



Policy

for the data library

PANGAEA® - Publishing Network for Geoscientific & Environmental Data

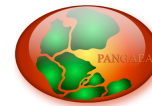
Open Access data archive and publishing system for earth system research

Alfred Wegener Institute for Polar and Marine Research (AWI), Bremerhaven
Center for Marine Environmental Sciences (MARUM), Bremen, Germany

The aim of this policy is to facilitate operation and use of the data library PANGAEA - Publishing Network for Geoscientific & Environmental Data by the research community. This policy recognises the benefits of providing free and open access to good quality data from earth and environmental sciences for future use in global change studies, research projects, and operational services such as portals, search engines and library catalogs. The operating institutes encourage the widest possible use of the Pangaea library, in order to best realise its potential value.

Principles

- The guiding principle of PANGAEA is Open Access to its content by research and education communities. This is in line with data policies of the IOC, the WDC System and the OECD.
- For any data, provided by Pangaea, the format and content of a data set must ensure its most widespread and easiest use by the scientific community.
- Pangaea is open to the community for data archiving. Any project and principle investigator (PI) is encouraged to integrate relevant data to the Pangaea library.
- Users of data from Pangaea are urged to properly use the data set citation or quote the related reference for supplements.



Data provision

- Data archiving includes:
 - 1 Metadata(*) of expeditions, stations and samples;
 - 2 Primary (factual) data from a) archives, b) expeditions, c) publications (supplements);
 - 3 Metadata related to the primary data of 2;
 - 4 Products resulting from compilations and interpretations of primary data.
- Metadata of expeditions/stations/samples (events) are submitted to the project management office. Labels must remain the same at any time when used in data submissions or publications.
- The data librarian maintains a dictionary of parameter definitions with unit, to be used as the agreed standard for all project data. Parameters are grouped into categories according to their related scientific field. Data submissions are required to use parameters and units as defined in the dictionary. New parameters are defined by the data librarian on request.
- Data are archived in a relational database, georeferenced in space and time; if a data set is very large or must have a proprietary format, it is archived as an *object* in a file system with a metadescription only, linked to the file.
- As soon as (validated) data become available, the PI is urged to submit the data in agreement with the import format. Any type of data must always be accompanied by a description (metadata) allowing future users to understand and process the data.
- The granularity and format of data sets is defined in agreement with the PI. The export format in principle is tab-delimited ASCII, headed by metadata fields according to international standards.

Quality assurance

- Data submitted for archiving have to be documented properly; documentation is archived together with each dataset.
- The scientific quality is always in the responsibility of the PI and all authors. Fields for its documentation like quality flags for single values, adjustable precision or documentation of methods are available.
- Technical quality control, i.e. completeness of metadata, consistence of formats and correctness of download is in the responsibility of the data managers.
- After import, the PI/authors are urged to proof-read data sets on the Internet and submit corrections to the data manager until final agreed publication.

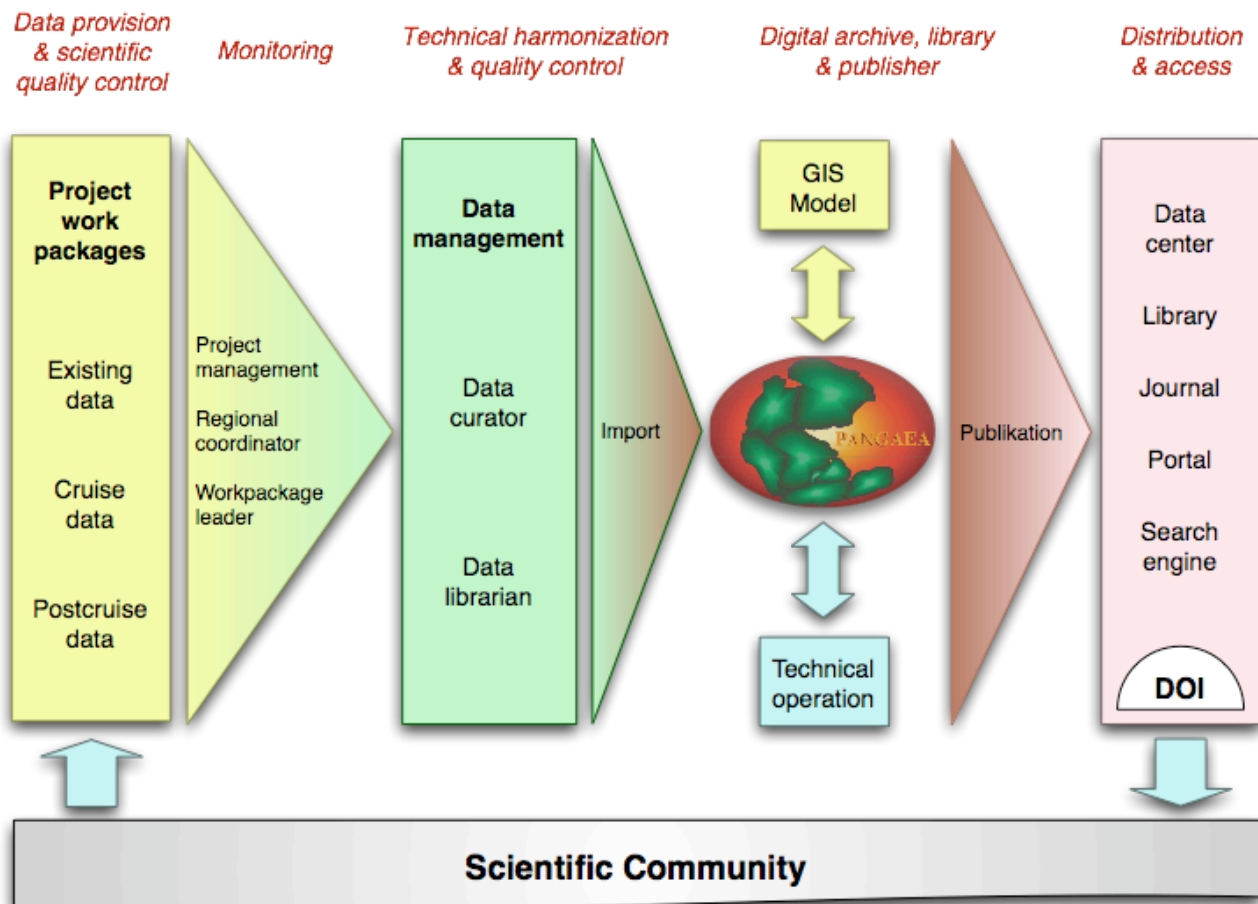


Access and Publication

- Any scientific primary data* related to a publication (supplement) shall be submitted to the data management at the same time as the manuscript is submitted to the editor. Authors will receive in return a DOI (Digital Object Identifier) to link to the data in the publication.
- Likewise, data sets can be made citable on its own. The citation is added to a public library catalog and will receive a DOI. Those data publications may also be added to personal or project publication lists.
- Higher level data products* can also be archived through Pangaea on request.
- Partner institutes and data providers agree, that data archived in Pangaea are made public available through appropriate technical setups on the Internet (e.g. portals, search engines, library catalogs, GIS) without further notification.
- Unpublished data are password protected by default; password protection for published data is set on request for a moratorium period. Providers may decide to withdraw data from the archive as long as it is not published. Metadata are always freely accessible.
- According to EU data policy all data collected during the lifetime of the project are made public two years after the termination of the project; regulations may differ in agreement between coordinator, partners and funding organization.
- Metadata are archived only in relation to available factual data. The metadata can be harvested and distributed by portals using the OAI-PMH standard.
- Data are made available under a Creative Commons Attribution license if not otherwise requested and outlined in the metadata.

Operation

- Long-term availability and operation of the system is ensured by the institutions AWI and MARUM, also responsible for the consistency of the content.
- The Backup of the data inventory is in the responsibility of the computer center at AWI with daily incremental backup and weekly full backup in two mirrored tape drive archives, located in different buildings.
- Data flow is organized by the data curation of the project/institute to the archiving facility, monitored by the project/institute management and supervised by the data librarian.
- Persistent identification, data publication and widespread distribution is performed by the networking functionality and webservices of Pangaea.



Data flow from project to the Pangaea library with Open Access distribution.

(*) Depending on the level of processing scientific primary (or factual) data can be differentiated between raw data, primary data and secondary data. Raw data are provided by a measuring system and are unprocessed; scientific primary data are resulting from the processing of raw data and are the basis for scientific interpretations and publications. Primary data have the highest priority for archiving; the related raw data files may be added if appropriate. Secondary data are higher level products resulting from compilations and interpretations of primary data, i.e. maps, profiles, statistics, graphics, models or any material produced for education and outreach. All information describing any of these three data types are metadata.