HERMES Data Policy

The aim of this policy is to facilitate operation and use of the information system PANGAEA - Publishing Network for Geoscientific & Environmental Data by the research community of HERMES. 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 and search engines. The scientific steering committee of HERMES encourages the widest possible use of the Pangaea library, in order to best realise the value of the HERMES data collection.

Principles

- The information system PANGAEA - Publishing Network for Geoscientific & Environmental Data is the central archive to all HERMES data. The system supports free and Open Access to its content by research and education communities in non commercial activities. This is in line with data policies and recommendations of the IOC, the WDC System and the OECD.
- For any data, provided by HERMES, the format, content and documentation of a data set must ensure its most widespread and easiest use by the scientific community.
- Users of data downloaded from Pangaea are advised to properly use the data set citation and/or quote the related reference.

Data provision

- In aggreement with all partners and following the workprogram and deliverables as outlined in the HERMES contract, any scientific primary data (1) already existing and beeing of relevance to HERMES or (2) produced through research within HERMES should be archived in the information system Pangaea. This includes data resulting from expeditions, on board or postcruise measurements, documentations and publications, formatted in machine readable form.
- Data archiving includes:
    1. Metadata(*) of expeditions, stations and samples;
    2. Scientific primary data from a) archives, b) expeditions, c) publications;
    3. Metadata related to the primary data (2);
    4. Products resulting from compilations and interpretations of primary data.

- Chief scientists are requested to send cruise reports including a station list to the project management office. Station labels as published in the cruise reports station list 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 aggreed standard for all project data. Parameter 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, for certain format reasons, must have a proprietary format, it is archived as a binary object in a file system with a metadescription in Pangaea, linked to the file.
- As soon as new data become available and are validated, the providers are requested to submit the data in agreement with the import format. Any type of data should always be accompanied by a description (metadata) allowing future users to understand and process the data at any time.
- Existing data will be integrated to the HERMES data archive if provided on request to former projects by the data management.
- The granularity and import format of data sets may be defined in agreement between the principle investigator (PI) and the data librarian.
- The export format in principle is tab-delimited ASCII, headed by metadata fields according to ISO19115, GCMD-DIF and DublinCore standards.

Quality assurance

- Data submitted for archiving should be documented properly; documentation is archived together with each dataset.
- The scientific quality is in the responsibility of the PI/author. The required fields for data documentation (quality flags, precission of values, documentation of methods) are available in the Pangaea data model.
- 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/author is requested to proof read data sets on the Internet and submit corrections to the data manager.

Access and Publication

- The project data management provides an up-to-date list of publications with links to data if archived.

- Any scientific primary data* related to publications 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 persistent identifier (DOI, Digital Object Identifier) for each data set that can be cited in the publication. The DOI is part of a full citation, accompanied with each data set.

- Likewise, one to many data sets can be made citable with a reference 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* in electronic form can also be archived through Pangaea and will receive a persistent identifer and citation on request.

- Data may also be submitted if validated but not yet published. Unpublished data are password protected by default and can be linked to references later on. 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 the status is ‘unpublished’. Metadata are always freely accessible.

- According to EU data policy all data collected during the lifetime of the project are made public three years after the termination of the project; regulations for certain data collections may differ in agreement between coordinator, partners and funding organization.

- Following recommendations of the EU (Colour of Ocean Data Symposium, Brussels 2002), metadata are archived only in relation to available factual data. The metadata solely may be mirrored to other metasystems like the Global Change Master Directory (GCMD).

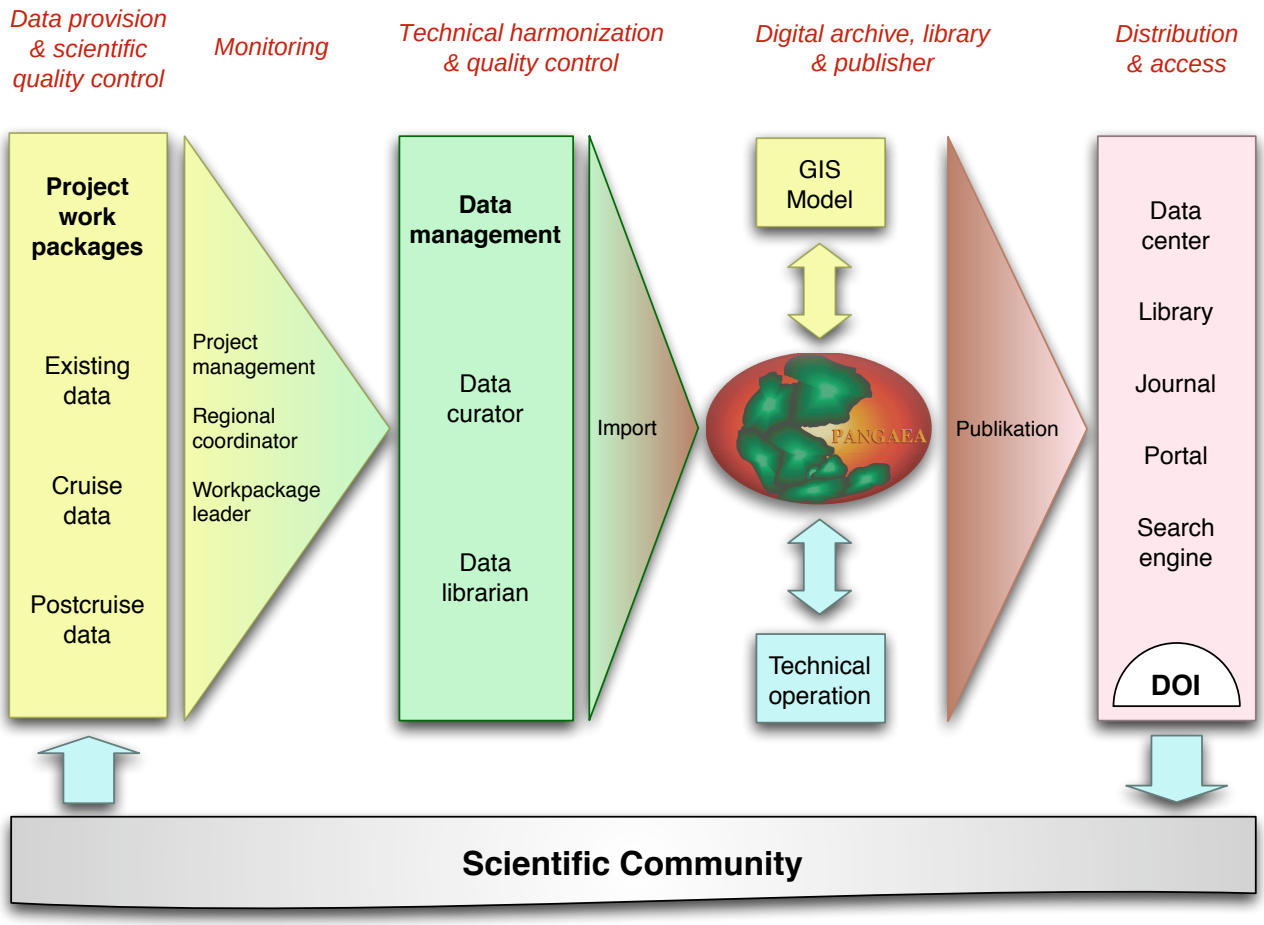
- Partner institutes and data providers aggree, that data archived in Pangaea are made public available on the Internet (e.g. portals, search engines, library catalogs, GIS) without further notification. Persistent identification, data publication and widespread distribution is performed by the networking functionality and webservices of Pangaea.

Operation

- Long-term availability of data is ensured by the institutions AWI and MARUM, responsible for the technical operation, the consistency of the content and the Internet connection of Pangaea.

- The Backup of the data inventory is in the responsibility of the computer center of the AWI with daily incremental backup and weekly full backup in two mirrored tape drives, located in different buildings.

- Data flow is organized from the workpackages via the data managers Ingo Schewe (Biology) and Veit Hühnerbach (Geosciences), to the archiving facility at AWI, monitored by the project management and supervised by the data librarian (Hannes Grobe).



Dataflow from the project to long-term archived and public available documented data.

Hannes Grobe, PANGAEA data librarian, *hgrobe@pangaea.de*

(*) 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 called metadata.