# Data Descriptor Template

**Scope Guidelines**

**Data Descriptors** submitted to *Scientific Data* should provide detailed descriptions of valuable research datasets, including the methods used to collect the data and technical analyses supporting the quality of the measurements. Data Descriptors focus on helping others reuse data, rather than testing hypotheses, or presenting new interpretations, methods or in-depth analyses. Relevant datasets must be deposited in an appropriate public repository prior to Data Descriptor submission, and their completeness will be considered during editorial evaluation and peer review. The data must be made publicly available without restriction in the event that the Data Descriptor is accepted for publication (excepting reasonable controls related to human privacy issues or public safety).

### Comprehensive dataset of shotgun metagenomes from oxygen stratified boreal and subarctic lakes and ponds

*110 characters maximum, including spaces*

### Authors

Moritz Buck1, Sarahi L. Garcia, Stefan Bertilsson, Leyden Fernandez Vidal, Gaëtan Martin, Jakob Zopfi, Antti Rissanen, Gustavo A. Martinez Rodriguez, Jatta Saarenheimo, Vicente Sedano, Sari Peura2

**Affiliations**

1. institution

2. institution

corresponding authors: Sari Peura ([sari.peura@slu.se)](mailto:sari.peura@slu.se)), Moritz Buck (moritz.buck@slu.se)

### Abstract

*170 words maximum*

Oxygen stratified lakes and ponds are significant sources of greenhouse gases and hotspots in the carbon cycle in the northern regions from boreal to arctic zone, but the microbial communities especially in their oxygen depleted bottom compartments are poorly characterized. In this paper, we are describing a comprehensive metagenome dataset including shotgun metagenomes from XX oxygen stratified lakes mainly residing in the boreal and subarctic regions, but also one reservoir from the tropical Puerto Rico and one lake from the temperate zone (Austria) are included. For most lakes and ponds the data includes a vertical sample series from the oxic surface to anoxic bottom layer. Further, the dataset includes environmental variables for the samples, such as oxygen, nutrient and carbon concentrations. The dataset is ideal for exploring the microbial taxonomic and functional diversity in key aquatic environments, critically impacted by the climate change.

### Background & Summary

*700 words maximum*

Oxygen stratified lakes are a typical feature of the northern regions and significant sources of greenhouse gas (GHG) emissions (REF). These lakes reside in the region critically impacted by the climate change (REF) and the contribution of these lakes to further enhancing the climate change via GHG emissions is dependent on the microbes inhabiting these waterbodies (REF). Especially important are the organisms in their anoxic compartment, as many of the prominent GHG are produced in this compartment (REFs). However, our knowledge regarding these microbial communities is still sparse, and only a few studies exist addressing the functional features of the microbial communities in the anoxic lake compartments (REFs). To address this gap in our knowledge, we have collected a comprehensive set of samples from XX lakes and ponds from oxygen stratified waterbodies from boreal and subarctic regions, and additionally a vertical sample series also from one tropical reservoir from Puerto Rico and a time series of vertical sample series from a temperate lake in Austria (Fig. 1, Table 1). For the majority of the lakes samples exist from throughout the water column, from oxic epilimnion, the oxygen transition zone (metalimnion) and from anoxic hypolimnion (Supplementary Table S1). For 12 of the boreal lakes only surface samples were collected. Additionally, for all samples also environmental conditions were analysed, including, but not limited to, oxygen, nutrient and carbon concentrations (Supplementary Table S1). For all samples metagenomes were sequenced using shotgun sequencing on Illumina NovaSeq platform at Science for Life Laboratory (Uppsala University, Uppsala, Sweden).

Our goal was to collect a comprehensive dataset that would allow broad analyses of the functioning of the microbial communities in oxygen stratified lakes representing high carbon concentration, and variable environmental conditions in regards of, for example, nutrient concentrations. With this data, we aimed at characterizing the taxonomic and functional identities of the microbes across boreal and subarctic lakes and ponds. This improves significantly our understanding of the functioning of these key environments in the elemental cycles in the northern landscapes. The data set can be used for assessing both general metabolic potential of the total lake communities, but also, via assembly and binning of metagenome assembled genomes (MAGs), the exploration of population genomes of individual community members for deeper understanding of the poorly known members of the communities.

### Methods

*Sample collection*

The samples were collected between 2009 and 2018 from XX locations expanding from the subarctic region to the tropics (Fig. 1, Supplementary Table S1). Majority of the samples were collected using Limnos sampler (Limnos pl., Poland), with the exception of the samples from La Plata (Puerto Rico), which were collected using XX and samples from Loclat, which were collected using XX sampler. For most lakes and ponds, multiple samples were collected, including samples from the oxic surface layer (epilimnion), the layer with steepest change in oxygen concentration and temperature (metalimnion) and from the layer where oxygen was below detection limit (hypolimnion). The exception to this were 12 lakes in Sweden, and five shallow ponds in Canada, for which only one sample from the oxic surface layer was taken (see Supplementary Table S1). Most of the lakes were sampled during the open water season, with the exception of 12 lakes with only surface samples, which were sampled during the ice cover period, as well as one Swedish lake (lake Lomtjärn), which was sampled twice during the ice cover period, and X samples in the time series of the Austrian lake (Loclat), which were taken from under the ice.

Simultaneous to collection of the DNA samples, also samples for environmental conditions were taken. These included temperature, pH, oxygen, total and dissolved nutrients (P and N species), CO2 and methane, total or dissolved organic carbon, iron, sulfate and chlorophyll a. As the samples were collected during multiple years and by different research groups, there was some variation for the procedures between the different sampling occasions, leading to variation in the final set of environmental data across the samples (Supplementary Table S1).

*Collection, extraction and sequencing of DNA*

Most of the DNA samples were collected on Sterivex filters (Millipore), except for the timeseries samples collected from Loclat, which were collected on XX filters, and time series samples from Finnish lake Mekkojärvi, which were collected on XX filters. In all cases, water was filtered through the filter until clogging. All filters were stored frozen (-20 to -80 ˚C) until the extraction of the DNA. For the majority of the samples, DNA was extracted using PowerSoil DNA extraction kit (XX) following the manufacturer´s instructions. Exception to this were samples included to Mekkojärvi time series, which were extracted using XX.

The quality check of the DNA prior to sequencing was done…

### Data Records

All sequences are deposited to XXX under accession number XXX. Etc….

### Technical Validation

### Usage Notes

*This section is optional*

### Code Availability

### Acknowledgements

Milla Rautio is acknowledge for her help to access the CEN research station in Whapmagoostui-Kuujjuarapik (Quebec, Canada) and Sergio Morales for his help obtaining the samples from Puerto Rico. We are grateful to Carla XX, Guys helping at KW, etc for their help with the sampling. The sequencing was funded by grant from Science for Life Laboratory biodiversity program. UPPMAX, grants

### Author contributions

### Competing interests

The authors declare no conflict of interest.

### Figures

Figure images should be provided as separate files and should be referred to using a consistent numbering scheme through the entire Data Descriptor. In most cases, a Data Descriptor should not contain more than three figures, but more may be allowed when needed. We discourage the inclusion of figures in the Supplementary Information – all key figures should be included here in the main Figure section.

For initial submissions, authors may choose to supply a single PDF with embedded figures.

Authors are encouraged to consider creating a figure that outlines the experimental workflow(s) used to generate and analyse the data output(s).

### Figure Legends

### Figure legends begin with a brief title sentence summarizing the purpose of the figure as a whole, and continue with a short description of what is shown in each panel and an explanation of any symbols used. Legends must total no more than 350 words, and may contain literature references. The first sentence of the legend will be used as the title for the figure. It should contain no references of any kind, including to specific figure panels, bibliographic citations or references to other figures or panels.

### Tables

Authors are encouraged to provide one or more tables that provide basic information on the main ‘inputs’ to the study (e.g. samples, participants, or information sources) and the main data outputs of the study; also see the additional information on providing metadata on page 6. Tables in the manuscript should generally not be used to present primary data (i.e. measurements). Tables containing primary data should be submitted to an appropriate data repository.

Authors may provide tables within the Word document or as separate files (tab-delimited text or Excel files). Legends, where needed, should be included in the Word document. Generally, a Data Descriptor should have fewer than ten tables, but more may be allowed when needed. Tables may be of any size, but only tables that fit onto a single printed page will be included in the PDF version of the article (up to a maximum of three).

Due to typesetting constraints, tables that do not fit onto a single A4 page cannot be included in the PDF version of the article and will be made available in the online version only. Any such tables must be labelled in the text as ‘Online-only’ tables and numbered separately from the main table list e.g. ‘Table 1, Table 2, Online-only Table 1’ etc.

### References

Bibliographic information for any works cited in the above sections, using the standard Nature referencing style.

In line with emerging [industry-wide standards for data citation](https://www.nature.com/articles/sdata2018259), references to all datasets described or used in the manuscript should be cited in the text with a superscript number and listed in the ‘References’ section in the same manner as a conventional literature reference. See ‘Citing Data’ below for further details.

### Additional Formatting Information

**Referencing Figures, Tables, and other content**

**The Word document may reference Figures (e.g. Fig. 1), Tables (e.g. Table 1), online-only tables (e.g. Online-only Table 1) and Supplementary Information (e.g. Supplementary Table 1, or Supplementary File 2, etc.). When information from metadata documents must be referred to, it should also be included in the main manuscript as Tables, and formatted in a way that suits human readability. To refer to the ISA-Tab metadata records within the manuscript, use the phrase “see associated Metadata Record”.**

**Citation format**

All references should be numbered sequentially, first throughout the text, then in tables, followed by figures and, finally, boxes; that is, references that only appear in tables, figures or boxes should be last in the reference list. Only one publication is given for each number. Only papers that have been published or accepted by a named publication or recognized preprint server should be in the numbered list; preprints of accepted papers in the reference list should be submitted with the manuscript. Published conference abstracts, numbered patents, and archived code with an assigned DOI may be included in the reference list. Grant details and acknowledgments are not permitted as numbered references. Footnotes are not used.

Scientific Data uses standard Nature referencing style. All authors should be included in reference lists unless there are six or more, in which case only the first author should be given, followed by ‘et al.’. Authors should be listed last name first, followed by a comma and initials (followed by full stops, '.') of given names. Article titles should be in Roman text; only the first word of the title should have an initial capital and the title should be written exactly as it appears in the work cited, ending with a full stop. Book titles should be given in italics and all words in the title should have initial capitals. Journal names are italicized and abbreviated (with full stops) according to common usage. Volume numbers and the subsequent comma appear in bold. The full page range should be given where appropriate. See the examples below:

**Journal Article**:

1. Schott, D. H., Collins, R. N. & Bretscher, A. Secretory vesicle transport velocity in living cells depends on the myosin V lever arm length. *J. Cell Biol*. **156**, 35‐39 (2002).

**Book** ‐ Book titles should be given in italics and all words in the title should have initial capitals:

1. Hogan, B. *Manipulating The Mouse Embryo: A Laboratory Manual* 2nd edn (Cold Spring Harbor Laboratory Press, 1994)

**Publicly available preprint:**

1. Babichev, S. A., Ries, J. & Lvovsky, A. I. Quantum scissors: teleportation of single-mode optical states by means of nonlocal single photon. Preprint at http://arXiv.org/quant-ph/0208066 (2002).

**Code:**

1. Gallotti, R. & Barthélemy, M. Source code for: The multilayer temporal network of public transport in Great Britain. *Figshare* https://dx.doi.org/10.6084/m9.figshare.1249862.v1 (2014).

**Online material** ‐ Stable documents hosted on the web may be cited in the main reference list, using the format below. Websites or dynamic web resources should be cited by embedding the URL in the main article text:

1. Manaster, J. Sloth squeak. *Scientific American Blog Network* http://blogs.scientificamerican.com/psi-vid/2014/04/09/sloth-squeak (2014).

**Technical or government report:**

1. Akutsu, T. *Total Heart Replacement Device.* Report No. NIH-NHLI-69 2185-4 (National Institutes of Health, 1974).

## Citing Data

In line with emerging [industry-wide standards for data citation](https://www.nature.com/articles/sdata2018259), references to all datasets described or used in the manuscript should be cited in the text with a superscript number and listed in the ‘References’ section in the same manner as a conventional literature reference.

An author list (formatted as above) and title for the dataset should be included in the data citation, and should reflect the author(s) and dataset title recorded at the repository. If author or title is not recorded by the repository, these should not be included in the data citation. The name of the data-hosting repository, URL to the dataset and year the data were made available are required for all data citations. For DOI-based (e.g. figshare or Dryad) repositories the DOI URL should be used. For repositories using accessions (e.g. SRA or GEO) an [identifiers.org](https://identifiers.org/) URL should be used where available. For first submissions, authors may choose to include just the accession number. Scientific Data staff will provide further guidance after peer-review. Please refer to the following examples of data citation for guidance:

1. Zhang, Q-L., Chen, J-Y., Lin, L-B., Wang, F., Guo, J., Deng, X-Y. Characterization of ladybird Henosepilachna vigintioctopunctata transcriptomes across various life stages. figshare <https://doi.org/10.6084/m9.figshare.c.4064768.v3> (2018).
2. NCBI Sequence Read Archive <http://identifiers.org/ncbi/insdc.sra:SRP121625> (2017).
3. Barbosa, P., Usie, A. and Ramos, A. M. Quercus suber isolate HL8, whole genome shotgun sequencing project. GenBank<http://identifiers.org/ncbi/insdc:PKMF00000000> (2018).
4. DNA Data Bank of Japan <http://trace.ddbj.nig.ac.jp/DRASearch/submission?acc=DRA004814> (2016).

**Manuscript metadata tables**

Authors are are asked to submit one or more tables presenting, at a minimum, the samples and subjects employed in the study, the experimental, observational and analytical manipulations performed on each, and the data outputs resulting from these manipulations with their manuscript. Please also see Data Records and Tables, above; please note that, should you wish to create more complex metadata records, template Study and Assay files are available either in a ZIP archive with this document or for separate download from the “[Submission guidelines](http://www.nature.com/sdata/publish/submission-guidelines)” page of the *Scientific Data* website.

Here, we provide four generic ‘Table 1’ examples, including two experimental study examples, one observational study example, and an example for an aggregated dataset of the type that may result from a meta-analysis.

**Experimental study Table 1 example**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Subjects** | **Protocol 1** | **Protocol 2** | **Protocol 3** | **Protocol 4** | **Data** |
| Mouse1 | Drug treatment | Liver dissection | RNA extraction | RNA-Seq | GEOXXXXX |
| Mouse2 | Drug treatment | Liver dissection | RNA extraction | RNA-Seq | GEOXXXXX |
| Mousen | Drug treatment | Liver dissection | RNA extraction | RNA-Seq | GEOXXXXX |

**Experimental study with replicates Table 1 example**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **Protocol 1** | **Protocol 2** | **Samples** | **Protocol 3** | **Data** |
| CellCulture1 | Drug treatment | RNA extraction | TechnicalRep1a | Microarray hybridization | GEOXXXXX |
| CellCulture1 | Drug treatment | RNA extraction | TechnicalRep2a | Microarray hybridization | GEOXXXXX |
| CellCulture1 | Drug treatment | RNA extraction | TechnicalRep3a | Microarray hybridization | GEOXXXXX |
| CellCulture2 | Drug treatment | RNA extraction | TechnicalRep1b | Microarray hybridization | GEOXXXXX |
| CellCulture2 | Drug treatment | RNA extraction | TechnicalRep2b | Microarray hybridization | GEOXXXXX |
| CellCulture2 | Drug treatment | RNA extraction | TechnicalRep3b | Microarray hybridization | GEOXXXXX |

**Observational study Table 1 example**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sample** | **Geographical location** | **Geoposition** | **Protocol** | **Data** |
| Body of water 1 | location name | latitude, longitude, altitude | Measurement of surface temperature | dataFile1 |
| Body of water 2 | location name | latitude, longitude, altitude | Measurement of surface temperature | dataFile2 |
| Body of water n | location name | latitude, longitude, altitude | Measurement of surface temperature | dataFile3 |

**Data aggregation study Table 1 example**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Source** | **Sample** | **Sample number** | **Temporal range** | **Protocol 1** | **Protocol 2** | **Data** |
| Database URL 1 | Dataset 1 | Number of samples in the dataset | Range of measurements reported in the dataset | Data assimilation procedure | Method to generate output data | dataFile1 |
| Database URL 1 | Dataset 2 | Number of samples in the dataset | Range of measurements reported in the dataset | Data assimilation procedure | Method to generate output data | dataFile1 |
| Database URL 2 | Dataset n | Number of samples in the dataset | Range of measurements reported in the dataset | Data assimilation procedure | Method to generate output data | dataFile2 |

**Machine accessible metadata records**

All Data Descriptor manuscripts published in *Scientific Data* are accompanied by detailed machine-accessible metadata files in ISA-Tab format depicting the workflow used to generate the accompanying datasets. In order to facilitate the creation of these records, authors may be asked to complete metadata templates for their data by the editorial office, if the Data Descriptor is of interest to the Editorial Board.

**Depositing your data to an appropriate repository**

Your *Scientific Data* manuscript will not be sent to review unless the dataset(s) described therein have been deposited in an appropriate public repository ([please see our list of recommended repositories](http://www.nature.com/sdata/policies/repositories)). Should a specific repository not be available for your field or data-type, or should the repository of your choice not permit confidential peer-review, you may upload your data to one of our [recommended generalist repositories](https://www.nature.com/sdata/policies/repositories" \l "general). Integrated submission systems are available for both figshare and Dryad.