TOAR Data Technical Guide #2

TOAR Data Input and Processing

Version 1.0 | 30 August 2021

Forschungszentrum Jülich GmbH
ESDE | JSC - FSD
52425 Jülich

Document Status

Created by:	Sabine Schröder, Niklas Selke, Jianing Sun	30 August 2021
Reviewed and approved by:	Martin Schultz	31 August 2021
Released by:	Mathilde Romberg	01 September 2021

Revision History

Version	Date	History
1.0	30 August 2021	Initial version

Content

List of Figures

List of Tables

1 Introduction

The Tropospheric Ozone Assessment Report (TOAR) activity of the International Global Atmospheric Chemistry (IGAC) organisation (see https://igacproject.org/activities/TOAR) is collecting surface ozone measurements and related data from all over the world in a central database at Forschungszentrum Jülich, Germany. It currently runs in its second phase, TOAR-II 2020-2024, https://igacproject.org/activities/TOAR/TOAR-II (in our context denoted as TOAR V2, [1]). The purpose of collecting this data is to provide globally consistent metrics for analyses of health, vegetation, and climate impacts from ozone air pollution. The database is exposed via a REST API and graphical web services which allow users to visualise data and compare them with other data sets and model observations. We collect data from cooperating data centres, harmonise it, and check its quality before adding it to the TOAR database.

The TOAR database infrastructure aims to provide complete coverage of worldwide surface ozone and related measurements. This means that data from several dozen sources (government agencies, research institutions, NGOs) are assembled together. Technically, we distinguish between individually contributed data, which is submitted to us by email or via a shared folder (usually a small number of files) and harvested data, which we obtain from publicly accessible sources (web services, data downloads, or database dumps made available to us). In a few cases in TOAR-I we also received large contributions (e.g. ~1000 files of Japanese ozone monitoring data). Such massive contributions are treated like harvested data in the workflows described below.

This technical guide describes the processing steps applied to the incoming data for the TOAR database. Note that not every processing detail can be described here, because the harmonisation of the many different data sources naturally means that many individual decisions must be taken on a day by day basis. While other data centres often enforce relatively strict rules for data providers and only accept data which has been processed according to their rules, it is a core objective of TOAR to collect data also from world regions with low data coverage and limited data processing capabilities. This implies that, especially for individually contributed data, a lot of communication takes place between the data providers and the TOAR data centre. Responsible persons and data formats may change, metadata profiles can be altered over time, and often very specific questions need to be sorted out with data providers before we can bring new data online. Nevertheless, we have tried to structure, organise and automate our data processing workflow as much as possible, not least to fulfil high standards with respect to the data documentation and reproducibility.

[1] TOAR phase I ran 2014-2019; in our context it is TOAR V1

2 TOAR Data Sources

Table 1 summarise sources of the data in the TOAR database. Note that it describes the data sources which provided surface ozone data to the first phase of the Tropospheric Ozone Assessment Report. While most of these data are re-used in TOAR-II, there may be some changes because of data licensing issues (TOAR-II declared to make all data available under a CC-BY 4 license, but not all data providers have granted permission to do so), or because the responsible agency for providing the data has changed. We hope and expect that several new sources of data can be added to this compilation over the next 1-2 years. A major new data collection effort for TOAR-II will occur during the year 2022, and this document shall be updated accordingly after completion[2].

Table 1: Summary of data sources in the TOAR database. This table only contains datasets encompassing about 10 stations. The TOAR database contains about 136 additional data sets which do not originate from one of the data providers listed below.

Data source	No. of stations	Variables	Type of submission	Upload frequency
Global				
GAW World Data Centre for Greenhouse Gases (WDCGG)[3]; GAW World Data Centre for Reactive Gases (WDCRG)[4] ()	218	benzene ch4 co ethane humidity irradiance ozone press propane relhum rn temp toluene totprecip	download individual files	manual update planned for 2022
Africa				
South African Network	24	ozone	submitted individual files	no updates
North America				
Government of Canada's CaPMon -Air quality monitoring networks and data[5]	19	ozone	download	no update loaded
U.S. Environmental Protection Agency (EPA), Clean Air Status and Trends Network (CASTNET)[6]	117	ozone	download	unregularly, on request

[2] More detailed documentation on the processing of data from specific sources is available to administrators on an internal Wiki page (https://gitlab.jsc.fz-juelich.de/esde/toar-data/toar-db-data/-/wikis/home#toar-database-new-version; internal use only).

[3] https://gaw.kishou.go.jp/

[4] https://www.gaw-wdcrg.org/

[5] www.canada.ca/en/environment-climate-change/services/air-pollution/monitoring-networks-data/canadian-air-precipitation.html

[6] https://java.epa.gov/castnet/reportPage.do

Data source	No. of stations	Variables	Type of submission	Upload frequency
U.S. Environmental Protection Agency (EPA), Air Quality Data (AQS)[7]	2963	u v humidity irradiance press totprecip temp ozone	download of database dumps	manual update; next planned for 2022/Q2
Environment and Climate Change Canada, National Air Pollution Surveillance (NAPS) program[8]	373	ozone	download of database dumps	manually; next update planned for 2022/Q1
South America				
Colombia air quality network	16	co irradiance no no2 ozone pm10 press relhum temp totprecip wdir wspeed	submitted	irregularly
Asia				
Acid Deposition Monitoring Network in East Asia (EANET)[9]	16	humidity irradiance ozone press temp totprecip u v	download individual files	manual update planned for 2022/Q1
Israeli air quality network (ISRAQN)	12	ozone	submitted individual files	no updates
South Korean National Institute Of Environmental Research (NIER)[10]	316	ozone	submitted individual files	No updates
National Institute for Environmental Studies Japan[11] (NIES)	1391	ozone ox	download individual files	manual update planned for 2022/Q1

[7] https://www.epa.gov/outdoor-air-quality-data

[8] https://www.canada.ca/en/environment-climate-change/services/air-pollution/monitoring-networks-data/national-air-pollution-program.html

[9] https://www.eanet.asia/

[10] https://www.nier.go.kr/eng/

[11] https://www.nies.go.jp/igreen/

Data source	No. of stations	Variables	Type of submission	Upload frequency
OpenAQ[12]	12078	co no2 ozone so2 bc	download	daily
Teheran air quality network[13]	20	ozone	submitted individual files	no updates
Australia				
Australian air quality network	56	ozone	contributed in batches via email	no updates
Europe				
European Environment Agency's Airbase[14]	6459	co humidity irradiance no no2 nox ozone press temp totprecip u v	download of database dumps	manually; next update planned for 2022/Q1
European Monitoring and Evaluation Programme (EMEP)[15]	196	irradiance ozone press relhum temp wdir wspeed	web harvesting	manual update, next planned 2022/Q2
German Environment Agency's Air Data (UBA)[16]	1004	benzene co no no2 ozone pm10 pm2p5 press relhum so2 temp toluene wdir wspeed	receipt of database dumps (validated data) and automated NRT harvesting	annual update (next: 2022/Q3) of validated data

[12] https://openaq.org/

[13] http://airnow.tehran.ir/home/DataArchive.aspx

[14] https://www.eea.europa.eu/data-and-maps/data/aqereporting-2

[15] http://ebas.nilu.no/DataSets.aspx

[16] https://www.umweltbundesamt.de/en/data/air/air-data

Data source	No. of stations	Variables	Type of submission	Upload frequency
Institute of Geosciences, Dept. Meteorology, University of Bonn, Germany (MIUB) [all european stations as of 2016]	Number of time series: 63369	cloudcover pblheight humidity press relhum temp totprecip u v	submitted csv data files; special processing not further described here.	no updates

3 The TOAR Data Processing Workflow

In general, our workflow for data ingestion consists of three major parts (Figure 1). The three building blocks are described in more detail in the following sub sections. The principal task of the data ingestion workflow is to convert raw data from different formats into harmonised data, load the data into the TOAR database, and publish the new data. The details of the different processing steps depend on the nature and format of the incoming data, especially during the manual pre-processing steps. The overall objective of this first step is to eliminate any formatting and other errors from the raw data, which would cause the main automated processing chain to fail, and to configure the subsequent automated processing steps. The goal of the automated processing is to store the new data in the database so that it can be reviewed and published. The actual data review (for individually submitted data only) is part of the semi-automated post-processing, which also includes publication of the data and an announcement in the TOAR data portal news.

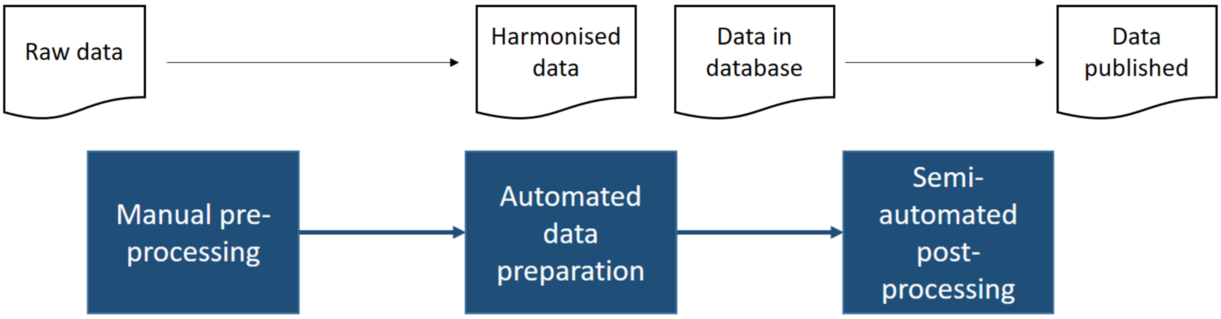


Figure 1: General structure of the TOAR data processing workflow

The basic data model of the TOAR database is the time series, which contains the recorded values of an atmospheric variable from one specific stationary location (a measurement station) during a certain time period. Each data record consists of a time series id, a timestamp, the measured, or in some cases simulated, data value, a data quality flag and a version tag (see TOAR_UG_Vol03_Database). Data records are bound together via the time series id, which unambiguously identifies one specific variable record at one geographical location. Note that it is possible to have more than one time series of the same variable at one station, for example if two different measurement techniques were used, if one of the records stems from a numerical simulation, if the same measurements were contributed through different monitoring networks[17], or if different filtering or interpolation algorithms have been applied to a time series, for example to eliminate polluted conditions at a clean air site. To identify time series and stations unambiguously, the TOAR database holds extensive metadata on the time series and the stations. The general procedure for importing external data is to map these fundamental parameters to the existing data in the TOAR database by applying several consistency checks and heuristic algorithms as described below.

The individual data processing scripts can be found at https://gitlab.jsc.fz-juelich.de/esde/toar-data/toar-db-data/-/tree/master/toar_v1/database_import/recently_used/INDIVIDUAL/scripts.

3.1 Manual pre-processing

Step 1: Archive raw data files

Before anything is done with the data received or downloaded, a backup copy of the data is made. Additional backup copies to a tape archive are made regularly as described in TOAR_TG_Vol01_Infrastructure.

[17] one should hope in this case, that all records would be identical, but this is unfortunately not the case in practice. See discussion in [1]

Step 2: Select Target Database

Depending on the type of data and the data provider (small number of individual data submissions versus database dumps or web-harvested data) the database for storing data is different – either data is stored temporarily in a staging database for review by the data provider or it is directly added to the TOAR database.

Step 3: Manual pre-screening

A quick visual inspection of the data files is made to ensure that the data formats and metadata content are compatible with the automated processing chain. Typically, the inspection includes aspects such as:

Test 3.1: Is the file in ASCII format (not zip, netCDF, image, etc.)?

Test 3.2: Does the file contain numbers roughly corresponding to expected values?

Test 3.3: Does the file contain date and time information in a legible format?

Test 3.4: Does the file contain the mandatory keywords (or is there a separate stations file with metadata on the measurement locations)?

Test 3.5: Is the naming of the file consistent with the TOAR file naming convention?

Depending on the nature of possible errors and the data provider, small corrections to the data formats or spelling of metadata will be made, or the data are sent back to the provider with a request to apply the necessary corrections. In case of changes in database dumps retrieved from the larger environmental agencies, the TOAR processing scripts or configuration files may need to be adapted before processing the data. This manual inspection step is skipped for near realtime data harvesting.

Step 4: Configure workflow

Some data providers use different names or spelling for the variable names (example: “ozone” is reported as variable “8” in EEA Airbase database dumps). It is therefore necessary to configure the automated workflow and provide a mapping from the variable names contained in the data to the standard names (controlled vocabulary) used in the TOAR database. In many cases this mapping is a trivial identity relation, i.e. the variable name is used without any change.

In this step we also gather all necessary information from the data file to identify the station and time series to which the data belong (see steps 8 and 11 below).

3.2 Automated data preparation

The main part of the data processing workflow is fully automated and summarised in Figure 3 below. In case of errors, the workflow is aborted. Where possible, manual fixes are then applied to the pre-screened data (see step 3 above) and the automated workflow is started again. In case of severe errors or ambiguities we will contact the data provider and try to identify solutions for the detected problems. These solutions may sometimes also include changes to our processing code, for example to accommodate new metadata elements or format changes in the data. If the automated workflow runs without errors, the end result will always consist of new entries in the target database identified in step 2 above. The workflow has been designed with several data quality tests, format tests and plausibility tests to avoid polluting the actual TOAR database with implausible or wrong data. Once, data are published in the TOAR database, changes will always be logged and the data will be versioned. This does not apply to the staging database, where data can be overwritten, for example if a data provider re-submits a dataset after inspection of the first processed results.

Step 5: Identify Variable

The variable name supplied with the data or through the configuration record (see step 4 above) is used to identify the variable in the TOAR database and a FastAPI query (db.query(variables_models.Variable).filter(variables_models.Variable.name == variable_name).first()) is issued to find out if the variable is already contained in the database.

Due to the prior harmonisation (step 4 above), it is ensured that the mapping to actual variable records in the TOAR database is unambiguous.

Step 6: Decision on further processing

If the variable that is named in the data file or configuration record is found, processing continues with step 8. Data records of variables which are not included in the TOAR database are ignored (see Step 7).

Step 7: Dismiss Record

If data files contain variables which are not part of the TOAR data curation efforts, these data files, rows or columns will be ignored. Where unknown variables are detected in individually submitted data files, a warning message will be displayed. During the processing of large-volume database dumps, unknown variables are silently ignored.

Step 8: Read Input Record

Depending on the data format, different data readers are invoked which govern the details of the subsequent processing steps inasmuch as loop order or other factors may vary. Often the data reader will consist of a simple pandas.read_csv() call (Python library), but in other cases, the data reader may involve an entire new workflow (e.g. OpenAQ data ingestion).

Step 9: Check Metadata and Harmonise it

If metadata such as station and time series information is provided within the data file (see TOAR standard format in TOAR_UG_Vol05_Data_Submission_Guide) then each metadata key is cross-checked with the TOAR standard metadata keys (controlled vocabulary) and in case of obvious errors, corrections will be applied (for example spelling errors such as “lngitude”). Where metadata values are also controlled (see document on controlled metadata), the correctness of the value and consistency of spelling is also checked. Unresolvable differences are reported back to the data provider and the workflow is aborted.

In the case of processing database dumps, metadata are often provided in separate tables and/or pasted together from various sources. Here, the script controls that all necessary metadata are provided for the subsequent processing.

In this processing step all metadata required for identifying or creating a station record or a time series record is collected. Subsequent tests that rely on such information (steps 10 and 14) make use of the metadata collected here. Only in case that new station records must be created there will be further metadata added to the processing, namely from the TOAR geolocation service (see description of step 12).

From individually contributed data files unknown but valid metadata key value pairs are collected and stored as “additional metadata”. In most cases, such additional metadata includes additional descriptions of the measurement (or, more generally, data generation method) and shall therefore be retained in the time series model of the database (see database model description in TOAR_UG_Vol03_Database). However, it is also possible that data providers include additional information about the measurement sites in their files. Therefore, if such additional metadata is found, the workflow is aborted and re-run with a new configuration, which describes how the additional metadata shall be processed.

Step 10: Identify Station

In order to ensure that data belonging to one measurement series are recorded as one time series at one station (and, conversely, data obtained at physically different locations are linked to different stations) the following set of rules has been implemented to decide if a new data record with station metadata information belongs to a station that is already recorded in the TOAR database. This seemingly easy problem is actually quite complicated in practice, because different monitoring networks may report station coordinates with different accuracy and sometimes the reported station coordinates are even wrong. Furthermore, there is no universal system of station identifiers established and in some cases, station identifiers are not even reported.

Rule 10.1: check the combination of role: resource_provider (organisation) and station code (db.query(models.Timeseries).filter(models.Timeseries.station_id == station_id).filter(models.Timeseries.variable_id == variable_id); role_num = get_value_from_str(toardb.toardb.RC_vocabulary,'ResourceProvider'); (contact.organisation.name == resource_provider) and (role_num == role.role)).
If a station with the same identifier and provider is found in the database, we can be very sure that the new data records belong to the same station. As an additional check, we then control whether the station coordinates from the new record are within 10 m of those stored in the database, and the database operator is informed about potential discrepancies.

Rule 10.2: if rule 10.1 did not lead to the identification of an existing station, the station coordinates are used as proxy for a station identifier. Based on a series of queries of the TOAR-1 database, where extensive work had been conducted to manually control station coordinates, we identified a threshold distance value of 10 m as most suitable criterion to decide if new data records should belong to an existing station or not (select * from stationmeta_core where ST_DistanceSphere(stationmeta_core.coordinates, ST_GeomFromText('POINT(lon lat)',4326)) < 10;).

Step 11: Decision on station

In the vast majority of cases, the application of rules 10.1 to 10.2 will lead to either one or no station identifier returned from the TOAR database. In the event that two or more station identifiers are returned, the data ingestion process will be aborted and the input data will be manually augmented to reach an unambiguous decision after consultation of further documentation and/or with the data provider. If no station id is found in the TOAR database, a new station entry will be created and filled with the metadata provided in the data record as well as additional metadata from our geospatial services (see Steps 12 and 13 below). The id from this new record is then used for further processing of the data.

Note that station coordinate tests only concern the latitude and longitude coordinates. There are situations where several independent measurements are taken at different heights at the same location (e.g. tower measurements). These will be added as individual time series at the same station (see Step 15 below).

Step 12: Collect Station Metadata from Global Earth Observation Data

A unique feature of the TOAR database is the rich metadata stored with each station, which is derived from various Earth Observation (EO) datasets and Open Street Map (for details, see TOAR_UG_Vol03_Database, section 4.2.1 station location). These metadata fields allow for globally consistent station classification schemes [1] and can be used in machine learning applications (e.g. [2]).

All of this additional metadata can be obtained via a special geolocation service (documentation in gitlab https://gitlab.jsc.fz-juelich.de/esde/toar-data/geolocationservices). The data ingestion workflow contains a list of template queries which are sent to the geolocation service for each new station that shall be added to the TOAR database. The metadata elements are described in TOAR_UG_Vol03_Database and at https://esde.pages.jsc.fz-juelich.de/toar-data/toardb_fastapi/docs/toardb_fastapi.html#stationmetaglobal . The geolocation service URLs that are invoked for each new station can be seen at https://esde.pages.jsc.fz-juelich.de/toar-data/toardb_fastapi/docs/toardb_fastapi.html#geolocation-urls.

Step 13: Create New Station

If a new data record has metadata information that suggests the data are coming from a station which is not yet included in the TOAR database (see Step 10 above), a new station id is generated and a new station record is created in the database. The metadata for stations is described at https://esde.pages.jsc.fz-juelich.de/toar-data/toardb_fastapi/docs/toardb_fastapi.html#stationmeta

At a minimum, the new station record must contain the mandatory fields (see TOAR_UG_Vol05_Data_Submission_Guide). If the metadata is insufficient to populate these fields, the data insertion process is aborted and the database operator is notified. Depending on the data source and the nature of the metadata error, different actions will be triggered from such notifications, i.e. either a communication thread with the data provider will be initiated, or we will investigate other measures to correct and/or complete the metadata information, such as document search, map inspections, etc.

If it is not possible to define a station with all necessary metadata, the data from this location will not appear in the TOAR database. Otherwise, the data insertion will be repeated for these data once the metadata has been completed.

Note that TOAR database users can comment on station metadata and suggest corrections or improvements. Ensuing metadata changes will be logged so that the history of station metadata can be followed from the user interface.

Step 14: Identify Time Series

In an ideal world, the information about the measurement location and the measured variable would suffice to uniquely identify a TOAR database time series and thus find out whether such a time series already exists in the database, or whether it has to be created as part of the data ingestion procedure. In reality, there are various confounding factors and therefore additional information is needed, before a time series can be unambiguously identified. The TOAR data ingestion procedure uses the following criteria:

Criterion 14.1: station_id (see Steps 10-12)

Explanation: the station id unambiguously defines a specific geographic location and ensures that all location-related metadata are available

Criterion 14.2: variable_id (see Step 4)

Explanation: the variable_id ensures that only data of known physical quantities are stored in the database and a description of these quantities is available

Criterion 14.3: roles

Explanation: if roles contains resource provider it has to be checked which organisation it is. Due to the reporting procedures for air quality data, the same original data records can be available from different providers (and they are not always identically stored). It is therefore important to differentiate between providers when assigning data to a specific time series in the TOAR database.

Criterion 14.4: sampling_frequency

Explanation: sampling_frequency (see: section 4.4.1 of TOAR_UG_Vol03_Database)

Criterion 14.5: version

Explanation: some datasets come with a version number issued by the provider. The TOAR database will typically feature the most recent version as the most relevant dataset, and it implements a thorough versioning scheme on the level of individual data samples (see TOAR_UG_Vol03_Database, section 4.3.1.3). Nevertheless, if a version number is provided with the data records, it will be used to unambigiuously identify a time series. If no version number is provided we set the version to NA.

Criterion 14.6: data_origin (measurement or model)

Explanation: data_origin is either “measurement” or “model”

Criterion 14.7: data_origin_type

(measurement method or model experiment identifier, e.g. COSMOS-REA6, COSMO-EPS, ECMWF-ERA5, etc.)

Explanation: this allows distinction between measurements with different techniques or data from different models or different model experiments

Criterion 14.8: sampling_height

Explanation: certain stations provide measurements at different altitudes (e.g. tower sites in the U.S.). As we define “measurement location” strictly by the latitude and longitude coordinates, a distinction between different altitudes must be possible. This is accomplished by using the sampling height as criterion. If no sampling height is provided, we implicitly assume a sampling height of 2 m above ground.

Criterion 14.9: data filtering procedures or other special dataset identifiers

Explanation: in some cases, for example at clean air sites, data filtering procedures are applied by the data providers on the sub-hourly time series to remove, for example, local pollution influences. These filters are named (for example “clean”), and this filter name is used as criterion. If no filter name is given, a blank string is used, which implicitly has the meaning “all data”.

Note: we are aware that there may occasionally be different instruments of the same type measuring at one location. Some providers would wish to interpret this as one time series, others as two distinct time series. For practical reasons it is currently impossible to curate such information (if available at all). Where available, information about specific instruments is captured in additional metadata, but its interpretation is up to the user.

Step 15: Decision on Time Series

If a time series record matching all criteria outlined above is found, processing will continue with insertion of data values into this time series (step 17). Otherwise a new time series will be created (step 16).

Step 16: Create New Time Series

If a new data record has metadata information that suggests the data are coming from a time series which is not yet included in the TOAR database (see Step 14 above), a new time series id and record are created in the database. The metadata for this record is described at https://esde.pages.jsc.fz-juelich.de/toar-data/toardb_fastapi/docs/toardb_fastapi.html#timeseries

Step 17: Prepare Time Series Data

Depending on the data source, the data records to be inserted can be formatted as a time series (one time-stamp per row), database dumps (one time-stamp per row, multiple variables and stations in one file), or as JSON records (one JSON dict per value denoting one time stamp of one variable at one station). Before the data values of a time series are added to the database, the data are reformatted so that they can be inserted directly with the SQL COPY command. This is much faster than INSERT.

The intermediary format for data insertion is:
timestamp, value, flag, timeseries_id, version label

At this ingestion stage, the flag is either the flag submitted by the data provider or a default flag (indicating that the data is expected to be correct).

For a description of the version label, see TOAR_UG_Vol03_Database, section 4.3.1.3 time series versioning.

Step 18: Automated Quality Control

If possible (time series contains enough time steps) an automated quality check will be performed on the data and the data flags will be modified. Some automated quality control tests are applied to ensure that time stamps and data values are reasonable. If less than 10% of the newly inserted data are flagged as questionable or erroneous in these tests, the new data are considered valid and are either published (harvested data and processing of database dumps) or made available to the data provider for review. If more than 10% of the new data records are flagged and the workflow is aborted. The workflow can be re-run after manual inspection and fixing of the problem.

Step 19: Store Time Series Records in Target Database

As final step in the general TOAR data ingestion workflow, the processed data records are inserted into the target database, i.e. either the staging database (for individually submitted data files) or directly the operational TOAR database (large file collections, database dumps and near realtime data). The database contains protection against overwriting of existing timestamps. If existing timestamps must be replaced, the new data records must be stored under a new version number.

3.3 Semi-automated post-processing

Bulk uploads from large file collections, database dumps or web services (including near realtime data) do not undergo any further post-processing once they have been inserted into the TOAR database. It is planned to develop some additional data quality and metadata consistency control software, so this may change in the future.

For individual data submissions, additional semi-automated post-processing needs to happen in order to approve of the uploaded data series and transfer the data from the staging database to the final TOAR database. Optionally, data providers are also asked if they wish to have their data published as B2SHARE record including a DOI.

Step 20: Data Review by Provider

The data review by providers is initiated through the automated generation of an email containing a summary report of the pre-processing and a standardised quality control plot. The email can be edited and must be sent by a human operator.

In this email we also ask if the data provider wishes to have the data published on B2SHARE.

The generated email to provider includes:

- A copy of the processed data file (or a link to file in a cloud)
- The results from toarqc and some more basic statistics as summary plots and yearly plots
- A short and concise summary of provenance and testing results as email text
- A request to double-check the evaluation and testing results and confirm that data appear in database as they should[18]
- An offer to publish the dataset as part of an existing or new collection in B2SHARE (with DOI); there should be one collection per station. Once a collection is established, files can be added but must not be changed – this is where the version number comes into play.
  The offer can be accepted by reply to email with sentence “Please publish my data”.

Step 21: Decision on Further Processing

Depending on the response by the data provider, the following options exist how the workflow may continue:

Option 21.1: The data provider approves of the data. Processing continues with step 22.

Option 21.2: The data provider requests specific corrections to the data or metadata. The corrections will be applied on the pre-screened data files and the workflow will be repeated. A new email will be sent to the provider to ask for confirmation of the changes and approval of the data.

Option 21.3: The data provider disapproves of the data. The workflow is abandoned and the data is deleted from the staging database. The raw data and pre-screened data files will be kept (unless deletion is specifically requested by the data provider) in

[18] this confirmation will be recorded in the database and made accessible as one data quality attribute (“dataset approved by provider”)

order to keep information for potential re-processing in case the data are re-submitted at a later stage.

Option 21.4: The data provider does not react to our email. In this case we first try to reach someone else from the same institution. If this also remains unsuccessful, the TOAR data centre team will decide on a case by case basis how to proceed.

Step 22: Transfer to TOAR Database and Publication

Data that has been approved by the data provider is transferred from the staging database to the final TOAR database. Through this step, the data is automatically made available under a CC-BY 4.0 license for re-use via the TOAR REST API (see https://toar-data.fz-juelich.de/api/v2).

Step 23: Create Data Publication on B2SHARE

If the provider has opted for a B2SHARE data publication, then a csv formatted copy of the time series including all metadata is automatically extracted and uploaded to B2SHARE (community TOAR). B2SHARE publications always include a DOI. The metadata for the B2SHARE data publication is automatically generated from the information stored in the TOAR database. Before publication on B2SHARE the record will be evaluated by a human operator.

In case the data to be published extends an already published time series a new version is added to that time series DOI (replacement of data file). Does the data constitute a new time series it gets a new DOI. In case this time series belongs to a not yet published station a station record will be generated and published resulting in a new station DOI. The DOI of a new time series is added to the station DOI it belongs to and vice versa.

Step 24: Archive Data Processing Log

All information that has been collected during the (semi) automated data processing (step 5 onwards) are collected in a file and stored together with raw and processed data in a file system. At certain events (e.g. new data added) these get compressed and archived.

4 TOAR Data Publications

The data which has been sent to TOAR by individual organisations and has been processed as described above can be published via the TOAR publication service. The preparation of the data to be published is done by the TOAR data centre team. Technically the data publication is done with EUDAT's B2SHARE service (https://b2share.fz-juelich.de/communities/TOAR). The TOAR data publication service prepares the metadata required by B2SHARE for the data set to be published. All TOAR data publications are released as open data under the Creative Commons - By Attribution (CC-BY 4.0) license (see https://creativecommons.org/licenses/by/4.0/).

As part of the data curation and ingestion workflow, we implemented tools which allow for automatic generation of a B2SHARE metadata record and the data upload. This will generate citable data publications, which can be used by the data providers or other users to unambiguously refer to individual data submissions or collections of measurements at individual observing sites.

There are two types of collections which can be published in this process, the "Single Station Collection" and the "Single Station Series":

1. Single Station Collection
   The publication contains a set of data files with multiple measured variables at one specific station during the time period the station was in operation.
   An example is available at https://b2share.fz-juelich.de/records/f4e3583024e84467aa7c2d24d5f4861d
2. Single Station Series
   The publication contains a set of data files containing a time series with one measured variable at one specific station during a given time period. It is part of a Single Station Collection.
   An example is available at https://b2share.fz-juelich.de/records/d5ae0c08311e4b51965fcb7a7a922850

TOAR data publications consist of rich metadata, which are a combination of generic B2SHARE metadata and TOAR-specific enhancements. For an up-to-date listing of the metadata refer to https://b2share.fz-juelich.de/communities/TOAR. The metadata of station time series data files attached to TOAR data publications are described in TOAR_UG_Vol05_Data_Submission_Guide.

5 TOAR Near Realtime Data Processing

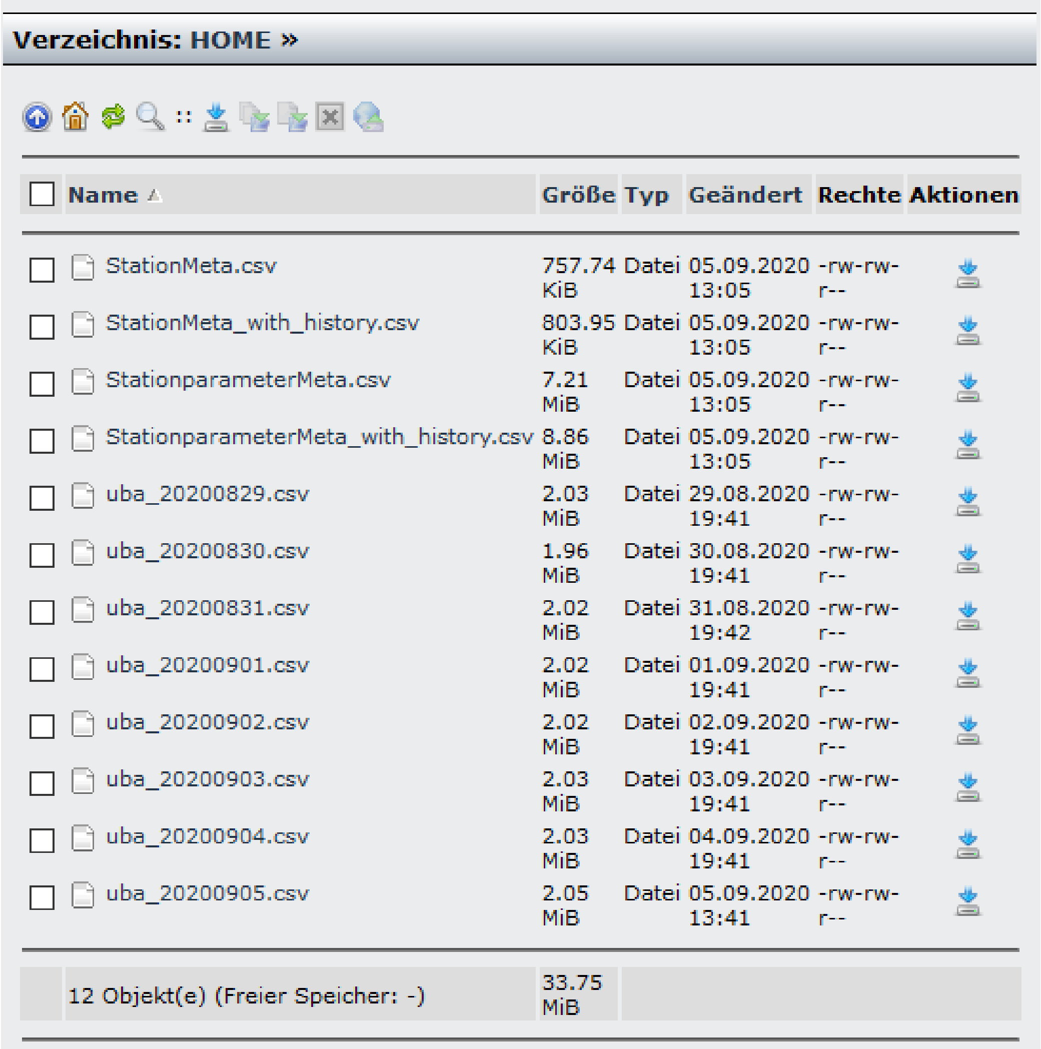
Currently we collect near real-time data from two data providers: UBA (https://www.umweltbundesamt.de/en: German Environment Agency) and OpenAQ (https://openaq.org: open air quality data). The corresponding data harvesting procedures are described below.

5.1 UBA Data Harvesting

Since 2001, the German Umweltbundesamt - UBA (www.umweltbundesamt.de) - provides preliminary data from a growing number (currently 1004) of German surface stations. Basis for the data exchange is the manual „Luftqualitätsdaten- und Informationsaustausch in Deutschland", Version V 5, April 2019 (in German).

At least ozone, SO2, PM10, PM2.5, NO2 and CO data for the current day are updated daily and provided continuously hourly up to a maximum of four previous days. Data is fetched from the UBA service 4 times per day (8 am,12 pm, 18 pm, and 22 pm (local time)).

The software for processing the data from UBA is available at https://gitlab.version.fz-juelich.de/esde/toar-data/toar-db-data/-/tree/master/toar_v2/harvesting/UBA_NRT . Data (StationparameterMeta.csv, StationMeta.csv, uba_%s.csv (%s denotes a date)) are harvested 4-times daily from http://www.luftdaten.umweltbundesamt.de/files/ (secured with access credentials).



Verzeichnis: HOME »

□	Name ▲	Größe	Typ	Geändert	Rechte	Aktionen
□	StationMeta.csv	757.74 KiB	Datei	05.09.2020 13:05	-rw-rw-r--	⬇
□	StationMeta_with_history.csv	803.95 KiB	Datei	05.09.2020 13:05	-rw-rw-r--	⬇
□	StationparameterMeta.csv	7.21 MiB	Datei	05.09.2020 13:05	-rw-rw-r--	⬇
□	StationparameterMeta_with_history.csv	8.86 MiB	Datei	05.09.2020 13:05	-rw-rw-r--	⬇
□	uba_20200829.csv	2.03 MiB	Datei	29.08.2020 19:41	-rw-rw-r--	⬇
□	uba_20200830.csv	1.96 MiB	Datei	30.08.2020 19:41	-rw-rw-r--	⬇
□	uba_20200831.csv	2.02 MiB	Datei	31.08.2020 19:42	-rw-rw-r--	⬇
□	uba_20200901.csv	2.02 MiB	Datei	01.09.2020 19:41	-rw-rw-r--	⬇
□	uba_20200902.csv	2.02 MiB	Datei	02.09.2020 19:41	-rw-rw-r--	⬇
□	uba_20200903.csv	2.03 MiB	Datei	03.09.2020 19:41	-rw-rw-r--	⬇
□	uba_20200904.csv	2.03 MiB	Datei	04.09.2020 19:41	-rw-rw-r--	⬇
□	uba_20200905.csv	2.05 MiB	Datei	05.09.2020 13:41	-rw-rw-r--	⬇
	12 Objekt(e) (Freier Speicher: -)	33.75 MiB				

Figure 2: Snapshot from 2020-09-05 17:00 CEST

Table 2: Mapping of data from daily files imported to the TOAR database variables

name of component in original file	name of component in TOAR database
Schwefeldioxid	so2
Ozon	o3
Stickstoffdioxid	no2
Stickstoffmonoxid	no
Kohlenmonoxid	co
Temperatur	temp
Windgeschwindigkeit	wspeed
Windrichtung	wdir
PM10	pm10
PM2_5	pm2p5
Relative Feuchte	relhum
Benzol	benzene
Ethan	ethane
Methan	ch4
Propan	propane
Toluol	toluene
o-Xylol	oxylene
mp-Xylol	mpxylene
Luftdruck	press

Table 3: Mapping of station type

term of station_type in original file	term of station_type in TOAR database
Hintergrund	background
Industrie	industrial
Verkehr	traffic

Table 4: Mapping of station_type_of_area

term of station_type_of_area in original file	term of station_type_of_area in TOAR database
ländlich abgelegen	rural
ländliches Gebiet	rural
ländlich regional	rural
ländlich stadtnah	rural
städtisches Gebiet	urban
vorstädtisches Gebiet	suburban

Table 5: Mapping of units and unit conversions

component	original unit	unit in TOAR database	unit conversion while ingesting
co	mg m-3	ppb	*858.95
no	ug m-3	ppb	*0.80182
no2	ug m-3	ppb	*0.52297
o3	ug m-3	ppb	*0.50124
so2	ug m-3	ppb	*0.37555
benzene	ug m-3	ppb	*0.30802
ethane	ug m-3	ppb	*0.77698
ch4	ug m-3	ppb	*1.49973
propane	ug m-3	ppb	*0.52982
toluene	ug m-3	ppb	*0.26113
oxylene	ug m-3	ppb	*0.22662
mpxylene	ug m-3	ppb	*0.22662
pm1	ug m-3	ug m-3	
pm10	ug m-3	ug m-3	
pm2p5	ug m-3	ug m-3	
press	hPa	hPa	
temp	degree celsius	degree celsius	
wdir	degree	degree	
wspeed	m s-1	m s-1	
relhum	%	%	

Validated data from the previous year is available at May 31st latest. This data is requested by email and then processed from the database dumps we receive. The validated data will supersede the preliminary near realtime data. The realtime data remains in the database but is hidden from the standard user access procedures via the data quality flag settings.

5.2 OpenAQ

OpenAQ[19] is collecting data in 93 different countries from real-time government and research grade sources. Starting on 26th November 2016, OpenAQ has already gathered more than one billion records, which has 306 Gigabyte in total size and covers air quality relevant variables BC, CO, NO2, O3, PM10, PM2.5 and SO2.

5.2.1 Data Provision

OpenAQ provides real-time meteorological data on Amazon Web Service[20] in daily directories. Data files composed of records of meteorological measurement values are put into the directory of the current day at irregular intervals. The directories with their data files are stored on Amazon Web Service permanently.

Each data file contains up to hundreds of thousands of records. Records are JSON[21] objects in the same structure throughout the entire life cycle. The task of our real-time data harvesting

[19] www.openaq.org

[20] https://openaq-fetches.s3.amazonaws.com/index.html

[21] www.json.org

procedure is to go through these records and save them into the TOAR database according to TOAR database scheme in about real time.

A key element for processing the OpenAQ data is a separate intermediate database, to help processing the data. Only after the data is ready to be stored in the TOAR database it will be uploaded.

The realised real-time data harvesting procedure consists of four steps, the first two download the data and store it in the intermediate database while the last two parse the fields and map them to the TOAR database scheme.

The first two steps (workflow1) are responsible for the action between Amazon Web Service and intermediate database, and the other two steps (workflow2) for the action between intermediate database and TOAR database.

Technically the open source software Apache Airflow is used for workflow automation, so that workflows are triggered in regular interval within a day.

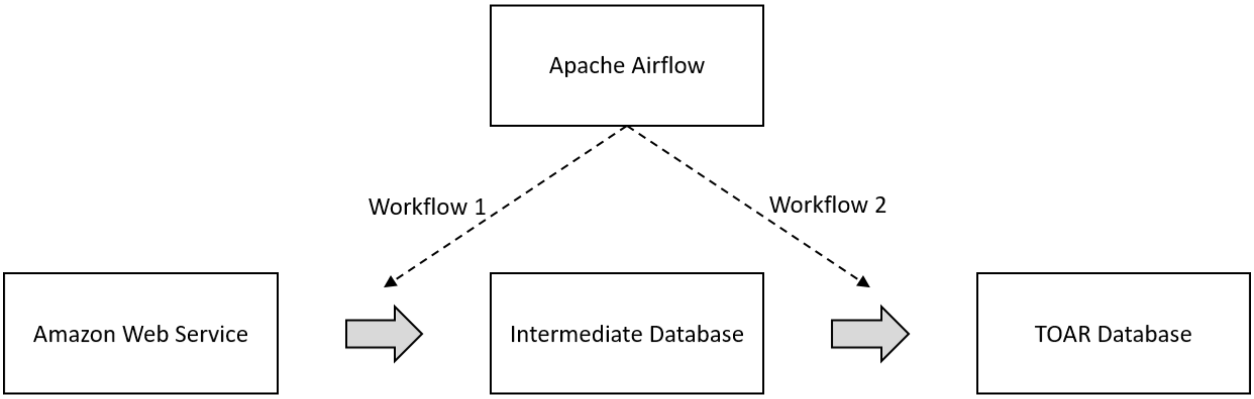


Figure 3: Overview of Processing Steps

5.2.2 The Intermediate Database

The reason for introducing an intermediate database is to make data parsing and mapping easier and to enable pre-evaluation, statistics, and visualisation. Thereby we flatten the long lists of tree-structured records into a two-dimensional table.

5.2.3 The Harvesting Workflows

- Workflow 1

  We use python and the boto3[22] python module for querying Amazon Web Service (AWS).

  First the newly created[23] sub directories on AWS have to be identified and retrieved which will then be inserted into the sub directory table and the data file table of the intermediate database.

  With that the current status of the intermediate database has been synchronised with the one of AWS and all unprocessed records are prepared in the intermediate database for importing into the TOAR database.

- Workflow 2

  The second workflow identifies the station and the timeseries in the TOAR database a new record belongs to in the way described in Steps 10 to 17 in section 3.2

  With the id of the identified time series, the value of the record will finally be inserted into the data table in TOAR database. In the end a record from the intermediate database is matched and saved into TOAR database (Figure 4).

[22] www.github.com/boto/boto3

[23] Compared to the directories retrieved in the last run

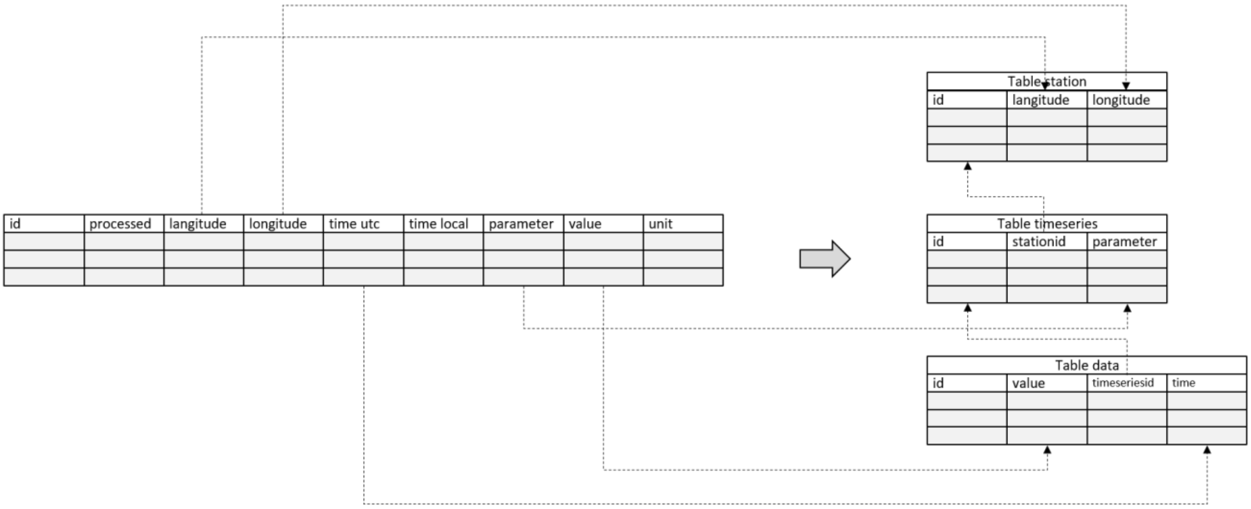


Figure 4: Simplified model of mapping records from intermediate database (left) into TOAR database (right).

5.2.4 Workflow Automation with Apache Airflow

The data harvesting process described in the last subsection can be executed in one batch or divided into two isolated workflows. In both cases it is desired to be scheduled, executed and monitored automatically. To this end we use the Apache Airflow workflow management software[24] installed on the same server as the intermediate database. Apache Airflow is registered as a system service, so that it will be started automatically on system boot. We define two separate workflows in Apache Airflow as depicted in Figure 3. Both workflows are scheduled hourly. On the web interface of Apache Airflow, we can monitor and manipulate the workflows with ease.

[24] airflow.apache.org

6 References

[1] Schultz et al., 2017
Tropospheric Ozone Assessment Report: Database and Metrics Data of Global Surface Ozone Observations,
Elementa Sci. Anthrop., https://doi.org/10.1525/elementa.244

[2] Betancourt et al., 2021
AQ-Bench: A Benchmark Dataset for Machine Learning on Global Air Quality Metrics,
Earth Syst. Sci. Data, 13, 3013–3033, 2021, https://doi.org/10.5194/essd-13-3013-2021