Milestone 4.3: Selection of web-based NF-TC-DC workflow management tool

Authors: ACTRIS Data Centre

Cathrine Lund Myhre (lead, & Head of DVAS), Patrice Henry (Deputy), Markus Fiebig (NILU, Head of In-Situ), Lucia Mona (CNR, Head of ARES), Ewan O'Connor (FMI, Head of CLU), Cathy Boonne (CNRS, Head of GRES), Benedicte Picquet Varrault (CNRS, Head of ASC), Claudio Dema (CNR, ARES), Anja Claude (CiGas TC), Thomas Müller (ECAC TC)

Work package no	WP4
Milestone no.	MS4.3
Lead beneficiary	NILU
Deliverable type	X R (Document, report) ☐ DEC (Websites, patent filings, videos, etc.) ☐ OTHER: please specify ....
Dissemination level	X PU (public) ☐ CO (confidential, only for members of the Consortium, incl. Commission)
Estimated delivery date	M12
Actual delivery date	29/06/2021
Version	Final

ACTRIS IMP (www.actris.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-INFRADEV-2019-2, Grant Agreement number: 871115

Reviewed by	Niku Kivekäs
Accepted by	Niku Kivekäs
Comments	

ACTRIS IMP (www.actris.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-INFRADEV-2019-2, Grant Agreement number: 871115

Contents

ACTRIS IMP (www.actris.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-INFRADEV-2019-2, Grant Agreement number: 871115

ACTRIS IMP (www.actris.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-INFRADEV-2019-2, Grant Agreement number: 871115

1 Background and purpose of this document

Production of fully quality assured ACTRIS data needs to follow a well-defined workflow involving national facilities (NFs), topic centres (TCs), and the Data Centre (DC). For observational platforms, the workflows have been discussed, and established between the ACTRIS TCs, DC and the NFs in ACTRIS PPP, to ensure cost efficient and clear roles and responsibilities and sharing of tasks. For exploratory platforms, the definition of precise workflows is in progress and involves NFs, TCs and DC. Roles and responsibilities of tasks for the data generation have been clearly defined but the definition of SOPs for each type of instrumentation is still under discussion. All workflows are available in ACTRIS Data Mananagement plan appendix 3-8[1], and also described in ACTRIS-IMP D4.1 "Descriptions of the workflows between ACTRIS components".

To guarantee traceability of data from the NFs, data need to be associated with instrument and data quality documentation produced at the responsible Central Facilities (CFs). The workflows will include steps distributed across NFs, TCs and DC, and web-based workflow management tool(s) will be selected and implemented to ensure full traceability of all data production steps (Deliverable due M12).

The purpose of this document is to assist and provide background for selection of workflow tools. A first step of selection of this tool is to clarify and assess the needs to be covered, and accordingly summarise the overview of the workflows and links between the CFs and NFs within ACTRIS.

Each DC unit describes the tools selected for their units to be implemented and tested. A report on the pilot tests will be delivered in the next step in *MS 21: Access to first version of web-based NF-TC-DC workflow management tool*, which is due in M24.

2 Workflow types in ACTRIS data production

2.1 General considerations on workflow tool selection

The concept of describing the tasks in a process, their order, connections, and dependencies, in a workflow is not unique to data management. Workflows can be used to systematically describe processes as diverse as assembling a car on an assembly line, purchasing an aircraft ticket involving a travel agency and an airline, or taking up a loan for a house with a bank. The tools suitable for executing a workflow will therefore depend on the entities handled or produced by the workflow, the entities needed for or involved in the process, and the question where the workflow will be executed.

For workflows in data management, the questions relevant for selecting tools for implementing a workflow are similar:

[1] https://github.com/actris/data-management-plan/blob/master/DMP/ACTRIS-DMP.md#8-Appendix

ACTRIS IMP (www.actris.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-INFRADEV-2019-2, Grant Agreement number: 871115

- What type of data items are handled? Specialised solutions exist e.g. for office documents in a business workflow, as opposed to data items processed through a scientific, single-user data analysis workflow.
- Who are the agents handling the data items? Examples include humans and pieces of software.
- Are data items handled by only one type of agent or several, e.g by humans (manual handling), software (automatic handling), or both (hybrid workflow)? In the latter case, the workflow tools need to be able to organise asynchronous handling, i.e. steps need to wait for the next agent to become available.
- Is the workflow executed in one place, or at several places? Does that mean executions of the same workflow in several places, or are the tasks within the same workflow execution distributed over several places?
- Do all entities involved in the workflow need to be identified by persistent identifiers? Do humans involved in task execution need to be authenticated?
- Does the tool need to document provenance, i.e. record which identified entities are involved in the execution of each workflow step as agents, input, or output?
- Is the whole use case similar to existing use cases, i.e. are there workflow tools that can be re-used?

In addition, aspects important for software development projects in general need to be considered:

- Is it a standard or a specialised use case?
- For a standard use case, which solutions exist already? If possible, standard solutions should be preferred since they can't be reproduced efficiently.
- For a use case requiring a customised solution, the software should be developed in-house. Customised solutions by commercial providers are expensive to maintain since provider charges hourly rates for all modifications.
- Customised solutions should be assembled from existing technologies in a modular way. With this architecture, it is easier to maintain and upgrade components individually. Open-source solutions can reduce cost.
- Choose technologies with a long expected lifetime and a large user community to optimise support.
- Choose technologies with in-house expertise available, or very mature ones with well-developed documentation ("black box").

An analysis of the data production workflows at all ACTRIS DC units resulted in a classification of 4 workflow types:

1. (Semi-)automatic data production for online instruments

Data production of instruments delivering a continuous data stream with little human interaction, e.g.only for execution of QC / QA measures (calibrations).

2. Data production for offline instruments

Workflow organized around handling a sample medium, e.g. a filter sample (preparing medium, exposing medium, extracting medium, analysing sample, producing result). Each sample handling step is documented by a machine-actionable protocol.

3. Documentation of QA / QC measures

QA / QC measures include the following cases: 1) Calibration of instruments on-site at the NF; 2) Calibration of instruments off-site at the TC; 3) Round-robin lab intercomparison organized by TC, i.e. a prepared sample or travelling standard is sent out to the NF for a lab comparison; 4) QA tests made routinely at NF and verified by TC. Each QA / QC measure execution is documented by a machine-actionable protocol, with a format specific to the QA / QC measure type.

4. Semi-manual data production for intermittent instrument operation

The instrument is only operated in periods, e.g.during chamber experiments or during a campaign. The data are uploaded semi-manually via a submission portal, which triggers a (semi-)automatic data curation workflow.

The workflow types and their use cases are described in more detail in the next section.

2.2 (Semi-)automatic data production for online instruments

2.2.1 Description of workflow type

The data are produced by a (mostly) continuously operating online instrument. The raw data are collected at the NF, joined with relevant discovery and use metadata, and uploaded to the topical DC unit. Production of the final data product needs to take into account QC (automatic or manual), instrument calibrations when transferring raw observations to targeted parameters, and a well-defined structure of data levels and processing steps between them to ensure traceable data production with documented provenance. Humans involved in manual workflow tasks need to be identified the same way as all other entities involved in workflow execution. The workflow is of hybrid type, i.e combines software controlled with manual tasks, which need to interact in an asynchronous manner.

This workflow type occurs at these ACTRIS DC units:

- ARES (production of aerosol profiling data)
- CLU (production of cloud profiling data)
- GRES (production of trace gas remote sensing data)
- IN SITU (production of online in situ aerosol particle & trace gas observations)

2.2.2 Recommendation of workflow tools

The following types of tools are recommended to handle this workflow:

- **Dedicated, instrument type & model specific data acquisition software:**
  Transfers data provided by instrument into well-defined format, attaches discovery and use metadata, stores data locally, displays data graphically for operator, and uploads data to DC unit in real-time. Instrument specific since parameters provided vary between instrument type and model.
- **Management platform for automated workflows**:
  Orchestrates software for automatic workflow tasks, i.e. starts a task when the result of the previous step becomes available, records provenance, interacts with tool for manual workflow tasks, informs operator about errors.
- **Issue tracker interacting with automated workflow, for handling human interaction if needed:**
  Informs humans responsible for manual workflow tasks about workflow executions waiting for their action, identifies human, records provenance of task execution, interacts with tool orchestrating automated workflow tasks.

Technologies recommended:

- **Management platform for automated workflows:**
  Apache Airflow is an open-source workflow management platform with a large user community. It is written in Python (modern and widely-used language) and uses the principle of “configuration by code”, which opens for use of extensions, e.g. for documenting provenance. It offers functions for upscaling workload.
- **Issue tracker for human interaction:**
  MantisBT and Bugzilla are free and open-source, web-based issue tracking systems with significant user communities and more than 20 years of project history. Both allow for customised workflows, and can interact with external systems, e.g. the automated workflow orchestration, via an API.

2.3 Data production for offline instruments

2.3.1 Description of workflow type

The data collection workflow is organised around handling a sample medium, e.g. a filter sample or an air sampling container. Handling the sample medium includes the steps of 1) medium preparation; 2) medium exposure; 3) sample preparation; 4) sample analysis, followed by steps for assembling and QC of the sample result. Each sample handling step is documented by an identified, machine-actionable protocol, followed by a hierarchy of data levels ensuring traceable data production with documented provenance. Humans involved in manual workflow tasks need to be identified the same way as all

ACTRIS IMP (www.actris.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-INFRADEV-2019-2, Grant Agreement number: 871115

other entities involved in workflow execution. The workflow is of hybrid type, i.e combines software controlled with manual tasks, which need to interact in an asynchronous manner.

This workflow type occurs at these ACTRIS DC units:

- IN SITU (offline aerosol & trace gas observations)
- ASC (sampled observations at chambers)

2.3.2 Recommendation of workflow tools

The following types of tools are recommended to handle this workflow:

- **Configurable, machine-readable data-serialization format for protocols, matching database for handling and archiving protocols**
  A data serialization format allows to structure pieces of information into defined concepts with defined vocabulary, including a hierarchy, all in a machine-actionable way. A collection of such items, e.g a protocol type, can be defined as a template, called a schema. For data production of ACTRIS offline observations, schemas are specific for each observation type and each handling step. A data serialization format comes with a matching database architecture.
- **Dedicated, instrument type & model specific data acquisition software:**
  Transfers data provided by instrument into well-defined data serialization format, including discovery and use metadata, stores data locally, displays data graphically for operator, and uploads data to DC unit. Instrument specific since parameters provided vary between instrument type and model.
- **Web-framework for manual authoring of sample handling protocols:**
  In case direct interaction with an instrument isn't possible for collecting the information for a sample handing protocol, it needs to be entered manually via a web-interface. The choice of the web-framework depends on the expertise available at the DC units or TC.
- **Management platform for automated workflows**:
  Orchestrates software for automatic workflow tasks, i.e. starts a task when the result of the previous steps, i.e. all protocols handling a sample, become available, records provenance, interacts with tool for manual workflow tasks, informs operator about errors.
- **Issue tracker interacting with automated workflow, for handling human interaction if needed:**
  Informs humans responsible for manual workflow tasks about workflow executions waiting for their action, identifies human, records provenance of task execution, interacts with tool orchestrating automated workflow tasks.

Technologies recommended:

ACTRIS IMP (www.actris.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-INFRADEV-2019-2, Grant Agreement number: 871115

- **Data serialisation format with matching database**
  Common data serialisation formats include the Extensible Markup Language (XML) and the JavaScript Object Notation (JSON). Matching databases include BaseX (for XML) and MongoDB (JSON).
- **Management platform for automated workflows:**
  Apache Airflow is an open-source workflow management platform with a large user community. It is written in Python (modern and widely-used language) and uses the principle of “configuration by code”, which opens for use of extensions, e.g. for documenting provenance. It offers functions for upscaling workload.
- **Issue tracker for human interaction:**
  MantisBT and Bugzilla are free and open-source, web-based issue tracking systems with significant user communities and more than 20 years of project history. Both allow for customised workflows, and can interact with external systems, e.g. the automated workflow orchestration, via an API.

2.4 Documentation of QA / QC measures

2.4.1 Description of workflow type

QA / QC measures occurring during ACTRIS data production cover 4 cases: 1) Calibration of instruments on-site at the NF; 2) Calibration of instruments off-site at the TC; 3) Round-robin lab intercomparison, sending prepared samples or travelling standards to NFs, organized by a TC; 4) QA tests made routinely at NF and verified by TC. Similar to documenting handling of offline samples, each execution of a QA/QC measure is documented by an identified, machine-actionable protocol Humans involved in executing the QA/QC measure need to be identified the same way as all other entities involved in workflow execution. The workflow is of semi-manual type. Protocols can be filled in manually via a web-interface, or by a software in case steps can be automated. Protocols are to be identified, archived in a suitable central database, and linked to the data. The protocol schema depends on type of the QA/QC measure and the instrument type.

This workflow type occurs at these ACTRIS DC units:

- IN SITU (calibration of online & offline in situ aerosol, trace gas, and cloud observations)
- ARES (calibration of aerosol lidars and photometers & QA tests made at NFs)
- CLU (calibration of cloud radars and microwave radiometers)
- ASC (calibration of chamber instruments)

2.4.2 Recommendation of workflow tools

The following types of tools are recommended to handle this workflow:

- **Configurable, machine-readable data-serialization format for protocols, matching database for handling and archiving protocols**
  A data serialization format allows to structure pieces of information into defined concepts with defined vocabulary, including a hierarchy, all in a machine-actionable way. A collection of such items, e.g a protocol type, can be defined as a template, called a schema. For data production of ACTRIS offline observations, schemas are specific for each observation type and each handling step. A data serialization format comes with a matching database architecture.
- **Dedicated, instrument type & model specific data acquisition software:**
  In case the QA/QC measure can be automated, the controlling software collects the relevant protocol items, displays data graphically for operator, and uploads protocol to the DC unit. Instrument specific since QA/QC measures vary between instrument type and model.
- **Web-framework for manual authoring of QA/QC measure protocols:**
  In case direct interaction with an instrument isn't possible for collecting the information for a QA/QC measure protocol, it needs to be entered manually via a web-interface. The choice of the web-framework depends on the expertise available at the DC units or TC.
- **Management platform for automated workflows**<:
  Controls automatic tasks such as linking of QA/QC measure protocols to data and taking into account QA/QC measure results in data production.
- **Issue tracker interacting with automated workflow, for handling human interaction if needed:**
  Informs humans responsible for manual workflow tasks about workflow executions waiting for their action, identifies human, records provenance of task execution, interacts with tool orchestrating automated workflow tasks.

Technologies recommended:

- **Data serialisation format with matching database**
  Common data serialisation formats include the Extensible Markup Language (XML) and the JavaScript Object Notation (JSON). Matching databases include BaseX (for XML) and MongoDB (JSON).
- **Management platform for automated workflows:**
  Apache Airflow is an open-source workflow management platform with a large user community. It is written in Python (modern and widely-used language) and uses the principle of "configuration by code", which opens for use of extensions, e.g. for documenting provenance. It offers functions for upscaling workload.
- **Issue tracker for human interaction:**
  MantisBT and Bugzilla are free and open-source, web-based issue tracking systems with

significant user communities and more than 20 years of project history. Both allow for customised workflows, and can interact with external systems, e.g. the automated workflow orchestration, via an API.

2.5 Semi-manual data production for intermittent instrument operation

2.5.1 Description of workflow

In these use cases, instruments deliver data only for dedicated periods. This may include observations which otherwise would be considers online or offline. This workflow covers observations conducted at atmospheric simulation chambers, but also on mobile platforms and during dedicated measurement campaigns. These data will be curated by feeding them into existing ACTRIS data production workflows. To this end, data have to be formatted by the data providers according to well-defined templates joining the data with discovery and use metadata. The data are submitted through a dedicated portal which gives the provider interactive feedback on syntax and semantic issues. Increased curation effort is expected for these data since many observations don't have established data reporting routines, and vocabulary describing them may need to be defined. QA / QC procedures will also follow existing routines wherever possible.

This workflow type occurs at these ACTRIS DC units:

- ASC (data collected at chambers by online / offline instruments)
- IN SITU (campaign service)

2.5.2 Recommendation of workflow tools

The following types of tools are recommended to handle this workflow:

- **Web-framework for interactive data submission portal:**
  The web-portal will provide the data submitter with interactive feedback on any syntax or semantic issues with the data submission. The choice of the web-framework depends on the expertise available at the DC units or TC.
- **Management platform for automated workflows**<:
  Controls automatic tasks such as picking up data from the data submission portal and feeding them into the curation and QA / QC procedure.
- **Issue tracker interacting with automated workflow, for handling human interaction if needed:**
  Informs humans responsible for manual workflow tasks, e.g. data curation and QA / QC, about workflow executions waiting for their action, identifies human, records provenance of task execution, interacts with tool orchestrating automated workflow tasks.

ACTRIS IMP (www.actris.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-INFRADEV-2019-2, Grant Agreement number: 871115

Technologies recommended:

- **Management platform for automated workflows:**
  Apache Airflow is an open-source workflow management platform with a large user community. It is written in Python (modern and widely-used language) and uses the principle of “configuration by code”, which opens for use of extensions, e.g. for documenting provenance. It offers functions for upscaling workload.
- **Issue tracker for human interaction:**
  MantisBT and Bugzilla are free and open-source, web-based issue tracking systems with significant user communities and more than 20 years of project history. Both allow for customised workflows, and can interact with external systems, e.g. the automated workflow orchestration, via an API.

The details of this workflow type need to be specified further during the implementation process.

ACTRIS IMP (www.actris.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-INFRADEV-2019-2, Grant Agreement number: 871115

3 Workflow tools in the ACTRIS DC units for pilot implementation and testing

3.1 In Situ data centre unit (In-Situ)

3.1.1 Workflow types

The workflow of the ACTRIS In Situ DC unit contains the following workflow types discussed above:

1. The online workflow branch of the ACTRIS In Situ DC unit corresponds to the "(semi-)automatic data production for online instruments" workflow type discussed above. It is used for instruments providing continuous data streams such as integrating nephelometer, filter absorption photometer, mobility particle size spectrometer, chemiluminescence photometer, but also observations of volatile organic compounds. ACTRIS In Situ online data production includes branches for providing real-time and fully, manually QCed data products.
2. The offline workflow branch of the ACTRIS In Situ DC unit corresponds to the "data production for offline instruments" workflow type discussed above. It is used for handling filter samples of aerosol particles for observations of organic and elemental carbon, organic tracers, but also elemental composition, but also chemical analysis of cloud water samples.
3. The QA / QC measures branch of the ACTRIS In Situ DC unit corresponds to the "documentation of QA / QC measures" workflow type discussed above. For ACTRIS In Situ, QA / QC measures occur for regular routine calibrations of instruments at NF stations, calibrations of instruments sent to the corresponding TC, and for NF laboratory intercomparisons organised by a TC sending out prepared samples.
4. The campaign service branch of the ACTRIS In Situ DC unit corresponds to the "Semi-manual data production for intermittent instrument operation" workflow type discussed above. It is used for all data reaching the ACTRIS In Situ DC unit outside the regular data production pathways.

3.1.2 Workflow tools

3.1.2.1 Online branch

The In Situ online data production starts with data acquisition at the NF station, conducted with a software package provided by the TC responsible for the observation formatting and uploading the data to In Situ DC . Solutions vary by instrument type, with a Python-based software used for aerosol physics and optics instruments, and a vendor provided software for particle chemical speciation. At In Situ DC, 2 data products are produced from the incoming data, a real-time product and a fully manually QCed data product. The individual workflow tasks are performed with software provided by the TC. So far, Python and R have been used as environment for this software. The execution of automatic workflow tasks will the orchestrated by Airflow operated at In Situ DC, including documentation of provenance using the Python PROV-O package. Manual workflow tasks needed for producing manually

QCed data are scheduled for the responsible human in a MantisBT issue tracker interacting with Airflow. Software tools assisting in manual QC are also provided by the TC, and orchestrated by Airflow.

3.1.2.2 Offline branch

The ACTRIS In Situ offline workflow branch will use a recommended data serialisation format with corresponding database for storing handling protocols of offline samples. The database will be located at In Situ DC unit. Both software packages interfacing instruments and web-interfaces will be used to produce the protocols. Software packages as well as protocol schemas are defined and provided by the TC, with advice by In Situ DC. A list of protocol schemas is provided in the Appendix. From the sample handling protocols, TC specified and provided software will produce the final product, automatically orchestrated by Airflow. Manual steps in this process, e.g. manual QC, is scheduled in the MantisBT issue tracker.

Due to their high level of maturity, observations of particulate organic and elemental carbon serve as pilot for setting up the ACTRIS In Situ offline data workflow. A series of protocols have been defined jointly by CAIS/ECAC and the DC to document the 4 (5) main processes that lead to the final data product: medium preparation, medium exposure, sample preparation, sample analysis (and data value calculation). These protocols were developed for OC&EC analyses as a pilot, but their applicability to other off-line analyses under ACTRIS was considered. They comprise all the information needed to assess the quality of the final data product, and its uncertainty. Some fields are to be filled by humans, others are to be filled automatically by reading files from e.g. the computer dedicated to the analytical instrument.

Further interaction between In Situ DC unit and responsible TCs is necessary for defining the content of the listed protocol templates.

3.1.2.3 QA / QC measures

Similar to the ACTRIS In Situ offline workflow branch, protocols for QA / QC measures will use a recommended data serialisation format with corresponding database for storing protocols of QA / QC measures. The database will be located at the In Situ DC unit. Both software packages interfacing instruments and web-interfaces will be used to produce the protocols. Software packages as well as protocol schemas are defined and provided by the TC, with advice by In Situ DC. A list of protocol schemas is provided in the Appendix.

Further interaction between In Situ DC and responsible TCs is necessary for defining the content of the listed protocol templates.

3.1.2.4 Campaign service

ACTRIS IMP (www.actris.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-INFRADEV-2019-2, Grant Agreement number: 871115

The ACTRIS In Situ campaign service will be based on the data submission and curation service used at ACTRIS In Situ today. The data, pre-formatted according to templates by the data provider, are received through a data submission portal, where the portal, written in Python Flask, gives interactive feedback on syntax and semantic issues. The subsequent data curation and QA/QC workflow, both within In Situ DC and towards NF station and TC, is handled and documented by a MantisBT issue tracker.

3.2 Aerosol remote sensing data centre unit (ARES)

3.2.1 Workflow types

The workflow of the ACTRIS ARES DC Unit contains the following workflow type discussed above:

1. **(semi-)automatic data production for online instruments**. It is used for lidar and photometer stations providing continuous data streams of observations of aerosol properties. ACTRIS ARES online data production involves lidar and photometer stations, the Single Calculus Chain for centralized data processing, CARS for approving configurations, and the NRT provision of fully QC lidar data products.
2. The QA / QC measures branch of the ACTRIS ARES unit corresponds to the "**documentation of QA / QC measures**" workflow type discussed above. For ARES, QA measures on lidars occurs routinely at the NF as local test. Results of the test are sent to the TC for their analysis, then the outcomes are inserted in the processing configurations. Additionally, calibration/intercomparison with reference TC systems are performed in case of need. QC procedures on the data are performed centrally and automatically at ARES. For photometer QA is done through calibration at TC. Consequently, the QC on the data is done keeping trace of it and it is done at DC level.

3.2.2 Workflow implementation

3.2.2.1 Standard processing

Lidar acquisitions at the NF stations are automatically submitted to the lidar processing part of ARES, namely the Single Calculus Chain. Submission can be done manually through the SCC Web Interface or automatically through a Python-based API. The submitted data should follow the data structure and format defined by SCC. Submitted data are processed by the SCC according to the specific lidar configurations defined by the stations and approved by CARS. Data products generated by the SCC are analysed with tools for automatic quality control and submitted to the ACTRIS Data repository.

At the same time, photometer data are sent automatically to ARES and processed. In case of aerosol optical properties obtained by lidar observations are simultaneous to photometer observations, the integration of lidar and photometer data is done at ARES DC unit.

ACTRIS IMP (www.actris.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-INFRADEV-2019-2, Grant Agreement number: 871115

Link with CLU processing is planned and partially already in place for the cloud masking and for atmospheric model use.

Currently, some of the described steps are not fully automated, but they will be to guarantee an efficient interaction between the parties. Tools are needed for:

- automatic submission of measurement data from NFs to ARES monitoring the performance of the instruments and automatic submission of the test data (NFs and CARS)
- procedures for approval of configurations (NFs, CARS and ARES)
- automatic quality control of the data products
- feedback to NFs and TC about QC

ARES is currently adopting Bugzilla as an issue tracker for human interaction.

3.2.2.2 QA/QC measures

QA procedures for the Aerosol Remote Sensing NFs can be listed as follows:

- monitoring tools of system performances at NF (for checking the stability of equipment) (to be developed)
- periodic submission of Quality Assurance test data to CARS
- intercomparison of systems with reference ones (periodic for photometers, and on base of specific and motivated needs for aerosol lidar)

Based on such tests/procedures, each NF proposes lidar configurations to be used for the processing. CARS (and ARES) provides feedback to the stations and approved lidar configurations for running the Single Calculus Chain.

QC on the data are done at ARES level in automatic way. Centralized QC are done for checking simultaneous different products provided by SCC. Further QC are done automatically during uploading on the ARES data repository.

For photometer QA is done through calibration at TC. Consequently the QC on the data is done keeping trace of it and it is done at DC level.

3.2.2.3 Campaign service

The campaign service for ACTRIS ARES follows the same procedures as described for NF stations; automated data transfer to ARES , centralized processing at ARES , and QA/QC applied following ARES and CARS standards.

ACTRIS IMP (www.actris.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-INFRADEV-2019-2, Grant Agreement number: 871115

3.3 Cloud remote sensing data centre unit (CLU)

3.3.1 Workflow types

The general workflow for the ACTRIS CLU DC Unit consists of the following workflow type discussed above:

1. (semi-)automatic data production for online instruments. All cloud profiling stations provide continuous observation data streams from cloud radar, lidar/ceilometer, microwave radiometer and disdrometer (together with Doppler lidar observations from certain stations). These data streams are then combined to create the cloud profiling products, with NRT provision from all stations. The data production for ACTRIS CLU involves the cloud profiling NFs (all data stream transfers are automated), CLU (centralised processing of all data streams) and CCRES (for providing calibration values).

3.3.2 Workflow implementation

3.3.2.1 Standard processing

For cloud profiling, the standard processing is fully automated with respect to data collection, data transfer, data processing and data production. The majority of QA/QC is also automated except for manual calibration of specific instruments, and final data product curation. After data transfer via API to CLU, all automated tasks are performed within CLU, under guidance from CCRES, whereas manual calibration tasks may be performed by NF or CCRES. Calibration values are stored in a timeseries database for extraction via API; entry may be automated or manual. Instrument performance monitoring is also being developed by CCRES, for implementation at CLU and at NFs. CLU, CCRES, and NFs are testing Apache Airflow together with in-house created workflow tools to manage these workflows.

An issue tracker software for human interaction will be explored based on experience in ARES and In-Situ, where this under implementation and in operation (In-Situ).

3.3.2.2 Instrument calibration

The calibration activities within the QA/QC process utilize three types of the calibration workflow described above. For cloud-profiling, it is more convenient to describe the workflow processes in terms of automated calibration, manual calibration, and instrument intercomparison as these give a more succinct description of the different workflows involved:

1. **Automatic calibration.** This activity is performed at CLU during centralised processing for all instrument types, although for certain instruments, the calibration values may be superseded by those provided via the manual calibration activities.
2. **Manual calibration**. Routine manual calibration at specified intervals, typically <6 months, is performed for the cloud radar and the microwave radiometer. The calibration values produced by these activities are uploaded manually to the calibration database by the NF, following the protocols developed by CCRES and CLU.

ACTRIS IMP (www.actris.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-INFRADEV-2019-2, Grant Agreement number: 871115

3. **Instrument intercomparison.** There are two forms of instrument intercomparison. For cloud radars, the CCRES facilities operate instruments alongside a suite of calibration equipment, including masts, drones, and reference radars. NF-operated instruments can visit one of the CCRES facilities in order to be calibrated using one of the standard techniques (mast, drone, etc.) and against the reference radar. CCRES can also utilise the reference radar as a travelling standard. The reference instrument will visit an NF and operate alongside the local instrument. This activity therefore combines attributes from both **Calibration of instruments off-site at the TC** and **round-robin instrument intercomparison**. Manual update of the calibration database will be scheduled in the same manner as for the workflow described for manual calibration activities performed by the NF.

3.3.2.3 Campaign service

The campaign service for ACTRIS CLU follows the same procedures as described for NF stations; automated data transfer to CLU, centralised processing at CLU, and QA/QC applied following CLU and CCRES standards. Note that manual calibration procedures may not be possible to follow in some locations due to logistical issues, and therefore, instrument intercomparison and manual calibration may take place before or after the instrument campaign deployment.

3.4 Trace gases remote sensing data centre unit (GRES)

3.4.1 Workflow types

The workflow of the ACTRIS GRES DC Unit (DMP document, Appendix 6) consists of the following workflow type discussed above:

1. **(Semi-) automatic data production for online instruments.** All trace gas remote sensing stations provide continuous observations data streams from FTIR, Lidar-DIAL, Pandora, Max-Doas and SAOZ instruments. Each data stream creates its own trace gas remote sensing products, with NRT provision from all stations. ACTRIS GRES online data production involves all trace gas remote sensing stations at NFs (all data stream transfers are automated), a central data processing system (one per data stream) for data production, a CREGARS check unit for the last final QA/QC check and the provision of full trace gas remote sensing products.

2. **Documentation of QA / QC measures**. For GRES, QA measured on the five instruments occurs routinely at the NF as local test. Results of the test are sent to the TC for their analysis and inserted in the processing configurations. In addition, calibration / intercomparison with reference TC systems are performed in case of need. QC procedures on the data are performed centrally and automatically at the CREGARS unit check. Consequently, the QC on the data is done keeping trace of it and it is saved at GRES DC level.

3.4.2 Workflow implementation

ACTRIS IMP (www.actris.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-INFRADEV-2019-2, Grant Agreement number: 871115

3.4.2.1.1 Standard processing

FTIR, Lidar-Dial, Pandora, MAX-DOAS and SAOZ acquisitions at the NF stations are automatically submitted using a “rsync” or “ftp” protocol to a dedicated central data processing system (CDPS): one per data stream. Note that Lidar Dial and SAOZ CDPS are a part of GRES. The data submitted are processed by the dedicated CDPS: FTIR and MAX-DOAS, BIRA in Belgium; PANDORA, Luftblick in Austria, Lidar-DIAL and SAOZ, GRES DC in France according to the instrument configuration defined by the stations and CREGARS. Data products (L0 to L2B data) are pushed from CDPS to GRES unit via a next cloud interface for data curation, the QC of the level 2B data are automatically done using the CREGARS check unit then the L2B data are submitted to ACTRIS Data repository.

Currently, some of the described steps are not fully automated, but we are going to improve it. However, tools are needed for:

- procedures for approval of configurations between NFs, CREGARS and GRES,
- feedback to NFs and TC about QC.

GRES has already adopted a next cloud interface to documentation and data exchanges and currently is going to test Apache Airflow to automate the data workflow submitted by the four CDPS.

An issue tracker software for human interaction will be explored based on experience in ARES and In-Situ, where this is under implementation and in operation (In-Situ).

3.4.2.2 QA/QC measures

3.5 Atmospheric simulation chamber data centre unit (ASC)

3.5.1 Workflow types

Simulation chambers are equipped with a high diversity of instrumentation in concordance with their scientific topics and a number of instruments coupled to chambers are home-made or highly customized. These instruments are called “specific” instrumentation, in opposition to “base” instrumentation which is often similar to the one used by observational platforms for *in situ* measurements and for which SOPs are generally provided by TCs. As a consequence, the data workflow differs for “base” and “specific” instrumentations.

For “base” and “specific” instrumentation, the workflow of the ACTRIS ASC DC unit contains the following workflow types discussed above:

1. The online workflow branch of the ACTRIS DC unit corresponds to the “Semi-manual data production for intermittent instrument operation” workflow type discussed above. It is used for data obtained from observations conducted at atmospheric simulation chambers with online instruments.

ACTRIS IMP (www.actris.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-INFRADEV-2019-2, Grant Agreement number: 871115

2. The offline workflow branch of the ACTRIS ASC DC unit corresponds to the “data production for offline instruments” workflow type discussed above. It is used for handling sample medium such as filter sample of aerosol particles or cartridge sample of VOCs, for observations of trace gas, aerosol and cloud chemical composition at elementary, functional or molecular scales.
3. The QA / QC measures branch of the ACTRIS ASC DC unit corresponds to the “documentation of QA / QC measures” workflow type discussed above. For ACTRIS ASC, QA / QC measures occur for regular routine calibrations of instruments at NF sites, calibrations of instruments sent to the corresponding TC (for base instrumentation only), and for NF laboratory intercomparisons organised by a TC sending out prepared samples (for base instrumentation only). It also includes “check for internal consistency” which includes check for carbon balance during the course of the experiments, check for consistency between various instruments measuring the same parameters, …

In addition to the instrumental part, the data workflow includes a description of the experimental protocol for chamber operation.

3.5.2 Workflow implementation

The definition of a precise workflow is ongoing and performed by working groups involving NFs, TCs and DC. Roles and responsibilities of tasks for the data generation have been clearly defined but the definition of SOPs for each type of instrumentation is still under discussion. One difficulty relies on the diversity of the instrumentation used for simulation chamber experiments that limits the establishment of common protocols and the automation of the data treatment and QA/QC procedures, in particular for specific instrumentation. It also prevents from centralized data treatment at the DC level. The data treatment is therefore performed at the NF level following TC recommendations (for base instrumentation).

3.5.2.1 Online branch

The ACTRIS ASC online workflow branch describes the protocols for instrument operation and data treatment of online instruments performed at the NF level. For “base” instrumentation, NFs will follow guidelines provided by TCs. It concerns fundamental environmental parameters such as temperature, pressure, relative humidity, and actinic flux but also concentration–time profiles of various species such as VOCs as well as basic atmospheric contaminants including NOx, aerosol number concentration and size distribution. Data and metadata (including all the information needed to assess the quality of the final data product, and its uncertainty) will be uploaded by NFs to ACTRIS ASC unit.

For “specific” instrumentation, as there is no SOPs provided by the TCs. ACTRIS Atmospheric Simulation Chamber staff will have to implement protocols enabling internal consistency of the measurements for a posteriori validations/reanalysis of the data produced by these high-technology instruments. In this context, full traceability of the methodologies, calibrations, algorithms, and software versions is implemented by the RPO operating the facility.

ACTRIS IMP (www.actris.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-INFRADEV-2019-2, Grant Agreement number: 871115

No tool has been defined so far to manage the data generation section of the workflow and discussions are in progress jointly with NFs, TDs and DC to define standard protocols for base instrumentation. Concerning the data provision by NFs to ACTRIS ASC unit, metadata provision interface is used (REST-API) and metadata are managed through a mongodb database.

3.5.2.2 Offline branch

The ACTRIS ASC offline workflow branch describes the protocols for storage, analysis and data treatment of offline samples performed at the NF level. For “base” instrumentation, NFs will follow guidelines provided by TCs. Data and metadata (including all the information needed to assess the quality of the final data product, and its uncertainty) will be uploaded by NFs to ACTRIS ASC unit.

No tool has been defined so far to manage the data generation section of the workflow and discussions are in progress jointly with NFs, TDs and DC to define standard protocols for base instrumentation. Concerning the data provision by NFs to ACTRIS ASC unit, metadata provision interface is used (REST-API) and metadata are managed through a mongodb database.

3.5.2.3 QA/QC

QA/QC includes calibration, instrument intercomparison and check for internal consistency (see above). It is performed mainly at the NF level, following guidelines provided by TCs for base instrumentation. Instruments can also be sent to TCs for calibration.

No tool has been defined so far to manage the QA/QC part of the workflow. An issue tracker software for human interaction will be explored based on experience in ARES and In-Situ, where this under implementation and in operation (In-Situ).

4 Appendices

4.1 List of In Situ protocol templates for on-site QC measures, off-site QC measures, and offline sample handling protocols

4.1.1 Aerosol observations

Table 5.1: QA / QC protocol types required for online aerosol observations.

Instrument	Protocol type	Responsible TC / TC unit	Data types contained in protocol
National Facility	Station audit report		
CPC	On-site calibration, Off-site calibration	ECAC-WCCAP	time series with unit
nano CPC	On-site calibration, Off-site calibration	ECAC-CCC	time series with unit string x/y data array, with units
MPSS	On-site calibration, On-site audit, Off-site calibration	ECAC-WCCAP	concentration over time with unit
nano MPSS	On-site calibration, On-site audit, Off-site calibration interlaboratory comparison	ECAC-CCC	concentration over time with unit string x/y data array, with units
CCNC	Off-site calibration	ECAC-WCCAP	time series with unit
Integrating nephelometer	On-site calibration, Off-site calibration	ECAC-WCCAP	time series with unit
Absorption Photometer	On-site calibration, Off-site calibration	ECAC-WCCAP	time series with unit
Extinction monitor	Off-site calibration	ECAC-WCCAP	time series with unit
ACSM	On-site calibration,	ECAC-ACMCC	time series with unit

Instrument	Protocol type	Responsible TC / TC unit	Data types contained in protocol
	Interlaboratory comparison		

Table 5.2: QA / QC protocol types and sample protocol types required for offline aerosol observations

Instrument	Protocol type	Responsible TC / TC unit	Data types contained in protocol
National Facility	Station audit report		
OC/EC analysis	Inter-laboratory comparison On-site calibration check	ECAC-OGTAC-CC (ERLAP)	OC, EC and TC sample loadings (µg/cm²)
Mass concentration of particulate organic tracers	Interlaboratory comparison	ECAC-OGTAC-CC	mass concentration

ACTRIS IMP (www.actris.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-INFRADEV-2019-2, Grant Agreement number: 871115

4.1.2 Cloud observations

Due to an extended schedule for setting up the cloud in situ TC, the content of these tables will be determined in the next iteration,

Table 5.3: QA / QC protocol types required for online cloud observations

Instrument	Protocol type	Responsible TC / TC unit	Data types contained in protocol

Table 5.4: QA / QC protocol types and sample protocol types required for offline cloud observations

Instrument	Protocol type	Responsible TC / TC unit	Data types contained in protocol

ACTRIS IMP (www.actris.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-INFRADEV-2019-2, Grant Agreement number: 871115

4.1.3 Trace gas observations

Table 5.5: QA / QC protocol types required for online trace gas observations

Instrument	Protocol type	Responsible TC / TC unit	Data types contained in protocol
National Facility	Station audit report	CiGas	String, picture, float, x/y data array, with units
GC-MS/FID	On-site calibration	NF	String, float Time series with unit
	Off-site calibration	CiGas	String, float, x/y data array, with units
	Interlaboratory comparison	CiGas	String, float, x/y data array, with units, Time series with units, picture
	Data workshop	CiGas	string
PTR-MS/PTR-TOF-MS/CI-APi-TOF	On-site calibration	NF	String; Time series with unit
	Off-site calibration	CiGas	String, float x/y data array, with units
	Interlaboratory comparison	CiGas	String, float, x/y data array, with units, Time series with units, picture
	Data workshop	CiGas	String
CLD/CAPS	On-site calibration	NF	String; Float, Time series with unit
	Off-site calibration	CiGas	String, Float, x/y data array, with units
	Interlaboratory comparison	CiGas	String, Float, x/y data array, with units, Time series with units, picture
	Data workshop	CiGas	String, picture

Table 5.6: QA / QC protocol types and sample protocol types required for offline trace gas observations

ACTRIS IMP (www.actris.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-INFRADEV-2019-2, Grant Agreement number: 871115

Instrument	Protocol type	Responsible TC / TC unit	Data types contained in protocol
National Facility	Station audit report	CiGas	String, picture, float, x/y data array, with units
VOC-Offline-sampling + GC-MS/FID	On-site calibration	NF	String; Time series with unit
	Off-site calibration	CiGas	String, float x/y data array, with units
	Interlaboratory comparison	CiGas	String, float, x/y data array, with units, Time series with units
	Data workshop	CiGas	String

4.2 List of Aerosol Remote Sensing protocol templates for QA/ QC measures

Table 5.7: QA / QC protocol types required for online aerosol remote sensing observations

Instrument	Protocol type	Responsible TC / TC unit	Data types contained in protocol
Aerosol Lidar	On-site test	CARS -AHL	Rayleigh fit measurement Trigger delay measurements Telecover test Dark measurements
Aerosol Lidar	Off-site calibration	CARS -AHL	Rayleigh fit measurement Trigger delay measurements Telecover test

ACTRIS IMP (www.actris.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-INFRADEV-2019-2, Grant Agreement number: 871115

Instrument	Protocol type	Responsible TC / TC unit	Data types contained in protocol
			Dark measurements Optical properties measurements
Photometer	Off-site calibration	CARS -ASPL	Direct Sun calibration Radiance calibration

ACTRIS IMP (www.actris.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-INFRADEV-2019-2, Grant Agreement number: 871115