TOAR Data User Guide #5

TOAR Data Submission Guide

toar-data.fz-juelich.de

Version 1.1.0 | January 17, 2023

The TOAR Data Team

Forschungszentrum Jülich GmbH
ESDE | JSC - FSD
52425 Jülich

CONTENTS:

LIST OF FIGURES

LIST OF TABLES

CHAPTER

ONE

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,[2]). 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.

This data submission guide is intended for individual data providers who wish their measurements to be added to the TOAR database. In general, the TOAR data team is always willing to discuss the modalities of data submission. If you are planning to deliver a large amount of data to TOAR, please get in contact with the data centre team[1] before preparing any data files as we may be able to accommodate other formats than the one described in this guide and can thus save you unnecessary work. Specifically, if you have your data already formatted as netcdf files or in NASA AMES format, we can easily accept these as well and would then only ask you to provide us with the additional metadata in a stations.csv file. If your data consists of a small number of files, we ask you to adhere to the guidelines described here as closely as possible. The more your data format differs from our template, the more likely we will not be able to process your data and it will thus be lost to the TOAR effort. In case of doubts or ambiguities we will get back to you to sort out any metadata or data issues.

The ASCII format described in this document is meant to facilitate processing of the submitted atmospheric data for inclusion in the TOAR surface station database at Jülich. But it is also designed to make the data submission as easy as possible for you as the data provider. The header will contain all necessary metadata information, and the data format shall make it easy for us to interpret your data with respect to the time of measurement and the measurand values (usually these will be ozone mixing ratios or concentrations). Please note that the scope of TOAR-II has been expanded to also include the analysis of a small set of ozone precursors and related variables (PM2.5 and meteorological information). If you have problems adhering to the TOAR data format, please contact the TOAR data team to find a solution.

The data can be submitted online at https://toar-data.org/contribute/#timeseries or you can send us suitably formatted data files as email attachment or deposit them in a shared folder. Note that we can also store additional documentation about the station(s) at which you perform your measurements in the TOAR database. If you wish to provide URLs, PDF documents or images which further document your measurement site(s), please let us know.

Please make sure to give us your explicit consent that you agree with publication of your data under the CC-BY 4.0 license (see section 2 for details) and that you allow us to curate your data as described in detail in The TOAR Data Processing Workflow.

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

[1] toar-data@fz-juelich.de

CHAPTER

TWO

DATA POLICIES

2.1 Data Submission Policies

All data in the TOAR database are openly shared under the CC-BY 4.0 license (see https://creativecommons.org/licenses/by/4.0/ for details). We must ask you to give us explicit consent that we are allowed to publish your data under this license. This also implies that we are allowed to curate your data in the manner that is described in TOAR Data Processing.

If you upload your data via the online portal https://toar-data.org/contribute/#timeseries please activate the radio button **"I understand that the data will be made freely available under a CC-BY 4.0 license and I confirm that I have the rights to grant such license to the TOAR data centre"**. If you submit your data by email (including the data files as attachment or a link to a shared folder), please copy the above sentence into your email so that we can document that you granted us permission to publish your data.

We must also ask you to confirm that you quality-controlled the data which you send to us. In the upload web portal, please activate the radio button **"I have read the TOAR data submission guide and confirm that the data submitted here has been obtained with commonly accepted measurement methods and quality controlled according to standard procedures."** Again, if you are submitting your data by email, please copy this sentence into the email text.

2.2 Data Privacy Rules

Please be assured that we take data privacy protection serious. As there are generally no ethical concerns involved in air quality or meteorological data, data privacy protection only applies to the personal information we obtain from you as data providers. Please note that we need to store your name, affiliation and email address in the TOAR database. This also applies to your co-workers who are named as contributors or in other roles (see description of roles in section 3.2 below). Optionally, you can also provide additional contact information such as your phone number, your Orcid id, and the address of your organisation. By default, this limited personal information will not be made available to TOAR database users. This implies that they cannot properly attribute their use of data to you as the data provider. Should you wish your personal information to be made available, please set the corresponding flag to false. All personal data that is processed by the TOAR data centre is handled according to the general data privacy rules (GDPR) of the European Union as expressed by the German Law (DA-GVO). In particular, this means that you have the right to request information about what information about you is stored and you have the right to order corrections or deletion of your personal data. See https://toar-data.fz-juelich.de/footer/privacy.htm for details.

CHAPTER

THREE

TOAR DATA SUBMISSION FORMAT

3.1 Filename

Please provide data files with a name containing the species, the station_id, and the period of the data record in the format:

```
{parameter}_{station_id}_{startyear}{startmonth}_{endyear}{endmonth}_{special}.
↪{extension}
```

Parameter should be written in lower case. *startyear* and *endyear* consist of 4 digits, *startmonth* and *endmonth* of 2 digits. The *special* tag is optional and can be used to identify, for example, wind sector filtered data or data from different sampling heights.

There is no need to break your data into individual years, but if it is more convenient for you to submit annual files, then please do so.

```
Examples:
1)   o3_UN4058943_201901_201912.dat
2)   wspeed_DEHR0003_200001_202012_cosmomodel.csv
```

3.2 File Header and Metadata Rules

Please provide as much of the following metadata information as possible. **Metadata keys that are marked with '*' are mandatory**; we will not be able to process your data if any of these elements is missing. It is possible, however, to provide the mandatory station metadata in a separate *stations.csv* file if this is easier for you.

Metadata shall be formatted as key-value pairs separated by a colon (example: `Station_id:  USH54S`). Line breaks in the metadata values are not allowed. You can start the header lines with metadata information with a comment symbol ('#', '*' or '!') or simply begin the line with the name of the metadata key. Starting the header line with any other character will make it invalid and prevents processing. Parsing of metadata keys is case-insensitive, so it doesn't matter if you use lowercase, uppercase or mixed-case characters. String formatted metadata values will however preserve their format. The order of the metadata elements does not matter, but we suggest that you stick to the order from Table 2 below. We suggest that you copy the template metadata header from the Annex of this document and edit the content of the values. If you cannot provide a given piece of information, simply delete the line.

You can also add additional metadata key value pairs in your files. These will be preserved in the TOAR database as additional_metadata in the station or timeseries records. Additional metadata can be retrieved from the database but is not available for data set searches. Some suggestions for recommended additional metadata are listed in Table 3. Please help us by starting any additional metadata key with station information with ‘station/_’. And please avoid lines starting with ‘Time,’ as we use this keyword to identify the start of the data section.

Empty lines in the file header are allowed and will be ignored.

Table 3.1: Valid header line **examples**

Header line	Comment
Station_name: Niederzier, Treibbachstraße	Valid key, separator ‚:‘, valid text, all in one line
# Station_name: Niederzier, Treibbachstraße	As alternative to the line above; it starts with a comment symbol
!station_lon : 6.469312	Correctly formatted, longitude given as decimal degrees east.
Station_geographic_context: mountain range	Valid key value pair. As there is no metadata element “station_geographic_context” in the TOAR database scheme, this metadata element will be saved as *additional_metadata*.

Table 3.2: TOAR file header and description of all metadata elements[3]

Metadata key	Data type / allowed values	Description
*Station_id (or station_code)	string	The station code as the station is registered in your network or as it shall be registered in the TOAR database. Don’t use blanks or special characters in station codes. Exception ‘-‘ (US AQS codes, for example). Example: `fr05237`

Additional information on “role codes” (Dataset_PI, Contributor, Collaborator, PointOfContact):

These terms are a subset of role codes that have been defined by ISO 19115 to standardise information processing. The explanations given for these role codes are rather vague. In the context of the TOAR data processing, we define these roles as follows:

Dataset_PI the principal investigator of a measurement. This is the person who is responsible for making the measurements and securing the quality of the data. In general, there should be exactly one Dataset_PI associated with every measurement. The Dataset_PI may delegate responsibilities, for example to technicians or postdoctoral researchers, and yet remain PI as the person overseeing the measurements and data distribution.

Collaborator a person who has been involved in making the measurements or processing the data, but who is either not part of the institution responsible for the measurement or who has “contributed” only temporarily. One situation we have encountered in TOAR, where nomination of collaborators makes sense is when university researchers assist government agencies in preparing their data for submission to the TOAR database.

Contributor this role applies to any person who is involved in making the measurements or processing the data. Normally, the Dataset_PI will decide who shall be listed as contributor. The distinction between contributor and collaborator is not very clear, but if you wish to

[3] Please see the Annex for a template header which you can copy into your data files and edit the values. We recommend that you also read the Metadata Reference to better understand the meaning of the various metadata attributes in this table. For an explanation of the meanings of Dataset_PI, Contributor, Collaborator, and PointofContact, please see the text below the table.

distinguish between people who were involved on a project level (collaborator) and those who work with you more permanently (contributor), then you can make use of these different roles.

PointOfContact one person dedicated to answer questions related to the dataset, either by the TOAR data team or by data users. The declaration as PointOfContact is independent from the role as Dataset_PI, Contributor, or Collaborator.

The TOAR database can distinguish between roles concerning the measurement station and roles concerning the measurement itself, but this is not reflected in the file header template. If you wish to provide us the information about the responsible persons for the operation of the station, then either send this information by email (for individual sites) or create a stations.csv file where you collect all the metadata concerning the site(s) of your measurements. In the stations.csv file you can use the same format to define roles as in the data file headers.

Table 3.3: Recommended key names for additional metadata[#f32]_

Metadata key	Data type / suggested values	Description
Sampling_type	string (one of: `continuous filter flask`)	Describes the sampling mode of your measurement device

3.3 Data Format

The data section of your data files shall always start with the **title line** `Time, [Variable name], Flag`. If your data doesn't contain data quality flags, we also accept the header `Time, [Variable name]`. Of course you should replace *Variable name* with the actual name of the variable you send to us. Please take the correct spelling of variable names from the REST API at https://toar-data.fz-juelich.de/api/v2/variables/ (e.g. 'o3' instead of 'ozone'). We recommend that you insert an empty line between the file header and the beginning of the data section to increase readability (see the example in the Annex Section 5).

Data should always be provided in chronological order. The actual data section should contain one line per "possible hour" in the year, i.e. a year with 365 days shall have 8760 data lines and a leap year 8784 data lines. Missing values should be coded with -9999. If it is easier for you to not report missing data at all, you can simply omit the lines with missing values, but then please don't report any missing data at all.

As stated in Section 3.1, you can either send individual files per year or combine the data from multiple years in one file.

Please stick to the formats described in Table 3.4 below.

Table 3.4: Formatting instructions for the data section in TOAR data files

Data	Format	Comments
Time	YYYY-MM-DD hh:mm	e.g. `2010-01-01 00:00` with hours starting at 00:00 and denote the **beginning** of the 1-hour averaging period
Value	Floating point number with greater equal 2 decimals	Make sure that the unit corresponds to the "Original_units" you specified in the header! Missing data should be labelled with a large negative number consisting of at least four '9's. We suggest to use -9999.

Flag	1-digit integer or other numeric value	The "flag" column is optional. We suggest to make use of the following flag values from WMO code table 0 33 020:
		`0: OK`
		`2: doubtful`
		`3: wrong`
		`7: missing value`
		If you use a different flagging scheme, please let us know the meaning of the flag values. We will generally be able to translate them.
		If you don't provide flag values, we will assume that all data except for values of -9999 are valid measurements.

CHAPTER

FOUR

PROVIDE CORRECTION FOR SUBMITTED DATA

In case you want to send updated or corrected data to data sets you've provided us earlier be aware of the following:

All data of the period defined in the submitted update will be overwritten with the newly sent data. This implies that you cannot sent only a subset of data for the period but the full (corrected) data set. Data points not covered in the update would get deleted from the TOAR database. The reason for this behaviour is that we cannot process updates manually but use the same automated workflow for the updated data as we've used for the original data set.

CHAPTER

FIVE

ANNEX: HEADER TEMPLATE

The following example describes an ozone monitoring time series from the TOAR V1 database and has been adapted to the new metadata format of TOAR V2. If you use this template to prepare your own data, please adapt or remove all information that does not apply to your time series. A suitable file name for the example below would be o3_CO001_200001-200012.dat. Send the data as well as the license and data privacy information to toar-data@fz-juelich.de.

You can download the file data_submission_template.txt[7].

In addition, please provide license and data privacy information, the template is available at https://toar-data.fz-juelich.de/documentation/DataPublicationQuestionnaire.docx.

```
##Station description, see :numref:`table-metadata-elements` (comment line - please␣
↪delete)
station_id: CO001
station_name: Gobernación de Caldas
station_country: Colombia
station_state: Caldas
station_lon: -75.5170
station_lat: 5.0684
station_alt: 2125
station_type: traffic
station_type_of_area: urban
timeshift_from_utc: -5

dataset_pi_name: Norbert Nobody
dataset_pi_email: n.nobody@nowhere.edu
dataset_pi_phone: +49 2461 6196870
dataset_pi_orcid: 0000-0000-0000-0000
dataset_pi_isprivate: false
dataset_pi_organisation_name: NU
dataset_pi_organisation_longname: Nowhere University
dataset_pi_organisation_country: somecountry

dataset_contributor_name: Norbert Nobody
dataset_contributor_email: n.nobody@nowhere.edu
dataset_contributor_organisation_name: NU
dataset_contributor_organisation_longname: Nowhere University

dataset_collaborator_organisation_name: collaboration company

```

(continues on next page)

[7] https://toar-data.fz-juelich.de/documentation/data_submission_template.txt

(continued from previous page)

```
dataset_pointofcontact_name: Norbert Nobody
dataset_pointofcontact_email: n.nobody@nowhere.edu

sampling_height: 15.
time_reporting: begin_of_interval
original_units: ppb
data_origin: Instrument
dataset_version:
measurement_programme_name: EMEP
measurement_programme_longname: European Monitoring and Evaluation Programme
measurement_programme_homepage: https://emep.int/
measurement_programme_description: a scientifically based and policy driven programme␣
↪under the Convention on Long-range Transboundary Air Pollution (CLRTAP) for␣
↪international co-operation to solve transboundary air pollution problems

##additional metadata, see :numref:`table-additional-metadata-key-names` (comment line -␣
↪please delete)
sampling_type: continuous
calibration_type: manual
calibration_frequency: 3 months
calibration_description: Multipoint calibration each three months approximately with␣
↪accredited laboratory (Claire), Zero Point Calibration with purified air (external␣
↪scrubber of silica gel, internal ozono scrubber of manganese dioxide), Span point with␣
↪gas generator (IZS-Span reference at 150ppb) (Note: this should be formatted without␣
↪line breaks in a real data file)
absorption_cross_section: Hearn1961
comments: This is only a template for a file header of a data file as it should be␣
↪submitted to the TOAR data centre for inclusion in the TOAR database. If we see this␣
↪comment in your data file, we will ask you if you are sure about the quality of your␣
↪metadata.

##Data section, see Table 4 (comment line - please delete)
time, o3, flag
2000-01-01 00:00, 15.24, 0
2000-01-01 01:00, 18.92, 0
...
```