Milestone M2 (M1.2)

Provision of data management and QA/QC tools for centralising, communicating and analysis

RI-URBANS

Research Infrastructures Services Reinforcing Air Quality Monitoring Capacities in European Urban & Industrial AreaS (GA n. 101036245)

By

NILU, TROPOS & INERIS

29 August 2022

RI-URBANS (www.RIURBANS.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-GD-2020, Grant Agreement number: 101036245

Milestone M2 (M1.2): Provision of data management and QA/QC tools for centralising, communicating and analyses

Authors: Markus Fiebig (NILU), Cathrine Lund Myhre (NILU), Richard Olav Rud (NILU), Thomas Müller (TROPOS), Jean-Eudes Petit (INERIS)

Work package (WP)	WP1: Novel AQ metrics and advanced source apportionment STs for PM and nanoparticles
Milestone	M2 (M1.2)
Lead beneficiary	NILU
Means of verification	Tools available for provision of data management and QA/QC tools for centralising, communicating and analyses
Estimated delivery deadline	M11 (31/08/2022)
Actual delivery deadline	31/08/2022
Version	Final
Reviewed by	WP1 leaders
Accepted by	RI-URBANS Project Coordination Team
Comments	Report describing ACTRIS resources for in situ data reporting in RI-URBANS.

RI-URBANS (www.RIURBANS.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-GD-2020, Grant Agreement number: 101036245

Table of Contents

RI-URBANS (www.RIURBANS.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-GD-2020, Grant Agreement number: 101036245

1. About this document

In RI-URBANS, the ACTRIS Research Infrastructure (RI) is responsible for the data management of all surface in situ data collected during the project. This milestone describes how ACTRIS meets the requirements of FAIR data management, how the data will be handled both for real-time and manual data submissions, and which software tools are available for RI-URBANS for conducting data curation and quality control.

This is a public document, available in the RI-URBANS website (https://riurbans.eu/work-package-1/#milestones-wp1). This document will be distributed to all RI-URBANS partners for their use and submitted to European Commission as an RI-URBANS milestone M2 (M1.2).

2. ACTRIS in-situ data management

The target of data management in ACTRIS (The Aerosol, Clouds and Trace Gases Research Infrastructure) is to comply with the requirements of data FAIRness to the best degree possible, where "FAIR" stands for *Findable, Accessible, Interoperable, Re-usable* as defined by the FORCE 11 interest group. The data FAIRness requirements are listed and defined below, along with the FAIR-enabling resources used by ACTRIS to implement the requirements:

- **Findable:**
  - *F1. (Meta)data are assigned a globally unique and persistent identifier:*
    All ACTRIS curated final data products will be identified by Digital Object Identifiers (DOIs), a service currently being implemented.
  - *F2. Data are described with rich metadata (defined by R1, see below):*
    ACTRIS metadata contain not only discovery metadata, but also metadata on operating procedures, data quality, and, in the future, provenance.
  - *F3. Metadata clearly and explicitly include the identifier of the data they describe:*
    The metadata contain one or several identified access points to the data..
  - *F4. (Meta)data are registered or indexed in a searchable resource:*
    The ACTRIS data centre collaborates with various data search portals which are operational or under development, e.g. the World Meteorological Organization (WMO) Information System (WIS), the Global Earth Observation System of Systems (GEOSS), and the European Open Science Cloud (EOSC).
- **Accessible**
  - *A1. (Meta)data are retrievable by their identifier using a standardised communications protocol:*
    - *A1.1 The protocol is open, free, and universally implementable:*
      ACTRIS offers standardised protocols such as OGC-WCS (Open Geospatial Consortium - Web Coverage Service) and OPeNDAP (Open-source Project for a Network Data Access Protocol) for data access, and OAI-PMH (ISO19115-2) and RestAPI for metadata access.
    - *A1.2 The protocol allows for an authentication and authorisation procedure, where necessary:*
      ACTRIS data are available without login by default, except e.g. real-time data.
  - *A2. Metadata are accessible, even when the data are no longer available:*
    The latest data version shall always be available, as well as versions previously publicly available.

RI-URBANS (www.RIURBANS.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-GD-2020, Grant Agreement number: 101036245

- **Interoperable**
  - *I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation:*
    ACTRIS uses, among others, XML schema for representing metadata, and the NetCDF file format to store data. Both are well described and widely used.
  - *I2. (Meta)data use vocabularies that follow FAIR principles:*
    ACTRIS has recently finished the first version of its own (meta)data vocabulary that meets itself the FAIRness criteria.
  - I3. (Meta)data include qualified references to other (meta)data:
    The ACTRIS vocabulary links to external reference vocabularies.
- **Re-usable**
  - *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:*
      ACTRIS data products use the Creative Commons by Attribution 4.0 (CC BY 4.0) license by default.
    - *R1.2. (Meta)data are associated with detailed provenance:*
      ACTRIS implements documentation of data provenance using the PROV-O ontology.
    - *R1.3. (Meta)data meet domain-relevant community standards:*
      ACTRIS includes the expert communities necessary for continuously improving metadata records.

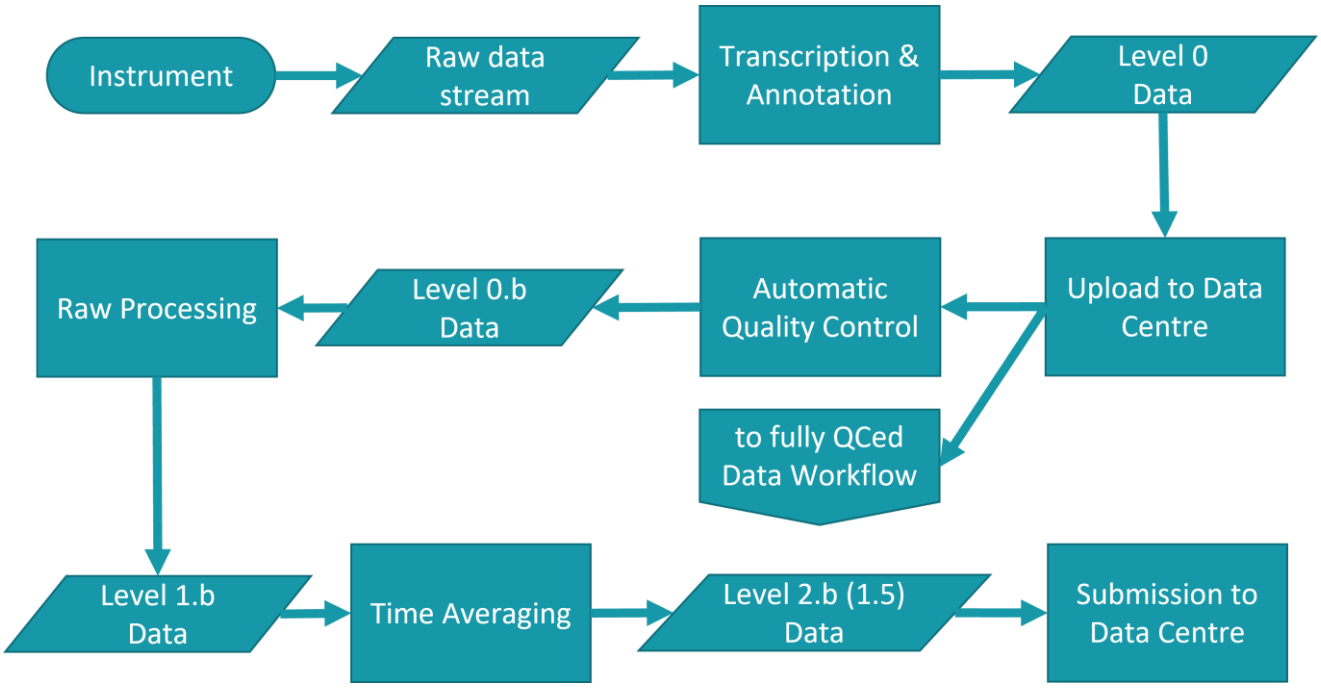

Figure 1: Abbreviated version of the ACTRIS in situ data production workflow for online observations. A well-defined data production workflow is a prerequisite for data traceability and documenting provenance of data production, which are both data FAIRness requirements.

RI-URBANS (www.RIURBANS.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-GD-2020, Grant Agreement number: 101036245

ACTRIS data production is described in detail in the ACTRIS data management plan (DMP). A short version of the data production workflow for ACTRIS in situ data, i.e. the branch for in situ real-time data production, is visualised in Figure 1. Online instruments produce a stream of raw data, which is transcribed into a standardised, widely used data format and annotated with metadata (level 0 data) by an ACTRIS provided software already at the station. The same software uploads the latest data package to the data centre at each turn of the hour. At the data centre, the data are processed to levels 1 and 2, where level 2 is published via the ACTRIS Data Centre data portal (for human interaction) and machine-to-machine interfaces. The hierarchy of data levels used for ACTRIS in-situ data is explained in **¡Error! No se encuentra el origen de la referencia.**.

Data Level	Manual QC	RRT/ auto QC	Description	
0	0a	0b	• Annotated raw data • format instrument specific • all data / information for processing to final value.	• contains all parameters provided by instrument as provided • ”native” time resolution
1	1a	1b	• processed to final parameter • invalid data removed • ”native” time resolution	• format property specific • STP correction if necessary
1.5 / 2	2 / 2a	1.5 / 2b	• aggregated to hourly averages • variability quantified	• format property specific

Table 1: Definition of data levels used in ACTRIS in situ data production.

Each higher level of data originates from the previous lower one according to the data production workflow, establishing traceability of the data back to the time of measurement. Level 0 data correspond to all parameters contained in the instrument raw data stream, plus information needed to process the data further, put into a standardised open data format. The time resolution matches the one defined as raw time resolution in the instrument. Level 1 data still has the same time resolution, but the data have been processed to the targeted parameter, e.g. from raw aerosol particle concentrations in level 0 to an aerosol particle number size distribution in level 1. Instrument status parameters are contained in level 0, but not anymore in level 1. In level 1, invalid data records are removed, and concentrations are reported for standard conditions of temperature and pressure (273.15 K, 1013.25 hPa). When producing level 2, data are averaged to hourly values, while the variability within the hour is quantified by percentiles. Data level identifiers containing an “a” denote data having undergone full, manual data quality control, while identifiers containing a “b” denote data having received automatic, real-time quality control.

Real-time data production is not the only in RI-URBANS for participating pilot stations. Also manually curated data are accepted by the ACTRIS in situ data centre unit at NILU. The tools for the manual data submission process are described in section 3.

RI-URBANS (www.RIURBANS.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-GD-2020, Grant Agreement number: 101036245

3. Tools for manual data submission

The ACTRIS In Situ data are hosted in the EBAS database operated at NILU. For manual data submissions, EBAS uses the NASA Ames 1001 data format with an EBAS specific profile. The NASA Ames 1001 specification ensures the files to be backwards compatible, while the EBAS profile meets today's requirements on data FAIRness. While there are more modern alternatives to NASA Ames 1001 available, it has the great advantage of being readable and editable by humans without any advanced software tools, which keeps the threshold low especially for less experienced users and data producers. Software libraries for handling NASA Ames 1001 are available in all major programming languages.

3.1. Instructions for assembling data submissions

The entry point to submitting data to the EBAS database is the EBAS data submission portal (Figure). The portal contains a section describing the process and workflow of submitting data to EBAS ("Getting started"). The main section ("Templates") contains a collection of instrument and variable specific templates for data submission where each entry especially in the metadata header is described individually. This applies to all 3 data levels (0, 1, 2). By definition, level 0 templates are instrument specific, while level 1 and 2 are specific for the observed parameter. To establish complete traceability, a data submission needs to contain all 3 data levels, from 0 via 1 to 2. While submitting all 3 levels is obligatory for ACTRIS stations, other frameworks might establish less strict policies, thus compromising on traceability.

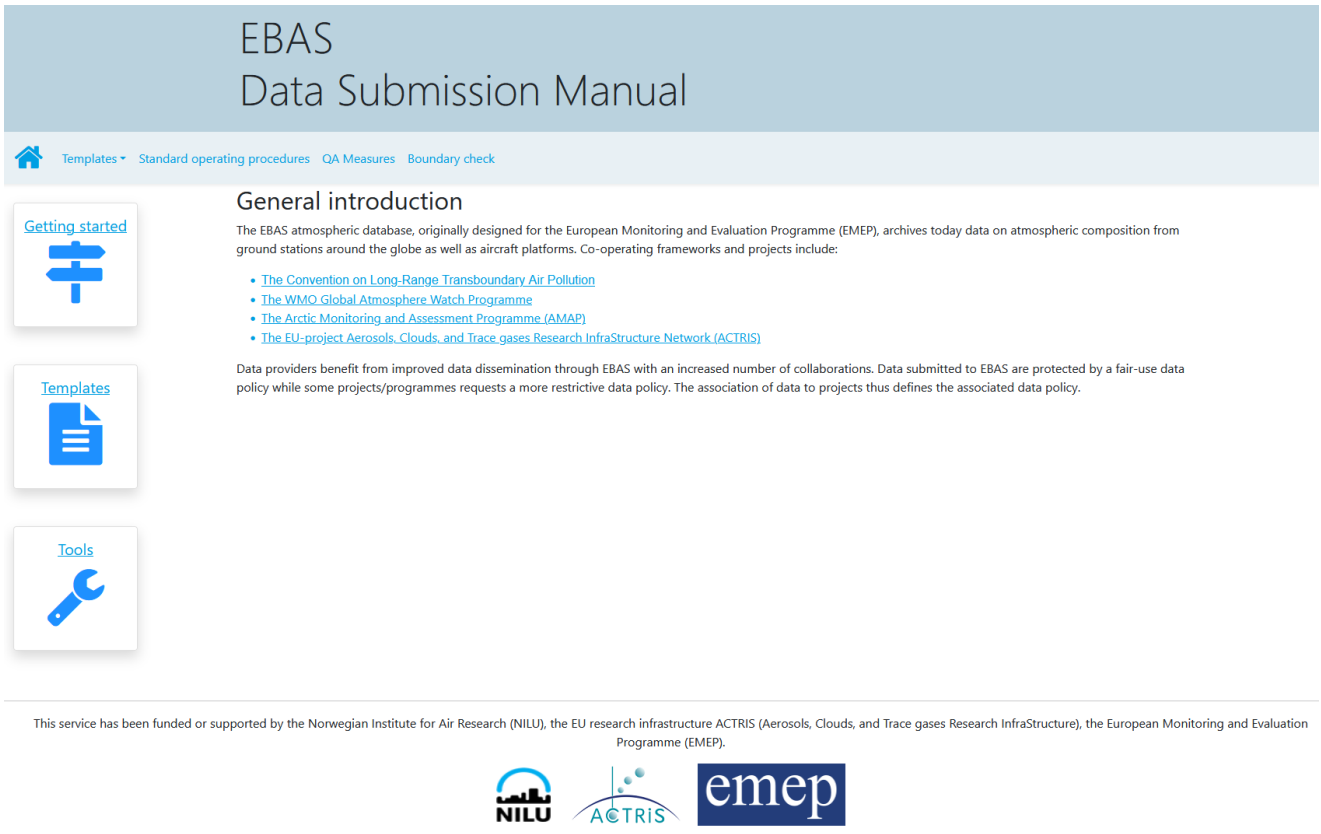

Figure 2: Screen shot of EBAS data submission portal describing all templates used for data submission to the EBAS database at NILU, EBAS hosts all ACTRIS In Situ data.

RI-URBANS (www.RIURBANS.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-GD-2020, Grant Agreement number: 101036245

The third section in the EBAS submission portal, “Tools”, collects software tools, services, and web resources around assembling and quality controlling data submissions, which partly are specific to certain variables. Two of these links on the tools page of the EBAS data submission portal link to the EBAS data submission tool and the Mantis issue tracker, which will be discussed in the next section.

3.2. Quality control of data submission and upload to data centre

Once a data provider has assembled a data submission with the tools of his preference, the file is submitted to the EBAS database using the EBAS data submission tool (Figure 3). The data file to be submitted is uploaded to the web-based tool for an initial quality check. The software performs checks on syntax and certain semantics, while giving the data provider immediate feedback on any errors or issues found. The provider can fix the errors in the tool, or upload new versions of the file until all issues are removed. This applies to all 3 data levels.

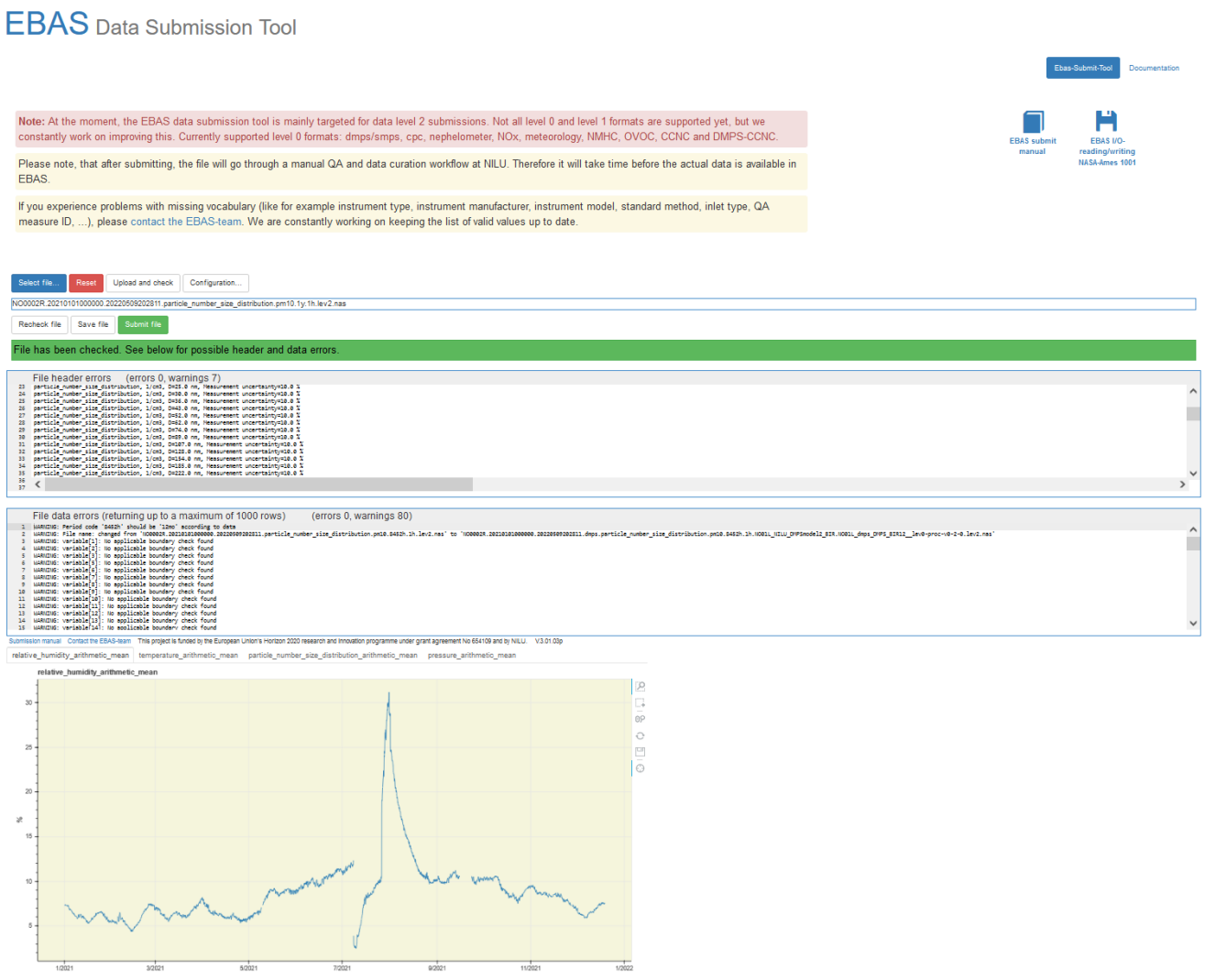

Figure *3*: Screen shot of EBAS data submission tool for interactive quality control of data files assembled for submission to the EBAS database at NILU.

RI-URBANS (www.RIURBANS.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-GD-2020, Grant Agreement number: 101036245

Once a data file is submitted via the submission tool, a member of the EBAS data curation team opens an issue for the file in the EBAS Mantis issue tracker (Figure 4). This issue tracker is the central tool for organising the communication between the data provider and the data curation group at the ACTRIS *in situ* DC node during the data curation process that follows the data submission. The issue tracker guides the data provider through the data curation workflow while documenting the process in a transparent way, making it possible for other group members on both ends to step in and contribute if needed. All users are identified by their own account on the issue trackers so tasks can be assigned to the responsible person.

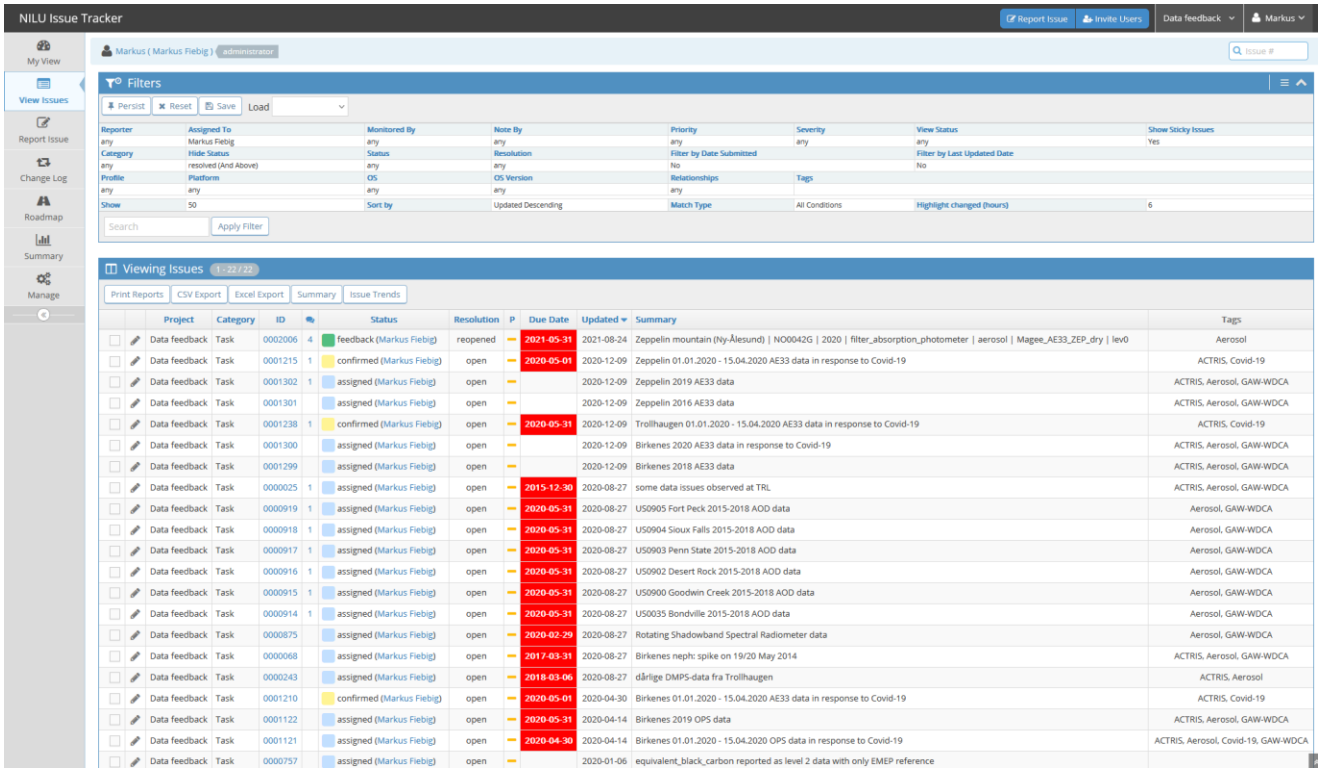

Figure *4*: Screen shot of the EBAS Mantis issue tracker tracing the communication between data provider and data center around a data submission in a transparent way, while guiding the data provider through the data curation workflow.

4. Tools for automatic (real-time) data submission

For ACTRIS in situ stations, the responsible Topic Centre (TC), the entity responsible for quality control of the observation, provides the station with software for data acquisition. This software collects the data from the instrument, stores it locally, visualises it on display, and uploads the data in real-time (at the turn of each hour) to the ACTRIS in situ data centre unit. In RI-URBANS, this service will be made available to stations participating in the project's pilot cities.

4.1. Software package for handling observations of particle number size distribution, aerosol particle light absorption and scattering

Within the CAMS21a project, a software package for data acquisition for absorption photometers, integrating nephelometers and particle number size spectrometers was developed. This package will be used for near real-time data acquisition at the RI-URBANS pilot sites. The software collects raw data and stores it in a local database (Figure 5). The database is the source for online visualisation of the data and the generation of hourly data files that are sent to the ACTRIS data centre where QC measures take place. The data comply with the ACTRIS in situ data

RI-URBANS (www.RIURBANS.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-GD-2020, Grant Agreement number: 101036245

management (see section 2). Raw data from all devices are stored in a local database (*influxDB*) on the data acquisition computer. Further data not part of the instrument raw data, e.g. relative aerosol humidity, temperature and air pressure can be stored in the same database. In this way, data files with all the necessary information for automatic quality control can be generated from one data source in a simple and transparent way. The data files are sent to the EBAS database every hour.

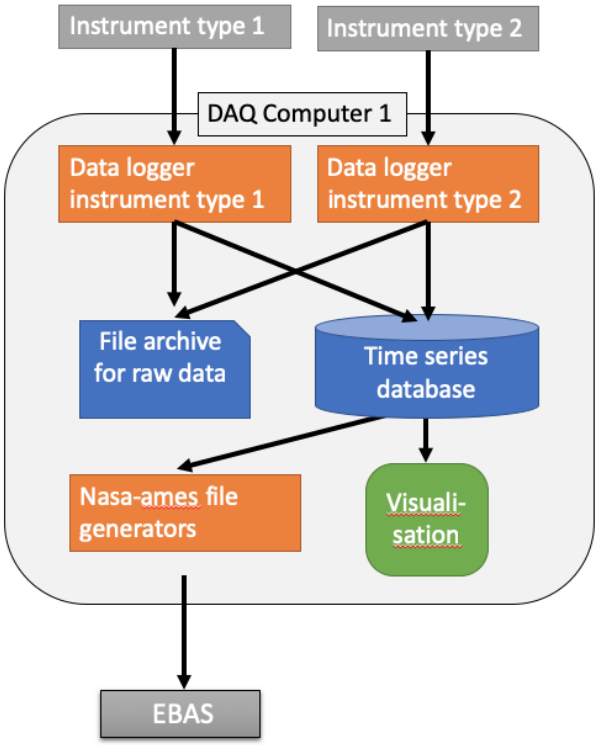

Figure 5: Diagram of the data flow on the DAQ computer.

The programs for instrument control and data recording for absorption photometers and integrating nephelometers can be controlled via a terminal to enable remote maintenance even with limited internet capacity. Optional graphical interfaces for these instruments are kept as identical as possible (Figure 6). Particle number size spectrometers are an exception, as they require special, in some cases proprietary software. These data are simply reformatted and transferred to the local database.

The visualisation of the data is realised exclusively via the free browser-based program *Grafana*, whereby the data is taken directly from the local database (Figure 7). The solution of browser-based visualisation and database offers the advantage that, with appropriate configuration, visualisation can take place on remote computers.

The data files are subjected to quality control after being transferred to the data centre. Automatically quality-assured data are stored as level 1b and level 2b data. The quality control software was developed and tested within the CAMS21a project.

RI-URBANS (www.RIURBANS.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-GD-2020, Grant Agreement number: 101036245

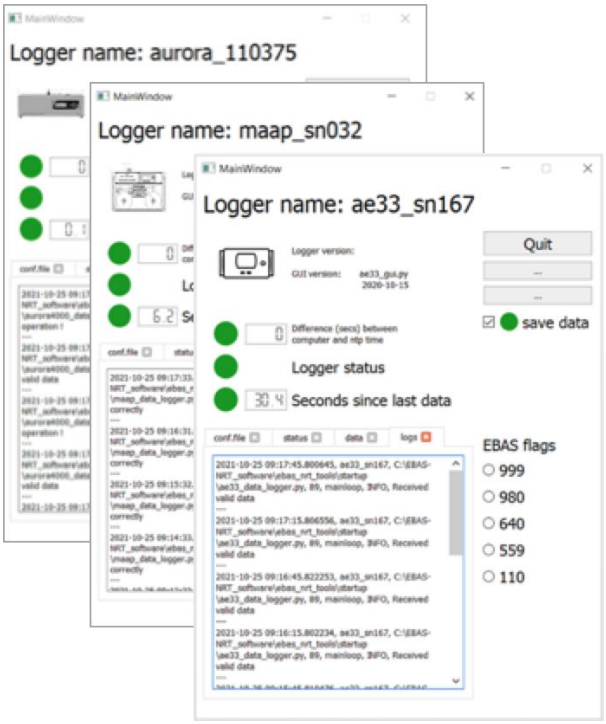

Figure 6: Screenshot of the graphical user interface facilitating control, data acquisition and processing.

Figure *7*: Screenshot of the real-time visualisation with access to data archive (local database).

RI-URBANS (www.RIURBANS.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-GD-2020, Grant Agreement number: 101036245

4.2. Software package for handling observations of aerosol particle chemical speciation by online mass spectrometer

For aerosol particle chemical speciation measurements, data acquisition software is provided by the manufacturer, since instruments are directly dependent of computer-driven calculations (on Igor, Wavemetrics©) in order to transform raw signals into atmospherically interpretable data. For that reason, a software package for data transfer only has been developed within CAMS21a project (Figure 8). It is totally independent of other tools (i.e. does not interfere with the acquisition) and runs as a background task. Complementary Python scripts have also been developed to ensure data submission in near real time.

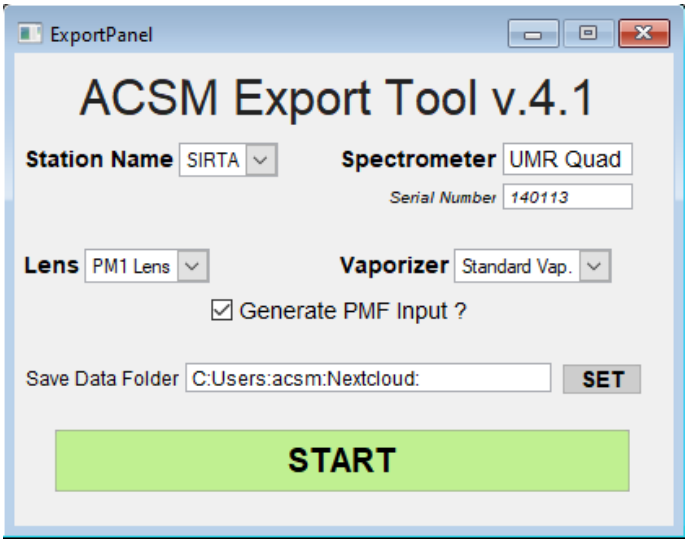

Figure *8*: Screen shot of interface of software facilitating instrument control, data collection, and real-time data submission for instruments measuring aerosol particle chemical speciation by online mass spectrometer

Input data for RT source apportionment are not yet part of the operational data reporting procedures of ACTRIS. For the pilot phase, a dedicated and direct submission to the source apportionment server has been set up. In the operational phase, production of source apportionment products will be transferred to the ACTRIS *in situ* DC unit in accordance with the ACTRIS in situ data workflow.

All RI-URBANS pilot sites will be equipped with the tool during a dedicated session of the ACMCC workshop this fall.

5. Outlook

Capacity building is a central part of RI-URBANS. In the first part of the project, partners have collected locally archived data of observation parameters handled by ACTRIS from air quality network across Europe. The partners have quality controlled and curated the data, making them ready for archiving and publication in the ACTRIS data centre. In the second part of the project, efforts will be placed in building capacity in the distributed air quality networks to continue data submission in the future, thereby creating a new resource of air quality data available while meeting the data FAIRness criteria.

RI-URBANS (www.RIURBANS.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-GD-2020, Grant Agreement number: 101036245