Open Data Guidelines

of the Potsdam Institute for Climate Impact Research

Digital data are an integral element of the modern research process. The Potsdam-Institute for Climate Impact Research (PIK) supports open access to research data by data publishing, defined as follows:

> *"Research data publishing is the release of research data, associated metadata, accompanying documentation, and software code (in cases where the raw data have been processed or manipulated) for re-use and analysis in such a manner that they can be discovered on the Web and referred to in a unique and persistent way. Data publishing occurs via dedicated data repositories and/or (data) journals which ensure that the published research objects are well documented, curated, archived for the long term, interoperable, citable, quality assured and discoverable – all aspects of data publishing that are important for future reuse of data by third party end-users."* (Austin et al., 2015)

PIK encourages modeling and project groups as far as applicable to publish their research data for several good reasons, among others:

- As a member of the Leibniz association PIK is a co-signer of the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (MPG, 2003).
- Research at PIK profits from the free availability of research data generated by third parties outside the institute. To a certain degree, this implies the responsibility to publish own research data.
- More and more journals request open access to data as part of the supplementary information of a paper (Nature Climate Change, 2016 and the references there). Additionally, there is a clear recommendation to publish data through dedicated data repositories and not as electronic supplement directly attached to the paper (COPDESS, 2015).

These guidelines comply with and complement the PIK Guidelines for Ensuring Good Scientific Modelling Practice.

Summary of recommendations

- Publish the data in a research data repository and assign a DOI
- Assign a license to the data
  - Recommended licenses without a copyleft clause:
    - Creative Commons Attribution International Public License CC BY 4.0
    - Open Data Commons Attribution License ODC-By v1.0.
  - Recommended licenses with a copyleft clause:
    - Creative Commons Attribution-ShareAlike International Public License CC BY-SA 4.0
    - Open Data Commons Open Database License ODbL v1.0

Steps to take

1. Discuss plans for data publishing with the PIK library
2. Choose a license and terms of use (ToU) for the data
3. Select a research data repository
4. Prepare data publication:
   data documentation, metadata for data discovery, interoperability of datasets
5. Receive official permission to publish the data
6. Document the publication and archive the used material

1. Discuss plans for data publishing with PIK library

The PIK library is the general contact point for data publishing. PIK library belongs to the Bibliothek des Wissenschaftsparks Albert Einstein - Campus Library. The Campus Library has a long standing competence in all steps for data publishing. If you see problems in publishing the data at all or with intellectual property rights also discuss your plans with your RD Co-Chair. In parallel you also can contact the Open Science advisory group of the Modelers' Council at opendata@pik-potsdam.de.

2. Choose a license and terms of use (ToU) for the data

When it comes to intellectual property rights (IPR) and licenses for data, the central notion is the database as "a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means" (European Community, 1996). For data generated by PIK and its scientific communities only a collection of data can be protected by property rights or a license, for a single data value this is impossible. The database notion is not restricted to a data collection stored in and individually accessible by a traditional database management systems such as Oracle or MySQL. It relates also to data stored in a file and organized in a well-structured manner.

2.1 Clearing of pre-existing intellectual property used in the data

Case 1: The data to publish is the result of a re-compilation of data from third parties and potentially of your own data

Then it has to be checked whether data from all sources can be merged and if so under which conditions. Normally, the conditions are specified in the licenses assigned to the individual data of the third parties.

Case 2: Data from third parties are used to compute new data to publish

A typical example is to use such data as driving forces of a simulation model and to publish its model output as new data. For this case the *sui generis* database right (Datenbankherstellerrecht) implies that such third party data can be used without having a license. Nevertheless it is good scientific practice to act also as for Case 1 and to check and respect the license of the data.

The license of a model as a software is independent of the license of the data computed with the model. Check the license of the model under which conditions you can use it.

Case 3: Third party data to be used in Cases 1 or 2 is a proprietary product or does not have a license Then it is recommended to address the third party directly and to negotiate under which circumstances the data can be used.

For all these cases check the licenses of all third party data on compatibility to each other and on compatibility with the license you plan to assign to the data to publish. License compatibility is a complex issue. As a rule of thumb,

- non-copyleft licenses (see Sect. 2.2) are compatible to each other
- copyleft licenses are not necessarily incompatible (e.g., CC licenses have a compatibility clause)
- copyleft and non-copyleft licenses are not necessarily compatible.

If a license is not compatible to others then ask the owner for providing a different license. Finally, it is always recommendable to contact the owner of third party data if there are any concerns about its right usage.

Set up a **contributors' license agreement** (CLA) to handle the intellectual property rights of all persons involved in developing the data ("inbound licensing"). Make sure that all contributors own all rights on their contributions. It is always recommended to have such an agreement to avoid conflicts in future. For more information check the related section of the PIK Open Source Guidelines and follow the recommendations made there, in particular, to use for the CLA http://contributoragreements.org.

2.2 Choose an appropriate license

Always assign a license to the data to publish and check its compatibility to licenses of used data as explained in Sect. 2.1 above.

An important difference among open source licenses is the inclusion or exclusion of a so called copyleft clause. Copyleft means that a derivative of a product which was published under a copyleft license has to be published under the same license. In contrast, a so-called permissive license does not have a copyleft clause and gives the licensee more freedom how to redistribute the data.

PIK recommends for data publishing the Creative Commons Attribution International Public License **CC BY 4.0** (https://creativecommons.org/licenses/by/4.0/) or later and the Open Data Commons Attribution License **ODC-By v1.0** (https://opendatacommons.org/licenses/by/) or later. Both licenses do not have a copyleft clause, are compatible and comparable. They allow to share (copy, redistribute and use the data), to create (produce works from the data) and to adapt (modify, transform and build upon the data) as long as you give an appropriate credit.

If you want to ensure copyleft then PIK recommends the Creative Commons Attribution-ShareAlike International Public License **CC BY-SA 4.0** (https://creativecommons.org/licenses/by-sa/4.0/) or later and the Open Data Commons Open Database License **ODbL v1.0** (http://opendatacommons.org/licenses/odbl/) or later. Both licenses are not compatible. For compatibility of ODbL v1.0 with other licenses check here. Please keep in mind that there should be good reasons for assigning a license with a copyleft clause to your data: The compatibility problem of

copyleft licenses (see Sect. 2.1 above) can prevent compiling and using data. When you plan to publish data in parallel with a text publication then some data journals do not accept data with a copyleft license due to the compatibility problem (see http://oaspa.org/information-resources/frequently-asked-questions/).

The licenses ODC-By v1.0 and ODbL v1.0 from opendatacommons.org fit better to data than the rather general licenses CC BY 4.0 and CC BY-SA 4.0 from creativecommons.org, although also the latter address databases in Section 4 of the license text. Make your decision which license scheme to use on the compatibility with the licenses of used data and anticipate how potential licensees of your data can cope with the license you assign to.

Please always discuss the choice of license with the PIK library. For information on licenses see also the PIK Open Source Guidelines.

2.3 Terms of use and authors' affiliations

Specify in the data header or as a separate file with the name COPYING the license with the data by a link to the license text and add the Terms of Use / the copying permission statement. This could read like

> This data is freely available under <full license name and its version> <link to the license text>. When using the data, cite it as follows: <citation as specified at the landing page of the data DOI> doi: <doi>.
> You can redistribute and/or modify the data under the terms of the above license, either the above version, or (at your option) any later version.
>
> This data is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the above license for more details.

If <citation as specified at the landing page of the data DOI> does not show the affiliations of the authors then specify with the data (in the data header or as a separate file with the name AUTHORS) the authors and their affiliations. For an example of all this information see Sect. 8.

3. Select a research data repository

Instead of storing the data at a PIK file system and referring to the data from a website, store the data always in a research data repository. Data repositories guarantee for long-term availability and permanent access to published datasets. Data should be submitted to domain-specific, community-recognized repositories where possible, or to general repositories if no suitable community resource is available. Often, data journals recommend special accredited repositories.

For an overview of research data repositories see http://www.re3data.org.

Examples for repositories are

- Bibliothek des Wissenschaftsparks Albert Einstein - Campus Library repository, http://dataservices.gfz-potsdam.de
  For more details see below.
- ESGF, http://esgf.llnl.gov/
  The Earth System Grid Federation ESGF is a Peer-to-Peer enterprise system collaboration that develops, deploys and maintains software infrastructure for the management, dissemination, and analysis of model output and observational data. To join ESGF for data publishing a local ESGF node has to set up. At PIK there is an ESGF node for ISIMIP data https://esg.pik-potsdam.de. ESGF does not support DOI assignment to the data.
- PANGAEA, https://www.pangaea.de/
  PANGAEA is a publishing repository for Earth system and environmental data. Data is stored with their spatial and temporal reference and a DOI is assigned.

4. Prepare data publication: data documentation, metadata for data discovery, interoperability of datasets

Document the data by specifying the original data, the model and its setting (for model output data), and all processing steps.

If you plan to publish the data …

- in a data journal (e.g., Earth and Space Science, Earth System Science Data, Geoscience Data Journal or Scientific Data)
  … then the paper is the documentation of the data. Before publishing the data please always discuss with the data journal the choice of license and how to access the data during the review process.
- as a data supplement to a publication in an other journal
  … then describe the data in the supplementary information, refer to the data by its DOI and write a detailed documentation as in the following bullet point.
- as a stand-alone publication, not related to a journal
  … then either supplement the data in the repository by a PDF file with the detailed documentation or describe the data in a PIK Report as a data publication supplement. A PIK Report can have a DOI.

If the documentation is in a separate document, an appropriate license should be assigned. A good choice is the Creative Commons Attribution International license CC BY 4.0 (see Sect. 2.2 above).

To make data **discoverable** on the web they have to be described by metadata. There are different metadata standards for Earth science data. The metadata standard is set by the repository, and the repository also publishes the metadata.

A DOI is an essential prerequisite for easy data identification and citation. Most publishing repositories have a DOI service and also host the web page where the DOI links to. This so-called DOI landing page

holds all information about the data and enables access to the data. The **citation** is provided by the data repository using the specified metadata after assigning a DOI to the data. It should comprise (i) the authors and their affiliations, (ii) the dataset name, (iii) the data repository name, and (iv) the publishing year (see Sect. 8 for an example).
Keep in mind that data that were published under a DOI cannot be changed anymore. If you expect to have several versions of the data discuss this with the data repository before submitting the first version and specify a version number in the metadata. Each version will be published with a new DOI.

Please always add the citation to the header of each dataset (see section 2).
The DOI of the data differs from the DOI of the related documentation paper or supplement.

Store the data in a format commonly used by the research community, e.g., NetCDF. Check the data repository for supported data formats. Ensure **interoperability** with other community data, e.g. by using standardized names for variables as in the NetCDF CF climate and forecast metadata conventions.

5. Receive official permission to publish the data

To clarify code ownership and publication rights the model has to undergo a formal application process. This helps PIK administration to evaluate risks and to double-check whether all preconditions for Open Data publication have been fulfilled. The corresponding checklist and application form can be found here. Receive the final permission to publish the data from your RD Co-Chair.

6. Document the publication and archive the used material

After publishing the data send the publication record to your RD Science Coordinator and/or Secretary. The record will be listed in the PIK Publication Database http://edoc.gfz-potsdam.de/pik/search.epl under the document type “Research data”. In parallel, create a publication record in the PIK Metadatabase http://metadata.pik-potsdam.de and archive all the digital material used and produced for the published data as well as the published data.

7. The Campus Library repository

The Bibliothek des Wissenschaftsparks Albert Einstein - Campus Library (CL) has a long-standing competence in all steps for data publishing. It operates the research data repository GFZ Data Services http://dataservices.gfz-potsdam.de, hosted at GFZ and guaranteeing long-term availability and permanent access to data and their description. The repository is accredited by journals (e.g., Earth System Science Data).

CL offers the following services for data publishing:

- Individual support of the whole data publishing workflow including personal consultations

- Assignment of DOIs for datasets, and publication supplements. CL is a DOI publication agent with the Technische Informationsbibliothek (TIB Technical Information Library) Hannover.
- DOIs from TIB/CL for PIK are free of charge
  The DOI scheme for PIK data is 10.5880/PIK.<year>.<sequence>
- Publication of the metadata
- Browsing and searching data and description of the data publication workflow ("about") from the data repository portal
- Tool to gather metadata for data discovery (necessary for data publication and search)
  Metadata Editor: http://dataservices.gfz-potsdam.de/panmetaworks/metaedit.
  Metadata formats: iso19115, datacite, dif, escidoc
- On-the-fly landing pages for data DOIs
- Data storage in the CL repository, hosted by GFZ. In exceptional cases, DOI links may be directed to external URLs.
- Temporary data access only for reviewers of a submitted paper in a data journal

All published PIK datasets at GFZ Data Services can be accessed by selecting datacenter PIK at the portal http://dataservices.gfz-potsdam.de/portal/.

8. Example for terms of use and authors' affiliations

The example relates to the doi 10.5880/PIK.2016.002

File COPYING or in the data header:

> This data is freely available under the Creative Commons Attribution-ShareAlike International Public License CC BY-SA 4.0 https://creativecommons.org/licenses/by-sa/4.0/. When using the data, cite it as follows: Cornford, Stephen; Asay-Davis, Xylar (2016): Ice-shelf surface, basal and bedrock topography data for the second Ice Shelf-Ocean Model Intercomparison Project (ISOMIP+). GFZ Data Services. doi: 10.5880/PIK.2016.002.
> You can redistribute and/or modify the data under the terms of the above license, either the above version, or (at your option) any later version.
>
> This data is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the above license for more details.

Only if affiliations are not specified under the doi:
File AUTHORS or in the data header:

> Cornford, Stephen – Centre for Polar Observation and Modelling, University of Bristol, Bristol, UK
> Asay-Davis, Xylar – Earth System Analysis, Potsdam Institute for Climate Impact Research, Potsdam, Germany