Enabling Canonical Analysis Workflows Documented Data Harmonization on Global Air Quality Data (Postprint)

Authors: Sabine, Schröder, Eleonora, Epp, Amirpasha, Mozaffari, Mathilde, Romberg, Niklas, Selke, Martin, G. Schultz, Sabine, Schröder

Date: 2022-11-28T00:00:00+00:00

Abstract

Data harmonization and documentation of the data processing are essential prerequisites for enabling Canonical Analysis Workflows. The recently revised Terabyte-scale air quality database system, which the Tropospheric Ozone Assessment Report (TOAR) created, contains one of the world's largest collections of near-surface air quality measurements and considers FAIR data principles as an integral part. A special feature of our data service is the on-demand processing and product generation of several air quality metrics directly from the underlying database. In this paper, we show that the necessary data harmonization for establishing such online analysis services goes much deeper than the obvious issues of common data formats, variable names, and measurement units, and we explore how the generation of FAIR Digital Objects (FDO) in combination with automatically generated documentation may support Canonical Analysis Workflows for air quality and related data.

Full Text

Preamble

RESEARCH PAPER

Enabling Canonical Analysis Workflows Documented Data Harmonization on Global Air Quality Data

Sabine Schröder†, Eleonora Epp, Amirpasha Mozaffari, Mathilde Romberg, Niklas Selke & Martin G. Schultz

Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

Keywords: FAIR; CWFR; FDO; Data Harmonization; TOAR

Citation: Schröder, S., et al.: Enabling canonical analysis workflows documented data harmonization on global air quality data. Data Intelligence 4(2), 259-270 (2022). doi: 10.1162/dint_a_{00130}

Received: September 18, 2021; **Revised:** December 17, 2021; **Accepted:** March 1, 2022

ABSTRACT

Data harmonization and documentation of the data processing are essential prerequisites for enabling Canonical Analysis Workflows. The recently revised Terabyte-scale air quality database system, which the Tropospheric Ozone Assessment Report (TOAR) created, contains one of the world's largest collections of near-surface air quality measurements and considers FAIR data principles as an integral part. A special feature of our data service is the on-demand processing and product generation of several air quality metrics directly from the underlying database. In this paper, we show that the necessary data harmonization for establishing such online analysis services goes much deeper than the obvious issues of common data formats, variable names, and measurement units, and we explore how the generation of FAIR Digital Objects (FDO) in combination with automatically generated documentation may support Canonical Analysis Workflows for air quality and related data.

1. INTRODUCTION

Canonical workflows consist of automated workflows or workflow fragments which allow for reusability of these snippets in different contexts. The development of re-usable workflows and software for scientific data analysis depends on re-usable data, which must be well enough described and standardized to ensure reliable and meaningful analysis results [1]. As workflows rely on a modular structure, concepts such as FAIR Digital Object (FDO) [2] and Research Object (RO) [3] can bring FAIRness to every piece of a workflow' s outputs.

In the field of global air pollution research, observational data are usually gathered by national and regional repositories operating under different legal frameworks and thus with different requirements concerning data formats, metadata standards, and quality control. In addition, research institutions around the world provide data to research repositories, which operate under varying funding levels and provide data at different curation levels. While there have been attempts to standardize air quality data (for example, through an air quality community of practice of the Group on Earth Observations (GEO) or a task team on atmospheric composition vocabulary of the World Meteorological Organisation (WMO)), no community-wide standard has emerged from these activities, presumably because air pollution is widely regarded as a local to regional scale problem and there are few incentives to harmonize data on a global level.

These circumstances describe the status quo in air quality data, when the TOAR activity started in 2013. TOAR is a global research effort to analyze the spatial distribution and temporal trends of ozone in the lower ~10 km of the atmosphere (i.e. the troposphere) and to provide data for assessing the impacts of ozone on human health, vegetation, and climate [4]. A basic element of TOAR phase I was the central database at Forschungszentrum Jülich and the infrastructure around it. The first phase of TOAR was concluded in 2019 with the TOAR Accomplishment Report [5] and a special issue of the Elementa: Science of the Anthropocene journal [6]. The second phase of the activity started in 2020 and is currently under development.

A major asset of TOAR is the unprecedented collection of harmonized observations of ground-level air pollution [7]. The TOAR-I database has assembled worldwide ozone observations from 15 different global, regional, and national monitoring networks, which together with data contributions from over 40 individual research groups comprise around 13,000 multi-year time series at ~7,000 independent measurement locations on all continents. The TOAR-I database is freely accessible via a graphical user interface [8] and a Representational state transfer (REST) API [9]. The REST API provides full access to all information in the database and has a variety of statistical aggregations built-in so that it is possible to construct standardized workflows for the analysis of global tropospheric ozone data. A few examples of such workflows have been made available on a git repository [10].

In TOAR phase II we have revised the database schema to further enhance FAIRness of this global air quality data holding and to align the TOAR data services with principles defined by the RDA [11] and EOSC-hub [12, 13]. The redesign of the TOAR data infrastructure also includes further standardization of the web services for data analysis and automated data quality control tools. These will lay the foundation to add more sophisticated workflows to the repository.

In this paper, we focus on two important, indispensable and inseparable prerequisites for workflow sharing: data harmonization and documentation. Canonical workflows need standardized data objects to be reusable in different contexts. The preparation of FDOs requires much more data harmonization and documentation than the obvious issues of common data formats, variable names, and measurement units might suggest. Besides the necessary semantic translations of quite different metadata descriptions, different ways of dealing with data quality are particularly relevant in this regard. A common understanding of metadata and data quality, a prerequisite for data re-use, can only be achieved through establishing a rigid workflow for the data harmonization accompanied with automatically generated documentation. It is the primary goal of the TOAR-II database to provide trustworthy data to its users while preserving all possibilities for new scientific analyses. In the following, we describe a typical use case of the TOAR database from a user' s point of view (section 2) and the resulting data harmonization and data quality control on our side (Section 3).

Section 4 describes some analysis workflows for TOAR data, developed during TOAR phase I. It also discusses the challenges of FDO generation for TOAR data in phase II. Section 5 gives a general summary and draws conclusions.

2. A USE CASE BASED ON THE TOAR DATABASE

An urgent question of our time is the analysis of the global surface ozone trend, as ozone is an air pollutant dangerous for humans and plants [14]. To provide scientifically sound answers to this question, some prerequisites on the data and the analysis tools must be met.

First, surface ozone measurements are available as time series at globally distributed stations. These time series must be long enough (covering more than 20 years) to allow for a statistically meaningful trend analysis. The data quality must be known so that specific time series can be selected or excluded from analysis as needed. A high temporal resolution (hourly) of the measurements, as well as a good global distribution of the stations are also required to derive general statements from the analysis.

Second, to assess the health and vegetation impacts of ozone, software is needed that calculates specific statistical metrics from the time series [15, 16, 17] and returns the results in a suitable form so that further evaluations can be carried out.

With the operational version 1 of the database of TOAR phase I, in the following referred to as TOAR-I database, and its associated web services a major part of the requirements listed above is already fulfilled. On the one hand, it is possible to quickly find suitable time series in the database, as the database provides rich metadata that can be filtered for. On the other hand, the REST API allows the calculation of a wide range of metrics on any interval of a time series. The result is provided in either json or csv format. Thus, users can perform the exact same calculations on every single time series without having to go through a major effort of collecting their own data and ensuring consistency within this collection as well as with other studies. The TOAR REST API can easily be integrated in user-specific analysis workflows, for example in Jupyter notebooks.

As long as the data in the underlying database are not changed, repeated calls to the same TOAR API endpoint will always generate the same response and return the same dataset. When new data are added to extend the time series in the database, REST API calls without a specified date range will automatically allow for an updated trend analysis. However, there is no mechanism in version 1 of the database that would allow the users to easily identify modified time series, for example after a data provider has sent us new files as a result of measurement re-calibrations or due to other circumstances. Furthermore, any modifications applied to original data as a result of the data harmonization and quality control procedures are not visible to users. Therefore, in version 1 of the TOAR data services, we cannot fully ensure reproducibility of results and the re-use of data may in some cases be limited because of incomplete documentation.

Addressing these issues in the re-design of the TOAR database can be motivated by the goal to enhance FAIRness. However, it is intriguing to investigate the impacts of these factors in the context of canonical workflows. This is the main objective of this paper.

3. DATA HARMONIZATION

As basis for the use case of a globally consistent trend analysis described above, the data from multiple sources needs to be harmonized and sufficiently equipped with metadata. In the new version of the TOAR database all data undergoes a clearly defined harmonization and documentation process before it is added. The overall process is automated to a high extent (Figure 1 [Figure 1: see original paper]).

Figure 1. TOAR-II data ingestion workflow. We established a common approach for data ingestion to ensure that data from different sources is handled in a defined and equal way and that all potential modifications are recorded in a provenance log.

As far as possible, these steps are carried out automatically. Besides allowing for standardized documentation, this automation prevents human processing errors, enables us to process huge datasets with acceptable effort and prevents the loss of information during the processing. In particular, the data harmonization reformats the data for its ingestion in the database. Clear rules are defined how the submitted metadata is mapped into the metadata schema used by the TOAR database (discussed in detail below). By harmonizing the very complex data in a rigid framework it is possible to automate the complex step of inserting data into the database, and thus allow users to understand how the data and metadata were generated. In addition to this data processing documentation, data quality reports and summary statistics are produced from the harmonized data to facilitate the selection of suitable data for the user's analysis. Workflows using this data benefit from the uniform data, metadata, and access schema.

3.1 Metadata

The structure of the database [18] has been set up to enhance data documentation especially with respect to traceability and provenance. Metadata standards such as ISO-19115 [19] (Geographic Metadata Information) build the basis for the TOAR-II database metadata catalogue. By blending controlled vocabulary and ontologies based on existing standards with less constrained full-text fields we are able to preserve any metadata supplied to us without the need to reach a broad community consensus about metadata standardization. Wherever possible suitable standard vocabulary is used, e.g., IPCC climate regions [20], MCD12C1 land cover [21], or ISO-3166 country codes [22].

One aspect of the TOAR database that substantially contributed to its adoption by the research community is the augmentation of provider metadata with globally consistent information derived from multiple Earth Observation data

products (see [7] and [23] for details). This adds additional context to the description of measurement locations and thereby enriches the analysis possibilities.

In the data harmonisation workflow we, for example, collect the station metadata to give a precise description of the station and its location consisting of codes, name, coordinates, country, state, coordinate_{validation}*{status}, type*{of}*{environment}, type*{of}*{area}, category, timezone, and additional*{metadata}. The WIGOS metadata elements category 3 for station/platform [24] build the basis for it. We make sure that all metadata fields contain reliable values, e.g. the coordinates are validated with the help of an external geolocation service and indicated by extra metadata items, and the information is consistent with the given country or type of area. If metadata is given by a data provider, lookup tables are created in close consultation with the provider to map the received metadata to the TOAR-II database schema. The data ingestion workflow automatically writes a processing log that is kept with the original data, while later changes made to data in the database automatically result in a log message that is inextricably linked to the data and metadata in the database.

All additional metadata supplied by a provider is preserved and stored in the additional_{metadata} field in a json structure. In this way, data is available through standardized access methods, but the provider' s own extensions can use different attributes.

Given the long-term vision of the TOAR data infrastructure, the challenge is to maintain content and understandability in case of changes in the vocabulary. Those might occur for example when countries split or join and new countries are established. To accommodate for such changes, we introduced versioning of the controlled vocabulary and preserve this information with the metadata. This allows for correct interpretation of outdated vocabulary even after decades.

3.2 Data Quality

To enable canonical workflows, data must be of known and documented quality. We strive for a high and unified standard so that users can trust the data they receive no matter where it initially came from. Data quality is a complex topic. There are many different views about what constitutes "good quality" data and this definition may depend on the application. That leads to the issue that even though most data we receive in the TOAR data infrastructure has already been quality controlled by the provider there are still varying degrees of quality.

To harmonize the data quality, we employ automated tests of different granularity using statistical methods and heuristics, which assign a score for each data point. Those scores are then translated into categorical data quality flags. They reflect our data curation procedures and document the data quality assessment of both the original provider and the TOAR data centre. Due to the naming convention of the data quality flags in the TOAR-II database, users can easily

see the quality status of all the data in the TOAR database and can refer to data with different quality levels in their analyses. The TOAR data quality control tool (published as Python package toarqc on gitlab [10]) can also be used on its own with different time series data and threshold values for the individual statistical test. This allows for usage in different workflows.

4. CANONICAL ANALYSIS WORKFLOWS AND THE TOAR DATABASE

The use case presented in section 2 shows the importance of data harmonization to ensure a known data quality and re-usable metadata for workflows within a specific scientific community. Canonical workflows, on the other hand, aim at re-using workflow elements in different contexts. This means that the TOAR workflows should be modular and sufficiently generic. In the following we discuss to what extent this can be achieved with help of the FDO and RO concepts. Conversely, we also consider the possibility to adapt other existing workflows to the analysis of TOAR data.

The concept of canonical workflows relies heavily on the use of persistent identifiers (PIDs) for saving and reproducing entire workflows or workflow elements. A particular challenge of working with TOAR data (and other atmospheric monitoring data) is the continuous extension of the data series as time progresses. As described in section 2, researchers will often want to repeat an analysis with the extended data rather than reproduce the exact same results which they obtained previously. Here, canonical workflows are used in two different contexts [25], which will be explained below.

While there are no standardized workflows using TOAR data yet, which could be labelled canonical, there is a relevant use case, which was developed in phase I of the TOAR activity, namely the analysis of health impact metrics for the 2017 Global Burden of Disease assessment [13]. This example prompted the development of new concepts to enhance FAIRness and create FDOs within the TOAR-II database and data service infrastructure.

The measures described in section 3 to harmonize the air quality data facilitates the definition of FDOs, from which canonical workflows can benefit. However, we still see a need for discussing the details for reproducibility of workflow results, which is not the subject of this study. For the TOAR database we intend to use a Research Object (RO). For this RO, "RO-Crate" could be a suitable tool for its management, as it is easy to implement and developed natively for Python [26]. The RO represents an aggregation of FDOs.

Such an FDO, in turn, includes the data together with its metadata, the link to the Git branch ID of the codes used to process this data, as well as the data storage location of the raw data [27]. This FDO has a unique PID [28] and is immutable. Its contents are traceable and reproducible [2].

Since the TOAR database is constantly growing and changing, it is important

that this behavior can be mapped by the RO to provide TOAR data that is usable in canonical workflows. Aside from some data which are automatically retrieved in near realtime, the majority of TOAR data is sent to us in (annual) batches on an irregular basis and inserted into the database. Thus, we propose that each of these data deliveries be created as an FDO after the harmonization and quality control has been completed and the data has been added to the TOAR database. This new FDO is registered with the database RO, which means that the RO and the database reflect the same state again.

Because the RO has an immutable PID, users can save their database requests and use the result to reuse existing workflows without further modification to repeat analyses on extended data sets. As new sets of FDOs are created over time, the RO will be updated with a pointer to the latest FDO and associated data while earlier FDOs remain accessible. In this way, the user is also given the opportunity to repeat the workflow on any snapshot of the database in the past by adapting the query to the RO.

It is important to point out that even though the RO concept allows for data aggregation that is growing incrementally over time, it is crucial to define the authority regarding the mutability and the authenticity of the RO, i.e. the institution who decides about versioning of the data and solves possible conflicts. A discussion on how such RO-crates can be used in canonical workflows can be found in [29].

5. CONCLUSION AND FUTURE WORK

The TOAR data infrastructure presents a relevant and interesting terabyte-scale use case for exploring FAIR data principles and canonical workflows in a dynamic setting with high demands on traceability and reusability. In this paper we discussed the challenges of harmonizing and documenting many different air quality and meteorological time series to allow for globally consistent analyses of ozone air pollution trends.

To further enhance the re-use of TOAR data and enable reusable workflows, we have developed a concept how FDOs and ROs can be integrated into the TOAR data infrastructure. Given the large amount of domain knowledge which is necessary to harmonize incoming data and make informed use of the air quality data stored in the TOAR database, it remains to be seen to what extent our workflows can be modularized and generalized so that other communities might benefit from the TOAR developments. Nevertheless, it appears sensible to develop the technical foundation for allowing canonical workflows through the creation of ROs as described in Section 4.

Adopting the Canonical Workflows for Research (CWFR) concepts will primarily enhance the consistency and documentation of the workflows employed within the TOAR initiative, although some steps of our data ingestion process can be easily transferred to other types of environmental data and perhaps even to data from other disciplines. For example, our data quality control tool can

be adopted to processing a large variety of time series data as the statistical tests and the order in which they are performed can be flexibly configured. Another element which could be adopted by other research communities dealing with geospatial data is our concept for extracting point information through web processing of a variety of Earth Observation data and use this information to enhance the metadata in the TOAR database.

CWFRs are a promising concept, and some common workflows, such as calculating averages on time series data, can be done well. However, we observe that there are many special analysis procedures associated with atmospheric data, which often differ in relatively small details, for example by considering a certain piece of metadata information or not. This makes it hard to generalize and re-use existing workflows for other purposes, since no common metadata standard for atmospheric data exists yet. The better a workflow is documented, and the more standardized elements are used, the easier it will be for the user to exchange elements, in the best case from a library of canonical workflows. Therefore, the technical introduction of such standardized elements for the TOAR database is a decisive step in making workflows reusable. Furthermore, trust in the provided data is essential and this requires clearly defined and well documented procedures for data quality control and assigning data quality flags. The experiences from the TOAR database show that data harmonization alongside with documentation constitutes a big step towards realizing the potential of canonical workflows at least within the community of atmospheric scientists.

AUTHOR CONTRIBUTIONS

S. Schröder (s.schroeder@fz-juelich.de) and M. G. Schultz (m.schultz@fz-juelich.de) developed the concept of the TOAR database and led its implementation. N. Selke (n.selke@fz-juelich.de) developed the data quality concept. E. Epp (e.epp@fz-juelich.de) and M. Romberg (m.romberg@fz-juelich.de) focused on metadata harmonization and documentation. A. Mozaffari (a.mozaffari@fz-juelich.de) established the link to the CWFR. All authors have made meaningful and valuable contributions to writing, revising and proofreading the manuscript.

ACKNOWLEDGEMENTS

The authors would like to thank Jianing Sun for providing them with Figure 1. They gratefully acknowledge the computing resources granted by Jülich Supercomputing Centre (JSC). M.G.S and S.S. acknowledge funding from ERC-2017-ADG#787576 (IntelliAQ).

REFERENCES

[1] Hardisty, A., Wittenburg, P. (eds.): Canonical Workflow Framework for Research (CWFR)—position paper—version 2, December 2020. Working paper. Available at: https://osf.io/9e3vc/. Accessed 9 December 2021

[2] De Smedt, K., Koureas, D., Wittenburg, P.: FAIR digital objects for science: From data pieces to actionable knowledge units. Publications 8(2), Article No. 21 (2020)

[3] Bechhofer, S., et al.: Research objects: Towards exchange and reuse of digital knowledge. Nature Precedings (2010). Available at: https://doi.org/10.1038/npre.2010.4626.1. Accessed 17 December 2021

[4] Tropospheric Ozone Assessment Report (TOAR): Global metrics for climate change, human health and crop/ecosystem research. Available at: https://igacproject.org/activities/TOAR. Accessed 17 December 2021

[5] The TOAR Steering Committee: IGAC' s Tropospheric Ozone Assessment Report. Available at: http://igacproject.org/sites/default/files/2019-11/TOAR_{accomplishments}{*September*}{2019}.pdf. Accessed 17 December 2021

[6] Lewis, A.: Tropospheric Ozone Assessment Report (TOAR): Global metrics for climate change, human health and crop/ecosystem research. Available at: https://online.ucpress.edu/elementa/pages/toar. Accessed 17 December 2021

[7] Schultz, M.G., et al. Tropospheric Ozone Assessment Report: Database and metrics data of global surface ozone observations. Elementa: Science of the Anthropocene 5, Article No. 58 (2017)

[8] The Jülich Supercomputing Centre. Available at: https://toar-data.fz-juelich.de/gui/v1/. Accessed 17 December 2021

[9] Git repository. Available at: https://toar-data.fz-juelich.de/api/v1/. Accessed 17 December 2021

[10] Toar-public. Available at: https://gitlab.jsc.fz-juelich.de/esde/toar-public. Accessed 17 December 2021

[11] Research Data Alliance. Available at: https://rd-alliance.org/. Accessed 17 December 2021

[12] EOSC-hub. Available at: https://www.eosc-hub.eu/news/guidelines-scientific-content-providers-eosc-hub. Accessed 17 December 2021

[13] Stanaway, J.D., et al.: Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. 392, 10159 (Nov 2018), 1923–1994. Appendix. https://doi.org/10.1016/S0140-6736(18)32225-6 (2018)

[14] Monks, P.S., et al.: Tropospheric ozone and its precursors from the urban to the global scale from air quality to short-lived climate forcer. European Geosciences Union 15(15), 8889–8973 (2015)

[15] Lefohn, A.S., et al.: Tropospheric ozone assessment report: Global ozone metrics for climate change, human health, and crop/ecosystem research. Elementa: Science of the Anthropocene 6, Article No. 27 (2018)

[16] Fleming, Z.L., et al.: Tropospheric ozone assessment report: Present-day ozone distribution and trends relevant to human health. Elementa: Science of the Anthropocene, Article No. 12 (2018)

[17] Mills, G., et al.: Tropospheric ozone assessment report: Present-day tropospheric ozone distribution and trends relevant to vegetation. Elementa: Science of the Anthropocene 6, Article No. 47 (2018)

[18] TOAR II Database. Available at: https://esde.pages.jsc.fz-juelich.de/toar-data/toardb_{fastapi}/docs/toardb_{fastapi}.html. Accessed 17 December 2021

[19] Geographic information–Metadata–Part 1: Fundamentals. Available at: https://www.iso.org/standard/53798.html. Accessed 17 December 2021

[20] IPCC: The intergovernmental panel on climate change. Available at: https://www.ipcc.ch/. Accessed 17 December 2021

[21] MCD12C1: The terra and aqua combined moderate resolution imaging spectroradiometer (MODIS) land cover climate modeling grid (CMG).

[22] ISO 3166 country codes. Available at: https://www.iso.org/iso-3166-country-codes.html. Accessed 17 December 2021

[23] GeoLocationServices. Available at: https://gitlab.jsc.fz-juelich.de/esde/toar-data/geolocationservices. Accessed 17 December 2021

[24] World Meteorological Organisation (WMO) Integrated Global Observing System (WIGOS). Available at: https://library.wmo.int/index.php?lvl=notice_{display}&id=20026. Accessed 17 December 2021

[25] Plesser, H.E.: Reproducibility vs. replicability: A brief history of a confused terminology. https://doi.org/10.3389/fninf.2017.00076. Accessed 17 December 2021

[26] Soiland-Reyes, et al.: Packaging research artefacts with RO-Crate. Available at: https://stain.github.io/ro-crate-paper/v/c497e5399a636c75e152ec212434f88d38901288/ (preprint). Accessed 17 December 2021

[27] Lannom, L., Koureas, D., Hardisty, A.R.: FAIR data and services in biodiversity science and geoscience. Data Intelligence 2(1-2), 122-130 (2020)

[28] Schwardmann, U.: Automated schema extraction for PID information types. In: 2016 IEEE International Conference on Big Data, Big Data, pp. 3036-3044 (2016)

[29] Mozaffari, A., et al.: HPC-oriented canonical workflows for machine learning applications in climate and weather prediction. Data Intelligence 4(2), 271-285 (2022)

AUTHOR BIOGRAPHY

Sabine Schröder is one of the main developers of the TOAR database. Before she joined the Earth System Data Exploration (ESDE) group at the Jülich Supercomputing Centre (JSC) as a scientific programmer in 2019, she worked as an application programmer for 12 years in the field of Atmospheric Chemistry with a focus on climate modelling on the HPC systems in Jülich. ORCID: 0000-0002-0309-8010

Eleonora Epp is a trained engineer and works as project manager at the Jülich Supercomputing Centre (JSC). She leads the certification process of the TOAR database and has contributed to the workflow design and documentation of the TOAR database. ORCID: 0000-0003-1266-0607

Amirpasha Mozaffari is the data manager of the ESDE group and responsible for developing the data management and workflows plans. Mozaffari has been trained in terrestrial earth system science and worked on numerical and statistical analysis of environmental data on supercomputers as well as numerical simulations and inversions of groundwater flow before he joined the ESDE group in June 2019. ORCID: 0000-0001-6719-0425

Mathilde Romberg develops the TOAR Ontology and supports the writing of TOAR documentation. Her background is in system administration, grid computing and workflow modelling. ORCID: 0000-0002-6135-3817

Niklas Selke is one of the developers for software accompanying the TOAR database. Before he joined the ESDE group at the Jülich Supercomputing Centre (JSC) as a scientific programmer in late 2020, he received a degree in Applied Mathematics and Informatics where his final thesis work was focused in an earth system science surrounding. ORCID: 0000-0002-9954-2250

Martin Schultz leads the ESDE group. He holds the ERC grant IntelliAQ and leads the activities on high-performance data provisioning. Schultz is a well-known expert in atmospheric science and numerical simulations. His research interests now focus on high performance data services and machine learning applications in the fields of meteorology and air quality. ORCID: 0000-0003-3455-774X

Note: Figure translations are in progress. See original paper for figures.