MOSAiC Data Policy

19.09.2019

The Multidisciplinary drifting Observatory for the Study of Arctic Climate (MOSAiC) is a collaborative, international project to address pressing scientific questions in the central Arctic. The project's success, and its ultimate impact on science and society, relies upon professional coordination and data sharing across the participants. A transparent Data Policy is essential to achieve MOSAiC science objectives, to facilitate collaboration, and to enable broad use and impact of the MOSAiC data legacy.

Executive Summary

This Data Policy regulates data management, access and release as well as authorship and acknowledgment. Signing this Data Policy is a pre-requisite for participation in MOSAiC field operations and being a member of the MOSAiC consortium.

Metadata Standards (for details see section 3)

Metadata shall make data findable and provide additional contextual information about measurement details, methods, relevance, lineage, quality, usage and access restrictions of the data. It shall allow coupling users, software and computing resources to the data. Hence, metadata must be machine-readable and interpretable as well as human understandable. Furthermore, metadata for each data set should follow the FAIR data principles in terms of fitness for purpose and fitness for re-use.

Data Ingest, Transfer, Storage and Archiving (for details see sections 5 and 6)

The MOSAiC Central Storage (MCS) aboard Polarstern is the basis for gathering data during the year of operation, offering near-real-time access and early processing of the data to the users underway. The land MCS provided by AWI is the central and reliable storage and working database of MOSAiC data within the AWI storage platforms.

Only MOSAiC consortium members with authentication/authorization will have access to the data prior to public release.

PANGAEA is the primary long-term archive for the MOSAiC data set and all primary data, with the exception of the subsequently mentioned cases, must be submitted to the PANGAEA data base for long-term archival. If this is not feasible due to the size of the data set or is not possible due to institutional data policies or commitments to other stakeholders, exceptions can be made if the data are stored in another long-term archive that provides unique and stable identifiers for the datasets and allows open online access to the data. These exceptions need to be documented in written agreements between the data provider and the MOSAiC Project Board and data manager.

Data Provision, Access and Sharing (for details see section 7)

Early access by the members of the MOSAiC consortium to the data is crucial for the successful collaboration within the consortium. Hence, all data must be made available to the consortium by the MCS as fast as possible. The following deadlines mark the latest points in time for transferring data to the MCS:

- All sensor data: Must be stored in the onboard MCS as fast as technically possible. Data that cannot be stored immediately in the on-board MCS have to be added as soon as possible or stored in the land MCS no later than 31 Jan 2021. Buoy data can be updated within one month

after the lifetime of the buoy if data are being collected beyond the end of the MOSAiC expedition.

- All fast analysis sample data: Must be stored on the land MCS no later than 31 Jan 2021.
- A primary subset of laboratory sample analysis data: Must be stored on the land MCS no later than 31 Jul 2021.
- Full collection of laboratory sample analysis data: Must be stored on the land MCS latest no later than 31 Jan 2022.

All MOSAiC raw and primary data are freely available to all MOSAiC consortium members as soon as they are stored in the on-board MCS or the land MCS.

For using data from the MCS for publications, the ***data provider*** **or** ***data PI*** **must be informed and offered collaboration** on the scientific analysis and must be offered co-authorship based on the principles described in section "Authorship and Acknowledgment" below. The *data provider* and/or *data PI* may object to the usage of data in a publication if that publication conflicts with his or her own publication strategy. Any such objection must be discussed and agreed upon in writing with the MOSAiC coordinator and data manager. The *data provider* and/or *data PI* may not object to the usage of data beyond the public release date.

Public Release of Data (for details see section 8).

MOSAiC data will be freely and publicly available on the open MCS or PANGAEA and/or alternate public archives on **1 Jan 2023**. From this date on there are no restrictions on data usage, but data users are encouraged to communicate with *data providers* or *data PIs* during early stages of all scientific analyses to ensure accurate usage and interpretation of data. The best practices on co-authorships described in the section "Authorship and Acknowledgment" below continue to apply.

Authorship and Acknowledgment (for details see section 9)

Generally, **co-authorship** on publications and other public documentation must be offered to those that have **made a substantial contribution** following the principles of good scientific practice. An inclusive co-authorship approach is encouraged.

Accordingly, co-authorship on publications and other public documentation must generally be offered to those that have made a substantial contribution to a) the intellectual conception or design of research; b) the acquisition, analysis, or interpretation of the data (i.e., including the *data provider* or *data PI*), or c) the drafting or significant revision of the work.

Lead authors have the ultimate decision authority and responsibility to identify and appropriately engage co-authors.

Contributors to the work that do not warrant co-authorship should be identified by name in the acknowledgments.

MOSAiC data must be acknowledged or referenced in publications and other public documentation, specifically including relevant digital object identifiers, data providers (if not co-authors), and funding agencies.

All publications and other public documentation using MOSAiC data must include a funding acknowledgment of MOSAiC in general in the following form:

"Data used in this manuscript was produced as part of the international 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.

Data Publication (for details see section 10)

The publication of MOSAiC data via data journals and data archives is strongly encouraged and will be facilitated by the MOSAiC Project Board and Data Group. The MOSAiC Project Board will centrally organize one or more special issues in a data journal, with an appropriate period for submission. These special issues will allow for linking all MOSAiC data sets and help to make data standards and procedures easily citable.

Responsibilities

Data Group Speaker

Stephan Frickenhaus

Data Manager (primary contact)

Antonia Immerz (Antonia.Immerz@awi.de)

Data Group

Atmosphere: Peter von der Gathen, Matthew Shupe (CU/NOAA), Sara Morris (CU/NOAA)

Ice/Snow: Marcel Nicolaus, Martin Schneebeli (WSL-SLF), Julia Regnery

Eco, Bio-Sampling: Allison Fong, Pauline Snoeijs-Leijonmalm (Se)

BGC: Walter Geibert

Ocean: Ben Rabe, Julia Regnery

Airborne: Andreas Herber

Remote sensing: Thomas Krumpen, Suman Singha (DLR)

Modeling: Ralf Jaiser

PANGAEA & data publishing: Daniela Ransby, Stefanie Schumacher, Amelie Driemel (info@pangaea.de)

Infrastructure Experts: Peter Gerchow, Angela Schäfer, Ingo Schewe, Mohammad Ajjan

Head of Data at AWI: Frank Oliver Glöckner

Head of Systems at AWI: Christian Schäfer-Neth

NSF Arctic Data Centre: Christopher Jones, Jesse Goldstein, Matt Jones

ARM: Giri Prakash

1. Objective

The purpose of this Data Policy is to codify the goals and principles of MOSAiC's research data life-cycle from production, documentation, sharing, usage and re-usage. This ensures that common procedures for data gathering, archiving and publication, as well as metadata and quality management are commonly implemented. By participating in the MOSAiC project, all members of the MOSAiC consortium agree to and comply with this Data Policy. By doing so, participants ensure that MOSAiC is a successful and resource-effective research project that also supports data accessibility, interoperability and re-usage following the FAIR data principles.

This policy aims to:

1. Ensure proper storage, backup and archiving of MOSAiC data in a central system.
2. Promote the visibility and accessibility of MOSAiC data for scientific and other applications.
3. Ensure the fair and equitable use of MOSAiC data and uphold the rights of individual scientists and institutions.

4. Enable the organized and timely analysis of the data.
5. Encourage the rapid publication and dissemination of scientific data, results and knowledge, to support the involvement of a broad user community.

2. Definitions

- **MOSAiC data**: Data collected aboard Polarstern, within the Central Floe Observatory, within the distributed network, and aboard Polar 5/6. This includes data from analyzed sample material and sample metadata and satellite data products.
- **Collaborating data**: Relevant data outside of MOSAiC data, brought to the MOSAiC consortium via the endorsement process (external aircraft data, re-supply vessel data, other coordinated activities). As defined by the endorsement, these data from collaborating partners are subject to the MOSAiC Data Policy.
- **External data**: Relevant data outside of the MOSAiC data and Collaborating data, but still of interest to the MOSAiC consortium and other users of MOSAiC data, including but not limited to operational model output, operational observations at other locations, etc. These data may be archived or cross-linked along with MOSAiC data at the discretion of the data provider but are not subject to the Data Policy and the provider is not entitled to the benefits of endorsement.
- **Data provider/PI**: All data streams must have a responsible party. The data provider is defined as the PI or institution that owns and/or operates an instrument, creates and analyzes samples, produces a model output, or otherwise produces a data set.
- **Consortium members**: Participants whose scientific activities are officially endorsed by the MOSAiC Science Board. Such participants are bound to the MOSAiC Data Policy and will have access to MOSAiC data as soon as they arrive at the MOSAiC Central Storage (MCS).
- **Public users**: Public users are those that use MOSAiC data or Collaborating data but are not part of the MOSAiC consortium.
- **Raw data**: Data directly produced by sensors, devices, or manual observation, prior to additional processing, calibration and quality assessment/control (never modified).
- **Primary data**: Processed data that modify a copy of the raw data, e.g., outliers removed, calibrated, quality controlled.
- **Value-added data/derived data product**: Products based on raw or primary data that may involve derivation of additional parameters or delayed-mode quality control using external data or post-use sensor calibration; model data or a combination with any external data, e.g., by data assimilation, visualization, classification, or clustering.
- **MOSAiC Central Storage (MCS)**: Connected central storage infrastructure that allows for the redistribution of data to consortium data users with authentication and authorization. Part of the MCS is aboard Polarstern for gathering and securing raw and/or primary data.
- **MOSAiC Standard operating procedures (MSOPs)**: MOSAiC teams specify procedures on how to handle devices, how to store samples, and how to process data. MSOPs are temporarily stored in the MCS. MSOPs document how data are processed from raw to primary and/or value-added data. They need to be published at the time the data are published in an open access format. When revised, MSOPs are subject to version control. MSOPs become, like data, open access and citable.
- **MOSAiC sensor and device registration**: Sensors and sampling devices are registered and managed centrally using the SensorWeb interface provided by AWI. The sensor registration is mandatory for controlling data streams through MCS and serve to augment data with metadata automatically. The combination of sensor registration and MSOPs will facilitate a high standard of quality management and documentation for referencing in publications.
- **MOSAiC Device ID (MDID)**: All sensors/instruments in MOSAiC have a unique ID and Uniform Resource Name (DeviceURN) in SensorWeb.

- **MOSAiC Sample ID (MSID)**: Physical samples or materials carrying physical or biological matter (e.g., filters) must have a unique ID.
- **MOSAiC Device Operation ID (MDOID)**: IDs registered in the Ship data system DShip, referring to coordinates and time. They can be recorded automatically, semi-automatically, or manually.

3. Metadata Standards

Metadata shall make data findable and provide additional contextual information about measurement details, methods, relevance, lineage, quality, usage and access restrictions of the data. It shall allow coupling users, software, and computing resources to the data. Hence, metadata must be machine-readable and interpretable as well as human-understandable. Furthermore, metadata for each data set should follow the FAIR data principles in terms of fitness for purpose and fitness for re-use. The metadata should be agreed on, listed, and explained within the MSOPs.

Specifically, within MOSAiC the following two general principles for providing metadata to MOSAiC datasets shall be endorsed:

- Metadata for sensors/devices must be registered in the SensorWeb. The derived DeviceURN from SensorWeb for each device should always be linked within the metadata for each data set ingested into the MCS as well as any derivate data to keep track of the available standardized meta data in SensorWeb.
- Specifically, all metadata necessary for archiving must be provided within the MCS at the moment data sets are ingested on board to ensure proper data sharing, findability, and re-usability during the expedition and later on. If this is not possible, e.g., due to technical limitations, all relevant data must be added latest until the public release date.

Recommendations for metadata and vocabularies

If further metadata are needed within the MSOPs we recommend using this collection of widely accepted metadata standards categorized by disciplines and communities to be adopted by MOSAiC sub teams.

Examples of standards are:

- **Oceanography, climatology, and modelling**
  - CF (Climate and Forecast) Metadata Conventions: The CF standard was framed as a standard for data written in netCDF format, with model-generated climate forecast data particularly in mind. However, it is equally applicable to observational datasets, and can be used to describe other formats. It is a standard for “use metadata” that aims both to distinguish quantities (such as physical description, units, and prior processing) and to locate the data in space and time.
  - ISO 19115: An internationally adopted schema for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data.
  - ISO 19115-2: Imagery and gridded data as an extension of ISO 19115 defining the schema required for describing imagery and gridded data.
- **Biology**
  - Ecological Metadata Language (EML): A metadata specification that is used to document environmental data from almost any scientific domain, and includes sections for describing spatial, temporal, thematic, and taxonomic coverage of datasets. Current release: EML 2.1.1.

- Darwin Core: A body of standards, including a glossary of terms (in other contexts these might be called properties, elements, fields, columns, attributes, or concepts) intended to facilitate the sharing of information about biological diversity by providing reference definitions, examples, and commentaries. Current Biodiversity Information Standards (TWDG) from October 2009.
- MIxS: Minimum Information about any (x) Sequence: The MIxS is a unified standard developed by the Genomic Standards Consortium (GSC) for reporting of minimum information about any (x) nucleotide sequence. It consists of MIGS, MIMS and MIMARKS standards and describes fourteen environments. MIGS, MIMS and MIMARKS share common mandatory core descriptors, differ in standard-specific elements and can be tailored to a particular environment by a subset of relevant environment-specific information components.

- **Provenance**
  - W3C Provenance Ontology (PROV-O): The PROV-O ontology provides terms that support the documentation of the lineage of activities (like data processing), used and produced resources (like data), and the agents (like scientists) associated with the activity. The DataONE ProvONE ontology extends the PROV-O ontology to explicitly capture lineage information for scientific workflows, and statements about data inputs, processing scripts, and data outputs can be expressed inside of DataONE packaging documents (OAI-ORE resource maps).

All variables and parameters (measurement attributes) must be documented with an attribute name and attribute definition that provides a human-readable context for the measurement. For numeric data, attributes must include the units of measurement using SI unit definitions. Where non-SI units are used, a mapping to SI units must be provided that includes a) a unit name, b) a unit definition, c) a unit notation abbreviation, d) the unit's parent SI unit name, e) a multiplier to the parent SI unit. For numerical data without a unit (e.g., percent, count x per count y, etc.), the unit should be noted as "dimensionless". For non-numeric, categorical data, coded values must be defined in a code/definition list, or be defined by an external, controlled vocabulary term. We recommend the **NERC Vocabulary Standard**, since registry of MOSAIC Sensors and devices via SensorWeb follows this vocabulary. The NERC Vocabulary Server (NVS) web service provides access to controlled vocabularies via an international, actively-contributing research community https://www.bodc.ac.uk/resources/vocabularies/. Any deviations from this recommendation must be individually discussed with the MOSAiC data manager. In case a specific vocabulary is agreed on, a mapping between the NERC vocabulary term and the term used in the metadata must be provided by the requesting party.

Recommendation for Processing Levels

Processing levels of all data stored in the MCS or published in PANGAEA or other certified repositories should be stated in the metadata. In general, the levels raw, primary and value-added/derived should be used (see definition above). If other conventions or standards for data levels exist these should be referenced in the metadata. Processed data in PANGAEA and other certified repositories should include the information how they have been derived from raw data (provenance). Additionally, the information how to gain access the raw data should be provided.

4. Metadata Registries

The purpose of metadata registries is to assemble provenance meta information for the discovery, quality assessment, interlinking, and assembly of otherwise disconnected data.

ActionLog – Actions are registered in the DShip system on board. Sampling, regular station visits, etc. can be recorded with an App on a specific MOSAiC tablet. The recorded logs are uploaded to DShip by the data support team aboard.

Devices registry – Sensors and sampling devices are registered in SensorWeb by the PIs with support from the data support team (on board, but mainly before expedition start). Configuration changes are registered in the same system.

5. Data Ingest and Transfer

The MCS aboard Polarstern is the basis for gathering data along the year of operation, offering near-real-time access and early processing of the data to the users underway.

The land MCS provided by AWI is the central, reliable storage and working database of MOSAiC data within the AWI storage platforms. It will furthermore serve to distribute data after the expedition, also for data publication in other repositories.

Raw data obtained during the MOSAiC expedition shall be stored in the MCS on Polarstern. Any deviations from this rule must be individually agreed upon with the data manager. The raw data are transferred to the on-board MCS semi-automatically. Additional data can be submitted manually to MCS via mobile external hard drives in 'delayed mode' by scientific cruise participants.

For the data ingest into MCS, the Raw Data Ingest Framework provided by AWI (RDIF/AWI) will be used. For this, sensor registration in SensorWeb is mandatory, as is naming a responsible person for data transfer to the MCS. A data set template is to be described for RDIF, implying a DeviceURN from SensorWeb, a filename filter as regular expression (RegEx), file format descriptions and additional metadata for PANGAEA (see annex).

The transfer of the raw data after each leg to the land MCS at AWI is organized centrally by the AWI data support team. Data transfer to the land MCS will be performed by means of mobile data storage mediums (hard disks) hereby also maintaining user rights. Data is then made accessible adhering to the specified user rights of all MOSAiC members. Furthermore, raw data transferred to the land MCS will be automatically archived in a WORM (write once, read multiple) system at AWI.

Primary data produced aboard Polarstern during the expedition can also be transferred to the land MCS at AWI via the centralized data transfer. User rights defined on the data will be maintained accordingly. Publication of primary data sets in PANGAEA or other recommended repositories is the responsibility of each scientist. Data copies will be made accessible to the participating institutes via the land MCS at AWI.

6. Data Storage and Archiving

The land MCS will store the data and metadata records during and beyond the duration of the MOSAiC project. It will serve as a working database for the early handling and exchange of data within the MOSAiC consortium. As stated in section 2, only consortium members with authentication/authorization will have access to the data until public release (see section 7 and 8).

The land MCS will be in operation and accessible until all pre-registered data from the expedition, and the associated derived and analyzed data and metadata are permanently archived and published.

PANGAEA is the primary long-term archive for the MOSAiC data set and all primary data, with the exception of the subsequently mentioned cases, must be submitted to the PANGAEA data base for long-term archival. If this is not feasible due to the size of the data set or not possible due to institutional data policies or commitments to other stakeholders, exceptions can be made if the data are stored in another long-term archive that provides unique and stable identifiers for the datasets

and allows open online access to the data. These exceptions need to be documented in written agreements between the data provider and the MOSAiC Project Board and data manager.

Metadata of primary data sets published in PANGAEA are provided in a machine-readable format via the website of PANGAEA and are harvestable. The completeness of the metadata is the responsibility of the data PI. This option to harvest the meta data enhances the global visibility of MOSAiC data.

In PANGAEA, data files are archived together with metadata. Its content is distributed via web services to portals, search engines, and catalogs of libraries and publishers. Each data set includes a bibliographic citation and it is persistently identified using a Digital Object Identifier (DOI). Interlinkage of MOSAiC IDs (links to, e.g., SensorWeb, sample IDs, Device IDs, Grant IDs) is possible and allows the clear identification of data, samples, methods and associated data flows. For a more detailed sketch of PANGAEA workflows and options see the annex.

Datasets stored in other well-established, long-term archives, e.g., due to requirements by national funding bodies, should nevertheless be reported to the data manager and PANGAEA to ensure long-term, robust linkage with and documentation of all data that are stored externally to PANGAEA.

Molecular data (DNA and RNA data) must be archived within one of the repositories of the International Nucleotide Sequence Data Collaboration (INSDC, www.insdc.org) comprising of EMBL-EBI/ENA, GenBank and DDBJ).

In any case, each data set must have a clearly identified primary archive. Any exceptions from the rules stated here need to be agreed on between the data provider and the MOSAiC Project Board and data manager.

7. Data Provision and Sharing among the MOSAiC Consortium Members

Early access by the members of the MOSAiC consortium to the data is crucial for the successful collaboration within the consortium. Hence, all data must be made available to the consortium by the MCS as fast as possible. The following deadlines mark the latest points in time for transferring data to the MCS:

- All sensor data: Must be stored in the onboard MCS as fast as technically possible. Data that cannot be stored immediately in the on-board MCS have to be added as soon as possible or stored in the land MCS no later than 31 Jan 2021. Buoy data can be updated within one month after the lifetime of the buoy if data are being collected beyond the end of the MOSAiC expedition.
- All fast analysis sample data: Must be stored on the land MCS no later than 31 Jan 2021.
- A primary subset of laboratory sample analysis data: Must be stored on the land MCS no later than 31 Jul 2021.
- Full collection of laboratory sample analysis data: Must be stored on the land MCS latest no later than 31 Jan 2022.

All MOSAiC raw and primary data are freely available to all MOSAiC consortium members as soon as they are stored in the on-board MCS or the land MCS.

For using data from the MCS for publications, the ***data provider*** **or** ***data PI*** **must be informed and offered collaboration** on the scientific analysis and must be offered co-authorship based on the principles described in section “Authorship and Acknowledgment” below. The *data provider* and/or *data PI* may object to the usage of data in a publication if that publication conflicts with his or her own publication strategy. Any such objection must be discussed and agreed upon in writing with the MOSAiC coordinator and data manager. The *data provider* and/or *data PI* may not object to the usage of data beyond the public release date.

8. Public Release of MOSAiC Data

Good progress of a highly collaborative and interdisciplinary project like MOSAiC requires open availability of data to a wide user audience as early as possible. At the same time, it is important to acknowledge the substantial work that goes into collecting, quality controlling, formatting, documenting, and releasing scientific data. MOSAiC policies pertaining to data use and acknowledgment aim to balance these two principles.

Data access and usage policies evolve in time according to a staged process outlined here, and in all cases the most data-restrictive approach is described while an accelerated publication of data is acceptable.

MOSAiC data will be **freely and publicly available** on the open MCS or PANAGEA and/or alternate public archives on **1 Jan 2023**. From this date on there are no restrictions on data usage, but data users are encouraged to communicate with *data providers* or *data PIs* during early stages of all scientific analyses to ensure accurate usage and interpretation of data. The best practices on co-authorships described in section 9 “Authorship and Acknowledgment” continue to apply.

9. Authorship and Acknowledgment

Authorship. Generally, **co-authorship** on publications and other public documentation must be offered to those that have **made a substantial contribution** following the principles of good scientific practice. An inclusive co-authorship approach is encouraged.

Accordingly, co-authorship on publications and other public documentation must generally be offered to those that have made a substantial contribution to: a) the intellectual conception or design of research, b) the acquisition, analysis, or interpretation of the data (i.e., including the *data provider* or *data PI*), or c) the drafting or significant revision of the work. Co-authors should understand the content of the work, be accountable for at least a section of the work and approve of the final draft. Additional standard guidelines for deciding on co-authorship on publications can be found via numerous on-line resources, such as http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html or https://www.dfg.de/sites/flipbook/gwp/files/assets/basic-html/page85.html.

Lead authors have the ultimate decision authority and responsibility to identify and appropriately engage co-authors.

Contributors to the work that do not warrant co-authorship should be identified by name in the acknowledgments.

Authorship conflicts may be resolved by the MOSAiC Project Board, possibly taking into consideration advice from further experts in the research field.

Acknowledging data usage. MOSAiC data **must be acknowledged or referenced in publications and other public documentation**, specifically including relevant digital object identifiers (DOI, see Section 7), data providers (if not co-authors), and funding agencies. A data acknowledgment or reference should also specify where the data was obtained, according to individual journal policies. A suggested format for acknowledging each data stream includes:

"*[Data descriptor]* data (*[Author name et al. (PubYear)]*) was provided by *[data provider, PI, and or Institution]* with support from *[Funding agency or institution]*.

The data has then to be cited in the References, e.g., as follows:

"Nicolaus, Marcel (2018): Shipborne visual observations of Arctic sea ice during POLARSTERN cruise PS106. PANGAEA, doi:10.1594/PANGAEA.889264, In: Hutchings, Jennifer K (2018): Shipborne visual observations of Arctic sea ice. PANGAEA, doi:10.1594/PANGAEA.889209."

Acknowledging MOSAiC in general. All publications and other public documentation using MOSAiC data must include a funding acknowledgment of MOSAiC in general in the following form:

"Data used in this manuscript was produced as part of the international 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.

Citing Research Platforms. All scientific and data publications must cite the article concerning the respective research platform:

"Polarstern: Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung. (2017). Polar Research and Supply Vessel POLARSTERN Operated by the Alfred-Wegener-Institute. Journal of large-scale research facilities, 3, A119. http://dx.doi.org/10.17815/jlsrf-3-163"

"Polar5 and Polar6: Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung. (2016). Polar aircraft Polar5 and Polar6 operated by the Alfred Wegener Institute. Journal of large-scale research facilities, 2, A87. http://dx.doi.org/10.17815/jlsrf-2-153"

10. Data Publication

Clear, consistent documentation of MOSAiC data will help to support a strong and lasting MOSAiC data legacy, promote the broad and appropriate use of MOSAiC data including the citation of data, and ensure proper acknowledgment of data creators. This documentation is particularly important for a large, inter-disciplinary, and international project like MOSAiC, which involves many disparate sources and providers of data. To this end, the publication of MOSAiC data via data journals and data archives is strongly encouraged and will be facilitated by the MOSAiC Project Board and Data Group.

- Data publication can take multiple forms such as data journals or data/metadata archives (potentially certified by WDS/CoreTrust). Data publications follow the FAIR data principles. The ultimate goals for data publication are to provide a clear description of the metadata and data, the specific instruments and measurements that created the data, the quality control procedures, the manner in which the data were processed, any embedded data dependencies (on other data sets), and any other special conditions or considerations for the data. To assist in data tracking and awarding of credit, it is important that data sets are given a digital object identifier (DOI). Additionally, associated data files, metadata description documents, and processing scripts and instruments should receive a persistent identifier (PID), which links to the datasets.
- Authorship on data publications should follow similar policies to authorship on scientific publications and must include those participants that have made substantial contributions to collecting the data, processing the data, and documenting the data (see Section 9). Each data publication needs a contact person and principle investigator (PI) who is familiar with and responsible for the scientific evaluation. This is especially relevant for "automated" measurements, where often the cruise scientist is chosen as PI, but was not involved in the data evaluation.
- The MOSAiC Project Board will centrally organize one or more special issues in a data journal, with an appropriate period for submission. These special issues will allow for linking MOSAiC data sets and help to make data standards and procedures easily citable. Each special issue will likely have an introductory manuscript that provides the context for the rest of the special issue. When organizing the special issues, the coordinator will specify a short list of recommendations for the information that should be specifically included in data publications. This process might involve specific MOSAiC formatting that will support consistency across the different publications.

- External Data: When used in a publication in the MOSAiC context, i.e., in combination with MOSAiC data, external data should be published in an appropriate open access data repository that also provides DOIs or at least persistently resolvable IDs.
- Synthesis Data: MOSAiC data may serve as a basis for synthesis data products, i.e., data from MOSAiC in combination with already published data or model data. Synthesis data should be published in the same manner as MOSAiC data. PIs working on synthesis data and related publications are encouraged to ensure that data from other sources becoming part of synthesis data are published.

11. Amendments

Variances
Any modifications to this policy that are needed on a case-by-case basis, i.e., conflicting requirements from a funding agency, must be endorsed by the MOSAiC Project Board.

Dispute resolution
Disputes on the Data Policy should be solved primarily by the involved individuals or MOSAiC team leaders. If resolution at this level is not possible the MOSAiC Project Coordination will act as a mediator in the conflict. If resolution cannot be achieved with the mediation of the Project Coordination, the MOSAiC Project Board will be engaged to resolve the dispute.
In case, the MOSAiC Project Board is not able to resolve the dispute amicably it will be referred to the competent German state court. German law under exclusion of its conflict of law regulation and under exclusion of the Convention on the International Sale of Goods (CISG) will be applicable.

MOSAiC Consortium
The term “MOSAiC Consortium” does not refer to a legal entity or institution. MOSAiC Consortium defines a scientific collaboration of many persons contributing scientific work to the project. Consequently, the term “Official Member” refers to the fact that the person signing the data policy will respect the Consortiums Data Policy and that he/she is registered for book keeping on a formal basis, and for realizing the technical basis of data sharing.

Signature

Name	
Institute	
e-Mail	

Hereby I declare that I fully consent to the MOSAiC Data Policy and become a registered MOSAiC Consortium Member.

Date, Signature

12. Annex

Requirements for MOSAiC Sample IDs (MSID)

Physical samples or materials carrying physical or biological matter (e.g., filters) must have a unique ID. Also, certain measurements and data products, such as photographs for instance must obtain a unique ID.

Creation of unique sample IDs is to be managed within the scientific teams.

The association with the device and its operation in which the sample was obtained must be documented. Therefore, the respective DeviceURN and DeviceOperation ID must always be related to a sample ID. This is achieved by annotating sampling log sheets enlisting sample-IDs with the DeviceURNs from SensorWeb of the involved devices and the DSHIP-DeviceOperation IDs in which the device was deployed. Storing the sampling log sheets in the respective directory of the MCS which reflects this structure exactly makes the metadata clear to the data user.

PANGAEA - sketch of workflows/options and metadata

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 can be submitted to and published in PANGAEA with access restrictions in place for a predefined period (until article publication, or during an embargo period). Metadata must be submitted together with the data (minimal requirements are dataset Author(s), PI, dataset title, MOSAiC ID(s), related institute(s) or publication(s)). Any documentation (e.g., MOSAiC Standard operating procedures, MSOPs) helping to understand the data can and should be linked to the dataset(s). 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. The MOSAiC Project ID (see Acknowledging MOSAiC in general, section 9) must be given in the Data Submission description. The MOSAiC Device ID(s) should also be provided. 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.

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 permanent DOI.

MOSAiC Grant IDs

MOSAiC grant-IDs are provided centrally by the MOSAiC science board via the MOSAiC Project Board. Grant-IDs are parse-able for analyzing citations within the Acknowledgments in papers referring to MOSAiC, see Acknowledging MOSAiC in general, section 9. Additional grant IDs from funding agencies might exist.