Research Data Directive of the Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research

This directive supplements the AWI research data guideline. In contrast to the long-term guideline, it can and will be regularly adapted to the developing state of the art and the changing scientific cultures of the disciplines.

Definitions:

Research data

The term ‘research data’ refers to all data and products listed below that represent added value for research. All research data must be provided with descriptive metadata.

Raw data

‘Raw data’ is data obtained directly from sensors, devices, or manually, unchecked, unencoded, uncompressed, unformatted and not subject to error correction.

Primary data

‘Primary data’ are processed data obtained from raw data, calibrated, quality- and error-tested, processed data which form the basis for the further research process.

Data products

‘Data products’ are data derived from raw or primary data, which may be combined with other data and/or have undergone further quality and calibration steps. These can be the source or end product of visualisations and/or classification and clustering procedures.

Metadata

‘Metadata’ is structured data about data or data blocks, i.e. information about characteristics of other data or descriptions of data. Metadata or meta-information enables the archiving and retrieval of data. All larger data collections and databases require metadata for structuring.

AWI data sets

Research data is considered an ‘AWI data set’ if it is collected, procured or prepared by AWI staff during their regular work, a project, training or further qualification at the AWI or by external users and associated partners and guests substantially using the resources/infrastructure of the Alfred Wegener Institute for the purpose of their own research. This concerns the entire data life cycle from raw data, including metadata, primary data to derived data products.

For the **re-use of AWI data sets**, it is necessary to collect metadata and to document the context of creation and the tools or software used. In this context, long-term access to and re-use of scientific software is essential. Producers and persons responsible for data, including methods, should be traceable along the data value chain.

The **AWI research data management** (e.g. O2A - Data Flow Framework) designs the handling of AWI data from planning, collection, quality control and analysis to archiving. This includes that AWI data sets are generally archived in long-term repositories (standard is www.pangaea.de) in accordance with the DFG Guidelines for Safeguarding Good Research Practice, Code of Conduct (4) and the FAIR data principles (12) and made citable by data publication. Furthermore, data and processes in the context of their collection must be documented in a subject-specific manner and provided with metadata. Clarity regarding the provenance of the data must be ensured by describing the data transformations. **Data management plans** (DMP) of research projects and infrastructure ensure the documentation of these processes and the description of the data.

Research Data Directive of the Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research

In line with established research data guidelines (6), AWI is striving to make the creation of a compact data management strategy per discipline-specific organisational unit or working group mandatory in applications or plans for research projects, or use of research infrastructure. This will also require the maintenance of a detailed data management plan in the further course of the project, which describes how the research data generated in the course of the research project or activity will be handled. This applies both to applications for external funding and all Helmholtz and centre-internal procedures, including programme-oriented funding and strategic investments. Applications for research projects should already include details of the resources required for this purpose and the copyrights, rights of use and access rights as well as the storage of the data, both during and after completion of the research project.

The AWI will provide the research projects with the possibility to make use of a standard plan of the center or an organisational unit. The data division, the research data officer and/or the permanent research data committee can be contacted in case of questions.

Responsibility

The responsibility for the scientific quality and the careful handling of the data of all research activities and infrastructure facilities at AWI lies with all persons involved in the value chain of research data. The guidelines and subject-specific recommendations of the DFG and the Alliance of German Science Organisations apply to the researchers.

In daily operations, responsibility for the quality, usability and preservation of research data is assumed by section leaders. This applies to all scientific projects, dissertations etc. in which Section staff are involved. In the case of projects involving more than one section, the assumption or distribution of this role must be clearly agreed between the heads of the sections. This applies in particular to cross-centre research within the framework of the PoF and in the case of externally funded projects. In these cases, however, it is still up to the heads of the sections to check whether the plans and actual actions in the projects are compatible with the AWI’s research data policy and guidelines.

If third parties use the scientific infrastructure of the AWI (without reimbursement of full costs), the scientific coordinators or heads of that infrastructure are responsible for ensuring that the users proceed in accordance with the research data guidelines and directives of the AWI. They must ensure that a data management plan is submitted prior to use and that compliance with it is checked and documented. Particular attention must be paid here to access to and preservation of the data. Access to infrastructures in responsibility of AWI can only be guaranteed to those users who have data management plans in place and implement them within a reasonable period (see research data guideline).

With the acceptance of data by relevant recognised long-term repositories (e.g. PANGAEA), the responsibility for their preservation can be passed on to them, provided that all dispositions regarding open access and, if necessary, deletion have been determined and the necessary financing has been secured. In the case of restricted access, the disposition/determination of the data is transferred from the submitting party to the AWI board of directors, represented by the data division and the research data committee.

Section heads and heads of the scientific (data) infrastructure participate in their disciplinary areas in the observance of good scientific practise, in particular, embargo periods and retention practise - what for and how long, as well as quality assurance and FAIRness - how data are received, described and formatted. They communicate their findings to the data division and the AWI research data officer. The relevant administrative offices are to be involved in this process, and information on the requirements of the research funding bodies is to be provided accordingly.

The AWI computing and data centre is responsible for the establishment or procurement and the secure operation of suitable IT systems or services that can be integrated into the scientific work with a manageable amount of effort, in which data, software and its documentation, as well as other documents that serve to describe the scientific work and its results, can be stored

and made accessible. It submits the plans and reports thereon to both the IT Board and the research data committee. It offers training courses for the handling of research data and research data infrastructure.

Users of the AWI IT systems and services are obliged to transfer all unpublished research data to directories accessible to project and working group members at least three months before they leave their employment. If research data remain in personal directories, the currently valid service agreement on the use of Internet, e-mail and storage systems shall apply.

In case of the collection or use of personal or personalisable data (e.g. surveys, detailed health statistics), advice must be obtained, and the AWI's data protection officer must be informed prior to collection/procurement. Plans for deviations from open access to data are - if necessary after consultation with the Technology Transfer Office – should be submitted to the research data officer and, if applicable, to the research data committee, which makes a recommendation to the Board of Directors

Archiving and Publication

All data must be deposited in a publicly accessible, citable long-term repository no later than two years after the survey, with a standardised licence (1). The deposited data may be subject to an embargo for a maximum of two further years. After expiry of the embargo period, the data must be made public immediately and actively using the FAIR data principles. This rule also applies, as far as possible, retroactively to all raw and primary data collected at AWI prior to the adoption of the research data guideline.

The board of directors reserves the right to constrain the future allocation of resources and the provision of resources from the programme and infrastructure if data cannot be found. The provision and archiving of data from projects in the qualification phase of scientists are part of the scientific output on which the qualification is based.

Deviating embargo periods from data collection for exclusive first use and for scientific validation and quality control can be set in special data management plans of the work areas with appropriate justification. This also applies if the confidentiality of research data, at least temporarily, is an indispensable prerequisite for later commercialisation by the centre. Details are regulated by the specifications of the respective funding programmes. Commercialisation of research data is the exception; it must be applied for at an early stage, and a utilisation plan must be submitted. When setting embargo periods, legal regulations are required, taking into account, in particular, the protection of personal data, as well as scientific interests and contractual agreements with cooperation partners and, where appropriate, exploitation interests. Embargo periods must be defined per section/area/working group/project in the DMP and approved by the research data committee.

The decision for an open handling of research results does not exclude their commercialisation. In accordance with the Helmholtz Association's mission in the field of Technology Transfer (9), commercial re-use of research results should also be made possible in principle. However - at least in the case of (partial) financing from public funds - the continued use of the original data by the scientific community must be ensured as well. An exclusive transfer of rights of use for an unlimited period must be ruled out.

However, the metadata concerning the existence of the data should, in any case, be published at the time of filing. The availability of the metadata necessary for subsequent use must be presented in a DMP corresponding to the raw data, primary data and derived data products. For all LKII infrastructure of AWI, the metadata publications on the existence of data are a mandatory part of the reports.

In determining and justifying the duration of the embargo periods, the AWI is guided by the corresponding specifications of professional societies, large research associations and research sponsors. Since such specifications are currently lacking for many fields of research, it should not lead to a waiver of the deadline being set, despite existing uncertainty. Instead,

the AWI, the research departments and working groups should actively participate in the determination process and review the appropriateness of decisions made at regular intervals by the research data committee. The PANGAEA data curators (https://pangaea.de/about/team.php) will assist with this.

Quality Assurance

For the traceability and re-use of research data, it is necessary to record metadata and to document the context in which the data was created, and the tools or software used. Accordingly, in data collection, in addition to the description of the procedure of data collection, framework parameters (as metadata) must also be recorded, which enable statements to be made about the origin, transformation and quality of the collected data in as standardised a form as possible (10). Which metadata must be collected for this purpose depends on the specific research plan/project in each case and should, in any case, comply with the FAIR data principles. Information for quality assurance and evaluation can be recorded in the form of laboratory books, but also in the documentation of data-generating processes in jointly created “Standard Operating Procedures” (SOPs) that are continuously coordinated within the discipline-specific research community also at the international level.

Within the framework of modern (“digital”) science, preference should be given to the explicit and machine-readable coding of this information, and this should be implemented - in theory, standardisation and efficiency-enhancing, non-encumbering practice - in accordance with the FAIR data principles. At the latest in connection with automated data analysis, this machine readability becomes necessary, as does the explicit and machine-readable specification of error variables (whether as part of the data or the metadata).

Information on the data formats used must be part of an extensive metadata set. If possible, open and free data formats should be used, since data should still be usable even after the originally used application has been discontinued. For subsequent use, especially with digital methods, care must be taken that quality-assuring metadata - if possible in digitised form and algorithmically accessible - is stored.

Scientific Recognition

Today, the generation of research data is still often considered to be secondary to their analysis. This differentiation is no longer appropriate in view of the skills needed to collect and process data. Improved recognition of scientific achievement, which is expressed in the collection of research data, therefore begins with the development of a new view of all work processes and the people involved in each of them, who can only realise excellent research results in a cooperative effort. The AWI recognises this additional effort as part of its research performance and is committed continuing to promote this at the national and international level.

When using data records of third parties, the obligation to cite and, if applicable, to offer co-authorship applies, based on the DFG rules of good scientific practice. Authors of scientific publications are always jointly responsible for their content. However, only those who have made a significant contribution to a scientific publication are considered to be authors. So-called “honorary authorship” is excluded in accordance with DFG rules of good scientific practice (8).

Long term Availability

AWI data sets must be archived and published in suitable, sustainably operated, trustworthy long-term repositories (2).

When archiving and, if necessary, making research data accessible, the implications arising from applicable law or contractually based rights of third parties must also be taken into account.

If for plausible reasons, the PANGAEA repository operated at AWI and MARUM cannot be used, the alternatively selected archives must be comparably qualified and interoperable with

the standards and practices of national and international research disciplines and communities. The basis for archiving should be certified data repositories that can guarantee citation via a complete data citation with Handle/DOI since this can be used as a standardised procedure for the application of one of the most important evaluation criteria in the scientific community. These repositories should comply with the FAIR principles. The research data committee provides a list of qualified archives and reviews and adds to it upon request.

In the computing and data centre, the “Data” and “Systems” divisions (which are also responsible for the AWI share of PANGAEA and other infrastructural data activities) represent the sustainably financed core of the infrastructure. Third-party funds from scientific projects refinance and supplement this core, but do not form the basis for financing research data management,

Qualification

Due to the high importance and the rapidly growing demand for qualified personnel for research data management and analysis in research and industry, the establishment of training courses at all AWI locations is urgently required, which may have to be provided in cooperation with universities and/or companies. Needs and corresponding offers must be regularly investigated and evaluated (7).

The departments and sections identify sufficiently qualified and experienced personnel (depending on the discipline Data Stewards/Data Scientists/Engineers, Research Software Engineers or other personnel with appropriate statistical, scientific or technical training) and explicitly indicate their data-related activities. The scientists and the science supporting staff inform themselves comprehensively about qualification and reputational aspects in relation to data and use the options of data publications and citable archives as a contribution to sustainable science in research, infrastructure, teaching and transfer

Legal questions

Both making research data accessible in the sense of Open Science and its commercial use require at least an examination of the necessary authorisation to dispose of the data[1]. In doing so, the labour law, the German Employee Invention Act and the Basic Law (Freedom of Science), among others, must be observed. Questions of ownership law may also include whether the content to be archived is protected by copyright, whether it is a trade or business secret, or with which licence content may be exclusively passed on or published. Further legal frameworks can result from fields as diverse as data protection or export controls. In general, there is an obligation to provide each data record with license information (licensing or usage regulations). It is recommended to use the Creative Commons licenses (3), and especially the CC0 license for metadata and CC-BY license for data.

The protection of personal data is a matter of course and is particularly important for (bio)-medical data. These are regulated in the European data protection basic regulation (5). The Council for Information Infrastructures has published recommendations on the subject (11).

V11, 15.05.2020

1 Here the term ‘power of disposal’ is used because the concept of ownership does not apply to non-material goods as it does to material goods. In the text, the term ‘property’ is nevertheless used in part because it conveys colloquially correctly that it is a matter of naming the natural or legal person who may permit an intended use.

Research Data Directive of the Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research

Referenzen

1. Alfred-Wegener-Institut, Leitlinien für verantwortungsvolle Wissenschaft am AWI, 2016, https://intranet.awi.de/fileadmin/Forschung/Risk_Assessment_Committee/20161013_Leitlinien_fuer_verantwortungsvolle_Wissenschaft_am_AWI_V8.pdf

2. Alfred-Wegener-Institut, Publikationsrichtlinie, 2014, https://intranet.awi.de/fileadmin/Dienste/ePIC/DE_awi-policy.pdf

3. Creative Commons Lizenzen, aufgerufen 2020, https://creativecommons.org/

4. Deutsche Forschungsgemeinschaft (DFG), Gute wissenschaftliche Praxis, aufgerufen 2019, https://www.dfg.de/foerderung/grundlagen_rahmenbedingungen/gwp/

5. EU-Datenschutz-Grundverordnung (EU-DSGVO), aufgerufen 2020, https://www.datenschutz-grundverordnung.eu/

6. forschungsdaten.org Data Policies, aufgerufen 2019, https://www.forschungsdaten.org/index.php/Data_Policies

7. Helmholtz-Gemeinschaft, Digitalisierungsstrategie, 2019, https://www.ufz.de/export/data/2/236513_2019-11-12_Digitalisierungsstrategie_DE_FF_klein.pdf

8. Leitlinien zur Sicherung guter wissenschaftlicher Praxis, Kodex, 2019, https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp.pdf

9. Mission der Helmholtz-Gemeinschaft, aufgerufen 2020, https://www.helmholtz.de/ueber_uns/die_gemeinschaft/mission/

10. Rat für Informationsinfrastrukturen (RfII), Herausforderung Datenqualität 2019, http://www.rfii.de/download/herausforderung-datenqualitaet-november-2019/

11. Rat für Informationsinfrastrukturen: Datenschutz und Forschungsdaten. Aktuelle Empfehlungen, 2017, http://www.rfii.de/download/rfii-empfehlungen-2017-datenschutz-und-forschungsdaten/

12. Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3:160018.