www.aquacosm.eu

WORKFLOW FOR ALLOCATING DOIs FOR TA DATASETS

Deliverable No: D4.8

PROJECT TITLE:

AQUACOSM-plus

Network of Leading Ecosystem Scale Experimental Aquatic Mesocosm Facilities Connecting Rivers, Lakes, Estuaries and Oceans in Europe and Beyond

PROJECT NUMBER:

871081

PROJECT TYPE:

Research and Innovation Action

WORK PROGRAM TOPICS ADDRESSED:

H2020-INFRAIA-2018-2020 / H2020-INFRAIA-2019-1

Deliverable title	**Workflow for allocating DOIs for TA datasets**
Deliverable number	D4.8
Deliverable version	1
Contractual date of delivery	30 November 2021 (M20)
Actual date of delivery	08 December 2021 (M20)
Dissemination level	Public
Nature of deliverable	Workflow
Work package	WP4
Lead Beneficiary	UKCEH
Author(s)	Heidrun Feuchtmayr (UKCEH), Andy Sier (UKCEH), Simon Keeble (BLIT), Johan Wikner (UMU), Lisette de Senerpont Domis (NIOO)
Editor	Katerina Symiakaki (FVB-IGB), Ayoub El Ghadraoui (FVB-IGB), Jens Nejstgaard (FVB-IGB)
EC Project Officer	Pierre QUERTENMONT

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 871081

Table of Content

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 871081

1. Abstract

This deliverable provides a guide for Transnational Access (TA) users, as well as other researchers and facility providers to understand the data structures available within mesocosm science and helps them to obtain a DOI (Digital Object Identifier) for their dataset(s) to ensure an open science workflow. Publishing data is a requirement for EU projects and one of the central aims of AQUACOSM-plus, ensuring good data management, lasting data curation, credits the data provider and, ultimately, will lead to increased research impact. The guide below will help all involved AQUACOSM-plus facility providers, as well as the TA user community to publish their data with an assigned DOI from the hundreds of projects conducted within AQUACOSM-plus. AQUACOSM-plus recommends the use of PANGAEA (www.pangaea.de) and we will outline in the below workflow the steps involved in publishing dataset(s) with PANGAEA in order to keep the time effort of uploading the data into a data repository to a minimum.

2. Overview of DOIs

2.1 What are DOIs?

DOIs (Digital Object Identifiers) are a persistent identifier used to identify objects uniquely, standardised by the International Organisation for Standardisation (ISO) (ISO 26324:2012). The ISO is an international standard-setting body composed of representatives from various national standards organizations. While most academics are familiar with DOIs being used to identify research journal articles, DOIs are also used to identify data sets once published. When a DOI is assigned to a dataset, the DOI is bound to metadata to give additional information about the dataset, e.g. data collection methods, experimental design, sampling regime, the nature and units of the recorded values, information on quality control, the details of the data structure or any other information useful to the interpretation of the data. It is important to note that the DOI for a document remains fixed over the lifetime of the document, i.e. no changes can be made to the data itself, however the metadata may change. The developer and administrator of the DOI system is the International DOI Foundation (IDF), which introduced it in 2000 (Wikipedia, 2021; Paskin, 2010). Within AQUACOSM-plus, we aim to promote the publication of datasets arising from experiments and TA users by raising awareness of the ease of DOI allocations for datasets and publishing by the below workflow for guidance.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 871081

2.2 Benefit of publishing data with DOIs

Apart from the mandatory open data policy within EU projects under the FAIR principles (findable, accessible, interoperable and reusable), there is an increasing need recognised within the research community to make results reproducible. More and more scientific journals and publishers are encouraging and mandating authors to provide data availability statements and, in some cases, do not publish the article before a link to the data repository is provided or the data is made available within the article or supplementary information. This has caused a huge change within some journals and publishers, with an increasing amount of data being made available (see Fig 1). Indeed, Colavizza et al (2020) found that on average, articles that included a link to data in a repository had a higher citation impact compared to articles which did not make their data available.

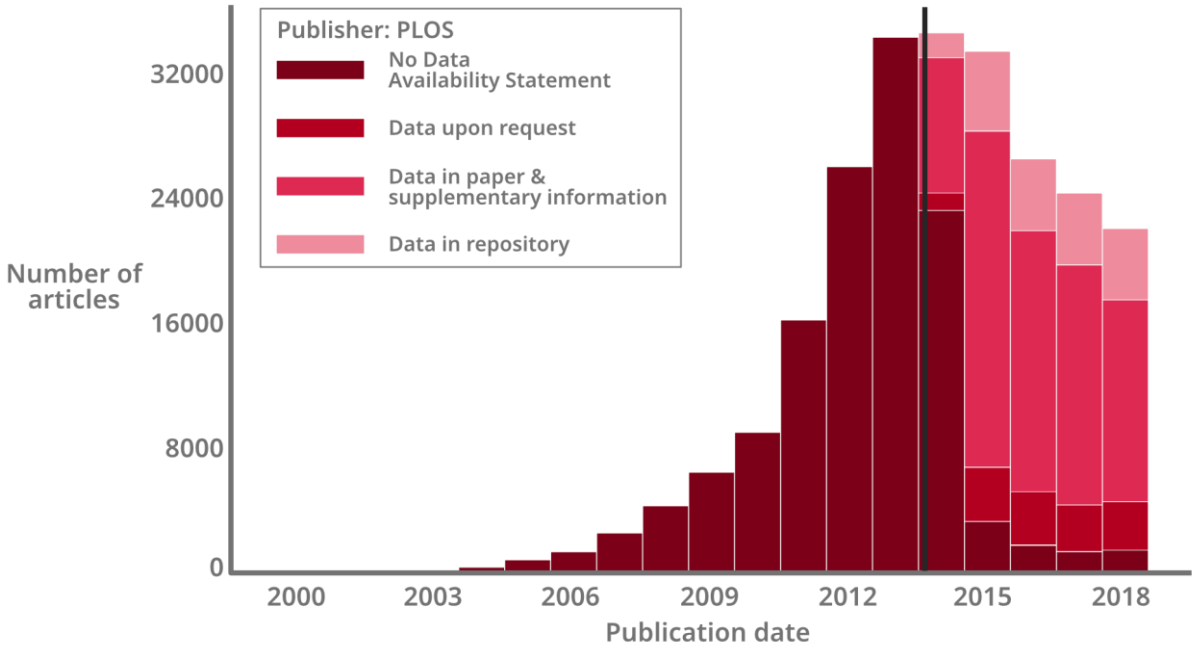


***Figure 1**: Redrawn from Colavizza et al (2020) and shows the number of articles in their dataset between the years 2000 and 2018. The solid vertical black line indicates the time the publisher (PLOS) introduced a data availability statement (DAS) mandate. The articles were classified into four categories: no DAS, data available on request, data contained within the article and supplementary information and a link to archived data in a public repository.*

Surprisingly, the highest number of articles published within the publisher PLOS make their data available with the paper/publisher or the supplementary information. This means that the data will not have a separate DOI assigned to it, and the citation (and DOI) of the paper and the data are the same. Publishing the data in a separate repository will provide the researcher with a DOI for the paper publication, as well as a DOI for the data publication, so has the

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 871081

potential to increase the number of citations and thus enables the researcher to get the deserved credit for their data sets and add them to their CVs. Additionally, datasets published with a repository is easier to find and independently discoverable by other researchers, thus easier to re-use and, in turn, will have a higher number of citations. Another advantage for researchers publishing their datasets is the increasing number of research institutes and universities recognising the importance of open data and its benefits. An increasing number includes published datasets with DOIs as a measured output in their researcher evaluation procedure.

Researchers need to understand that opening up data is not a chore, but indeed has been shown to yield many benefits (Popkin, 2019):

- Catalyse new collaborations
- Increase confidence in findings
- Generate goodwill among researchers
- Satisfaction of contributing to the scientific enterprise
- Giving back something of value to the taxpayers.

Sharing data can be very beneficial, e.g. by offering the chance to be part of a global dataset publication. Those publications bringing together openly shared data sets are often a) well cited and b) published in high profile journals. Usually, everyone who shared their data has the chance to be a co-author of the study, e.g. Pilla et al (2021) published a global data set of long-term summertime vertical temperature profiles in 153 lakes in the Nature journal Scientific Data with all data contributors as co-authors.

However, many researchers worry that by sharing their data, they will not be the first ones to publish a paper on this data (Popkin, 2019). Rest assured, most data repositories allow the data provider to set embargo periods, i.e. the data will only be publicly available on a certain date. For researchers involved in AQUACOSM-plus, it will be of huge benefit to publish their metadata, as well as their datasets as soon as possible after the experiment has ended, as:

1) This will ensure that the data is dealt with shortly after the experiment is ended, and data collection is still fresh in everybody's mind and data quality control is high.
2) It will take less time to write metadata, clean up data and quality control the data if these are done soon after data collection. Furthermore, in the long run, this will be beneficial to the whole team who was involved in the mesocosm experiment.
3) A lot of TA users are PhD students and postdocs, likely to move on within a short period of time after the experiment and data could potentially be lost.
4) Experimental mesocosm data needs to be shared between TA users/researchers from different countries after the experiment has ended, and data is likely being kept in web-

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 871081
Responsibility for the information and views set out in this report lies entirely with the authors.
The European Commission is not responsible for any use that may be made of the information it contains.

based interactive tools (e.g. google docs) where errors could be introduced by accident as there is no version control.

In short, it will benefit the TA user community itself (as well as the global research community, regulators and agencies interested in water quality, in the long run) if the dataset(s) are published as soon as possible.

3. TA Data Management within AQUACOSM-plus

As stated in the grant agreement, primary data (this excludes metadata) should be openly accessible within six months after completion of the publishable dataset, with reasons given if this is not possible. The publishable data set is defined as a dataset that has been subject to processing routines aimed at e.g. Quality Assurance and Quality Control. Reasons for not making the publishable dataset openly accessible may include competitive advantages such as the completion of a PhD thesis, in which case a moratorium of up to three years is granted.

3.1 Metadata Portal and Project Tracker

AQUACOSM-plus collects metadata from TA users via a web-interface after an experiment has ended, as part of the TA requirements. This metadata is publicly available. Both the Metadata Portal and the Project Tracker (which monitors the progress of TA projects from proposal to completion) have the ability to record the DOIs created. As stated in the data management plan (D4.4), WP4 will develop a common strategy on best practices in reducing (meta) data heterogeneity, including a controlled vocabulary for metadata. This will ensure a smooth adaptation into a data repository and thus reduce the time effort needed by the TA users for publishing their dataset with, e.g. PANGAEA, and obtain a DOI for the dataset. The metadata portal will also contain the DOI (if available) and an email will be sent to the data owner every six months to encourage them to revisit the metadata and add any DOIs obtained.

3.2 Centralised Primary Data Portal

AQUACOSM-plus will develop a primary data shell managing primary data directly into an existing database (e.g., PANGAEA, Task 4.3). This will enable project partners to directly archive, quality assure, calculate and investigate their own experimental data in the first place. In addition, systematically collecting data from different experiments and projects in an international data base will promote homogenous quality across data sets and allow cross-ecosystem and multidisciplinary research to be readily performed (i.e., reuse of data sets). Metadata for the data entry will automatically connect to the metadata portal. While this

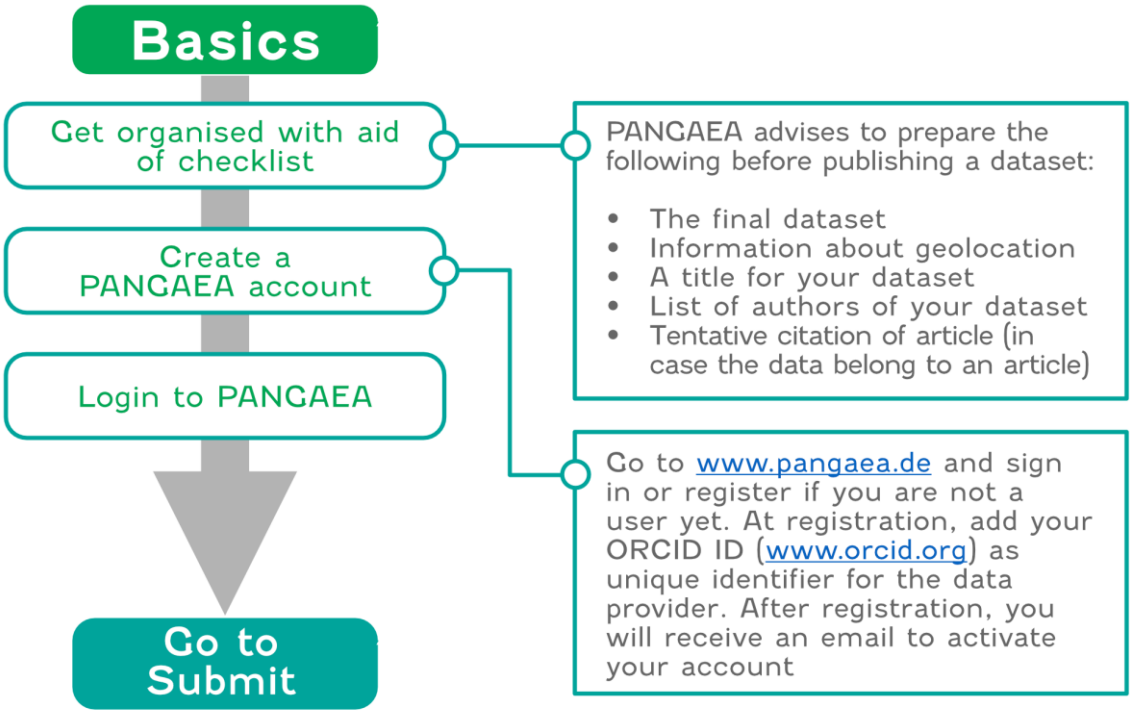
This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 871081

database will eventually lead to a centralised mesocosm open data repository and is of immense importance to project partners and open science, it is important to note that no DOIs will be allocated within this process, but DOIs can be linked to a dataset within the primary data portal.

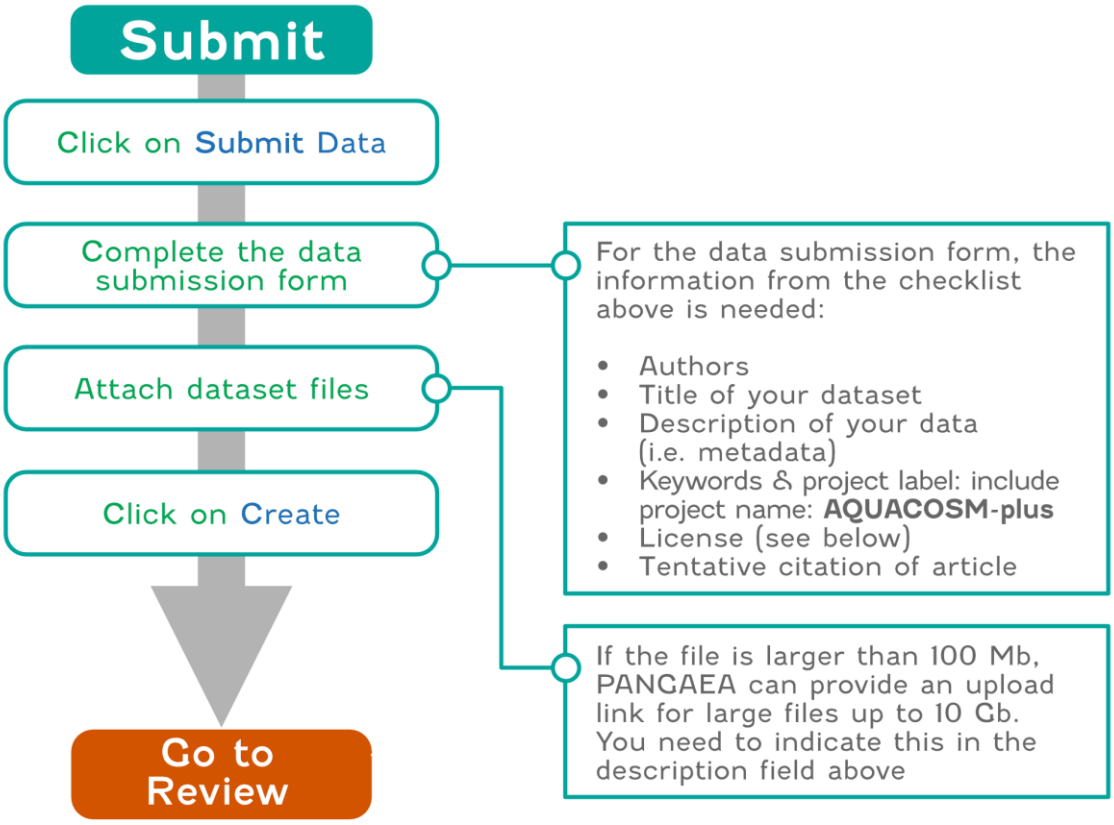
This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 871081

4. Workflow of publishing TA datasets with PANGAEA

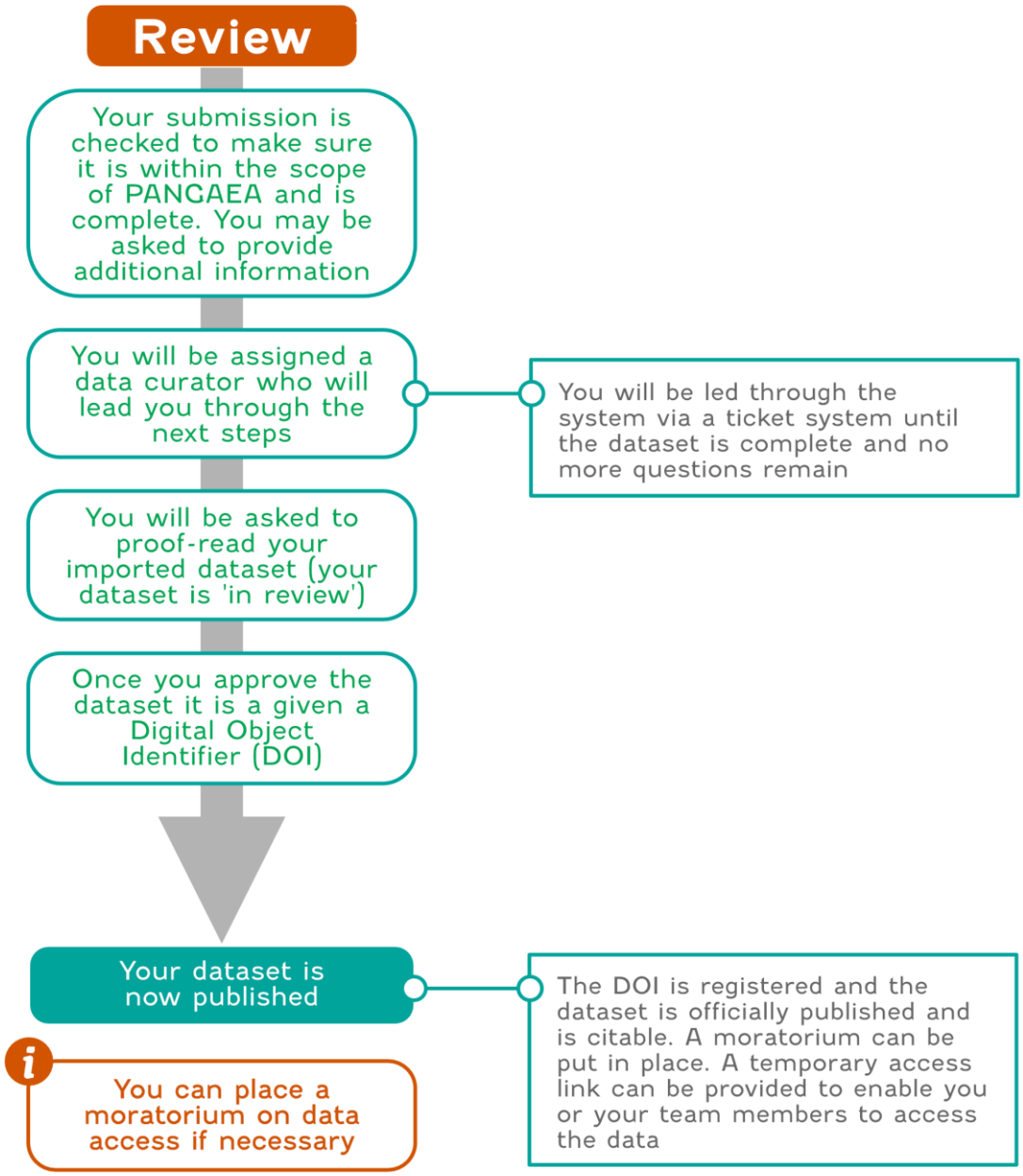
PANGAEA offers a long-term data archiving service as a fully curated data repository and is situated in the EU (Germany). The process of publishing your data with PANGAEA and getting a DOI assigned is straightforward. The figures below outline every step along the way. PANGAEA also produced a video (https://www.youtube.com/watch?v=5bJfSuAukTQ) explaining the steps involved in how to publish FAIR Data and obtain a DOI for a dataset; concurring with the steps outlined below. Metadata should already be available to the data provider from the entry in the project tracker (see above). Whenever data from the AQUACOSM-plus project is published with PANGAEA, the data provider needs to ensure that AQUACOSM-plus, together with the TA project acronym, are mentioned as the project name. This will make it easier to find all project related data submissions using the query function in the search URL on the PANGAEA webpage.



This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 871081
Responsibility for the information and views set out in this report lies entirely with the authors.
The European Commission is not responsible for any use that may be made of the information it contains.



This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 871081
Responsibility for the information and views set out in this report lies entirely with the authors.
The European Commission is not responsible for any use that may be made of the information it contains.



This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 871081

There are different types of open access licenses. PANGAEA recommends the use of the CC-BY license “creative commons attribution 4.0 international” as it makes the data freely available to everyone but usage requires to cite/attribute the original author(s). There are other options available, for further information, please see https://wiki.pangaea.de/wiki/License. The maximum moratorium duration PANGAEA offers is two years, however, after this period there are options for further extensions if the necessity is sufficiently explained and communicated to PANGAEA. The moratorium can be given for various reasons, e.g. a dataset linked to a paper publication, or a dataset from a PhD student to ensure the data is not published before the completion of a PhD thesis. Some AQUACOSM datasets have already been published with PANGAEA and so they have experience with datasets originating from mesocosm experiments
(e.g. https://doi.pangaea.de/10.1594/PANGAEA.928922 or https://doi.pangaea.de/10.1594/PANGAEA.872440).

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 871081
Responsibility for the information and views set out in this report lies entirely with the authors.
The European Commission is not responsible for any use that may be made of the information it contains.

5. References

Colavizza G, Hrynaszkiewicz I, Staden I, Whitaker K, McGillivray B (2020) The citation advantage of linking publications to research data, *PLoS one*, 15(4):e0230416, https://doi.org/10.1371/journal.pone.0230416

Paskin N (2010) Digital Object Identifier (DOI) System, *Encyclopedia of Library and Information Sciences* (3rd ed.), Taylor and Francis, pp. 1586–1592

Pillla R, Mette E, Williamson C, Adamovich B, Adrian R, Anneville O, Balseiro EG, Ban S, Chandra S, Colom-Montero W, Devlin S, Dix M, Dokulil M, Feldsine N, Feuchtmayr H, Fogarty N, Gaiser E, Girdner S, González M, Hambright KD, Hamilton D, Havens K, Hessen D, Hetzenauer H, Higgins S, Huttula T, Huuskonen H, Isles P, Jöhnk K, Keller W, Klug J, Knoll L, Korhonen J, Korovchinsky N, Köster O, Kraemer B, Leavitt P, Leoni B, Lepori F, Lepskaya E, Lottig N, Luger M, Maberly S, Macintyre S, McBride C, McIntyre P, Melles S, Modenutti BE, Müller-Navarra D, Pacholski L, Paterson S, Pierson D, Pislegina H, Plisnier PD, Richardson D, Rimmer A, Rogora M, Rogozin D, Rusak J, Rusanovskaya O, Sadro S, Salmaso N, Saros J, Sarvala J, Saulnier-Talbot E, Schindler D, Shimaraeva S, Silow E, Sitoki L, Sommaruga R, Straile D, Strock K, Swain H, Tallant J, Thiery W, Timofeyev M, Tolomeev S, Tominaga K, Vanni M, Verburg P, Vinebrooke R, Wanzenböck J, Weathers K, Weyhenmeyer G, Zadereev E & Zhukova T (2021) Global data set of long-term summertime vertical temperature profiles in 153 lakes, *Scientific Data*, 10.1038/s41597-021-00983-y

Popkin G (2019) Setting your data free, *Nature*, vol 569, pp 445-447, https://www.nature.com/articles/d41586-019-01506-x

Wikipedia (2021) Digital object identifier, https://en.wikipedia.org/wiki/Digital_object_identifier

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 871081
Responsibility for the information and views set out in this report lies entirely with the authors.
The European Commission is not responsible for any use that may be made of the information it contains.