ACTRIS compoment, the data products are stored in different way. For instace, **E ARLINET** uses a file based archive, but Near-surface aerosol component (**EBAS**) uses a relational database. Furthermore, ACTRIS is currently working on the combination of them to have combined products.

Near-surface aerosol component (**EBAS**) has full support for versioning. In this case, versions are time stamped, allowing retrieving previous versions even if sub-sets of a dataset have been updated at different times. The other ACTRIS components do not support versioning yet.

For ACTRIS is important that every digital data object is tagged with a unique & persistent digital identifier (PID), for tracking data use as payback for true open access. Near-surface aerosol component (**EBAS**) has been using data center issued URI for a couple of years. However, it needs to be implemented for other data center components. **LIDAR** do not uses PIDs. However, they have a nomenclature for naming the files, so they know exactly which kind of information is stored in each file. For example, for **EARLINET** the name of the files are: *Station's code+ date+ parameter+time.*

The metadata contains checksum information for the objects. And some information about the metadata is available from the ACTRIS's website. For example, for the stations they have a list with the code of stations (two characters identifiers). Besides, in the description of the databases, ACTRIS has stored the metadata, which is also available in the website. The data and metadata can be linked (but no with PIDs) by the name of the files.

ACTRIS delegates to the data generators the task of storing and archiving the data, so they have the decision to use a standardized system or not. In the case of the physical samples, This depends on the component. For example, **CLOUDNET** uses the instruments serial number. Near surface (**EBAS**), where they have filters for collecting the particles, they don't have any tracking number. While for **LIDAR**, they use a handbook, where they storethe characterization of all the channels of the instruments, etc., which is very important for the analysis of the data. This information is stored in EXCEL files, with all the information inside. For data undergoing processing, QA and QC procedures are applied /to be developed for the three components. The 3 components have a processing chain which foresees the release in a 2nd step of QC data. And the quality checked data is available with a DOI. Finally, the publishable data is linked with a DOI and stored in the CERA database.

ACTRIS uses 'station code' and DOI as identifier systems. Those DOIs are maintained by an external organization called CERA [1], based at Hamburg.

ACTRIS does not allocate budget for this. The target is to find a solution without charges for scientific use.

Citation

ACTRIS scientific community use the data differently depending if they belong to the RI or not:

- Internal users (coming from inside their community), for understanding what is happening, some times, they need to study the data coming from other stations.
- External users which are:
    - Modelling community, for evaluating models.
    - Satellite community, for validation of satellites.
    - People from different communities.
    - User community it is very wider, in fact global. Collaboration with WMO etc. for all Data Center components.

For referring to datasets in different publications, users normally cited them by using DOI in the Reference section (preferred) and/or by providing the data as supplementary information via a link. Furthermore, sometimes authors offer co-authorship of papers. It is important to mention that, when a ACTRIS dataset is cited by using DOI, all the DOI's authors receive the credits.

For some ACTRIS data users, it is important to refer to specific subsets. And ACTRIS recommends to use DOI, since it can refer to some subsets. For example, for calibration purposes they have a subset called Calipso, and users can refer to that dataset. They also have a subset for Volcanic Eruptions. However, it is important to mention that granularity of citations needs to be fine enough to allow this option.

ACTRIS has already a DOI collection of all data for **EARLINET** [2]. However, **CLOUDNET** and **EBAS** components are working for having DOI related with their dataset.

The ACTRIS' strategy for collecting information about the usage of their data products is based on counting the number of download/access, references in scientific literature, and by measuring the scientific impact.

ACTRIS considers important to have a good quality check of the data (which is currently time consuming) before obtaining the DOI. Therefore, the implementation of some automation of quality check will improve the process of getting DOI (planned within ACTRIS2 project). Besides, for getting DOI, they need to accomplish some standards, and sometimes those standards changes, making difficult to follow them.

Formalities (who & when)

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