TOAR Database Infrastructure

Notes Before Completing the Application

We have read and understood the notes concerning our application submission.

True

Reviewer Entry

Reviewer 1

Comments:
I can confirm it

Reviewer 2

Comments:

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background & General Guidance

Glossary of Terms

BACKGROUND INFORMATION

Context

R0. Please provide context for your repository.

Repository Type. Select all relevant types from:

Domain or subject-based repository, Research project repository

Reviewer Entry

Reviewer 1

Comments:
accepted

Reviewer 2

Comments:
Accept

Brief Description of Repository

The TOAR database infrastructure supports the “Tropospheric Ozone Assessment Report” (TOAR) and the Earth system science community with a database of harmonised multi-annual time series of atmospheric constituent measurements and meteorological variables. TOAR project Phase II is an activity of the International Global Atmospheric Chemistry (IGAC) organisation . Version 2 of the TOAR database infrastructure has been developed based on the requirements of TOAR phase II and is committed to the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles. The TOAR data infrastructure is operated at Forschungszentrum Jülich embedded in the infrastructure at the Jülich Supercomputing Centre.
Data is harvested or received from measurement stations collecting surface ozone data and networks of such stations from all over the world. The data accepted for the TOAR database are global air quality data with the focus on ozone data and related information. Data is received from various providers and is openly accessible (CC BY4.0). The data is detailed in section 4.1 of https://toar-data.fz-juelich.de/documentation/TOAR_UG_Vol03_Database.pdf.

Reviewer Entry

Reviewer 1

Comments:
it is clear about the necessary information summried.

Reviewer 2

Comments:
Accept

Brief Description of the Repository’s Designated Community.

The designated community are the researchers of the TOAR activity, an open international community of scientists investigating topics related to air quality and atmospheric chemistry, including climate change, impacts on human health and ecosystems, carbon and nitrogen cycles, biodiversity, and other environmental issues. The community is actively collaborating across geographical boundaries and disciplines in order to address the most pressing global change and sustainability issues through scientific research. Also all scientists who are interested in the research of climate data belong to the designated community.

Reviewer Entry

Reviewer 1

Comments:
the user community is focused and cleared definated.

Reviewer 2

Comments:
Accept

Level of Curation Performed. Select all relevant types from:

A. Content distributed as deposited, B. Basic curation – e.g. brief checking; addition of basic metadata or documentation, C. Enhanced curation – e.g. conversion to new formats; enhancement of documentation

Reviewer Entry

Reviewer 1

Comments:
clear and accpeted

Reviewer 2

Comments:
Accept

Comments

All data that is received for being stored in the TOAR database is stored as is together with the curated data and metadata. The metadata is either received with the data or generated. Plausibility checks are made for the metadata to identify inconsistencies such as the location of a station which is classified as urban but the coordinates point to a rural place etc. During data curation we harmonise the data to use the same formatting style and follow our controlled vocabulary. Details are provided in https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol02_Data_Processing.pdf.

Reviewer Entry

Reviewer 1

Comments:
clear and accpeted

Reviewer 2

Comments:
Accept

Insource/Outsource Partners. If applicable, please list them.

Outsource partners are

• Data archival: as part of multicopy redundancy we store daily backup-copies of the database at the IT Centre of the RWTH Aachen University, about 27 km away from Jülich. The partner has not undertaken any trustworthy repository assessment but they comply with the ITIL 'Best Practices' for the design of their internal processes especially for backup

and restore (https://help.itc.rwth-aachen.de/en/service/t4ctl3msqrmt/article/00e9f3476f244892867895d58aa3a91d/) Forschungszentrum Jülich and RWTH Aachen have a general contract for collaboration on scientific and technical level (JARA, https://www.jara.org/en/).

• Data publication: we are using a B2Share instance maintained by another group at the Jülich Supercomputing Centre for publishing data sets to have them referenceable with a DOI. A trustworthy repository assessment has not been undertaken yet.

B2Share is a service developed by EUDAT. Forschungszentrum Jülich is a member of the EUDAT CDI (https://eudat.eu/eudat-cdi/members) and thereby is entitled to run instances of the services. The relationship to the group running the service at JSC is organisational; there is no SLA.

• Software maintenance for operating systems and basic software on the server machines used for the TOAR database infrastructure is provided by other groups at the Jülich Supercomputing Centre. This includes handling of backups and caring of long-term availability of data.

This is an organisational relationship, there is no SLA.

• Hardware maintenance is outsourced to the manufacturers.

These are contractual relationships with SLA covering on-site 9x5 next business day. In case of hardware failures Virtual Machines can be moved as a first measure to get back into production. Customer replaceable parts will only be delivered.

Insource partners are

• There are about 40 different data providers with whom we have non-contractual agreements on individual basis. The major data providers are listed in section 2 of https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol02_Data_Processing.pdf.

Reviewer Entry

Reviewer 1

Comments:
Thank you for the explanation. It is clear.

Reviewer 2

Comments:
Accept

Summary of Significant Changes Since Last Application (if applicable).

not applicable

Reviewer Entry

Reviewer 1

Comments:
none

Reviewer 2

Comments:

Other Relevant Information.

The Tropospheric Ozone Assessment Report (TOAR) is an activity by the International Global Atmospheric Chemistry (IGAC) initiative, it receives financial and logistical support from IGAC, Forschungszentrum Jülich GmbH, the US National Oceanic and Atmospheric Administration, and the World Meteorological Organization (an agency of the United Nations). It is a project driven activity by IGAC and progresses in a sequence of phases with a duration of five years each.
IGAC, founded in 1990, has a long-term vision. It operates under the umbrella of Future Earth and is sponsored by the international Commission on Atmospheric Chemistry and Global Pollution (iCACGP), a commission of IAMAS-IUGG (see https://igacproject.org/mission). IGAC‘s International Project Office (IPO) is located in the U.S. and funded by NSF, NOAA, and NASA, and renews its funding on a three-year grant application cycle (currently approved through 2024).
In TOAR phase II more than 270 scientists from 37 countries collaborate to produce a scientific assessment of tropospheric ozone changes. A series of publications as well as published data sets have been produced during the first phase of the assessment (http://igacproject.org/sites/default/files/2019-11/TOAR_accomplishments_September_2019.pdf). At the core of this activity is the TOAR database to make ozone data and related information available to academic users and other stakeholders. The TOAR database infrastructure has been developed by Forschungszentrum Jülich GmbH and is maintained and continuously improved by the research group Earth System Data Exploration within the division Federated Systems and Data of the Jülich Supercomputing Centre. The TOAR Database has been registered with re3data (http://doi.org/10.17616/R3FZ0G).
The size of the database is about 3 TB and will be increased by about 150 GB/year.
While the first instance of the TOAR database was specifically tailored towards the first TOAR assessment, this data collection and data service infrastructure has received great attention by the atmospheric science community and other stakeholders. This has motivated us to develop the original project database into a fully functional TOAR database infrastructure which will exist in the long-term independently of the current project support. We therefore apply for certification under the rules of the CoreTrustSeal.

Reviewer Entry

Reviewer 1

Comments:
Thank you for the explanation. It is clear.

Reviewer 2

Comments:
Accept

ORGANIZATIONAL INFRASTRUCTURE

1. Mission/Scope

R1. The repository has an explicit mission to provide access to and preserve data in its domain.

Compliance Level:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase
agree

Reviewer 2

Comments:
3 – The repository is in the implementation phase
Accept

Response:

The mandate for operating the TOAR Database Infrastructure is given by the TOAR phase II activity of IGAC. Its mission is to act as the central hub for data access in support of research assessing the impacts of ozone air pollution on human health, vegetation, and climate. Besides maintaining a data portal with links to ozone data sets from research organisations all over the world, we operate a database of harmonised surface ozone measurements and related data. This is one of the largest collections of quality controlled air pollution measurements in the world. All data in the database are easily accessible through open, freely available and well documented web services. The TOAR data team is committed to the FAIR principles and aims to achieve the highest standards with respect to data curation, archival, and re-use (see https://toar-data.fz-juelich.de/).
The Jülich Supercomputing Centre (JSC) at Forschungszentrum Jülich has accepted the responsibility for this repository and is committed to its long-term operation. JSC has been operating the first German supercomputing centre since 1987. Data management is an important part of it. About 250 experts and contacts for all aspects of supercomputing and simulation sciences work in JSC. JSC is part of the large-scale national research facility, Forschungszentrum Jülich GmbH, which is part of the Helmholtz Association of German Research Centres. In May 2020 the director of JSC signed a letter to guarantee the operation of the TOAR Database Infrastructure for at least 10 years.

Reviewer Entry

Reviewer 1

Comments:
clear and accepted

Reviewer 2

Comments:
Accept

2. Licenses

R2. The repository maintains all applicable licenses covering data access and use and monitors compliance.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
agree

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

All data in the TOAR version 2 database is publicly accessible under Creative Commons License (CC BY 4.0) "Attribution". TOAR makes use of open data only and provides attribution information with the queried data. The data policy is available at https://toar-data.fz-juelich.de/footer/terms-of-use.html. There is no monitoring of compliance.

Reviewer Entry

Reviewer 1

Comments:
clear and accepted

Reviewer 2

Comments:
Accept

3. Continuity of access

R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.

Compliance Level:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase
agree

Reviewer 2

Comments:
3 – The repository is in the implementation phase
Accept

Response:

The TOAR Database Infrastructure is hosted by the Division Federated Systems and Data of the Jülich Supercomputing Centre at the Forschungszentrum Jülich GmbH. Every five to seven years the Forschungszentrum Jülich applies for its base funding for the next period in the so called Program-oriented Funding (PoF, https://www.helmholtz.de/en/research/program-oriented-funding/) process. The current program (the program period is 2021-2027) contains Earth System Science with the research group Earth System Data Exploration and the TOAR activity as integral part. That gives the funding for continued operation.
The long-term plans to ensure continued operation are to apply for further funding in the next PoF period. This will depend on the strategic plans of the JSC and Forschungszentrum Jülich and on the future of the TOAR-II activity. If at some point the operation of the TOAR database infrastructure will terminate, we will seek partners in the community to take over with due preparation time. We are optimistic to find a partner from the Earth System Science to take over the TOAR Database Infrastructure in the unlikely event that Forschungszentrum Jülich cannot continue hosting it but we cannot give any guarantee. For enabling another institution to take over hosting and operating the TOAR database infrastructure, all software and data as well as the documentation are designed as open source and open data (see https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol01_Infrastructure.pdf, .../TOAR_TG_Vol02_Data_Processing.pdf, and the TOAR gitlab repository at https://gitlab.jsc.fz-juelich.de/esde/toar-data).
In our current effort, JSC will do everything possible to secure the TOAR database and its data for at least 10 years in a long-term archival system. The director of JSC guarantees the technical conditions for the continuous access to the TOAR database.

Reviewer Entry

Reviewer 1

Comments:
clear and accepted

Reviewer 2

Comments:
Compliance level 3 because there is now public continuity plan and because there is no formal written agreement between the repository and another organization that would guarantee taking over in the case of discontinuity.

4. Confidentiality/Ethics

R4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
agree

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

The TOAR Database does not host sensitive personal data nor any other delicate or confidential data (https://toar-data.fz-juelich.de/footer/terms-of-use.html). The infrastructure supports the data producers to comply with rules of good scientific practice as called for by the WDS Data Policy, DFG (German Research Foundation), HGF (Helmholtz Association) and others.
There are no ethical or legal issues involved in air quality and meteorological data except ownership. As we establish direct communication with all data providers, we make sure that they provide only data to the TOAR database that they are entitled to pass on (see section 2.1 of https://toar-data.fz-juelich.de/documentation/TOAR_UG_Vol05_Data_Submission_Guide.pdf; section 2 of https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol02_Data_Processing.pdf).
Before a data publication, the proper attribution of roles (according to ISO19115 and DataCite) is discussed with the provider. In the case of OpenAQ, which itself is a data collector, all data they provide, to the best of their knowledge, has been made available for free redistribution and use throughout the world (see https://github.com/openaq/openaq-info/blob/master/DATA-POLICY.md).

Reviewer Entry

Reviewer 1

Comments:
clear and accepted

Reviewer 2

Comments:
Accept

5. Organizational infrastructure

R5. The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
agree

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

The TOAR Database Infrastructure is hosted by Forschungszentrum Jülich GmbH, a governmental research institute which is funded by the Federal Republic of Germany and the State of North Rhine-Westphalia. Every seven years the strategic research policy is adapted when applying for base funding in the PoF process described in R3. With this funding and third-party funding (about 40% of total budget) for research projects the experienced staff can be kept and new staff trained to the needs. The Tropospheric Ozone Assessment Report (TOAR) is an activity sponsored by the International Global Atmospheric Chemistry (IGAC) initiative. About five Full Time Employees develop and maintain the TOAR Database Infrastructure. With the in kind hosting of the TOAR Database Infrastructure by the Jülich Supercomputing Centre the TOAR Database as integral part of it has robust funding commitments.
About 250 experts and contacts for all aspects of supercomputing and simulation sciences work in JSC. Staff members regularly take part in training courses on, e.g., system administration, data management, programming or web security. As part of its company policy Forschungszentrum Jülich offers training programs as well as support for individual training (external classes). By career, the majority of the staff members come from computer science and physics. They take part in and make contributions to international conferences and some are involved in projects like Earth System Grid Federation (ESGF), Research Data Alliance (RDA), Helmholtz Data Federation (HDF), European Open Science Cloud (EOSC), European Network for Earth System Modelling (ENES) and Coupled Model Intercomparison Project (CMIP).
The TOAR data infrastructure is governed by IGACS's TOAR-II activity and its steering committee. Advice and feedback is received from the user forum. In addition, Forschungszentrum Jülich carries out the operation of the TOAR Data Centre through JSC's division Federated Systems and Data (see Figure 1 of

https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol01_Infrastructure.pdf and section ‘Governance and Procedural Documents’ at https://igacproject.org/activities/TOAR/TOAR-II).

Reviewer Entry

Reviewer 1

Comments:
clear and accepted

Reviewer 2

Comments:
Accept

6. Expert guidance

R6. The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either inhouse or external, including scientific guidance, if relevant).

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
agree

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

The TOAR-II steering committee leads the entire Tropospheric Ozone Assessment Report activity and ensures that all developments within TOAR-II, including the TOAR database infrastructure, are in line with the scope and objectives of TOAR-II. Specific user feedback and user requests are gathered through the TOAR Data User Forum, which has no formal structure and membership. The steering committee and the user forum comprise experts from the discipline and some people with a data science background. The user forum convenes once a year and began its operation in January 2021 on day 2 of the TOAR-II Kick-Off Workshop (https://igacproject.org/activities/TOAR/TOAR-II).
The steering committee cooperates directly with the TOAR Data Infrastructure team to prioritize the advice and requests

from the User Forum. The TOAR Data Infrastructure team reports on data status and data usage and regularly prepares a summary of issues and a roadmap for further development. The communication is mostly done via e-mail and web meetings.
Further communication with data curation and data science experts happens on a regular basis due to the strong involvement of the Jülich Supercomputing Centre in European research data management activities (e.g. European Open Science Cloud). TOAR database infrastructure staff also attends relevant fora on open data and repository management, such as Research Data Alliance (https://www.rd-alliance.org/) and ENVironmental Research Infrastructure-FAIR (https://envri.eu/home-envri-fair/).

Reviewer Entry

Reviewer 1

Comments:
clear and accepted

Reviewer 2

Comments:
Accept

DIGITAL OBJECT MANAGEMENT

7. Data integrity and authenticity

R7. The repository guarantees the integrity and authenticity of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
agree

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

The curation lifecycle starts with the data ingestion workflow. We have two categories of data sources, existing databases from larger networks which are harvested and data submissions from individuals or individual networks which are processed in a largely automated workflow (see https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol02_Data_Processing.pdf).
Submitted data is stored as is before it is pre-processed and harmonized. In case of errors or questions direct feedback communication with the provider is established. Most changes done during the harmonization are kept in a text file with the data. Harvested data, for example in the format of database dumps, are treated with special versions of the pre-processing scripts. Before the data is published the data providers are given the opportunity to review the data and request corrections. All communication in this regard is archived.
Once inserted into the database all changes, for example due to recalibration on the provider side, occur with specific software tools and are logged in the database itself (see section 4.3 on provenance in https://toar-data.fz-juelich.de/documentation/TOAR_UG_Vol03_Database.pdf). There is also a strict data versioning schema in place. Currently, there are no tools in place to directly compare different dataset versions. These will be developed as the need arises.
A snapshot of the database is produced and published whenever we freeze a certain state of the database for assessment purposes and at regular intervals. The integrity of these snapshots can be checked with MD5 checksums.
The completeness of metadata in the sense of “complete description of the data” depends on the user community and specific use case. All metadata that are sent to us by providers are preserved and made available to users. Essential metadata, such as station location, variable name, and physical unit of the data, are always available - a dataset will not be ingested into the TOAR database and it will not be published without these metadata. The metadata is fully documented in the reference guide (see https://toar-data.fz-juelich.de/documentation/TOAR_UG_Vol05_data_submission_guide.pdf).
Due to the nature of the measurements, missing data values are common and cannot be avoided. Upon data extraction from the database it is possible to obtain statistics about the data completeness ("coverage") for a given period (see Table 6: Definition of metrics used in the TOAR analyses. of https://ucp.silverchair-cdn.com/ucp/content_public/journal/elementa/5/10.1525_elementa.244/3/elementa-5-244-s1.pdf).
For TOAR data publications, according to the DOI minting in B2SHARE, new versions of data publications from individually submitted data will result in a new publication record/version with a new DOI. In such cases, cross references between the dataset versions are maintained.
Where possible, ontologies from the ISO 19115 standard for geographic information (https://def.isotc211.org/ontologies/iso19115/) and controlled vocabulary from the Climate and Forecasting (CF) Conventions (http://cfconventions.org/) are used. Further, we adopt terminology from other sources. By blending controlled vocabulary and ontologies with less constrained full text fields we are able to preserve any metadata supplied to us.
Provenance information are generated as metadata in the data management workflow and are subsequently maintained. These metadata are stored together with the data and made available to the users (see section 4.3 in https://toar-data.fz-juelich.de/documentation/TOAR_UG_Vol03_Database.pdf). All data depositors are personally known to the database maintainers or to members of the TOAR steering group. In rare cases where no previous contact existed we establish credibility of the data providers through questioning peers, for example in the IGAC community (https://www.igac.org) or one of the Scientific Advisory Groups of the Global Atmosphere Watch Programme of the World

Meteorological Organization (https://public.wmo.int/en/programmes/global-atmosphere-watch-programme). Typically, new data providers will also come with peer reviewed journal articles describing their datasets.
For data collected from other large air quality measurement archives the provenance information contains the original data source. Links to external metadata are not provided.

Reviewer Entry

Reviewer 1

Comments:
clear and accepted

Reviewer 2

Comments:
Accept

8. Appraisal

R8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
agree

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

The scope of the TOAR database infrastructure's data collection effort has been decided in consultation with the TOAR-II steering committee (see Governance section of https://igacproject.org/activities/TOAR/TOAR-II) and is well defined through its mission to support the Tropospheric Ozone Assessment Report with global surface observations of air pollutants and related meteorological variables.
The TOAR database infrastructure staff will reject submissions of data which do not fall in the scope of the database. If in doubt, consultation with the TOAR-II steering committee will be sought.
https://toar-data.fz-juelich.de/documentation/TOAR_UG_Vol05_Data_Submission_Guide.pdf contains a format

description for data submissions and a comprehensive checklist that helps data providers to verify and organize their data and metadata submissions. The first screening for completeness of data and metadata is done manually. We communicate directly with data providers to ensure completeness and correctness of metadata and data. The validation procedures for submitted as well as harvested data are described in R11 and include automated quality screening of metadata and data.
Deviations from the TOAR data format are handled gracefully: where possible, automated corrections are applied as part of the data harmonization step (see https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol02_Data_Processing.pdf). We also accept database dumps or other formats from data providers, who send us large volume data.
We work together with the data providers to ensure utmost completeness and correctness of the metadata. Especially in case of data providers, who offer large data volumes, we have defined specific metadata and data mapping procedures, so that all metadata in the TOAR database are harmonized. All of these procedures are automated to the extent possible (see https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol02_Data_Processing.pdf) and made available through a git repository (https://gitlab.jsc.fz-juelich.de/esde/toar-data).
Currently there is no policy for removing data from the TOAR database and we don't expect that this will be necessary. The only items with a DOI are data publications from data providers who submitted their data to us. These data sets are published with an external provider (B2SHARE instance at Jülich Supercomputing Centre). In case changes are made to such a data set the updated data set is published and the DOI received in return is stored with that version of the data in the TOAR database.

Reviewer Entry

Reviewer 1

Comments:
clear

Reviewer 2

Comments:
Accept

9. Documented storage procedures

R9. The repository applies documented processes and procedures in managing archival storage of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
agree

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

All storage systems used are located at the JSC with the exception of the backup copy which resides at the RWTH Aachen University. The storage facilities of the TOAR database infrastructure are managed by professional staff at JSC. There are dedicated contact persons for all hardware and operating system issues.
The TOAR database is regularly secured using a combination of full backups and write ahead logs (see https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol01_Infrastructure.pdf). Multiple and physically separated copies of these dumps including checksums are stored in Jülich and are mirrored at RWTH Aachen University.
The storage strategy has been developed in collaboration with the experienced HPC system administrators at JSC and follow standard risk management strategies for scientific computing centres. Risk assessments are regularly performed (refer to section 4 of https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol01_Infrastructure.pdf). The TOAR data infrastructure team regularly discusses possible risks and their potential implications and will adapt the data storage procedures accordingly so that these risks are minimized.
At JSC all storage resources are under active support with appropriate support contracts in place. For rotating disks, devices exhibiting errors are replaced. The redundancy levels of the systems allow to tolerate more than one disk failure without implications on the data integrity and availability. Tape media are replaced on a regular basis whose cadence is determined by data growth and technology availability. The health, functionality and performance of all storage resources are actively monitored by a dedicated team of operators.

Reviewer Entry

Reviewer 1

Comments:
clear and accepted

Reviewer 2

Comments:
Accept

10. Preservation plan

R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
agree

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

The procedures for TOAR data preservation and the long-term data archival strategy are described in section 3 of https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol01_Infrastructure.pdf. We distinguish two elements of data preservation here:
(i) data preservation during the lifetime of the TOAR initiative and the operation of the TOAR database infrastructure,
(ii) data preservation beyond the operation of the TOAR data centre at JSC.
The steps taken are
• The operation of the TOAR database infrastructure and its data services including curation, quality control etc.) is presently secured until the end of 2027 (end of current PoF period). Since JSC has a long-term strategic mission to develop the “data centre of the future”, there is very little risk that funding of the TOAR database infrastructure operation will not be covered by the subsequent PoF period. In the worst case scenario, we will take measures to at least ensure the availability of data in archived form (e.g. TOAR database dumps) and we will seek out another institution to become the new host for the TOAR data centre. There is a commitment by the director of JSC to ensure that at least one copy of all TOAR data is kept and made available to the user community until May 2030.
• While in operation the availability of TOAR data is ensured through the backup and archival strategy (see R9). Backup copies are preserved across system changes and actively copied when new hardware replaces existing systems. The consistency and functionality of these copies will be tested when changes occur.
• The database infrastructure and its services will be regularly updated and migrations will be performed as necessary. This will ensure that TOAR data remain discoverable (see R13), accessible and (re)usable (see R14). Beyond the TOAR database infrastructure operations we must rely on backward compatibility of the employed file formats, i.e. PostgreSQL database dumps, csv, and netCDF files.
• The long-term preservation of the TOAR database content is achieved through providing open access to regular database dumps on a data publication service with DOI together with the TOAR database software on gitlab. This ensures longevity of the data even beyond any possible termination of the TOAR activity. While in operation, the database software will be regularly updated to keep up with technology development and any potential changes of data formats will be implemented.
• Regular user forums and user interaction allow to foresee specific format and data aggregation requests. They ensure that code development can be planned early and prevent file formats from becoming obsolete. Data are not stored as files

but in a database, i.e. conversion tools for outputting data in various formats can easily be made available. Durable and widely accepted standard file formats have been chosen for the database (see section 4.1 of https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol01_Infrastructure.pdf).
• The TOAR database infrastructure team has defined clear responsibilities for each of the procedures outlined above (see author list of https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol01_Infrastructure.pdf).
The data providers (depositors) have no formal contract with the TOAR database infrastructure but all aspects of long-term preservation that are relevant for the depositors are described in https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol02_Data_Processing.pdf and https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol01_Infrastructure.pdf. As we require all data providers to agree on having their data published under a CC-BY 4.0 license it is ensured that the repository has the rights to copy, transform, and store the data and make it accessible.
A mapping of the TOAR database infrastructure to OAIS terms is provided in https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol03_OAIS_Mapping.pdf.

Reviewer Entry

Reviewer 1

Comments:
clear and accepted

Reviewer 2

Comments:
Accept

11. Data quality

R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
agree

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Accept

Response:

There does not exist any universal standard for assessing the quality of air pollution data. However, the TOAR Data Infrastructure is an integral part of the TOAR activity. This ensures that the necessary expertise to assess the quality of data and metadata is available.
The quality of data and metadata are checked for each data entry via a combination of automated and manual tools as described in https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol02_Data_Processing.pdf. In case of individual time series submissions, a first manual inspection is followed by an automated workflow to check for metadata errors or inconsistencies and screen the data values for possible errors with the help of statistical methods. For bulk uploads the data are checked periodically by re-running the statistical tests. Metadata resulting from these bulk uploads is controlled on occasion, especially when the dataset for the next TOAR assessment is prepared. Users can report data and metadata issues at any time. Automated checks for compliance with controlled vocabulary are implemented. A strong element of quality control is the TOAR assessment itself, when scientists from the TOAR community exploit the entirety of observational datasets to provide an updated state of the science estimate of ozone's global distribution and trends relevant to climate, human health and vegetation (https://igacproject.org/activities/TOAR/TOAR-II).
Air quality monitoring is organized regionally. There is no authoritative global community of practice for air quality data. We mostly follow the comprehensive WIGOS metadata standard (https://library.wmo.int/?lvl=notice_display&id=19925#.XvYToOfgphE) of the World Meteorological Organisation which provides a rather complete description of stationary ground-based atmospheric measurements as we store them in the TOAR database.
Data with incomplete, erroneous or inconsistent metadata information or datasets with clear quality problems are not inserted into the TOAR database. In such cases we communicate with the data providers to remedy the issues.
The TOAR web services provide a user feedback function to easily report any data or metadata issues. Furthermore, the TOAR User Forum discusses all aspects of the TOAR data and metadata including their quality and regularly provides feedback.
The TOAR database schema allows to store related works in the form of links or pdf copies. However, this feature has so far not been used by the user community.

Reviewer Entry

Reviewer 1

Comments:
clear and accepted

Reviewer 2

Comments:
Accept

12. Workflows

R12. Archiving takes place according to defined workflows from ingest to dissemination.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
agree

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

We have mapped the TOAR Database Infrastructure to the OAIS terms (see https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol03_OAIS_Mapping.pdf).
Data is ingested into the TOAR database according to the workflows described in https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol02_Data_Processing.pdf. When data from a new network of measurement stations is added to the collection, the TOAR community is informed through news entries on the TOAR data portal (https://toar-data.org/news/).
After ingestion of new time series received from individual providers, a standardised quality control plot is generated, inspected by the TOAR Database Infrastructure team and shared with the data provider.
All output from the TOAR database is generated through automated scripts, which have been verified by the TOAR user community via independent analyses. Code changes are tested for consistency via unit tests and new statistical methods are verified by the TOAR statistics team. An important quality check occurs during preparation of the TOAR assessment, because the dataset is then scrutinized by many critical scientists.
All software used in the TOAR data workflows is managed through gitlab (https://gitlab.version.fz-juelich.de/esde/toar-data and https://gitlab.version.fz-juelich.de/esde/toar-public).

Reviewer Entry

Reviewer 1

Comments:
clear and accepted

Reviewer 2

Comments:
Accept

13. Data discovery and identification

R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.

Compliance Level:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase
clear

Reviewer 2

Comments:
3 – The repository is in the implementation phase
Accept

Response:

The TOAR database infrastructure version 2 offers a REST web service for searching and accessing data in the TOAR database. A graphical user interface building on version 1 is in planning and the search capabilities shall be further enhanced in the future. The database has been registered in re3data (http://doi.org/10.17616/R3FZ0G).
The metadata is provided through the REST API of the TOAR database (https://toar-data.fz-juelich.de/api/v2/). A searchable catalogue is under development. This JSON interface is (technically) interoperable, but the metadata are not fully compliant to internationally agreed standards, because these standards are not fully developed and harmonised in the field of air quality monitoring.
Datasets resulting from a search in the TOAR database, which are relevant to the TOAR activity, can be published in the external B2SHARE service (https://b2share.fz-juelich.de/communities/TOAR) and receive a DOI (section 4 of https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol02_Data_Processing.pdf). For this publication, the TOAR database infrastructure team maps the relevant TOAR metadata to the DataCite metadata standard.

Reviewer Entry

Reviewer 1

Comments:
clear and accepted

Reviewer 2

Comments:
Accept

14. Data reuse

R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
agree

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

Data reusability depends on interpretable metadata, documented change processes, and usable data formats. A large part of the metadata is controlled and managed as controlled vocabulary. We make an effort to make all metadata in the TOAR database easily understandable. All metadata are defined in an ontology (https://toar-data.fz-juelich.de/documentation/ontologies/v1.0/). The underlying standard for all metadata in the TOAR data infrastructure is ISO 19115 „Geographic Information – Metadata“ and we adopt standards from related infrastructures such as the WIGOS metadata standard from the World Meteorological Organisation where appropriate. Detailed documentation of the TOAR metadata is given by the gitlab pages documentation (https://esde.pages.jsc.fz-juelich.de/toar-data/toardb_fastapi/docs/toardb_fastapi.html) and https://toar-data.fz-juelich.de/documentation/TOAR_UG_Vol03_Database.pdf also provides background information. The controlled vocabulary and associated ontologies are available through the TOAR REST API (https://toar-data.fz-juelich.de/api/v2).
All changes to the database schema and other metadata are documented and can be tracked through the gitlab versioning system containing the database and web services source codes (https://gitlab.jsc.fz-juelich.de/esde/toar-data). Major changes will lead to a new version of the TOAR database infrastructure (our URLs contain a version number).
The most ubiquitous formats in the designated user community are ASCII, CSV, netCDF, and JSON (or GeoJSON). Currently, data are delivered as either JSON or CSV. Metadata queries return JSON output, and the ontology is delivered as XML (OWLDOC). A netCDF converter and other metadata formats are planned. As we are closely involved in the TOAR activity and have good links into the wider atmospheric science community, we will pick up new requirements and

enhance our data services accordingly.

Reviewer Entry

Reviewer 1

Comments:
clear and accepted

Reviewer 2

Comments:
Accept

TECHNOLOGY

15. Technical infrastructure

R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
agree

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

The setup of the infrastructure and its maintenance is described in https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol01_Infrastructure.pdf. The infrastructure is using Virtual Machines (VM) in an OpenStack (https://www.openstack.org/) or a VMWARE (https://www.vmware.com/) cloud environment with Ubuntu (https://ubuntu.com/) 20.04 LTS operating system. The setup of the service is well documented

in https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol01_Infrastructure.pdf, such that it can be rebuild by other system administrators. This is also the case for the documentation of the backup procedures, which additionally includes a section on the recovery procedures that has been verified to work appropriately.
The TOAR database infrastructure is monitored for problems on hardware as well as on services level and a ticket system (Znuny with email interface) in combination with support shifts is employed to track and solve problems.
The TOAR database infrastructure is available 24/7, experience so far shows a general availability of 98%. Its network configuration provides data rates of 1GB/s.
Along the lines of the ISO Standard 14721:2012 (OAIS Reference Model), the high-level services are realized in the following way:
Ingest: An elaborate workflow of ingestion steps has been implemented in python as described in https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol03_Data_Processing.pdf.
Archival Storage: TOAR database is a PostgreSQL 13 database (software provided by the Ubuntu distribution) with the PostGIS and toar_controlled_vocabulary extensions installed. It runs on a VM in the OpenStack cloud environment. The database is backed up incrementally on a daily basis.
JSC as host of TOAR database infrastructure takes the specific precautions of supercomputing centres against loss of infrastructure and data in case of local-scale emergencies, including a comprehensive backup strategy with copies at a remote location. The hardware used is closely monitored and a hot-swap strategy is in place. Also the regular acquisition of new hardware and re-installation of database and services help to prevent data loss.
Data Management: The data is managed via mechanisms of the standard PostgreSQL tools (eg. vacuum), the corresponding reports are automatically created. Management of metadata, versioning, and data quality is given in https://toar-data.fz-juelich.de/documentation/TOAR_UG_Vol03_Database.pdf.
Preservation Planning: The procedures for TOAR data preservation and the long-term data archival strategy are described in https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol01_Infrastructure.pdf (see also R10).
Access: For accessing the data a REST interface is provided and a graphical user interface will be programmed in python which uses the REST API to access the database. The interface is running on a VM in a VMWARE cloud environment at JSC. Higher-level software making use of the database is part of the git repository and publicly accessible (see https://gitlab.jsc.fz-juelich.de/esde/toar-public).
Administration: The OpenStack environment uses Kubernetes (https://kubernetes.io/) for container deployment as well as puppet (https://puppet.com/) for configuration and change management. The systems are continuously monitored using Icinga (https://icinga.com/) and Nagios (https://www.nagios.org).

Reviewer Entry

Reviewer 1

Comments:
clear and accepted

Reviewer 2

Comments:
Accept

16. Security

R16. The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
agree

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

The TOAR database infrastructure is operated at Forschungszentrum Jülich embedded in the infrastructure at the Jülich Supercomputing Centre which maintains a high level of protection. The physical facilities are protected on campus level by a barbed wire fence with camera surveillance. Access for staff and visitors is only possible by controlled gates. A security service monitors the campus; at night security guards patrol the buildings.
The machine rooms within JSC have locked doors with a small number of people having access keys (electronic key system) to enter the rooms.
JSC's security system relies on
• restricted access to the hardware: only entitled administrators have an access key to the machine rooms;
• hardware monitoring to get early warnings of defecting hardware;
• firewalls: access from outside is blocked at a firewall guarding internal systems, publicly accessible systems are protected by a firewall monitoring the traffic and raising alarms when detecting unusual access patterns;
• service monitoring with icinga to automatically detect malfunction;
• automated test procedures for testing the accessibility of the services are in place;
• software management with version control and unit testing according to good programming practices.

Section 4 of https://toar-data.fz-juelich.de/documentation/TOAR_TG_Vol01_Infrastructure.pdf describes the risk assessment and the measures taken. The risks are evaluated on the basis of the risk assessment matrix from Matthew S. Mayernik (https://datascience.codata.org/articles/10.5334/dsj-2020-010/).

The TOAR database infrastructure supports two security levels: the standard JSC level for all user accessible services and a higher level for the TOAR database itself. Data read access requires no authentication or authorisation. Any write operations and system maintenance operations are restricted to few selected persons and are protected with ssh, firewall

rules and the database user management.

Reviewer Entry

Reviewer 1

Comments:
clear and accepted

Reviewer 2

Comments:
Accept

APPLICANT FEEDBACK

Comments/feedback

These Requirements are not seen as final, and we value your input to improve the CoreTrustSeal certification procedure. Any comments on the quality of the Requirements, their relevance to your organization, or any other contribution, will be considered as part of future iterations.

Response:

Reviewer Entry

Reviewer 1

Comments:
this applictaion has been well prepared. I am satisfied with it.

Reviewer 2

Comments:
A solid application, and R13 might even be on level 4. I recommend approving this application.