TOAR Data User Guide #3

TOAR Database

toar-data.fz-juelich.de

Version 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 TOAR database supports the Tropospheric Ozone Assessment Report activity (https://igacproject.org/activities/TOAR) through a uniform provision of harmonised long-term measurement series of ground-level (aka “surface”) ozone concentrations. TOAR has started its second phase (TOAR-II) in 2020 and we, the TOAR data team at Forschungszentrum Jülich, have developed a version 2 of the TOAR database to support TOAR-II. Unless explicitly noted, all information in this document applies to version 2 of the TOAR database and the associated web services. The TOAR-II activity is expected to end in 2024 and the majority of data gathering will take place in 2022. Version 1 of the database[1] will be operated in parallel until further notice. Note that there may be differences in the data series between versions 1 and 2 of the database due to updated information (e.g. new data submissions) or because of data license issues[2] .

Besides its main focus on ground-level ozone measurement series, the TOAR database also contains datasets of ozone precursors and of meteorological variables which can be used in the interpretation of the ozone concentrations and their changes in time. The data in the TOAR database is collected from several different sources (for details see TOAR Data Sources). Most of these data sources are public data archives and repositories. Some data stems from real-time or near-real time sources (OpenAQ initiative and the German Federal Environmental Agency, UBA). However, the TOAR database also functions as primary repository for some datasets which are not curated elsewhere.

Datasets (“series”) in the TOAR database are limited to ground-level measurements at stationary locations (“stations”). While the database contains some records where sampling occurred at higher altitudes (e.g. towers), vertical profile measurements or measurements from moving platforms (e.g. ships, aircraft) are out of scope for the TOAR database.

The TOAR-II activity pledges to adhere to the principles of COPDESS (https://copdess.org/) and the TOAR data infrastructure has been designed to support the emerging best practices for data sharing in the Earth and Space Sciences. The TOAR data team strives to operate its services including the TOAR database at the highest possible level of FAIRness (see https://www.force11.org/group/fairgroup/fairprinciples). A detailed assessment of the TOAR data service FAIRness can be found in Section 6 of this document.

In order to serve the database’s main purpose to provide “easily accessible, documented data on ozone mixing ratios, exposure and dose metrics at thousands of measurement sites around the world freely accessible for research on the global-scale impact of ozone on climate, human health and crop/ecosystem productivity”, all data in the TOAR database version 2 are openly accessible and can be used, modified and re-distributed under the Creative Commons (CC) BY license (i.e. “by attribution”; see https://creativecommons.org/licenses/by/4.0/)[2] .

[1] TOAR V1 is described in Schultz, M. G. et al. (2017) Tropospheric Ozone Assessment Report: Database and Metrics Data of Global Surface Ozone Observations, Elem Sci Anth, 5, p.58. DOI: http://doi.org/10.1525/elementa.244

[2] Version 1 of the TOAR database operated under a different license model and contained embargoed data, which could not be distributed for research without explicit consent by the dataset providers. This “mixed-license” operation made it very difficult to further enhance the TOAR data services and we therefore adopted a fully open data policy for TOAR-II.

Access to TOAR data is provided through one of three main channels:

- a Representational State Transfer (REST) Application Programming Interface (API) at https://toar-data.fz-juelich.de/api/v2/[3] ,
- a graphical web interface at https://toar-data.fz-juelich.de/gui/v2/[4] ,
- TOAR data publications on https://b2share.fz-juelich.de/communities/TOAR.

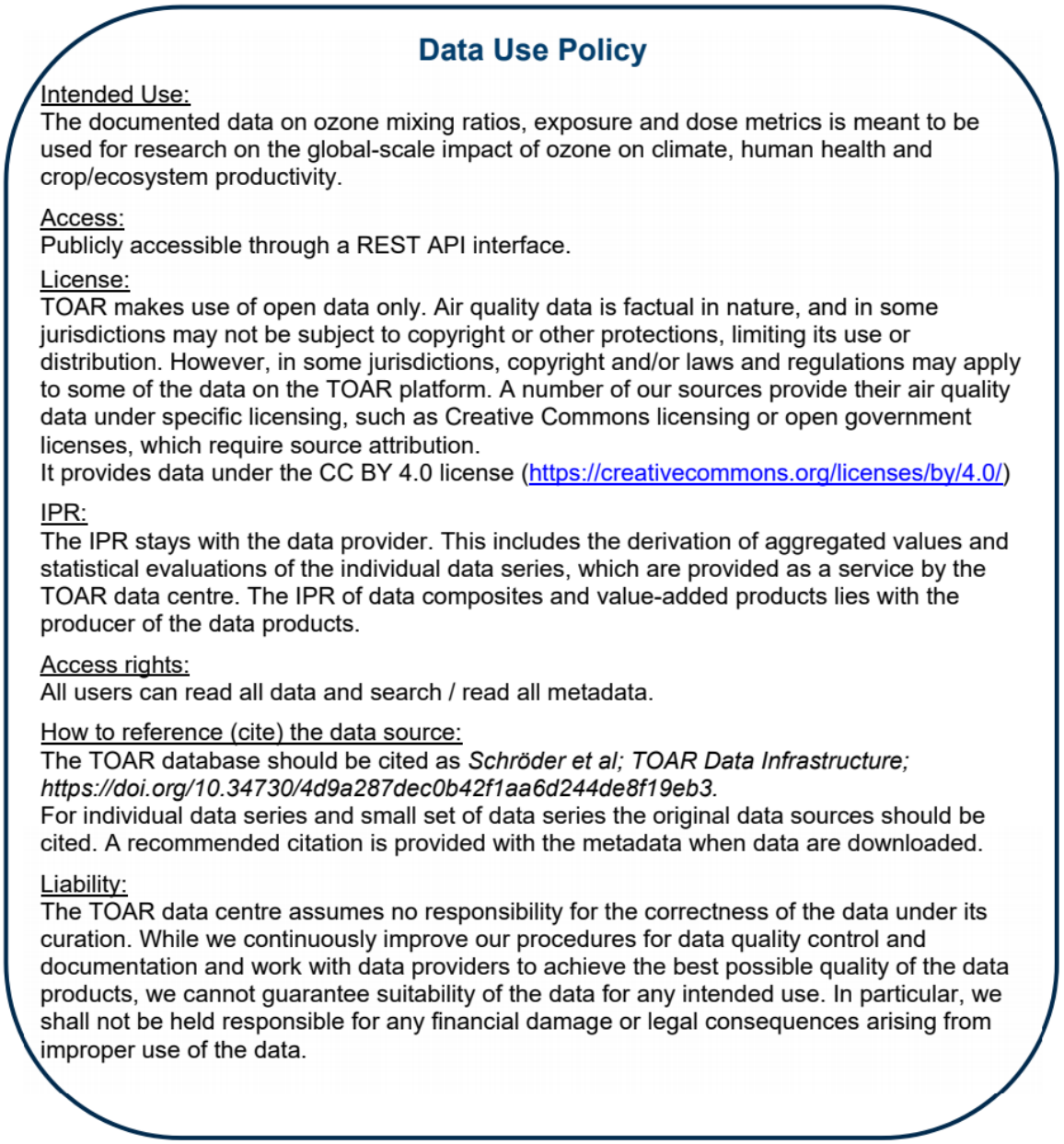
Beginning with version 2, the first two channels allow direct access to the hourly-resolved ozone (precursor and meteorological) data. The third channel, the TOAR data publications, provide on the one hand access to harmonised data deposits of contributed data[5] and on the other hand pre-compiled aggregated datasets supporting the TOAR assessment papers.

If you are using or re-distributing data from the TOAR database, please adhere to the TOAR data use policy defined in Fig. 1.1 below and inform yourself about the terms and conditions of the CC-BY 4.0 license under which TOAR data are distributed.

[3] The version 1 REST API at https://join.fz-juelich.de/services/rest/surfacedata/ should now be accessed via https://toar-data.fz-juelich.de/api/v1/.

[4] At the time of writing the GUI to access data from the TOAR database version 2 is still under development. Version 1 of the GUI, i.e. the JOIN web interface, can be reached at https://toar-data.fz-juelich.de/gui/v1/

[5] The primary data provided by individual research teams or air quality agencies. B2SHARE data publications include a DOI which shall be used to properly cite such datasets.



Data Use Policy

Intended Use:
The documented data on ozone mixing ratios, exposure and dose metrics is meant to be used for research on the global-scale impact of ozone on climate, human health and crop/ecosystem productivity.

Access:
Publicly accessible through a REST API interface.

License:
TOAR makes use of open data only. Air quality data is factual in nature, and in some jurisdictions may not be subject to copyright or other protections, limiting its use or distribution. However, in some jurisdictions, copyright and/or laws and regulations may apply to some of the data on the TOAR platform. A number of our sources provide their air quality data under specific licensing, such as Creative Commons licensing or open government licenses, which require source attribution.
It provides data under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/)

IPR:
The IPR stays with the data provider. This includes the derivation of aggregated values and statistical evaluations of the individual data series, which are provided as a service by the TOAR data centre. The IPR of data composites and value-added products lies with the producer of the data products.

Access rights:
All users can read all data and search / read all metadata.

How to reference (cite) the data source:
The TOAR database should be cited as *Schröder et al; TOAR Data Infrastructure; https://doi.org/10.34730/4d9a287dec0b42f1aa6d244de8f19eb3.*
For individual data series and small set of data series the original data sources should be cited. A recommended citation is provided with the metadata when data are downloaded.

Liability:
The TOAR data centre assumes no responsibility for the correctness of the data under its curation. While we continuously improve our procedures for data quality control and documentation and work with data providers to achieve the best possible quality of the data products, we cannot guarantee suitability of the data for any intended use. In particular, we shall not be held responsible for any financial damage or legal consequences arising from improper use of the data.

Fig. 1.1: TOAR Data Use Policy

CHAPTER

TWO

ACCESSING DATA THROUGH THE GRAPHICAL USER INTERFACE

The graphical user interface (Dashboard) for the TOAR phase 2 database is currently under development and will be described here as soon as it is available. For the time being, data from the TOAR database version 2 can only be accessed via the REST API (see Section 3).

Access to version 1 of the database (from TOAR-I) is available through the GUI at https://toar-data.fz-juelich.de/gui/v1/ which redirects to https://join.fz-juelich.de. This web interface requires registration and is described at https://join.fz-juelich.de/static/documentation/JOIN_FAQ.pdf.

CHAPTER

THREE

ACCESSING DATA THROUGH THE REST APPLICATION PROGRAMMING INTERFACE

A Representational State Transfer (REST) service allows querying all metadata and data products from the TOAR database of surface ozone observations. This API can be used in a web browser or from within a program, from a Unix shell, or in a graphical web application. This section describes the URL structure and sample queries of the TOAR V2 REST interface. For general information on REST, please consult other resources[8] .

3.1 General Information

3.1.1 Base URL

https://toar-data.fz-juelich.de/api/v2/

Response: Description and documentation of the available REST services.

3.1.2 Services

The following information services are available and described individually below. Each service is invoked by appending its name and possible query arguments to the base URL.

- stationmeta: query station ids, station names, and station location from the database
- timeseries: query the data series id and specific metadata of a series from the database
- data: get timeseries data from the database
- variables: query information on variables
- contacts: query information on contacts
- controlled_vocabulary: query the controlled vocabulary and their description from the database
- ontology: query the used ontology of the database
- database_statistics: only provides number of users, stations, time series and data records, there is nothing from any kind of statistical product
- analysis: get bulk time series data and aggregated time series data
- geolocation_urls: query information about geolocation urls
- stationmeta_changelog: query information on stationmeta

[8] e.g. https://restfulapi.net/ or https://mlsdev.com/blog/81-a-beginner-s-tutorial-for-understanding-restful-api

- timeseries_changelog: query information on timeseries
- search: query for stations / time series with certain metadata
- statistics: a separate package providing various statistics on TOAR data

3.1.3 Query Arguments

In order to control the database queries and hence the response of the TOAR REST service, you can add arguments to the service URL. These arguments must adhere to the format `argumentname=value`. The first argument is prepended by a **?** character, all other arguments are separated by **&** characters.

3.1.4 Response Format

The default response format is json. You can control the format with the `format=` option in the data and ontology queries. Currently, json[6] and csv[7] are supported.

3.2 Description of the Services

For all services the default for the number of returned entries is 10, in case you want to see more entries use the query option `?limit=<integer: count>`

3.2.1 Stationmeta

```
Query:
https://toar-data.fz-juelich.de/api/v2/stationmeta/[id/][?QUERY-OPTIONS]

where QUERY-OPTIONS are:
limit=<integer: count> (examples: 10)
bounding_box=<min_lat>,<min_lon>,<max_lat>,<max_lon>
country=<country code>,<country code>, ...   (country code defined in ISO-3166 ALPHA-2)
htap_region_tier1_year2010=<htap region number>
...

Response:
Each query result consists of all fields of station metadata.
If no QUERY-OPTIONS are given, the complete set of stations will be returned.

Example:
https://toar-data.fz-juelich.de/api/v2/stationmeta/CPT134S00/

Further query items are:
* /stationmeta/{station_code}
* /stationmeta/id/{station_id}
* /stationmata/?<any station metadata field as defined in https://toar-data.fz-juelich.
↪de/api/v2/#stationmeta>
*
* /stationmeta_changelog/{station_id}
```

[6] https://www.json.org/json-en.html
[7] https://en.wikipedia.org/wiki/Comma-separated_values

3.2.2 Time Series

```
Query:
https://toar-data.fz-juelich.de/api/v2/timeseries/[?QUERY-OPTIONS]

where QUERY-OPTIONS are:
limit=<integer: count>
station_code=<station code1>,....
variable_id=<integer: variable identifier in TOAR BD>
format=<string> (json|csv)

Response:
Each query result consists of all fields of time series metadata.
If no QUERY-OPTIONS are given, the complete set of time series will be returned.

Example (1), query the first time series:
https://toar-data.fz-juelich.de/api/v2/timeseries/?limit=1

Example (2), query the time series with id 25:
https://toar-data.fz-juelich.de/api/v2/timeseries/25

Example (3), query the timeseries of ozone measurements at the three listed stations␣
↪(variable_id 5 = ozone)
https://toar-data.fz-juelich.de/api/v2/timeseries/?station_code=MX_PED,CPT134S00,CH0001G&
↪variable_id=5

Further query items are:
* /timeseries/{timeseries_id}
* /timeseries/id/{timeseries_id}
* /timeseries/unique/
* /timeseries/citation/{timeseries_id}
* /timeseries/?<any timeseries metadata field as defined in https://toar-data.fz-juelich.
↪de/api/v2/#timeseries>
*
* /timeseries_changelog/{timeseries_id}
```

3.2.3 Data

```
Query:
https://toar-data.fz-juelich.de/api/v2/data/timeseries/[?QUERY-OPTIONS]

where QUERY-OPTIONS are:
format = <string> (json|csv)
flags = <string> (see controlled vocabulary for data-flags: https://esde.pages.jsc.fz-
↪juelich.de/toar-data/toardb_fastapi/docs/toardb_fastapi.html#data-flag)

Response:
Each query result consists of the fields that are specified in the columns argument. If␣
↪columns are not specified, the output of each record will consist of the fieldsseries_
↪id, network_name, station_id, parameter_label as the series query.
If no QUERY-OPTIONS are given, the complete set of stations will be returned.
```

(continues on next page)

(continued from previous page)

```
Example (1), query data of time series with id "52":
https://toar-data.fz-juelich.de/api/v2/data/timeseries/52

Example (2), query data of time series with id "52" and return the result as comma␣
↪separated list:
https://toar-data.fz-juelich.de/api/v2/data/timeseries/52/?format=csv

Further query items are:
* /data/{timeseries_id}
* /data/id/{timeseries_id}
* /data/timeseries/{timeseries_id}?flags={flag_name}
* ...
```

3.2.4 Variables

```
Query:
https://toar-data.fz-juelich.de/api/v2/variables/[id/][?QUERY-OPTIONS]

where QUERY-OPTIONS are:
limit= <integer: count> (default: 10)

Response:
Each query result consists of a list of variables  with name, longname, dispalyname, cf_
↪standardname, units, chemical-formular, and its internal id, which can be used to␣
↪directly query that specific variable.

Further query items are:
* /variables/{name}   or /variable/?name={name}
* /variables/id/{variable_id}   or /variables/?id={variable_id}
```

3.2.5 Contacts

```
Query:
https://toar-data.fz-juelich.de/api/v2/contacts/[persons/|organisations/|id/][?QUERY-
↪OPTIONS]

where QUERY-OPTIONS are:
limit= <integer: count> (default: 10)

Response:
Each query result consists of a list of contacts, either all kinds, persons,␣
↪organisations, or the information for a specific id.

Further query items are:
* /contacts/persons/id/{person_id}
* /contacts/persons/{name}
* /contacts/organisations/id/{organisation_id}
```

(continues on next page)

(continued from previous page)

```
* /contacts/organisations/{name}
* /contacts/id/{contact_id}
```

3.2.6 Controlled Vocabulary

```
Query:
https://toar-data.fz-juelich.de/api/v2/controlled_vocabulary/

Response:
List of the complete vocabulary in json (raw) format.

Further query items are:
* /controlled_vocabulary/{name}
```

3.2.7 Database Statistics

```
Query:
https://toar-data.fz-juelich.de/api/v2/database_statistics/

Response:
The database statistics is given: number of users, number of stations, number of time␣
↪series, and the number of data records. You can also query for only one of these␣
↪numbers by using its name.

Further query items are:
* /database_statistics/{name}
```

3.2.8 Ontology

```
Query:
https://toar-data.fz-juelich.de/api/v2/ontology/[?QUERY-OPTIONS]

where QUERY-OPTIONS are:
format = <string> (xml|owl|doc)

Response:
By default, the query will return the ontology in xml format.

Example:
https://toar-data.fz-juelich.de/api/v2/ontology/?format=xml
```

3.2.9 Search

As basis for formulating searches use https://toar-data.fz-juelich.de/api/v2/#stationmeta, https://toar-data.fz-juelich.de/api/v2/#timeseries to list all metadata fields and their definitions as well as the controlled vocabulary defined for a subset of the metadata fields. All metadata fields can be used in searches and combined in one query with **&**:

```
Query: https://toar-data.fz-juelich.de/api/v2/search/[?QUERY-Options]

where QUERY-OPTIONS are
any metadata field = value (or comma separated list of values)

Response:
all metadata of all stations and timeseries fullfilling the criteria.

Example:
https://toar-data.fz-juelich.de/api/v2/search/?bounding_box=49,7,50,8&variable_id=5,4␣
↪will get you all the time series of stations within an area between 49°N 7°E and 50°N␣
↪8°E that record ozone or pm1

https://toar-data.fz-juelich.de/api/v2/search/?name=Aachen  will provide all stations␣
↪located in the town of Aachen (done via similarity search)
```

3.2.10 Analysis

The base URL for the analysis package, a web-service for the calculation of various anaylsis on TOAR data, is https://toar-data.fz-juelich.de/api/v2/analysis . There the API is documented, especially all available analysis methods are listed with their definitions.

```
Query: https://toar-data.fz-juelich.de/api/v2/analysis/[ENDPOINT]/[?QUERY-OPTIONS]

where ENDPOINT is:
data/timeseries: bulk time series download
statistics: aggregated time series download

where QUERY-Options are
any metadata field = value (or comma separated list of values)
flags = (see controlled vocabulary for data-flags: https://esde.pages.jsc.fz-juelich.de/
↪toar-data/toardb_fastapi/docs/toardb_fastapi.html#data-flag)
sampling = temporal aggregation (only for endpoint=statistics)
metrics = statistical aggregation (only for endpoint=statistics)


Response:
A zip file with the requested data in csv format.

Example:

Example (1), query data of all German time series:
https://toar-data.fz-juelich.de/api/v2/analysis/data/timeseries/?country=DE&limit=None

Example (2), query annual median values of all German time series:
```

(continues on next page)

(continued from previous page)

```
https://toar-data.fz-juelich.de/api/v2/analysis/statistics/?sampling=annual&
↪metrics=median&country=DE&limit=None
```

CHAPTER

FOUR

METADATA REFERENCE

The following sub sections describe the metadata of the TOAR V2 database following the structure of high-level criteria of FAIR data management. For a detailed description of metadata attributes of the individual database tables and a list of all controlled vocabulary definitions, see https://esde.pages.jsc.fz-juelich.de/toar-data/toardb_fastapi/docs/toardb_fastapi.html. There you will always find the up to date information.

4.1 Variables

While the main purpose of the TOAR V2 database is to provide ground-level ozone concentration time series, the database also contains data for several ozone precursor variables and meteorological information. Table 4.1 below provides a summary of the variables in the TOAR database including their short name, long name and physical units. Available variables can be queried as described in Section 3.2.4.

Table 4.1: Variables in the TOAR database

Variable Name	Variable long name	Units
albedo	albedo	%
aswdifu	diffuse upward sw radiation	W/m**2
aswdir	direct downward sw radiation	W/m**2
bc	black carbon	nmol mol-1
benzene	benzene	nmol mol-1
ch4	Methane	nmol mol-1
cloudcover	total cloud cover	%
co	carbon monoxide	nmol mol-1
ethane	Ethane	nmol mol-1
humidity	atmospheric humidity	g kg-1
irradiance	global surface irradiance	W m-2
mpxylene	m,p-xylene	nmol mol-1
no	nitrogen monoxide	nmol mol-1
no2	nitrogen dioxide	nmol mol-1
nox	reactive nitrogen oxides (NO+NO2)	nmol mol-1
o3	ozone	nmol mol-1
ox	Ox	nmol mol-1
oxylene	o-xylene	nmol mol-1
pblheight	height of PBL	m
pm1	particles up to 1 µm diameter	µg m-3
pm10	particles up to 10 µm diameter	µg m-3
pm2p5	particles up to 2.5 µm diameter	µg m-3
press	atmospheric pressure	hPa

continues on next page

Table 4.1 – continued from previous page

Variable Name	Variable long name	Units
propane	Propane	nmol mol-1
relhum	relative humidity	%
rn	radon	mBq m-3
so2	Sulphur dioxide	nmol mol-1
temp	atmospheric temperature	degC
toluene	toluene	nmol mol-1
totprecip	total precipitation	kg m-2
u	u-component (zonal) of wind	m s-1
v	v-component (meridional) of wind	m s-1
wdir	wind direction	degree
wspeed	wind speed	m s-1

Within the TOAR V2 database we store the following information about each variable:

- Variable Name: a short name to identify the variable (see Table 4.1, left column)
- Variable long name: a more descriptive name of the variable (see Table 4.1, middle column)
- Displayname: a variant of the variable name that is recommended for plotting
- Cf_standardname: a standardized description of the variable quantity (see http://cfconventions.org/standard-names.html)
- Units: a string defining the physical units in which the variable data are stored in the TOAR database. Note that we apply unit conversion in case we receive data in different units (see Table 4.1, right column)
- Chemical_formula: variables which express mixing ratio or concentration values are sometimes named by their chemical formula and sometimes as clear names. This depends on common practice. This field will always contain the chemical formula of such variables (e.g. C6H6 for the variable benzene).

4.2 Station Characterisation

Air pollution levels are controlled by several factors. Among the most important factors are the proximity to emission sources and the geographic environment around a measurement site. As a user you may often want to stratify air pollution data with respect to certain site characteristics, e.g. „urban“ or „rural“. There are numerous ways in which environmental agencies around the world define metadata attributes to describe stations in a standardised way. However, these standardisations differ widely across regions. Furthermore, data contributed from individual research groups often do not follow the standardised terminology of environmental agencies, because the employed terms do not seem to be appropriate for the description of the specific site which is operated by the research group. The problem of labelling stations as “urban” or “rural” is quite complex as can be demonstrated with using population density as proxy. “Built-up areas” which constitute major cities in Europe may be regarded as relatively small villages in other parts of the world, e.g. in East Asia, South Asia, or Africa. Even if population density (and total number of people) in such a “village” in India, for example, may be much larger than in, say, a German city, the air pollutant emissions (with respect to ozone precursors at least) may be much greater in the small city compared to the large village. Therefore, the use of simple proxy variables will generally not lead to a meaningful separation between (ozone) air pollution regimes.

The TOAR database offers various ways for the characterisation of measurement stations and we try to harmonise the employed terminology to the extent possible. There are four different approaches to station

characterisation implemented in the TOAR database and its corresponding web services. These are described below in the order of increasing complexity and decreasing level of harmonisation. For analyses supporting the TOAR-II assessment, we recommend the use of the TOAR station characterisation (Section 4.2.2), perhaps augmented with information from specific global metadata fields (Table 4.6) and, for individual sites and where available, with detailed station descriptions (Section 4.2.5).

4.2.1 Station Location

The locations of measurement sites are stored in the TOAR database with at least 4 decimals. In theory, this allows the pinpointing of stations within 12 m or less. However, in reality, the coordinates may not be as precise as this, because the inlet of the air quality measurements may be located away from the station building, or station locations have been reported with wrong or imprecise coordinates. We therefore perform some coordinate validation of the metadata in the TOAR database (details given in[9]) and document any changes that are applied to station coordinates in the metadata changelog (see Section 4.3.2).

Geographical coordinates are saved as a PostGIS POINT location with lat and lon given in degrees_north and degrees_east, respectively, using the World Geodetic System (WGS) 84 coordinate reference system. Station altitudes are given in metres. Note that the station altitude value refers to the ground-level altitude of the measurement site. Air sampling inlets are typically at 10-15 m above ground. Where available, the sampling height is stored in the metadata of each measurand's time series as the sampling heights may differ between species.

Table 4.2: country, state, and timezone

Name	Description
country	The country, where the station resides, or which operates the station (e.g. in Antarctica) (see controlled vocabulary: Country Code)
state	The state or province, where the station resides
timezone	Station timezone (see controlled vocabulary: Timezone)

4.2.2 TOAR Station Characterisation

For the analysis of ground-level ozone monitoring data in the first TOAR assessment, a globally applicable station characterisation scheme was defined based on several geospatial datasets[9] . Four categories of stations were defined, which were expected to yield different patterns of ozone pollution and allow for some separation of ozone trends and their causes. The main goal was to find a distinction between "urban" and "rural" sites, i.e. sites which exhibit clear pollution signatures from either category. Due to the different ozone patterns at high altitude stations, a third category "rural, high elevation" was added. To enhance the separation between the "urban" and "rural" classes, threshold values for population density and other parameters were defined relatively rigidly. As a result, about 50% of all stations were not associated with either class and were therefore labelled as "unclassified".

The table below summarizes the criteria which we employed in the "toar1_category" (this is the name of the corresponding metadata field in the TOAR database and REST API). It should be noted that the definition of the threshold criteria in Table 4.3 was somewhat ad-hoc and based on a somewhat subjective analysis.

[9] TOAR V1 is described in Schultz, M. G. et al. (2017) Tropospheric Ozone Assessment Report: Database and Metrics Data of Global Surface Ozone Observations, Elem Sci Anth, 5, p.58. DOI: http://doi.org/10.1525/elementa.244

Table 4.3: Summary of criteria for the toar1_category (see[9]). For details on the specific geospatial variables, see Section 4.2.4

toar1_category value	geospatial criteria
Urban	is defined as: station_population_density >= 15000 and station_nightlight_1km >= 60 and station_max_nightlight_25km == 63
RuralLowElevation	station_omi_no2_column <= 8 and station_nightlight_5km <= 25 and station_population_density <= 3000 and station_max_population_density_5km <= 30000 and station_google_alt <= 1500 and station_etopo_relative_alt < 500
RuralHighElevation	station_omi_no2_column <= 8 and station_nightlight_5km <= 25 and station_population_density <= 3000 and (station_google_alt > 1500 or (station_google_alt > 800 and station_etopo_relative_alt < 500))
Unclassified	no classification given

We are planning to use cluster techniques to define a more objective set of station classes for the second TOAR assessment. First, preliminary results appear promising, but it should be noted that even with such techniques there will always be some subjective moment regarding, for example, the number of clusters that are “meaningful”, or the evaluation of the separation, i.e. the criteria used to measure “success”. Depending on the outcomes of this effort, a “toar2_category” may be added to the TOAR database at a later stage.

4.2.3 European Station Characterisation Scheme

Since 2018, the rules for reporting air quality data including the metadata describing the site locations, have been laid out in the “Member States’ and European Commission’s Common Understanding of the Commission Implementing Decision laying down rules for Directives 2004/107/EC and 2008/50/EC of the European Parliament and of the Council as regards the reciprocal exchange of information and reporting on ambient air[10] ”. Annex II of this document describes the terms used in the European air quality database (Airbase).

[10] DIRECTIVE 2008/50/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 21 May 2008 on ambient air quality and cleaner air for Europe, available from https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32008L0050, last accessed: 11 Jul 2022

Table 4.4: Station classification in relation to prominent emission sources (Decision Annex II D(ii), item 22) (see also: http://dd.eionet.europa.eu/vocabulary/aq/stationclassification for an electronic version)

station_type	description
Traffic	Located in close proximity to a single major road.
Industrial	Located in close proximity to a single industrial source or industrial area. A wide range of industrial sources can be considered here, including - thermal power generation - district heating plants - refineries - waste incineration/treatment plants, dump sites - mining, including gravel, oil, natural gas - airports - ports.
Background	Any location with is neither to be classified as “traffic” or “industrial”. Located such that its pollution levels are representative of the average exposure of the general population (or vegetation and natural ecosystems) within the type of area under assessment. The pollution level should not be dominated by a single source type (e.g. traffic), unless that source type is typical within the area under assessment. The station should usually be representative of a wider area of at least several square kilometres.

Table 4.5: Classification of the Area (Decision Annex II D(ii), item 28) (see also the electronic version of this vocabulary at http://dd.eionet.europa.eu/vocabulary/aq/areaclassification/view)

station_type_of_area	description
urban	Continuously built-up urban area meaning complete (or at least highly predominant) building-up of the street front side by buildings with at least two floors or large detached buildings with at least two floors. With the exception of city parks, large railway stations, urban motorways and motorway junctions, the built-up area is not mixed with non-urbanised areas.
suburban	Largely built-up urban area. 'Largely built-up' means contiguous settlement of detached buildings of any size with a building density less than for 'continuously built-up' area. The built-up area is mixed with non-urbanised areas (e.g. agricultural, lakes, woods). It must also be noted that 'suburban' as defined here has a different meaning than in every day English i.e. 'an outlying part of a city or town' suggesting that a suburban area is always associated to an urban area. In our context, a suburban area can be suburban on its own without any urban part.
rural	All areas, that do not fulfil the criteria for urban or suburban areas, are defined as "rural" areas. There are three subdivisions in this category to indicate the distance to the nearest built-up urban area: * Rural – near city: area within 10 km from the border of an urban or suburban area; * Rural – regional: 10-50 km from major sources/source areas; * Rural - remote: > 50 km from major sources/source areas.

While the use of these categories may be useful for the analysis of European air quality data, we note that non-European data providers generally use different categories and definitions to label their measurement sites. While we try to harmonize the values of this attribute, these labels remain somewhat subjective for non-European data.

4.2.4 Station Characterisation Through Geospatial Data

The “toar1_category” (Section 4.2.2) offers an easy-to-use classification scheme that can be universally applied to air quality stations worldwide. Often, this crude classification will be insufficient to capture important air pollution features at specific site types so that typical statistical properties of air quality time series from such sites will get lost in the mixture of sites subsumed in the broader classification. For example, coastal and island sites often exhibit typical diurnal cycles of ozone concentrations which differ markedly from stations further inland.

To allow for more refined analyses of air quality data, version 2 of the TOAR database offers an extended variety of metadata elements to characterize stations. These metadata elements have been derived from several geospatial datasets at spatial resolutions from 90 m to 10 km. As air quality data analyst you may often be more interested in the area around a measurement station than in the geospatial properties at the site location itself. Therefore, in addition to the pixel value at the location of the measurement site, we often provide aggregated values of the geospatial data within distances of 5 and 25 km to the site location. The aggregation method depends on the geospatial field. For example, we will report “max_population_density_25km_year2015” and “mean_nightlights_5km_year2013”.

Table 4.6 lists the geospatial field names, that are available for the TOAR station characterisation. Detailed descriptions and service URLs can be found at https://esde.pages.jsc.fz-juelich.de/toar-data/toardb_fastapi/docs/toardb_fastapi.html#stationmetaglobal and https://esde.pages.jsc.fz-juelich.de/toar-data/toardb_fastapi/docs/toardb_fastapi.html#geolocation-urls respectively.

Table 4.6: StationmetaGlobal - TOAR database fields of geospatial information for the characterisation of measurement sites

Name	Type	Description
mean_srtm_alt_90m_year1994	number	mean value within a radius of 90 m around station location of the following data of the year 1994: {‘units’: ‘m’, ‘data_source’: ‘NASA Shuttle Radar Topographic Mission (SRTM)’, ‘citation’: ‘Jarvis, A., H.I. Reuter, A. Nelson, E. Guevara, 2008, Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database (http://srtm.csi.cgiar.org).’}
mean_srtm_alt_1km_year1994	number	mean value within a radius of 1 km around station location of the following data of the year 1994: {‘units’: ‘m’, ‘data_source’: ‘NASA Shuttle Radar Topographic Mission (SRTM)’, ‘citation’: ‘Jarvis, A., H.I. Reuter, A. Nelson, E. Guevara, 2008, Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database (http://srtm.csi.cgiar.org).’}

continues on next page

Table 4.6 – continued from previous page

Name	Type	Description
max_srtm_relative_alt_5km_year1994	number	maximum value within a radius of 5 km around station location with relative altitude of the following data of the year 1994: {'units': 'm', 'data_source': 'NASA Shuttle Radar Topographic Mission (SRTM)', 'citation': 'Jarvis, A., H.I. Reuter, A. Nelson, E. Guevara, 2008, Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database (http://srtm.csi.cgiar.org).'}
min_srtm_relative_alt_5km_year1994	number	minimum value within a radius of 5 km around station location with relative altitude of the following data of the year 1994: {'units': 'm', 'data_source': 'NASA Shuttle Radar Topographic Mission (SRTM)', 'citation': 'Jarvis, A., H.I. Reuter, A. Nelson, E. Guevara, 2008, Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database (http://srtm.csi.cgiar.org).'}
stddev_srtm_relative_alt_5km_year1994	number	standard deviation within a radius of 5 km around station location with relative altitude of the following data of the year 1994: {'units': 'm', 'data_source': 'NASA Shuttle Radar Topographic Mission (SRTM)', 'citation': 'Jarvis, A., H.I. Reuter, A. Nelson, E. Guevara, 2008, Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database (http://srtm.csi.cgiar.org).'}
climatic_zone_year2016	string	value for the year 2016 of the following data: {'units': 'None', 'data_source': 'University of East Anglia Climatic Research Unit; Harris, I.C.; Jones, P.D. (2017): CRU TS4.00: Climatic Research Unit (CRU) Time-Series (TS) version 4.00 of high-resolution gridded data of month-by-month variation in climate (Jan. 1901- Dec. 2015). Centre for Environmental Data Analysis, 25 August 2017 (http://dx.doi.org/10.5285/edf8febfdaad48abb2cbaf7d7e846a86)', 'citation': 'University of East Anglia Climatic Research Unit; Harris, I.C.; Jones, P.D. (2017): CRU TS4.00: Climatic Research Unit (CRU) Time-Series (TS) version 4.00 of high-resolution gridded data of month-by-month variation in climate (Jan. 1901- Dec. 2015). Centre for Environmental Data Analysis, 25 August 2017 (http://dx.doi.org/10.5285/edf8febfdaad48abb2cbaf7d7e846a86)'} (see controlled vocabulary: Climatic Zone 2019)
htap_region_tier1_year2010	string	value for the year 2010 of the following data: The 'tier1' region defined in the task force on hemispheric transport of air pollution (TFHTAP) coordinated model studies according to figure 4 of https://publications.jrc.ec.europa.eu/repository/bitstream/JRC102552/lbna28255enn.pdf (see controlled vocabulary: Station HTAP Region)

continues on next page

Table 4.6 – continued from previous page

Name	Type	Description
dominant_landcover_year2012	string	value for the year 2012 of the following data: {'units': 'no unit', 'data_source': 'ESA 2017 and UCLouvain', 'citation': 'ESA. Land Cover CCI Product User Guide Version 2. Tech. Rep. (2017). Available at: http://maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf'} (see controlled vocabulary: Station Landcover Type)
landcover_description_year2012	string	description of the values for the year 2012 within a radius of 25 km around station location of the following data: {'units': 'no unit', 'data_source': 'ESA 2017 and UCLouvain', 'citation': 'ESA. Land Cover CCI Product User Guide Version 2. Tech. Rep. (2017). Available at: http://maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf'} (see controlled vocabulary: Station Landcover Type)
dominant_ecoregion_year2017	string	value for the year 2017 of the following data: {'units': 'None', 'data_source': 'RESOLVE Biodiversity and Wildlife Solutions', 'citation': 'Eric Dinerstein, David Olson, Anup Joshi, Carly Vynne, Neil D. Burgess, Eric Wikramanayake, Nathan Hahn, Suzanne Palminteri, Prashant Hedao, Reed Noss, Matt Hansen, Harvey Locke, Erle C Ellis, Benjamin Jones, Charles Victor Barber, Randy Hayes, Cyril Kormos, Vance Martin, Eileen Crist, Wes Sechrest, Lori Price, Jonathan E. M. Baillie, Don Weeden, Kieran Suckling, Crystal Davis, Nigel Sizer, Rebecca Moore, David Thau, Tanya Birch, Peter Potapov, Svetlana Turubanova, Alexandra Tyukavina, Nadia de Souza, Lilian Pintea, Jose C. Brito, Othman A. Llewellyn, Anthony G. Miller, Annette Patzelt, Shahina A. Ghazanfar, Jonathan Timberlake, Heinz Klöser, Yara Shennan-Farpon, Roeland Kindt, Jens-Peter Barnekow Lilleso, Paulo van Breugel, Lars Graudal, Maianna Voge, Khalaf F. Al-Shammari, Muhammad Saleem, An Ecoregion-Based Approach to Protecting Half the Terrestrial Realm, BioScience, Volume 67, Issue 6, June 2017, Pages 534–545, https://doi.org/10.1093/biosci/bix014'} (see controlled vocabulary: Station ECO Region Type)

continues on next page

Table 4.6 – continued from previous page

Name	Type	Description
ecoregion_description_year2017	string	description of the values for the year 2017 within a radius of 25 km around station location of the following data: {'units': 'None', 'data_source': 'RESOLVE Biodiversity and Wildlife Solutions', 'citation': 'Eric Dinerstein, David Olson, Anup Joshi, Carly Vynne, Neil D. Burgess, Eric Wikramanayake, Nathan Hahn, Suzanne Palminteri, Prashant Hedao, Reed Noss, Matt Hansen, Harvey Locke, Erle C Ellis, Benjamin Jones, Charles Victor Barber, Randy Hayes, Cyril Kormos, Vance Martin, Eileen Crist, Wes Sechrest, Lori Price, Jonathan E. M. Baillie, Don Weeden, Kieran Suckling, Crystal Davis, Nigel Sizer, Rebecca Moore, David Thau, Tanya Birch, Peter Potapov, Svetlana Turubanova, Alexandra Tyukavina, Nadia de Souza, Lilian Pintea, Jose C. Brito, Othman A. Llewellyn, Anthony G. Miller, Annette Patzelt, Shahina A. Ghazanfar, Jonathan Timberlake, Heinz Klöser, Yara Shennan-Farpon, Roeland Kindt, Jens-Peter Barnekow Lilleso, Paulo van Breugel, Lars Graudal, Maianna Voge, Khalaf F. Al-Shammari, Muhammad Saleem, An Ecoregion-Based Approach to Protecting Half the Terrestrial Realm, BioScience, Volume 67, Issue 6, June 2017, Pages 534–545, https://doi.org/10.1093/biosci/bix014'} (see controlled vocabulary: Station ECO Region Type)
distance_to_major_road_year2020	number	value for the year 2020 of the following data: {'version': 0.6, 'generator': 'Overpass API 0.7.55.9 ab41fea6', 'copyright': 'https://www.openstreetmap.org/copyright', 'timestamp': ''}
mean_nightlight_1km_year2013	number	mean value within a radius of 1 km around station location of the following data of the year 2013: {'units': 'None', 'data_source': 'NOAA National Centers for Environmental Information (NCEI)', 'citation': 'None'}
mean_nightlight_5km_year2013	number	mean value within a radius of 5 km around station location of the following data of the year 2013: {'units': 'None', 'data_source': 'NOAA National Centers for Environmental Information (NCEI)', 'citation': 'None'}
max_nightlight_25km_year2013	number	maximum value within a radius of 5 km around station location of the following data of the year 2013: {'units': 'None', 'data_source': 'NOAA National Centers for Environmental Information (NCEI)', 'citation': 'None'}
max_nightlight_25km_year1992	number	maximum value within a radius of 25 km around station location of the following data of the year 2013: {'units': 'None', 'data_source': 'NOAA National Centers for Environmental Information (NCEI)', 'citation': 'None'}

continues on next page

Table 4.6 – continued from previous page

Name	Type	Description
mean_population_density_250m_year2015	number	mean value within a radius of 250 m around station location of the following data of the year 2015: {'data_source': 'The European Commission, Joint Research Centre', 'citation': 'Schiavina, Marcello; Freire, Sergio; MacManus, Kytt (2019): GHS-POP R2019A - GHS population grid multitemporal (1975-1990-2000-2015). European Commission, Joint Research Centre (JRC) [Dataset] doi:10.2905/0C6B9751-A71F-4062-830B-43C9F432370F PID: http://data.europa.eu/89h/0c6b9751-a71f-4062-830b-43c9f432370f'}
mean_population_density_5km_year2015	number	mean value within a radius of 5 km around station location of the following data of the year 2015: {'data_source': 'The European Commission, Joint Research Centre', 'citation': 'Schiavina, Marcello; Freire, Sergio; MacManus, Kytt (2019): GHS-POP R2019A - GHS population grid multitemporal (1975-1990-2000-2015). European Commission, Joint Research Centre (JRC) [Dataset] doi:10.2905/0C6B9751-A71F-4062-830B-43C9F432370F PID: http://data.europa.eu/89h/0c6b9751-a71f-4062-830b-43c9f432370f'}
max_population_density_25km_year2015	number	maximum value within a radius of 25 km around station location of the following data of the year 2015: {'data_source': 'The European Commission, Joint Research Centre', 'citation': 'Schiavina, Marcello; Freire, Sergio; MacManus, Kytt (2019): GHS-POP R2019A - GHS population grid multitemporal (1975-1990-2000-2015). European Commission, Joint Research Centre (JRC) [Dataset] doi:10.2905/0C6B9751-A71F-4062-830B-43C9F432370F PID: http://data.europa.eu/89h/0c6b9751-a71f-4062-830b-43c9f432370f'}
mean_population_density_250m_year1990	number	human population on a square of 250 m for the year 1990 (residents km-2)
mean_population_density_5km_year1990	number	mean value within a radius of 250 m around station location of the following data of the year 1990: {'data_source': 'The European Commission, Joint Research Centre', 'citation': 'Schiavina, Marcello; Freire, Sergio; MacManus, Kytt (2019): GHS-POP R2019A - GHS population grid multitemporal (1975-1990-2000-2015). European Commission, Joint Research Centre (JRC) [Dataset] doi:10.2905/0C6B9751-A71F-4062-830B-43C9F432370F PID: http://data.europa.eu/89h/0c6b9751-a71f-4062-830b-43c9f432370f'}

continues on next page

Table 4.6 – continued from previous page

Name	Type	Description
max_population_density_25km_year1990	number	maximum value within a radius of 25 km around station location of the following data of the year 1990: {'data_source': 'The European Commission, Joint Research Centre', 'citation': 'Schiavina, Marcello; Freire, Sergio; MacManus, Kytt (2019): GHS-POP R2019A - GHS population grid multitemporal (1975-1990-2000-2015). European Commission, Joint Research Centre (JRC) [Dataset] doi:10.2905/0C6B9751-A71F-4062-830B-43C9F432370F PID: http://data.europa.eu/89h/0c6b9751-a71f-4062-830b-43c9f432370f'}
mean_nox_emissions_10km_year2015	number	mean value within a radius of 10 km around station location of the following data of the year 2015: {'units': 'kg m-2 s-1', 'data_source': 'https://atmosphere.copernicus.eu/sites/default/files/2019-06/cams_emissions_general_document_apr2019_v7.pdf', 'citation': "Granier, C., S. Darras, H. Denier van der Gon, J. Doubalova, N. Elguindi, B. Galle, M. Gauss, M. Guevara, J.-P. Jalkanen, J. Kuenen, C. Liousse, B. Quack, D. Simpson, K. Sindelarova The Copernicus Atmosphere Monitoring Service global and regional emissions (April 2019 version) Report April 2019 version null 2019 Elguindi, Granier, Stavrakou, Darras et al. Analysis of recent anthropogenic surface emissions from bottom-up inventories and top-down estimates: are future emission scenarios valid for the recent past? Earth's Future null submitted 2020"}
mean_nox_emissions_10km_year2000	number	mean value within a radius of 10 km around station location of the following data of the year 2000: {'units': 'kg m-2 s-1', 'data_source': 'https://atmosphere.copernicus.eu/sites/default/files/2019-06/cams_emissions_general_document_apr2019_v7.pdf', 'citation': "Granier, C., S. Darras, H. Denier van der Gon, J. Doubalova, N. Elguindi, B. Galle, M. Gauss, M. Guevara, J.-P. Jalkanen, J. Kuenen, C. Liousse, B. Quack, D. Simpson, K. Sindelarova The Copernicus Atmosphere Monitoring Service global and regional emissions (April 2019 version) Report April 2019 version null 2019 Elguindi, Granier, Stavrakou, Darras et al. Analysis of recent anthropogenic surface emissions from bottom-up inventories and top-down estimates: are future emission scenarios valid for the recent past? Earth's Future null submitted 2020"}
wheat_production_year2000	number	no wheat production metadata stored yet

continues on next page

Table 4.6 – continued from previous page

Name	Type	Description
rice_production_year2000	number	no rice production metadata stored yet
omi_no2_column_years2011to2015	number	no OMI NO2 column metadata stored yet
toar1_category	string	The station classification for the Tropsopheric Ozone Assessment Report based on the station proxy data that are stored in the TOAR database (see controlled vocabulary: Station TOAR Category)

Note that the geospatial data that are incorporated in the TOAR database may not always be accurate at the local scale. Most of these data have been derived from satellite measurements of various physical properties (e.g. reflectance) of the Earth surface, and measurement errors or imperfect retrieval algorithms may lead to occasional errors. Note also that the "geospatial settings" around a measurement station can change with time. For example, in rapidly developing regions a station which had been located in a rural setting when it was established might be completely surrounded by buildings and roads a few years later. We therefore store geospatial data of different years in our backend services and in some cases we calculate the metadata values for at two different years, so that you can use this information as an indication for the change in the drivers of air pollution trends.

4.2.5 Individual Station Description

While the station information provided through methods 1-3 (Section 4.2.1 to Section 4.2.3) is largely consistent across the globe, there may be additional, relevant information about measurement sites that cannot be captured by the metadata elements described so far. For this reason, the TOAR V2 database allows storage of additional information which can help to characterise a measurement station and thus guide the analysis of air pollution data from that site.

Three types of auxiliary data can be submitted to the TOAR data centre as supporting information about stations:

1. URLs to web sites with detailed station information,
2. StationmetaAuxDoc - PDF documents with station descriptions (any language, but English would be preferred),
3. Photographs of the station buildings and facilities.

Download links for this information can be obtained together with all other station metadata from the REST API query stationmeta (see Section 3.2.1).

Finally, any other information about a station can be provided in the form of a structured JSON string ("additional_metadata" field). This feature is used to capture station metadata information from different data providers which cannot be mapped directly to the metadata fields defined in the TOAR database. Such information is extracted from the submitted data files when the data are uploaded into the database. We ask data providers to begin such metadata elements with "station_" (see TOAR Data Submission Format). An example is given below.

```
additional_metadata =
  {'station_environment': 'situated in a forest clearing near a small lake',
   'station_year_of_construction': 1954
  }
```

Fig. 4.1: Example of additional station metadata elements as they can be extracted from submitted data files

4.3 Provenance Information

Provenance is the chronology of the ownership, custody or location of a historical object (Wikipedia, 2021, citing the Oxford English Dictionary). In FAIR data management, provenance is important to trace the ownership of a data record and possible modifications which were applied to data and metadata after the data record has been created. Ideally, all data should have a complete track record from the measurement to the data analysis or visualisation in a scientific article, on a web page, etc. For air quality data, this is rarely possible up to now, because most data providers don't maintain complete records of their data processing or because such records are not published in machine-readable digital format. In the TOAR database, we try to capture all provenance information that is made available to us by the data providers and we have implemented several measures to ensure that all modifications applied to data and metadata which we apply as part of the data curation process are captured and documented. This comprises the preservation of information about the institution and/or person who has done something with the data (so-called role codes), the archival of any changes applied to the metadata after initial screening of the data we receive[11] , a versioning scheme for data sets (i.e. time series), and the inclusion of provenance information in our data quality flags (see Section 5.1). The following sub sections describe these elements in more detail.

4.3.1 Role Codes

Different people and/or institutions are involved in the processing of a dataset from the original measurement to the provision of the data via files or a web service. Likewise, as part of the data curation performed at the TOAR data centre, some metadata elements or data values may be modified, for example in order to harmonize the metadata elements ("controlled vocabulary"), or during quality control of time series. Role codes define specific actions or responsibilities of people or organisations so that it becomes traceable who has done what with the data. The ISO19115[12] Standard defines a set of 20 role codes. We adopted a subset of these role codes for the TOAR database to maximize interoperability. However, as the definitions of the role codes provided by ISO are very abstract, we have extended the role codes table with our own definitions of the roles as we understand them in the context of air quality data management. Table 4.7 lists the role codes which are used in the TOAR database and their extended definition strings.

Table 4.7: The role codes of ISO19115 and their definition in the TOAR database

Internal Number	*Role Code*	*Role Code Definition*
0	Point of Contact	Party who can be contacted for acquiring knowledge about or acquisition of the resource

[11] It happens sometimes that we must manually correct spelling, date formats or other information, before we can submit new data to our automated data ingestion workflow, which keeps track of all modifications. In these cases, not all changes made to the data are preserved, but the raw data files will be archived and can be made available for comparison.

[12] https://standards.iso.org/iso/19115/resources/Codelists/gml/CI_RoleCode.xml

Roles are documented for station metadata and for time series metadata and data (Fig. 4.2). More than one role can be defined for each station or time series record. According to the ISO definition, role codes can be assigned to an institution or to a person or to both. In the TOAR database this is handled via the generic Contact model, which has one field for person and one field for organisation. Fig. 4.3 provides an example for the definition of roles in the metadata of an ozone measurement time series.

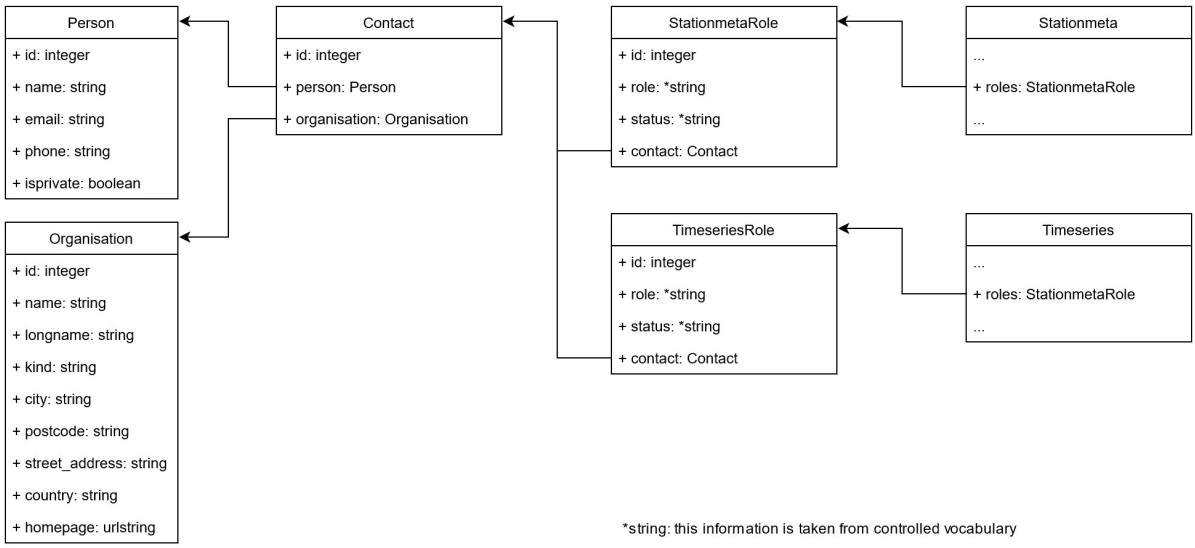


Fig. 4.2: TOAR database model for recording roles of people and organisations in the data creation and curation process

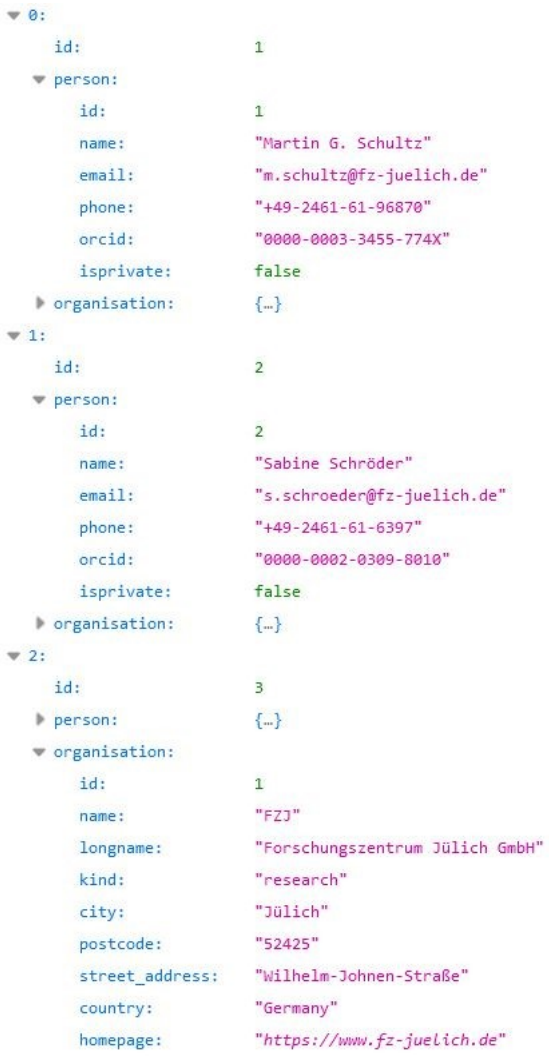


Fig. 4.3: Example metadata describing the roles of people and organisations involved in the creation and storage of an ozone time series from the German Umweltbundesamt

4.3.2 Metadata Change Logs

All station and time series metadata records are associated with a changelog table which may contain 1..N change records for every specific station and timeseries entry preserving any modifications applied to the metadata. Figure 5 shows the structure of the StationmetaChangelog and TimeseriesChangelog records. Both structures record the date and time when the modification was made, a free text description of the applied change, a JSON formatted string with the old and new values, a reference to the station or time series, the numerical id of the author who applied the change, and a change type field, which uses controlled vocabulary (see Table 4.8). The changelog of a time series is not only used to save modifications of the metadata, but they normally also contain a summary of modifications applied to the data values of this time series. Exceptions are made for near realtime data streams where new data records are not monitored via the changelog mechanism to avoid the excessive creation of trivial metadata. To allow for the tracking of data changes, the TimeseriesChangelog structure contains the additional fields period_start, period_end, and version. The latter refers to the version number after the change has been applied (see Section 4.3.3).

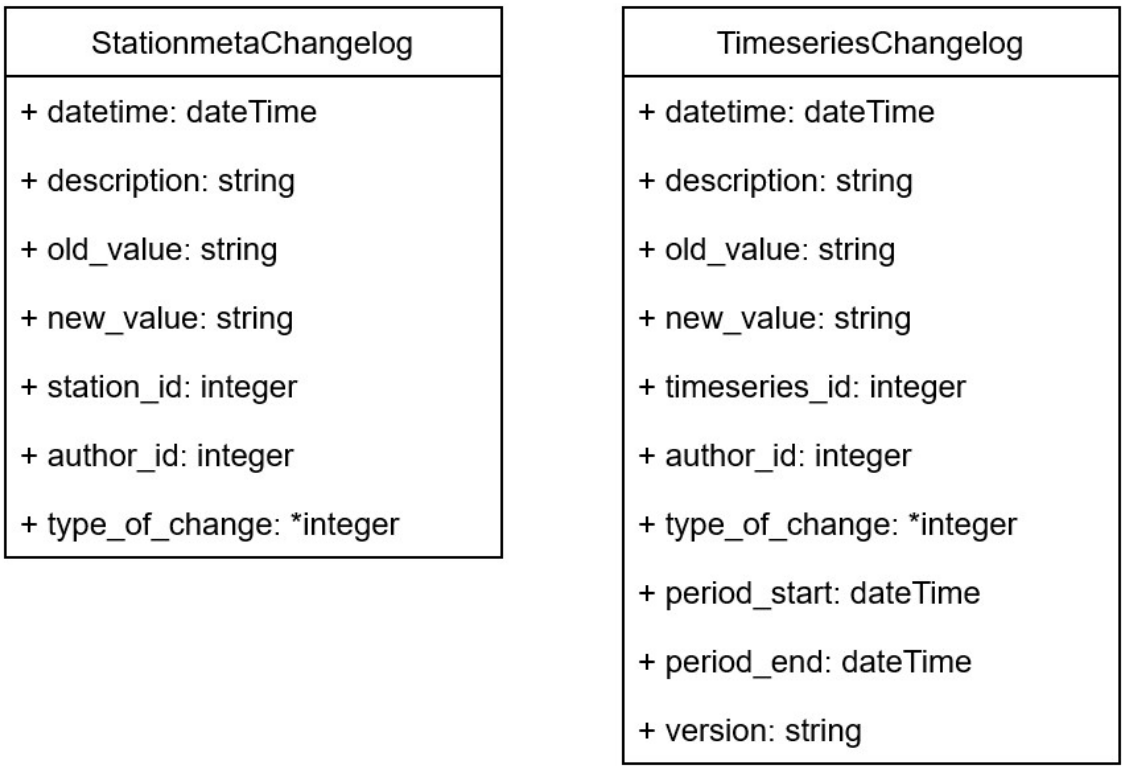


Fig. 4.4: Structure of StationmetaChangelog and TimeseriesChangelog records. Each Stationmeta or Timeseries entry may contain 1..N Changelog entries.

Table 4.8: List of change types for StationmetaChangelog and TimeseriesChangelog. Change types 4-6 only apply to TimeseriesChangelog records.

Value	*Name*	*Description*
0	Created	created
1	SingleValue	single value correction in metadata
2	Comprehensive	comprehensive metadata revision
3	Typo	typographic correction of metadata
4	UnspecifiedData	typographic correction of metadata
5	Replaced	replaced data with a new version
6	Flagging	data value flagging

4.3.3 Time Series Versioning

Any modification to the data values of a TOAR time series leads to a new time series version number. Furthermore, as described above, all changes (except for the addition of near realtime data) are documented in a corresponding changelog entry.

The version numbers of TOAR time series follow the common triple notation major.minor.micro (see for example PEP440 of Python). For technical reasons, version strings are internally stored in a fixed length format (example 000001.000001.20200911100000). The TOAR REST API and web interfaces will display the version numbers in a truncated user-friendly form (1.1.2020-09-11T11:10:0000). As the example shows, we use the micro number to store a date label. This facilitates the handling of near realtime data, because it allows to preserve the information when the last modification was made to the time series without having to add a changelog entry for each value addition.

Preliminary data will always have a major version number of 0. Once data have been approved (or "validated") by the data provider, the version number is at least 1. Any change in the major version number implies that at least 25% or one full year of the data were modified or replaced (this includes changes in the data quality flags). In practice, this occurs if we receive updates of entire time series or several years, or if data need to be re-calibrated. If new data are appended to an existing time series as a result of a new data submission, only the minor version number will be increased and the micro version number will be set to the modification date, regardless of the length of the new data fragment. As mentioned above, the addition of new near-realtime data samples only changes the micro version number. Changes to the version number occur automatically as part of the data ingestion workflow (see Automated Data Preparation). However, it is also possible that the TOAR data curators manually increase a time series version, for example after a thorough evaluation and data quality flagging exercise.

The data values of deprecated versions are preserved in a special table named "data_archive". There is currently no interface planned to allow users the reconstruction of time series corresponding to a specific version number. This requires manual intervention of the TOAR database curators. However, the main purpose of the time series version number is to allow comparisons between data downloaded at different times: if the version number has changed between two downloads, users can use the changelog information to find out what happened in the meantime and decide which version they should use for their analysis.

4.3.4 Provenance in Data Quality Flags

The TOAR data quality flags are explained in Section 5.2. In the context of provenance, it is only relevant to highlight the fact that the names of the quality flags contain a statement of what we as TOAR data curators have done to the data quality status (e.g. "_confirmed"). Table 5.2 in Section 5.2 contains detailed definitions of the data quality flags which explicitly describe whether a flag value has been set by the original data provider or by the TOAR data curators and document if the data quality flag value has been changed as a result of the TOAR data quality control procedures. We note that the flagging scheme allows the reconstruction of the original provider flagging with one exception: if validated data sent to us contains no flagging information, we first assume that all data are OK and modify the data quality flag only if our automated quality control routine detects suspicious or clearly erroneous features. It is thus not possible to reconstruct from the data in the database whether data was explicitly flagged as OK or simply not flagged at all.

4.3.5 Description of the Data Origin

The TOAR database contains air quality and meteorological observations as well as meteorological values from numerical weather models to allow for more elaborate analyses of ozone variability and changes. In the future, we may also add time series to the database which are generated through machine learning, for example to fill gaps in the measurement time series. It is therefore important to preserve information about the data source, i.e. whether data comes from a measurement, a numerical model, or a machine learning model. This is expressed in the metadata element data_origin_type, which can assume the values 'measurement' or 'model'.

For the measurement of air pollutant concentrations and meteorological variables, many different methods exist. Air pollution experts are often interested in the details of the measurements, down to the specification of instrument manufacturer and model number. While such information is sometimes available from the data providers, there is no harmonisation of such metadata and we don't have the resources to harmonize hundreds or thousands of individual instrument specifications. However, through use of the additional_metadata fields, it is possible to preserve any such information which is given to us. See the Annex: Header Template for an example how such information can be provided.

As there (at least so far) is less variation in the names of numerical models from which we extract data, the field data_origin will contain the name of the numerical model for such data. Currently, the allowed values for data_origin are thus 'Instrument' (for all kinds of measurements), 'COSMOREA6', and 'ERA5'. Additional information, such as a model version number, may again be placed in the additional_metadata field of the time series metadata.

Other aspects of data origin, i.e. references to the data provider, are described in the section on role codes (Section 4.3.1).

4.4 Other Aspects of Time Series Metadata

4.4.1 Sampling Frequency and Aggregation

The primary sampling frequency of data in the TOAR database is hourly. However, the database allows to store data with other sampling frequencies to enable the inclusion of historic data, for example. The allowed values of the metadata field sampling_frequency in the time series description are:

Table 4.9: allowed values of the metadata field sampling frequency in the timeseries description

Number	*Description*	*Description 2*
0	Hourly	hourly
1	ThreeHourly	3-hourly
2	SixHourly	6-hourly
3	Daily	daily
4	Weekly	weekly
5	Monthly	monthly
6	Yearly	yearly
7	Irregular	irregular data samples of constant length
8	Irregular2	irregular data samples of varying length

As part of the data harmonisation performed by the TOAR data centre staff, data values may be processed to yield one of the data frequencies listed in Table 4.9 above. For example, the German UBA reports their data as 30-minute averages and there are other data providers who submit data at 15-minute intervals. When aggregation is performed as part of the data ingestion process, this is noted in the metadata field

aggregation of the time series metadata. The default value for aggregation is None, i.e. (hourly) data have been inserted as they were provided. The pre-defined aggregation values are:

Table 4.10: Pre-defined data aggregation values

Number	*Description*	*Description 2*
0	Mean	mean
1	MeanOf2	mean of two values
2	MeanOfWeek	weekly mean
3	MeanOf4Samples	mean out of 4 samples
4	MeanOfMonth	monthly mean
5	None	none
6	unknown	unknown

Note that most data values are in fact aggregates of values which were originally sampled with higher frequency. For example, ozone measurements are typically performed once per minute and the data are averaged over the reporting interval chosen by the data provider. The aggregation field of the TOAR database only describes any aggregation performed by the TOAR database team and provides no information about any data processing done by the provider.

4.4.2 Handling of Time / Time Zones

All timestamps in the database are stored in UTC. During the data ingestion process the timezone at source is converted to UTC. The support for extraction in local timezones is planned for the future.

CHAPTER

FIVE

DATA QUALITY

All data and metadata in the TOAR database have been subject to some quality checks. Nevertheless, nobody is perfect and therefore it is not unlikely that you may identify errors, inconsistencies or „weird looking“ data if you only dig deep enough. Most of the data that are kept in the TOAR database originate from quality-controlled repositories, which are maintained by professional data managers. Other data come from resources with fewer resources or potentially less knowledge about the many complex facets of providing FAIR[14] data. Finally, there are data sources, which provide „preliminary“ data in near real-time and such data can obviously not be checked by trained human experts before they are posted.

The TOAR database has been designed with the primary objective to support the Tropospheric Ozone Assessment Report, and therefore our focus lies on providing the data which are most useful for scientific analyses of global air quality and reflect our best knowledge about global air pollutant concentrations. Due to the data curation procedures described below, the data you obtain from the TOAR database may not always be completely identical to data from the same measurements which you might get from the original data providers. Therefore, TOAR data are not suitable for legal purposes, such as the initiation of law suits because of non-attainment of air quality standards.

The TOAR data centre developed a largely automated workflow to process and add new data into the TOAR database (see The TOAR Data Processing Workflow). One step in this workflow is the execution of automated scripts for checking the metadata which describes a measurement site and each individual time series. There is also an automated quality control tool, which performs some basic statistical tests on new data to ensure that at least gross errors are captured and that no „garbage“ enters the database. We are continuously working to improve this quality control tool and plan to add more sophisticated tests in the future. As part of our responsibilities in the TOAR assessment, we will double-check as much data as we can and perform several manual checks through database queries and visualisations at the time when the phase II assessment will be prepared. As TOAR database user you can help us by keeping an eye on the data you download and by informing us about any data or metadata issues you encounter when using the data from the TOAR database. We will try our best to follow your leads and inform the original data providers about any issues that can be confirmed.

During the first phase of TOAR, a semi-quantitative analysis was performed to determine the fraction of erroneous and questionable data among all ground-level ozone time series which are stored in the TOAR database (see[13]). In general it was found that over 95 % of all data points can be regarded as „trustworthy“ in the sense that they exhibit „typical“ behaviour of ozone time series and show no obvious anomalies. Through the creation of animated maps and trend plots of the TOAR data it could be confirmed that the vast majority of data „fits together“ nicely, which means that errors in the aggregated ozone statistics are likely smaller than 5 parts per billion and trend estimates should be „reasonably accurate“[15] . As the TOAR database allows downloads of hourly values including the data quality flags, you can always re-assess the

[14] Findable, Accessible, Interoperable and Re-usable. For details see https://www.force11.org/group/fairgroup/fairprinciples and the TOAR data FAIRness assessment in Section 6 below.

[13] TOAR V1 is described in Schultz, M. G. et al. (2017) Tropospheric Ozone Assessment Report: Database and Metrics Data of Global Surface Ozone Observations, Elem Sci Anth, 5, p.58. DOI: http://doi.org/10.1525/elementa.244

[15] In the second phase of TOAR, a dedicated statistics working group will explore more quantitative ways of assessing the accuracy and robustness of ozone trends.

quality of the data you obtain from us. You can also re-run our automated quality control tool, which is available from https://gitlab.jsc.fz-juelich.de/esde/toar-public/toarqc.

5.1 Data and Metadata Curation

Data quality is a complex topic and there are many different views about what constitutes „good quality data“. With respect to the metadata describing stations and time series we aim to achieve the best possible consistency through the use of controlled vocabulary (see https://esde.pages.jsc.fz-juelich.de/toar-data/toardb_fastapi/docs/toardb_fastapi.html#controlled-vocabulary) on the one hand, and by performing some algorithmic tests on the other hand. For example, we will compare reported station altitudes with the altitude returned from a fine resolution digital elevation model at the given latitude and longitude coordinates. A warning will be raised if the results differ too much. The development of such algorithmic tests is ongoing and will be documented at a later stage.

The quality of the actual data values can never be assessed with full certainty, but experience and statistical methods can at least provide good clues. In the current version of our automated quality control tests, we check the data ranges and test for outliers as well as unrealistically long periods of constant values and significant step changes. Thresholds for these tests have been developed based on sample data which have been determined to be of high quality due to

(i) trust in the data providers, and

(ii) visual inspection of the time series and various descriptive statistics.

The automated quality control tool will not delete any data, but instead change the data quality flag (see Section 5.2). Any such changes applied to the data will be recorded and are made accessible through the time series' „change log“.

There is some debate in the scientific community of environmental observers and database managers about the roles they have in the data curation procedures and about the respective rights and duties. As a general guiding principle it is often stated that only the first-hand data providers are allowed to make changes to their data and metadata, because they are the only ones who have the full insight into the measurement conditions. On the other hand, many modern data collection efforts place more responsibility on the data curators in the data centres, because only there it is possible to assess different data sets with common standards and to apply additional tests, which involve comparisons with neighbouring sites or with numerical model data. Best practice suggests that the results from such tests are communicated back to the data providers and they are then charged with the task to correct the data and re-send to the data centre. In practice, we have found that it is often more efficient to suggest specific corrections to the data providers and ask for their approval, because this means less work for them. In rare cases, the TOAR data centre may also modify data values without the approval of providers; for example, if the data come from a large monitoring network and there are no direct communication channels with the providers, or if we are convinced that data are erroneous, but the data provider will not react to our inquires. Such changes will only be applied if the correction is obvious. A typical example are unit conversions, which may be necessary if the metadata in the submitted file header is inconsistent with the data values. In any case will we document all of these changes and make this information available to you.

5.2 Data Quality Flags

As described above, the quality of TOAR data is documented via so-called data quality flags. There are numerous flaging schemes in use around the world with varying level of detail. Some of the datasets which we receive for inclusion in the TOAR database provide quality information with their data, others don't.

We define four possible status code ranges to indicate whether a given data value is appropriate for use or not. In addition, code values greater 100 can be used for aggregated queries (Table 5.1).

Table 5.1: status code range for data quality

Status code range	Data quality
0 - 9	OK
11 - 19	questionable
20 - 29	erroneous
90 - 99	missing or unknown status
100 - 140	combination of specific data quality flags

Normally, you will be interested in "OK" data only, which means that you can filter data with quality flag < 10. However, in this case it is easier to request 'AllOK' data (flag value 100, see Table 5.2).

As mentioned above, all data are subjected to some automated tests before inclusion in the TOAR database. These tests can only lower the level of confidence in the data, but never change data that were labelled as questionable or erroneous by the data provider into OK values.

The second aspect that might be relevant for assessing the data quality is whether these data have been validated by the provider or not. While in the first phase of TOAR the database only accepted validated data, the expansion to previously uncovered world regions with help of OpenAQ necessitated the inclusion of realtime data, which are never thoroughly validated, although they might have passed some automated quality control checks.

To facilitate the selection of data with a specific quality status, we defined two sets of quality flags. The first set consists of aggregate flags, which allow you to easily select data according to their status as OK, questionable, or erroneous, and to distinguish between validated and preliminary data if you wish to do so (Table 5.1). The second set of flags preserves the information of the original quality assessment by the provider as well as any possible modification introduced through our automated quality control procedures (Table 5.2). These more detailed flag values are the values that are actually stored in the database. You can use both flag sets in the REST interface.

Table 5.2: aggregated data quality flags of the TOAR database[Page 41, 16]

Flag value	Flag name	Description	Combination of original flag values (Table 5.3)
100	AllOK	Data values were deemed OK by the provider and the TOAR quality control tool did not find any obvious errors. Note that validated data with no explicit quality information is treated as "provider OK", whereas preliminary data with no explicit quality information is treated as "not checked by provider". This status also covers data values which had been erroneous at first but were corrected by the provider or based on feedback by the provider.	OKValidatedVerified, OKValidatedQCPassed, OKValidatedModified, OKPreliminaryVerified, OKPreliminaryQCPassed, OKPreliminaryModified, OKEstimated

continues on next page

Table 5.2 – continued from previous page

Flag value	Flag name	Description	Combination of original flag values (Table 5.3)
101	ValidatedOK	Data were sent by provider as validated data, data values were deemed OK by the provider and the TOAR quality control tool did not find any obvious errors.	OKValidatedVerified, OKValidatedQCPassed, OKValidatedModified
102	PreliminaryOK	Data were sent by provider as preliminary (or realtime) data, data values were deemed OK by the provider (usually no explicit quality information is given with realtime data) and the TOAR quality control tool did not find any obvious errors.	OKPreliminaryVerified, OKPreliminaryQCPassed, OKPreliminaryModified
103	NotModifiedOK	Similar to AllOK, but modified data values are not included.	OKValidatedVerified, OKValidatedQCPassed, OKPreliminaryVerified, OKPreliminaryQCPassed
104	ModifiedOK	Data values had been erroneous at first but were corrected by the provider or based on feedback by the provider.	OKValidatedModified, OKPreliminaryModified, OKEstimated
110	AllQuestionable	Data were labelled as questionable by provider or marked as questionable by the automated TOAR quality control test.	QuestionableValidatedConfirmed, QuestionableValidatedUnconfirmed, QuestionableValidatedFlagged, QuestionablePreliminaryConfirmed, QuestionablePreliminaryUnconfirmed, QuestionablePreliminaryFlagged, QuestionablePreliminaryNotChecked
111	ValidatedQuestionable	Validated data that were labelled as questionable by provider or marked as questionable by the automated TOAR quality control test.	QuestionableValidatedConfirmed, QuestionableValidatedUnconfirmed, QuestionableValidatedFlagged
112	PreliminaryQuestionable	Preliminary (realtime) data that were labelled as questionable by provider or marked as questionable by the automated TOAR quality control test.	QuestionablePreliminaryConfirmed, QuestionablePreliminaryUnconfirmed, QuestionablePreliminaryFlagged, QuestionablePreliminaryNotChecked

continues on next page

Table 5.2 – continued from previous page

Flag value	Flag name	Description	Combination of original flag values (Table 5.3)
120	AllErroneous	Data were labelled as erroneous by provider or marked as erroneous by the automated TOAR quality control test.	ErroneousValidatedConfirmed, ErroneousValidatedUnconfirmed, ErroneousValidatedFlagged1, ErroneousValidatedFlagged2, ErroneousPreliminaryConfirmed, ErroneousPreliminaryUnconfirmed, ErroneousPreliminaryFlagged1, ErroneousPreliminaryFlagged2, ErroneousPreliminaryNotChecked
121	ValidatedErroneous	Validated data that were labelled as erroneous by provider or marked as erroneous by the automated TOAR quality control test.	ErroneousValidatedConfirmed, ErroneousValidatedUnconfirmed, ErroneousValidatedFlagged1, ErroneousValidatedFlagged2
122	PreliminaryErroneous	Preliminary (realtime) data that were labelled as erroneous by provider or marked as erroneous by the automated TOAR quality control test.	ErroneousPreliminaryConfirmed, ErroneousPreliminaryUnconfirmed, ErroneousPreliminaryFlagged1, ErroneousPreliminaryFlagged2, ErroneousPreliminaryNotChecked

continues on next page

Table 5.2 – continued from previous page

Flag value	Flag name	Description	Combination of original flag values (Table 5.3)
130	AllQuestionableOrErroneous	Data were labelled as questionable or erroneous by provider or marked as questionable or erroneous by the automated TOAR quality control test.	QuestionableValidatedConfirmed, QuestionableValidatedUnconfirmed, QuestionableValidatedFlagged, QuestionablePreliminaryConfirmed, QuestionablePreliminaryUnconfirmed, QuestionablePreliminaryFlagged, QuestionablePreliminaryNotChecked, ErroneousValidatedConfirmed, ErroneousValidatedUnconfirmed, ErroneousValidatedFlagged1, ErroneousValidatedFlagged2, ErroneousPreliminaryConfirmed, ErroneousPreliminaryUnconfirmed, ErroneousPreliminaryFlagged1, ErroneousPreliminaryFlagged2, ErroneousPreliminaryNotChecked
131	ValidatedQuestionableOrErroneous	Validated data that were labelled as questionable or erroneous by provider or marked as questionable or erroneous by the automated TOAR quality control test.	QuestionableValidatedConfirmed, QuestionableValidatedUnconfirmed, QuestionableValidatedFlagged, ErroneousValidatedConfirmed, ErroneousValidatedUnconfirmed, ErroneousValidatedFlagged1, ErroneousValidatedFlagged2
132	PreliminaryQuestionableOrErroneous	Preliminary (realtime) data that were labelled as questionable or erroneous by provider or marked as questionable or erroneous by the automated TOAR quality control test.	QuestionablePreliminaryConfirmed, QuestionablePreliminaryNotChecked, ErroneousPreliminaryConfirmed, ErroneousPreliminaryUnconfirmed, ErroneousPreliminaryFlagged1, ErroneousPreliminaryFlagged2, ErroneousPreliminaryNotChecked

continues on next page

Table 5.2 – continued from previous page

Flag value	Flag name	Description	Combination of original flag values (Table 5.3)
140	NotChecked	Preliminary (realtime) data on which no automated quality control procedure has been run due to, for example, an incomplete time series. Note that a simple range check with bounds defined per variable is normally run anyhow, but this simple test cannot lead to the result “QC passed”.	OKPreliminaryNotChecked, QuestionablePreliminaryNotChecked, ErroneousPreliminaryNotChecked

Table 5.3: the specific flag values defined in the TOAR database

Flag value	Flag name	Description
0	OKValidatedVerified	Data was received from provider as final validated data and passed the automatic quality control tests of the TOAR data centre. In addition, the data was subjected to manual inspection of the data summary plots.
1	OKValidatedQCPassed	Data was received from provider as final validated data and passed the automatic quality control tests of the TOAR data centre.
2	OKValidatedModified	Data was received from provider as final validated data and did not pass the automatic quality control tests of the TOAR data centre in the first pass. The data value was changed according to feedback from the data provider or if an obvious correction was possible.
3	OKPreliminaryVerified	Data was received from provider as preliminary or near realtime data and passed the automatic quality control tests of the TOAR data centre. In addition, the data was subjected to manual inspection of the data summary plots.
4	OKPreliminaryQCPassed	Data was received from provider as preliminary or near realtime data and passed the automatic quality control tests of the TOAR data centre.
5	OKPreliminaryModified	Data was received from provider as preliminary or near realtime data and did not pass the automatic quality control tests of the TOAR data centre in the first pass. The data value was changed according to feedback from the data provider or if an obvious correction was possible.
6	OKEstimated	Data value derived from an interpolation or modelling tool to fill a data gap. Note: you will never find this flag value in any “original” time series, but the name of the time series will indicate clearly if it contains estimated values. Some statistics may be more reliable if they are based on complete time series and thus avoid sampling biases.
7	OKPreliminaryNotChecked	Data was received from provider as preliminary or near realtime data and no QC test was run, for example because of an incomplete time series.

continues on next page

[16] These flags allow for convenient selection of data with the most relevant quality criteria, i.e. OK, questionable, or erroneous on the one hand and validated or preliminary on the other hand. The flags are composites of more specific flag values which are listed in Table 5.5.

Table 5.3 – continued from previous page

Flag value	Flag name	Description
10	QuestionableValidatedConfirmed	Data was received from provider as final validated data with a quality flag indicating potential problems with the data value. The data value was also flagged as suspicious by the automatic quality control tests of the TOAR data centre.
10	QuestionableValidatedConfirmed	Data was received from provider as final validated data with a quality flag indicating potential problems with the data value. The data value was also flagged as suspicious by the automatic quality control tests of the TOAR data centre.
12	QuestionableValidated-Flagged	Data was received from provider as final validated data with no indication of potential problems. However, the data value was flagged as suspicious by the automatic quality control tests of the TOAR data centre.
13	QuestionablePreliminaryConfirmed	Data was received from provider as preliminary or near real-time data with a quality flag indicating potential problems with the data value. The data value was also flagged as suspicious or erroneous by the automatic quality control tests of the TOAR data centre.
14	QuestionablePreliminaryUnconfirmed	Data was received from provider as preliminary or near real-time data with a quality flag indicating potential problems with the data value. However, the data value was not flagged as suspicious or erroneous by the automatic quality control tests of the TOAR data centre.
15	QuestionablePreliminaryFlagged	Data was received from provider as preliminary or near real-time data with no indication of potential problems. However, the data value was flagged as suspicious by the automatic quality control tests of the TOAR data centre.
16	QuestionablePreliminaryNotChecked	Data was received from provider as preliminary or near real-time data with a quality flag indicating potential problems with the data value. No QC test was run, for example because of an incomplete time series.
20	ErroneousValidatedConfirmed	Data was received from provider as final validated data with a quality flag indicating an erroneous data value. The data value was also flagged as suspicious or erroneous by the automatic quality control tests of the TOAR data centre.
21	ErroneousValidatedUnconfirmed	Data was received from provider as final validated data with a quality flag indicating an erroneous data value. However, the data value was not flagged as suspicious or erroneous by the automatic quality control tests of the TOAR data centre.
22	ErroneousValidated-Flagged1	Data was received from provider as final validated data with no indication of potential problems. However, the data value was flagged as erroneous by the automatic quality control tests of the TOAR data centre.
23	ErroneousValidated-Flagged2	Data was received from provider as final validated data flagged as questionable values. However, the data value was flagged as erroneous by the automatic quality control tests of the TOAR data centre.

continues on next page

Table 5.3 – continued from previous page

Flag value	Flag name	Description
24	ErroneousPreliminaryConfirmed	Data was received from provider as preliminary or near realtime data with a quality flag indicating an erroneous data value. The data value was also flagged as suspicious or erroneous by the automatic quality control tests of the TOAR data centre.
25	ErroneousPreliminaryUnconfirmed	Data was received from provider as preliminary or near realtime data with a quality flag indicating an erroneous data value. However, the data value was not flagged as suspicious or erroneous by the automatic quality control tests of the TOAR data centre.
26	ErroneousPreliminaryFlagged1	Preliminary or near realtime data was received from provider with no indication of potential problems. However, the data value was flagged as erroneous by the automatic quality control tests of the TOAR data centre.
26	ErroneousPreliminaryFlagged1	Preliminary or near realtime data was received from provider with no indication of potential problems. However, the data value was flagged as erroneous by the automatic quality control tests of the TOAR data centre.
28	ErroneousPreliminaryNotChecked	Data was received from provider as preliminary or near realtime data with a quality flag indicating an erroneous data value. No QC test was run, for example because of an incomplete time series.
90	MissingValue	The data provider reported a missing value at this time stamp. Generally, the TOAR database will not explicitly store missing values but instead simply leave out the data value at that timestamp. However, there are situations when missing values are coded in the time series, for example if a new version of a dataset replaces formerly valid values by missing values.
91	UnknownQualityStatus	Also known as „not checked“. Technical flag to allow setting a quality status to unknown. The data provider did not report the data quality status and no QC test was run, for example because of an incomplete time series. This flag value can only be seen for realtime data, because all validated data are assumed to be OK by default.

The following two tables summarise how flag values may be modified as a result of the automated quality control tests which are run during data ingestion or as part of a data inspection.

Table 5.4: possible flaging states of **validated** data depending on the data quality status offered by the data provider and the result of our automated QC tests

	toarqc		
provider	OK	Questionable	Erroneous
OK	OKValidatedQCPassed	QuestionableValidatedFlagged	ErroneousValidatedFlagged1
Questionable	QuestionableValidatedUnconfirmed	QuestionableValidatedConfirmed	ErroneousValidatedFlagged2
Erroneous	ErroneousValidatedUnconfirmed	ErroneousValidatedConfirmed	ErroneousValidatedConfirmed

Table 5.5: Possible flaging states of **preliminary** data depending on the data quality status offered by the data provider and the result of our automated QC tests

	toarqc			
provider	OK	Questionable	Erroneous	NotChecked
OK	OKPrelimi-naryQCPassed	QuestionablePre-liminaryFlagged	ErroneousPrelimi-naryFlagged1	OKPrelimi-naryNotChecked
Questionable	QuestionablePre-liminaryUncon-firmed	QuestionablePre-liminaryConfirmed	ErroneousPrelimi-naryFlagged2	Question-ablePrelimi-naryNotChecked
Erroneous	ErroneousPrelimi-naryUnconfirmed	ErroneousPrelimi-naryConfirmed	ErroneousPrelimi-naryConfirmed	ErroneousPrelimi-naryNotChecked

In some situations of realtime data processing the only automated test that can be run is a crude range test (for example if many values from different stations at one specific time step are inserted). This situation does not qualify as full QC test. Therefore, values are only flagged as erroneous (26, 27, or 24 depending on the provider flag) or as not checked (7, 16, 28).

CHAPTER

SIX

FAIR DATA

This section provides a self-assessment of the level of FAIRness that has been accomplished by the TOAR data infrastructure and services. The main components of the TOAR data infrastructure are a relational database housing the data together with its metadata, a REST API and a graphical user interface to access the data, and a publication service preparing data sets to be published in the B2SHARE service.

The FAIRness requirements are taken from GO FAIR (https://www.go-fair.org/fair-principles/) and the assessment is influenced by the common set of core assessment criteria for FAIRness developed by the RDA FAIR data maturity model Working group (https://www.rd-alliance.org/groups/fair-data-maturity-model-wg).

6.1 Overview

FAIRness evaluates openness and interoperability of data according to the four main criteria “findable”, “accessible”, “interoperable”, and “re-usable”. The following table lists the GO FAIR requirements and summarizes our self-assessment how far the TOAR data infrastructure is matching these criteria.

Table 6.1: FAIRness Self Assessment

To Be Findable	
F1. (Meta)data are assigned globally unique and persistent identifiers | 100%	
F2. Data are described with rich metadata	100%
F3. Metadata clearly and explicitly include the identifier of the data they describe	100%
F4. (Meta)data are registered or indexed in a searchable resource	75%
To Be Accessible	
A1. (Meta)data are retrievable by their identifier using a standardised communication protocol | 75%	
A1.1 The protocol is open, free and universally implementable	75%
A1.2 The protocol allows for an authentication and authorisation where necessary	75%
A2. Metadata should be accessible even when the data is no longer available	75%
To Be Interoperable	
I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation | 75%	
I2. (Meta)data use vocabularies that follow the FAIR principles | 75%	
I3. (Meta)data include qualified references to other (meta)data	75%
To Be Resuable	
R1. (Meta)data are richly described with a plurality of accurate and relevant attributes	75%
R1.1. (Meta)data are released with a clear and accessible data usage license	75%
R1.2. (Meta)data are associated with detailed provenance	75%
R1.3. (Meta)data meet domain-relevant community standards	75%

6.2 Discussion

In the following we discuss the FAIRness requirements one by one.

F1: (Meta)data are assigned globally unique and persistent identifiers

The database itself is registered with re3data.org and with that has a globally unique DOI provided by DataCite (https://www.datacite.org/, TOAR: http://doi.org/10.17616/R3FZ0G). The metadata describing the database is available with the same DOI.

Data with its metadata from individual data providers, which are published on B2SHARE have globally unique DOIs from DataCite assigned to them. Every instrument time series is published as an individual data record, and all time series belonging to one station are grouped as a collection. The DOI of the collection shall be used as the primary DOI to identify and reference a dataset.

Currently, the data contained in the TOAR database as well as in the published data at B2SAHRE are time series data. Once other datasets (vertical profiles, satellite retrievals, model (gridded) data) are added, a similar concept will be applied.

Data retrieved from other sources, e.g. data replicated from large environmental data archives, are assigned a unique identifier within our database. These data can be unambiguously identified through a combination of human-readable metadata attributes (station_id, variable_id resource_provider, version, data_origin, measurement_method or model_experiment_identifier, sampling height, data_filtering_procedures (processing step 14, Criterion 14.1 - Criterion 14.9).

The original unique identifiers of replicated datasets are preserved as metadata attributes in the TOAR database if they are available and accessible. This allows for back-referencing to the original data source.

F2: Data are described with rich metadata

The metadata describing the TOAR database in the re3data.org registry follows the re3data requirements while the metadata of data publications in B2SHARE complies with the requirements of B2SHARE and DataCite.

The data in the TOAR database has a rich metadata profile covering most aspects of provider information, location description, instrument description, data quality and version information. A highlight of the TOAR database is the ability to preserve additional metadata information from providers, which cannot be mapped to the harmonised TOAR metadata profile. For details see TOAR metadata documentation: Section 4 above and http://esde.pages.jsc.fz-juelich.de/toar-data/toardb_fastapi/docs/toardb_fastapi.html#models.

F3: Metadata clearly and explicitly include the identifier of the data they describe

The metadata provided for the TOAR database at re3data.org contains the link to the user interfaces of the database. The metadata available for data publications of the TOAR community in B2SHARE contain the links to the data sets contained in the data collection in the form of DOI of the collection/PID of the data set.

The TOAR database's data and metadata are never separated, ensuring a clear mapping of the metadata to the data they describe.

F4. (Meta)data are registered or indexed in a searchable resource

Through the registration in re3data.org the TOAR database is indexed and thereby searchable. TOAR data publications on B2SHARE are indexed in b2find.eudat.eu and with that searchable.

A1: (Meta)data are retrievable by their identifier using a standardised communication protocol

We use https (with REST) for (meta)data retrieval, which is a standardized communication protocol. The REST-API allows for data being accessed automatically.

A1.1 The protocol is open, free and universally implementable

https (with REST) is open, free and universally implementable.

A1.2 The protocol allows for an authentication and authorisation where necessary

https allows for an authentication and authorisation where necessary.

A2. Metadata should be accessible even when the data is no longer available

Metadata of the TOAR database in re3data.org as well as those of data publications in B2SHARE / B2FIND will be kept persistently according to the respective policies of the service organisations. In the TOAR database itself, data and metadata are contained in the same physical space. Efforts are taken to keep the (meta)data persistently.

I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

B2SHARE data publications use an extension of the Dublin Core Schema for the metadata, while DataCite developed a custom metadata scheme[17].

The TOAR metadata uses

1. commonly used controlled vocabularies (e.g. adapted from IPCC[18], MODIS CMG[19], HTAP[20], ...), represented in an ontology and

[17] http://schema.datacite.org/meta/kernel-4/doc/DataCite-MetadataKernel_v4.4.pdf
[18] Intergovernmental Panel on Climate Change
[19] Moderate Resolution Imaging Spectroradiometer (MODIS) Land Cover Climate Modeling Grid (CMG) (MCD12C1) Version 6 data product (https://lpdaac.usgs.gov/products/mcd12q1v006/)
[20] Task Force on Hemispheric Transport of Air Pollution (TF HTAP)

2. a good data model (a well-defined framework to describe and structure metadata).

The TOAR ontology uses OWL and SKOS and can also be provided as RDF or JSON-LD. The TOAR REST API provides data and metadata within a JSON structure, that is broadly usable in python scripts.

I2: (Meta)data use vocabularies that follow the FAIR principles

The TOAR metadata scheme has been built from existing standards (e.g. ISO 19115 "geographic information- metadata") and is accessible at http://esde.pages.jsc.fz-juelich.de/toar-data/toardb_fastapi/docs/toardb_fastapi.html. The ontology can be browsed at https://toar-data.fz-juelich.de/api/v2/onloglogy

Currently, the controlled vocabulary used in the metadata fields has been defined and is covered by the ontology, e.g. the terms for the type of area a station is located in which are urban, suburban, rural and unknown. They are not published and accessible through a globally unique identifier but accessible from the webpage given above. The identifiers of the metadata have been defined with the TOAR metadata scheme at http://esde.pages.jsc.fz-juelich.de/toar-data/toardb_fastapi/docs/toardb_fastapi.html.

I3: (Meta)data include qualified references to other (meta)data

Within the TOAR data publications on B2SHARE, metadata on individual time series are linked to the respective collections and vice versa, given their unique DOI.

Currently it is planned to link the TOAR metadata for contact persons with their ORCID and organisations with their web link. The development is ongoing. The ontology already links term definitions to their source and where data are replicated from other repositories, the metadata includes a reference to the original data repository, pointing specifically to the original metadata. Further links can be stored in the auxiliary metadata.

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes

Besides the general metadata provided with re3data.org for the TOAR database the database has a rich metadata profile covering most aspects of provider information, location description, instrument description, data quality and versioning information. A highlight of the TOAR database is the ability to preserve additional metadata information from providers, which cannot be mapped to the harmonized TOAR metadata profile. The metadata profile is available at http://esde.pages.jsc.fz-juelich.de/toar-data/toardb_fastapi/docs/toardb_fastapi.html.

R1.1. (Meta)data are released with a clear and accessible data usage license

TOAR data publications on B2SHARE always come with a CC-BY (4.0) license. Clear display and easy access to this license is a feature of B2SHARE.

Replicated data (or other datasets which are not published on B2SHARE) from TOAR data providers are also available under the CC-BY license.

R1.2. (Meta)data are associated with detailed provenance

The TOAR data ingestion and data publication workflow is clearly documented (refer to TOAR Data Processing). The source of the data is part of the metadata as detailed in Section 4.3 above.

All processing steps from receipt of the original data to the data publication in the TOAR database and/or as B2SHARE record are documented and could be made available on request. Changes to the data in the TOAR database are automatically logged in the changelog which is part of the metadata.

R1.3: (Meta)data meet domain-relevant community standards

As discussed above (I1 and I2), we use ontologies and controlled vocabulary based on ISO-19115 and the WIGOS standard wherever possible. A standard which covers all necessary aspects of the TOAR-II activity does not exist yet. The TOAR data team follows the developments / refinements of community metadata standards as undertaken for example by the German national research data infrastructure (NFDI) initiative or the the European ENVRI-FAIR project.

The data is provided in csv, html, and json format; a NetCDF output format will also soon be available.