TOAR Data Technical Guide #3

Mapping of TOAR Data Centre components to OAIS Model

toar-data.fz-juelich.de

Version 1.0 | February 1, 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 Infrastructure is a hub for tropospheric ozone data rather than a classical data archive for such data. Times series data from established networks of stations as well as data from individual measuring stations is collected, curated, augmented with quality indicators, and stored permanently in the TOAR database. In case networks are later changing their data due to their curation process, it will again be collected, run through the ingestion process and changed in the TOAR database. Version numbers attached to the data and metadata as well as earlier published database dumps still reflect the previous status.

The individual measuring stations together with the managed and unmanaged networks of stations form the group of data producers. The TOAR management is done by the IGAC project activity TOAR (https://igacproject.org/index.php/activities/TOAR), specifically by the Steering Committee of TOAR Phase-II, and the designated community is the TOAR community, more general all researchers who analyse tropospheric ozone data with respect to the global-scale impact of ozone on climate, human health and crop/ecosystem productivity.

In the following the key components of the TOAR Data Centre are defined:

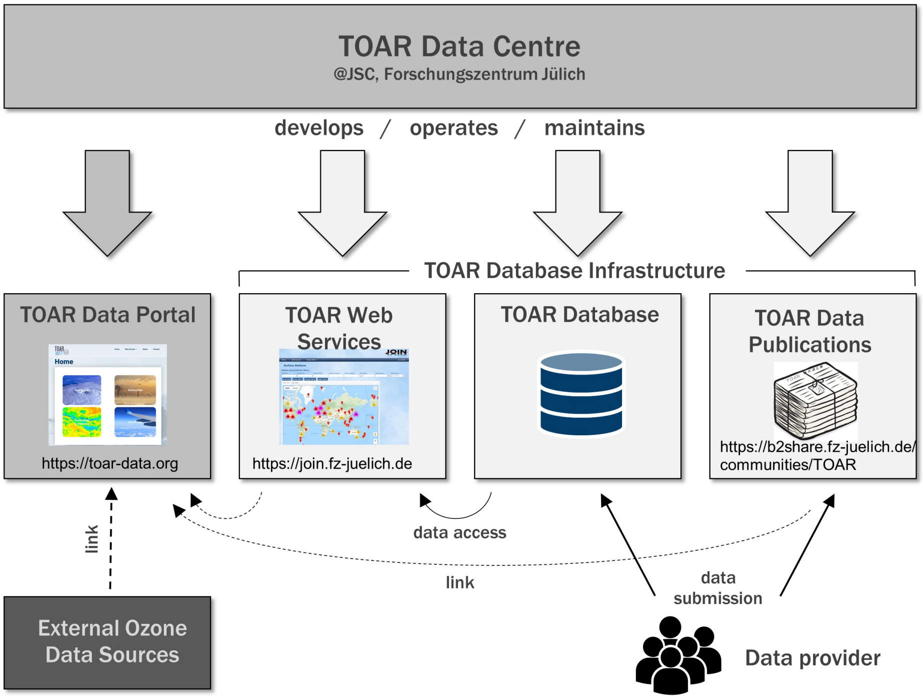
TOAR Data Centre
@JSC, Forschungszentrum Jülich
develops / operates / maintains
TOAR Database Infrastructure
TOAR Data Portal
https://toar-data.org
TOAR Web Services
https://join.fz-juelich.de
TOAR Database
TOAR Data Publications
https://b2share.fz-juelich.de/communities/TOAR
link
data access
link
data submission
External Ozone Data Sources
Data provider

Fig. 1.1: Overview of the TOAR Data Centre Components

The **TOAR Data Centre** (TOAR DC) comprises services and data in support of the Tropospheric Ozone

Assessment Report activity. Its main goal is to provide access to tropospheric ozone data for research on the global-scale impact of ozone on climate, human health and crop/ecosystem productivity. A subset of its components build the TOAR Database Infrastructure.

The **TOAR Database** gets data from individual stations, which are curated before stored, and harvests data from measurement station networks. It provides an API for (data ingest and) data access. Publicly available data from relevant networks such as UBA (German Environment Federal Office) or OpenAQ (https://openaq.org) are downloaded, harmonised, and quality checked before they are stored in the TOAR database. These data might change over time at their origin because of the providers' own curation processes, which means that they will subsequently be changed in the TOAR database. The data from individual providers are formatted according to the TOAR submission guidelines and contain the requested metadata. In the ingestion process the data are curated and enriched by other metadata. The curated data and metadata are sent to the providers for approval before they are stored and published in the TOAR database.

The **TOAR Data Portal** provides access to various external ozone data sources and to the TOAR database by linking to the respective web services.

The **TOAR Web Services** provide the REST interface for accessing the TOAR database.

The **TOAR Data Publication Service** enables the TOAR data curators to publish data from individual providers to an external service B2SHARE at Forschungszentrum Jülich. In this process the necessary metadata for the publication is generated automatically through a query to the TOAR database and mapped to the B2SHARE metadata fields. (B2SHARE is the EUDAT user-friendly, reliable and trustworthy service for researchers, scientific communities and citizen scientists to store and publish research data from diverse contexts).

All TOAR Database Infrastructure services are run on hardware and software which is installed and maintained at JSC (Jülich Supercomputing Centre, www.fz-juelich.de/ias/jsc), a section of the Institute for Advanced Simulation (IAS) at Forschungszentrum Jülich. They make use of general computer centre tools and services such as Tivoli Storage Manager (TSM) for backup, archiving and hierarchical storage management, which are implemented to serve all systems in the supercomputer centre and on Forschungszentrum Jülich campus.

CHAPTER

TWO

MAPPING TOAR DATA CENTRE TO THE OAIS MODEL

In this section we will map from a high-level perspective relevant components of our archival system to the corresponding features of the OAIS model. The archival information system in the TOAR Data Centre context is the TOAR database with its API, its web service and its publication service.

2.1 The OAIS Functional Model

OAIS [1][2] is a reference model for ‘Open Archival Information Systems’. It has been accepted as a de facto standard for organisations with digital archiving requirements.

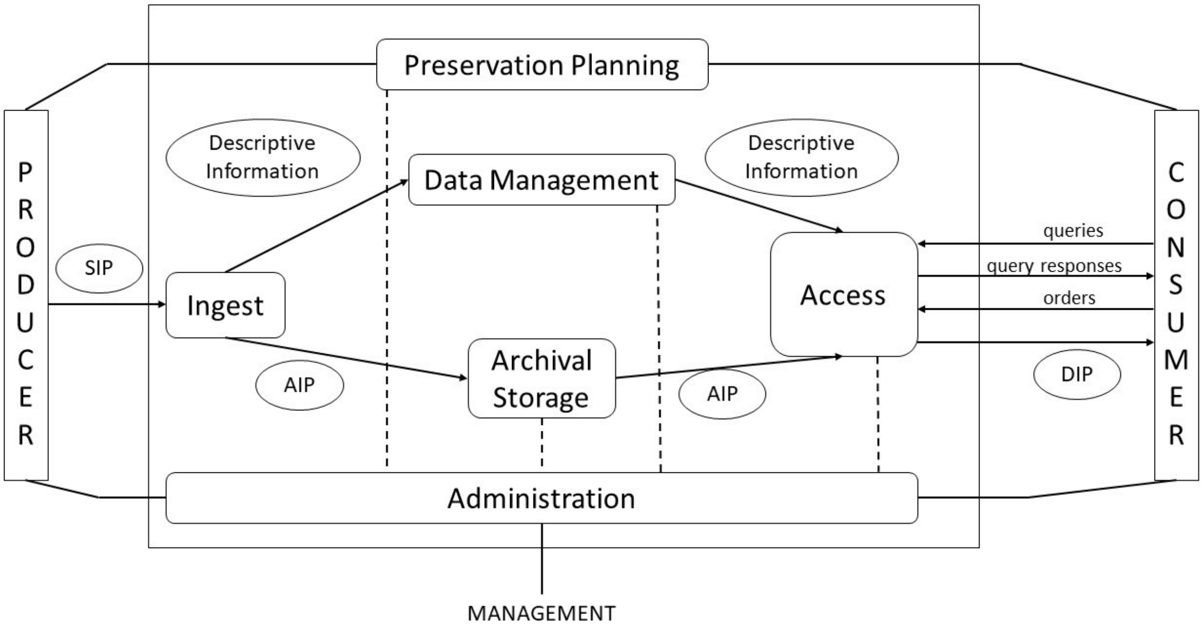

Fig. 2.1: OAIS Functional Entities

2.2 Functional Components

2.2.1 Ingest

In the ingestion process the data is curated and enriched by other metadata. Data is either sent to us by individual measurement stations or harvested from networks of measurement stations. The curated data and metadata of the individual data providers is sent back to the providers for approval before it is stored in the TOAR database. The data downloaded from networks is stored as is with the available metadata and enriched by quality indicators. Details of the ingestion process are documented in the Data Processing technical guide.

2.2.2 Archival Storage

The data is stored into a PostgreSQL (version 13) database. Write-Ahead Logging (WAL) is a standard method for ensuring data integrity. The build-in WAL mechanism of Postgres V13 is used. The TOAR database is backed up incrementally on a daily basis. Once a month, a base backup is created, on top of which a point-in-time recovery can be performed starting from the base backup up to the last backed up daily increment. The last two monthly base backups as well as the respective increments are kept on the system itself. The last deleted base backup as well as the corresponding increments are available through the backup system for another 30 days after deletion. A copy of the database is hold at RWTH Aachen (multicopy redundancy).

2.2.3 Data Management

The data is managed via mechanisms of the standard PostgreSQL tools (eg. vacuum). Reports are automatically created.

Additionally, workflows are in place for updating data and metadata in the database itself. Versioning of data and metadata is included in the processes to make the entries unique.

A REST API is available for accessing the data in the database. The requests are checked periodically to ensure satisfactory accessing times, which may result in optimising the database with PostgreSQL means (eg. indexing).

2.2.4 Preservation Planning

We distinguish two elements of data preservation here: (i) data preservation during the lifetime of the TOAR initiative and the operation of the TOAR DC, (ii) data preservation beyond the operation of the TOAR DC at JSC. The procedures for TOAR data preservation and the long-term data archival strategy are described in the TOAR_TG_Vol01_Infrastructure document.

2.2.5 Access

The TOAR Web Service provides access to the data in the TOAR database making use of the TOAR database API. The TOAR database API is accessible to all users.

The data in the TOAR database are freely available for read access. Authentication and authorization are only necessary for data ingestion (write access) and is provided through PostgreSQL roles.

(A higher-level access to TOAR data and other ozone data from external sources is provided through the TOAR data portal which links to the specific portals or access points of the different data sources.)

2.2.6 Administration

The TOAR Data Centre including its archival information system and its services are managed by the Jülich Supercomputing Centre (JSC) at Forschungszentrum Jülich.

All systems, networks, processes and performance are constantly monitored with Nagios. Operating system and software updates are regularly implemented.

2.2.7 Common Services

The TOAR database server is located on a virtual machine (VM) running Ubuntu inside an OpenStack cloud environment at Jülich Supercomputing Centre (JSC). It is guarded by a firewall and cannot be directly reached from the outside but only via the API service. The physical location of the data (tablespace) is mounted via NFS over a 10 GBit/s connection.

2.3 Information Packages

2.3.1 Submission Information Package

Data from individual providers must contain a specific set of metadata, which is documented in TOAR_UG_Vol05_Data_Submission_Guide. Data harvested from measurement station networks are taken as they are.

2.3.2 Archival Information Package

During the ingestion (or an update) process data and metadata are curated and quality controlled. They are stored in the database with version number and the history of changes.

2.3.3 Dissemination Information Package

TOAR's API provides the extraction of data and metadata with flexible search options. Output formats are available in json and html. Derived data products can be generated via the API. The API also includes access to descriptive information from the ontology maintained for TOAR.

In case of data publication, the extracted data is augmented with the metadata required by the external B2SHARE service. The API URL to reproduce the dataset published in B2SHARE is stored in the data file's metadata in the history tag and in the abstract of the B2SHARE record.

2.3.4 DOI Data Publication at TOAR Data Centre

In addition to the tasks as defined by OAIS a data publication workflow has been setup at TOAR DC. Data publication is the act of releasing data for (re)-use by others. The main objective is to elevate data to be first class research outputs. (https://en.wikipedia.org/wiki/Data_publishing)

For data from individual data submissions the 'Ingest' functional entity provides the data publication process at TOAR DC with DOI assignment. For more detail see TOAR_TG_Vol02_Data_Processing.

In this section we define the roles which are not described in OAIS in order to discriminate between the TOAR data publication process and the OAIS functions and roles.

Creator: Responsible for quality assurance of data and metadata with review criteria defined by publisher and community/project. **Publisher**: Responsible for editorial functions, including definition and assembly of metadata elements. Responsible for registration of DOI. **Custodian**: Party that accepts accountability and responsibility for the resource and ensures appropriate care and maintenance of the resource.

This describes our responsibilities as TOAR data centre team. Figure 3 below shows the ISO and the corresponding Datacite/B2SHARE role codes during the data processing workflow.

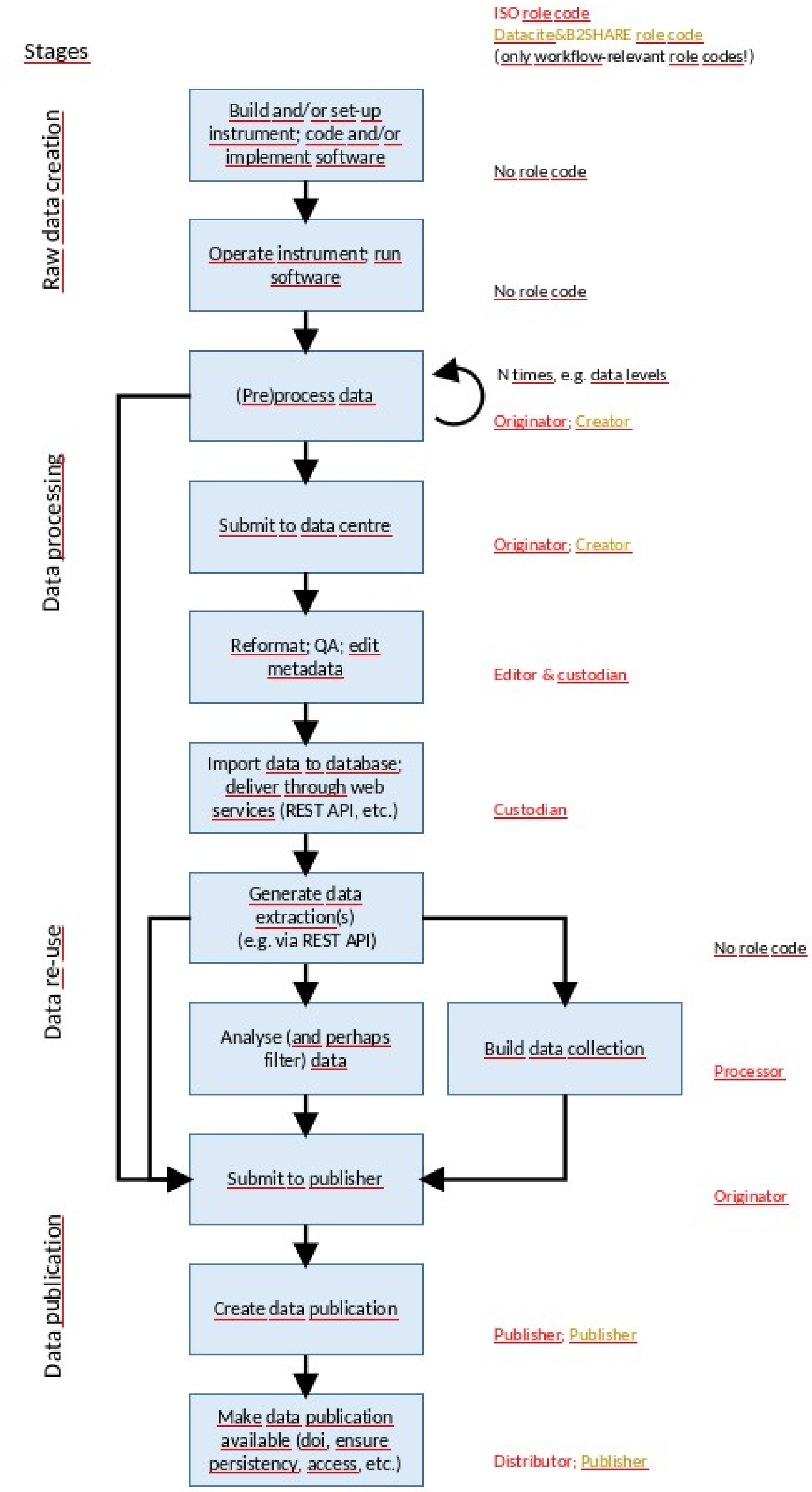
ISO role code
Datacite&B2SHARE role code
(only workflow-relevant role codes!)
Stages
Raw data creation
Build and/or set-up instrument; code and/or implement software
No role code
Operate instrument; run software
No role code
Data processing
(Pre)process data
N times, e.g. data levels
Originator; Creator
Submit to data centre
Originator; Creator
Reformat; QA; edit metadata
Editor & custodian
Import data to database; deliver through web services (REST API, etc.)
Custodian
Data re-use
Generate data extraction(s) (e.g. via REST API)
No role code
Analyse (and perhaps filter) data
Build data collection
Processor
Submit to publisher
Originator
Data publication
Create data publication
Publisher; Publisher
Make data publication available (doi, ensure persistency, access, etc.)
Distributor; Publisher

Fig. 2.2: TOAR data workflow and role codes

CHAPTER

THREE

REFERENCES

[1] Lavoie, B.F., 2004. The Open Archival Information System Reference Model: Introduction Guide. DPCTechnology Watch Series Report 04-01. Available at: http://www.dpconline.org/docs/lavoie_OAIS.pdf, last accessed 25 June 2021

[2] CCSDS Recommended Practice for an Oais Reference Model. Consultative Committee for Space Data Systems (CCSDS), Magenta Book 2012. Available at http://public.ccsds.org/publications/archive/650x0m2.pdf, last accessed 25 June 2021