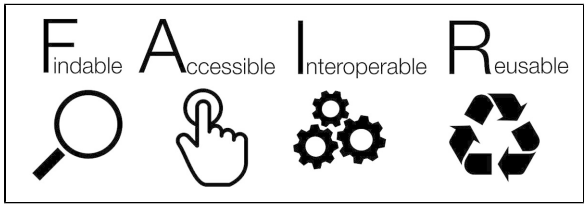
In detail: MOSAiC data publication guideline

[Why do I have to publish “my” MOSAiC data?] [What is data publication?] [Where do I publish “my” MOSAiC data?] [When do I submit and publish “my” MOSAiC data?] [How do I approach data publication?] [How do I share and publish my code/software related to the data?]

Why do I have to publish “my” MOSAiC data?

- Each PI / lead author assures both raw and primary (processed) data are published, which will guarantee their Findability, Accessibility, Interoperability and Reusability (FAIR data principles).
- MOSAiC participants by agreeing to the MOSAiC Data policy ensure that MOSAiC is a successful and resource-effective research project. With that, each PI / lead author assures both raw and primary (processed) data are published. Data publication in a dedicated data repository will guarantee their Findability, Accessibility, Interoperability and Reusability (FAIR data principles).
- Benefits of data publication:
  - Visibility – more citations
  - Credibility – more credits
  - Exchange – improve accessibility



Fair data principles by SangyaPundir from Wikimedia Commons

Resources

- MOSAiC Data Policy
- FAIR data principles

A digital object identifier (DOI) is a persistent identifier used to identify objects uniquely

Metadata: Data about Data

- What? Parameter, unit
- Who? Author(s), PI, Institute(s), Articles
- Where? Latitude, longitude, depth in ice/snow/water/soil, altitude, elevation, distance along a profile, ...
- When? Date/time, age, ...
- How? method

Resources

- FAIR data principles
- DFG Guidelines for Safeguarding Good Research Practice
- PANGAEA

What is data publication?

Why making a distinction between data publication and a manuscript?

- Authorship and acknowledgement:
  - Data authors can be, but doesn't have to be equal to the paper authors
  - Acknowledging contributions of scientists, technicians, students, who generated the data, but did not contribute to the interpretation or manusc
  - Authors of datasets: those who contributed to collection and processing of data
  - Follow general rules of good scientific practice
- FAIR data principles
  - Findability, Accessibility, Interoperability and Reusability: these can be guaranteed by certified data repositories
  - Data publishers / repositories: focus on metatadata

What isn't data publication?

- Data publication isn't adding a data table as a supplement to a published scientific paper. Often these are in the form of xlsx file or table in pdf. If the not open access, the supplement isn't open access either. The dataset isn't citable!
- Data publication isn't sharing data at MOSAiC Central Storage (MCS). MCS is not a long term storage, or an archive.

What is data publication?

- Data publication is a published data set or data collection equipped with a complete set of metadata. It is fully citable by having a title, authors, abst persistent identifier (usually DOI). It can have but need not have a reference to a scientific paper publication.

How to cite a data set correctly?

- Correct citation: Authors (YYYY) Title. PANGAEA, DOI. (not only DOI)
- Manuscript's Data availability statement example: "Data for this study were published open access (Authors, YYYY)." followed by a data set full cita List of references.
- Example of a full citation of a data set published in PANGAEA: *Timofeeva, Anna; Smolyanitsky, Vasily; Bessonov, Vladimir; Petrovskiy, Tomash (20 Special sea ice observations aboard Akademik Fedorov MOSAiC leg 1, 2019-09-25 to 2019-10-20. PANGAEA, https://doi.org/10.1594/PANGAEA.9*
- At PANGAEA web interface, each published data set contains buttons in the header to copy citations or export the citations in the preferred format.

Where do I publish "my" MOSAiC data?

- The default repository for MOSAiC is PANGAEA.
- My national funding agency requires depositing data in a special national repository. What should I do? These cases are handled as exceptions (see Data Policy) and a legitimate reason for not publishing the data in PANGAEA. At the moment, written agreements have been signed with several repositories:
  - Arctic Data Center (ADC) - see information for data submission,
  - Atmospheric Radiation Measurement (ARM) data center,
  - British Oceanographic Data Centre (BODC),
  - UK Polar Data Centre and
  - Centre for Environmental Data Analysis (CEDA).

- When archiving data in these repositories, it is always important to acknowledge the MOSAiC project (see Data Policy). The agreements assure FAIR data publication and future findability of MOSAiC data from a single access point (portal to project and data).
- Other exceptions are possible for special data types (e.g., genomics, source code, high volume model data), for which PANGAEA is not a suitable repository and a dedicated community repository exists which fullfills the FAIR criteria. When archiving data in these repositories, it is always important to acknowledge the MOSAiC project (see Data Policy). Otherwise, findability of MOSAiC data from a single access point (portal) cannot be assured in the future. If you are unsure if this applies, contact the PANGAEA team.
- PANGAEA does not assign or link own DOIs to data sets published elsewhere.

Resources

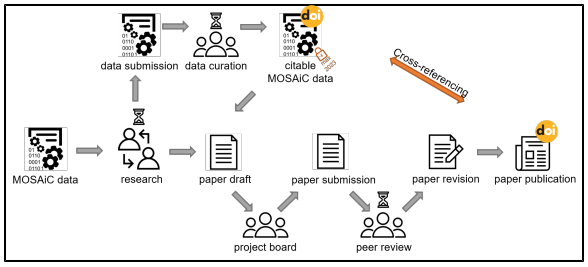
- MOSAiC datasets collection at Marine Data Portal
- PANGAEA
- Arctic Data Center (ADC) data submission information and published data at DataONE portal
- Atmospheric Radiation Measurement (ARM) data center
- British Oceanographic Data Centre (BODC)
- UK Polar Data Centre
- Centre for Environmental Data Analysis (CEDA)

When do I submit and publish “my” MOSAiC data?

- **Submit** your quality controlled data sets (primary data) **as early as possible and before they are used for a paper**. Including reference of data set already at the paper submission stages enables considering the data for peer review, which for some journals is compulsory.
- Like paper publication, data publication involves editorial work, which requires time (sometime up to several weeks). Do not wait with the submission of data for publication for the last minute. No data citation will be possible before the data are actually archived in the repository.
- In PANGAEA, during "in review" status, the data can be **password protected**. This means only metadata are accessible, but data itself cannot be viewed or downloaded by other users (except of authors with an associated PANGAEA account). The data set can also be open access at this stage.
- PANGAEA editors can already provide a **temporary access key for reviewers** or colleagues.
- While data set status is “in review”, the content can still be changed, but the citation including the future DOI is provided already for use in your manuscript.
- After final publishing, followed by the DOI registration, no changes to the data sets are possible. New version need to be archived instead and linked to the previous version.
- Even published MOSAiC data sets can remain under password protection until the publication of the associated paper or until the end of MOSAiC **moratorium in January 2023**. These data sets are only accessible to the authors of the data sets with an associated PANGAEA account. Access rights to password protected MOSAiC data can be additionally granted to all MOSAiC members, who signed the data policy. These users need to have a PANGAEA account and can be added to a MOSAiC user group upon their request. To request access, please write an e-mail to info@pangaea.de.
- Cite your data sets in any paper which is using them. Remember, data sets have full citations which can be used just like any other references (see above).

Data publishing workflow for primary (quality controlled) data

The workflow drafted below includes publishing data sets in a data repository before the paper submission. Like that, the paper reviewers have access to the underlying data and can fully evaluate the conclusions of the study. This also enables correct cross-referencing of both data and manuscript.



group, peoples icon made by icon king from www.freeicons.io; password icon made by icon king from www.freeicons.io; document, content, article, letter, paper icon made by BECRIS from www.freeicons.io; edit, document, note, writing, review icon made by BECRIS from www.freeicons.io; engagement, customer, user, interaction, branding icon made by BECRIS from www.freeicons.io; essentials, sand, clock, time icon made by byandriy matviychuk from www.freeicons.io; Data Icon #135890 from https://icon-library.com

A digital object identifier (DOI) is a persistent identifier used to identify objects uniquely

Resources

- PANGAEA

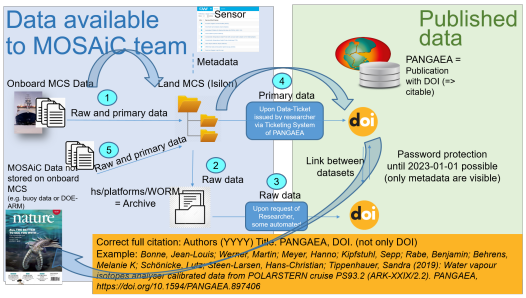
How do I approach data publication?

Raw data / Primary data publication workflow

- Raw data publishing will be a semi-automatic process and the responsible PIs should provide necessary metadata and initiate the data archivation from MCS to PANGAEA. The raw data publication can be initiated once the sensor description and the sensor raw data description had been completed by the PIs.
- Primary data publication (calibrated data, data ready made for a paper publication) needs to be always initiated by the authors by opening a data submission ticket in PANGAEA or other designated repository, if exceptions apply.
- If the raw data wasn’t published with PANGAEA at the time of primary data publication yet, and is needed, contact the PANGAEA team for initiating the raw data publication.
- During data publication instruct the editors in your data repository to create links to other versions of data (e.g., raw data), especially when they were or are being published in another repository.
- All published data must include a funding acknowledgment of MOSAiC in the following form: "Multidisciplinary drifting Observatory for the Study of the Arctic Climate (MOSAiC)" with the tag "MOSAiC20192020". Additionally, the Project ID given for specific expedition must be mentioned. For the Polarstern expedition this is "AWI_PS122_00". Additional attributions like specific award /grant numbers might be added.

Resources

- Data submission to PANGAEA
- Data submission (PANGAEA documentation)
- Event lists at PANGAEA
- SENSOR information system
- Raw data publication



MOSAiC device operation ID / Event

- The MOSAiC **Device operation ID**(s) registered by D-ship action log called **Events** in PANGAEA (further only Events) are important identifiers, that link the data with metadata stored in SENSOR. Therefore, each data set submitted to PANGAEA needs to be equipped with the by relevant Events.
- The Events list is available after the end of each leg from PANGAEA page https://www.pangaea.de/expeditions/byproject/MOSAiC. It can be found for viewing or download under "*Event list:* " links.
- Errors cannot be corrected in D-ship log, but can be corrected in PANGAEA Event list & SENSOR records
- Also information, that wasn't available at the time of D-ship log finalizing at the end of each leg, e. g. end position and Date/time of a buoy record, can be added using this procedure.
- Follow instructions for submitting an event correction sheet to PANGAEA either as a separate submission or in connection to related data.

Primary data submission to PANGAEA

Publication of primary data sets in PANGAEA or other recommended repositories is the responsibility of each scientist (MOSAiC data policy).

The data can be submitted via https://www.pangaea.de/submit/. Sign in with your user name. If you are not a PANGAEA user yet you need to register with your name and E-Mail address, or ideally your ORCID ID.

This login can later also be used to access your password protected dataset during the review of your data publication. After signing up, you will receive a link via email to activate your account and sign in.

Once logged in, click on “Submit data” and a data submission form will open.

Metadata must be submitted together with the data. Minimal requirements are:

- dataset Author(s), PI (each parameter can have a separate PI)
- dataset title: should reflect what has been measured, observed or calculated, when, where and how
- MOSAiC device operation ID(s) / Event labels associated with individual data / data files
- related institute(s) or publication(s)
- description of your data (dataset abstract)

Finally you need to attach your dataset files. If your data files are larger than 100 MB we can provide an upload link for large files up to 10 GB per file. Please indicate this in the Description-field.

Any documentation (e.g., MOSAiC Standard operating procedures, MSOPs) helping to understand the data can and should be linked to the dataset(s). This can be done in the form of text or pdf-file, or hdl (link to Epic). If no persistent link to the documents can be provided, PANGAEA can archive the documents permanently alongside the data.

The granularity of the data is up to the author(s) of the dataset. Lower-granularity datasets can be combined in a (time-)series collection dataset as in https://doi.org/10.1594/PANGAEA.873032.

During submission (https://www.pangaea.de/submit/), the connection with MOSAiC has to be clearly stated in the Label Field of the Data Submission ("**MOSAiC**"). The MOSAiC Project ID (for the Polarstern expedition this is "AWI_PS122_00") is internally associated as a grant number of the MOSAiC project and does not have to be inserted in the submission form additionally.

Once all necessary information is entered and data files uploaded, the submission is finalized by pressing the “Create” button. Do not worry if you are uncertain about some fields or content, you will still be able to make final adjustments during the following steps and with support of your data curator. First the submission passes a brief editorial review to make sure the data submission is complete. If questions remain, we will get back to you. Then the submission will be assigned to a data curator who will lead you through the further process.

Good to know

MOSAiC Device operation ID (registered in D-ship Action log) **= station Label** (in SENSOR) **= Event** (in PANGAEA)

The list of MOSAiC Events can be found at https://pangaea.de/expeditions/byproject/MOSAiC.

ORCID: Open Researcher and Contributor ID is a nonproprietary alphanumeric code to uniquely identify scientific and other academic authors and contributors.

After we've imported your data to PANGAEA, you'll be asked to proofread your dataset, which is then "in review". We lead you through this iterative process with our ticket system until the data submission is complete and approved by you. Once the dataset is approved, the digital object identifier, DOI, is registered and with that the dataset is officially published and citable.

If the related manuscript has not been published yet, a moratorium on access and publication can be put in place. At the same time, PANGAEA can provide a temporary key to enable access to your datasets for example for anonymous reviewers. In general, moratorium on MOSAiC data is possible until 2023-01-01.

If a published dataset needs to be updated, PANGAEA will upload a new version of this dataset, with new documentation and complete metadata (clearly providing information on the changes between the versions). Both versions can be linked but will have their own DOIs.

Datasets in PANGAEA may be archived as stand-alone publications of data (e.g., https://doi.org/10.1594/PANGAEA.753658) or as supplements to an article (e.g., https://doi.org/10.1594/PANGAEA.846130).

Data submission to PANGAEA

Within the data table, parameters (table header) should be submitted with full names and units. Data submitted in the form of videos, photos, geoTIFF, shape files, netCDF, sgy, etc. will be archived as is (e.g., https://doi.org/10.1594/PANGAEA.865445).

More information on data submission can be found in https://wiki.pangaea.de/wiki/Data_submission.

Preparing tabular data submission to PANGAEA

Data submitted as TAB-delimited text data files or in excel-format

- **Georeference is mandatory** (latitude/longitude in decimal degree): for all samples, observations and measurements made somewhere on Earth
- Third dimension: water depth, altitude, depth in ice, …
- Date/Time: ISO-format (e.g. 2020-04-07T13:34:11)
- For each observation provide **Event (Device operation ID)** in the first column
- **Parameters** are always accompanied by a **unit**
- Abbreviations should be explained
- A separate metadata table can be added, with short name / long name / PI / method / comment for each parameter

Keep in mind when submitting in excel-format:

- Remove **formulas**
- Use the correct number of **decimals** you want to use (we cannot add any after import) – one data series has fixed number of decimals
- Use only one decimal separator (don't mix "." and ",")
- Don't use colors, fonts, comments
- No value? Keep cell empty. No "NaN", "NA", "-", " ", ...
- Don't combine cells
- One value – one cell (no ranges "-", or errors / uncertainties "±": these need to be entered in separate columns
- One table – one sheet (file) (no multiple tables with a different structure in a single sheet)
- Other excel pains: date formats, character encoding (UTF-8), hidden columns, cell comments ... - try to spare us

See also: https://wiki.pangaea.de/wiki/Data_submission for additional information

Preparing binary files submission to PANGAEA

Binary files with specific formats (e.g. shape files, netCDF, segy, images, videos, ...) will be archived as links to files. For >20 files or >100 MB: ask us upload link in the ticket.

Please prepare a file list including:

- Event / Device operation ID
- Latitude, Longitude
- Date/Time
- Data description (readme file)
- File names should not contain spaces and special symbols

How do I share and publish my code/software related to the data?

- All MOSAiC members are strongly encouraged to publish their scripts and codes for data processing along with their data. It is a proper way to enhance data provenance and foster FAIR principles of data management.
- Most important is a unique handle or DOI for your code to link and cite it in your analyses or published datasets. Software can be published for example at crossref.org, figshare.com, or zenodo.org. MOSAiC promotes the latter one, since it is hosted at and operated by CERN (yes, the one with the Large Hadron Collider), part of OpenAIRE, free and open to all, and offers usage statistics. If you work with git, the zenodo.org version feature can be easily fed by your Github account, so every release of your software can get a separate DOI. It still can be integrated manually, for the case you are using a different flavor of git (e.g. Gitlab, Bitbucket, ...).
- For NSF-funded MOSAiC members the Arctic Data Center provides a special page dealing with MOSAiC related topics. There is also a special section in their data submission guidelines regarding software.

Resources

- Arctic Data Center (ADC) data submission information
- zenodo.org