# **mARS tutorial**

September 2018

The microbial Antarctic Resource System (mARS) is an online platform where Antarctic (or other polar/alpine) microbial datasets of research projects are archived and made discoverable for the scientific community.

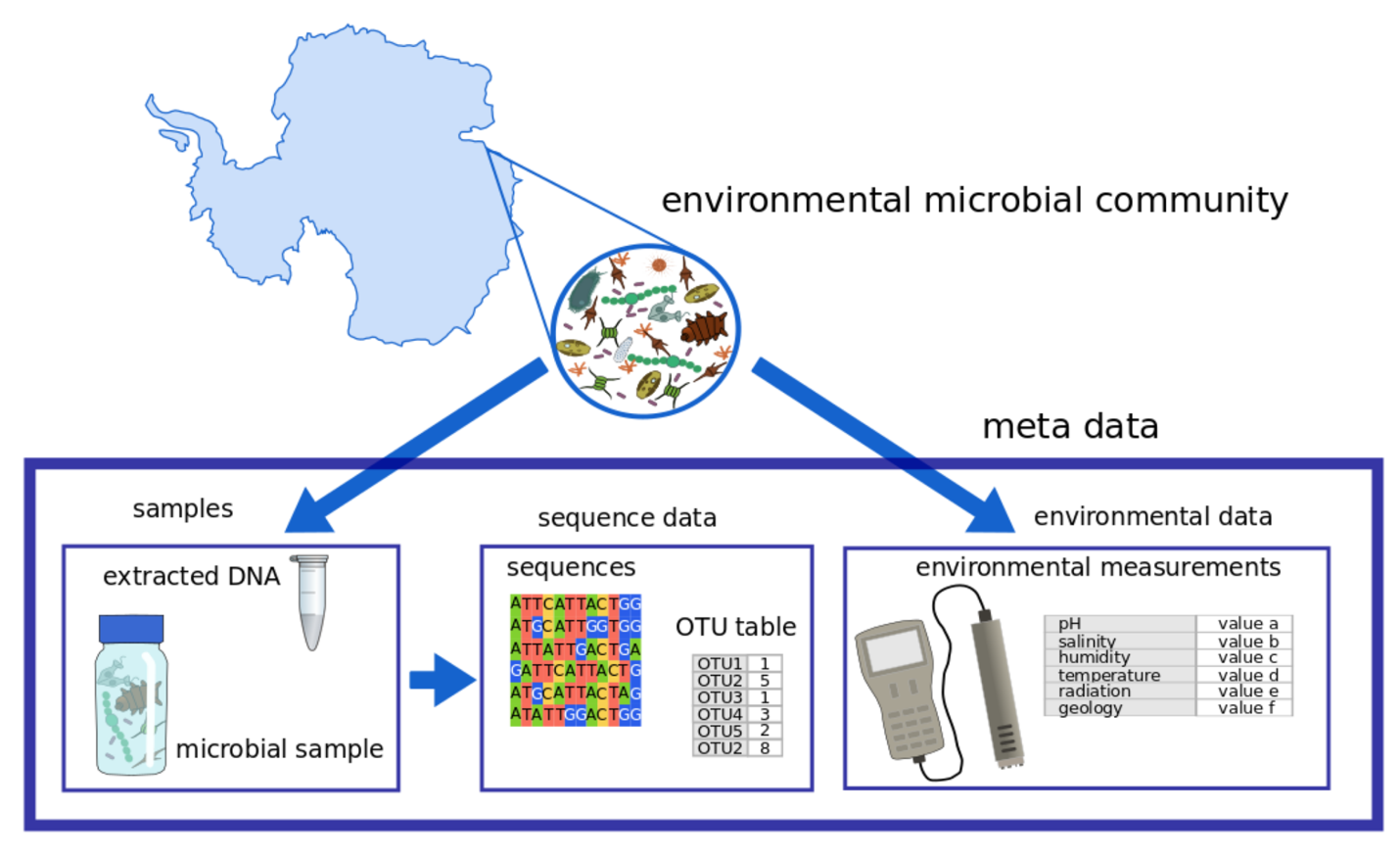
**How to use these tutorials**

The best way to use these tutorials, is to follow them step by step while simultaneously running through the protocol with your own dataset (i.e. the metadata and environmental data).

**Before you start**

* Send an email to request a username and password from the [IPT administrator](mailto:%20nyoudjou@gmail.com)
* Get a google login if you don’t already have one - to access and copy the two template files that you need. The templates - and precise data formats that are required are defined and explained in these templates.

**mARS data structure**

****

The information that is gathered during a scientific project can be divided into different classes of data, each having their own digital requirements, utilities and limits. In the case of microbial datasets, three classes of data are relevant for the mARS platform, and all originate as information extracted from a physical sample (e.g. a soil sample, enrichment culture,…)

In the case of microbial datasets, three classes of data that are ultimately gathered from a physical sample (e.g. a soil sample, enrichment culture,…) are relevant for the mARS platform. First, there is the meta data. This is all contextual information surrounding the sample, such as the expedition when the sample was taken, the person who took it, and for what project. Second, the environmental data, encloses all chemical and physical measurements that were taken to characterize the conditions under which the sample was originally found. These can include geographic coordinates, temperature, pH, ion measurements, etc. Finally, the sequencing data encloses all DNA, RNA or protein sequences (i.e. the DNA, RNA or protein molecules who’s composition was read using laboratory analyses), which typically serves as a proxy for the community composition or genomic/transcriptomic/proteomic makeup of the microbial organisms that were found in the sample.

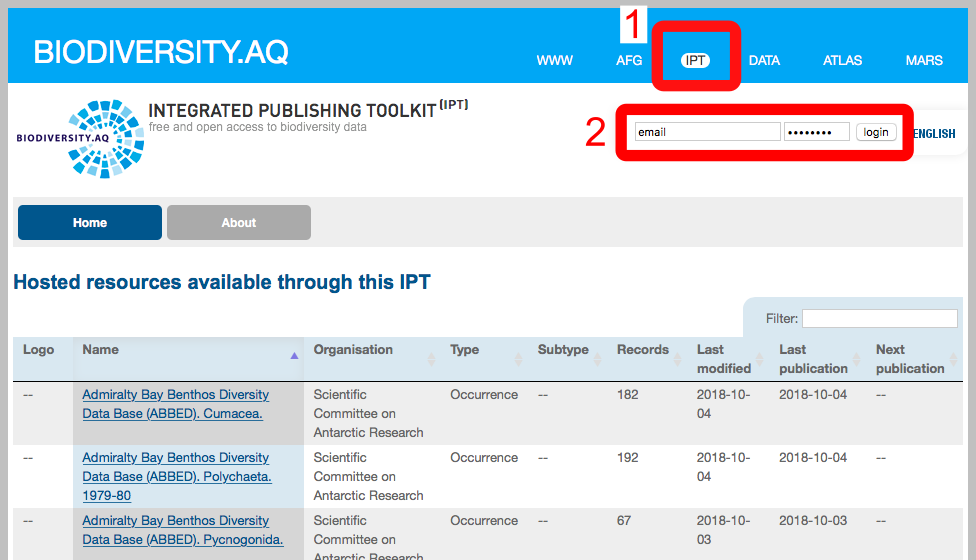
**Tutorial 1: metadata documenting your project**

The aim of this tutorial is to document your dataset, by providing all the necessary metadata in the DarwinCore standardized terminology framework.

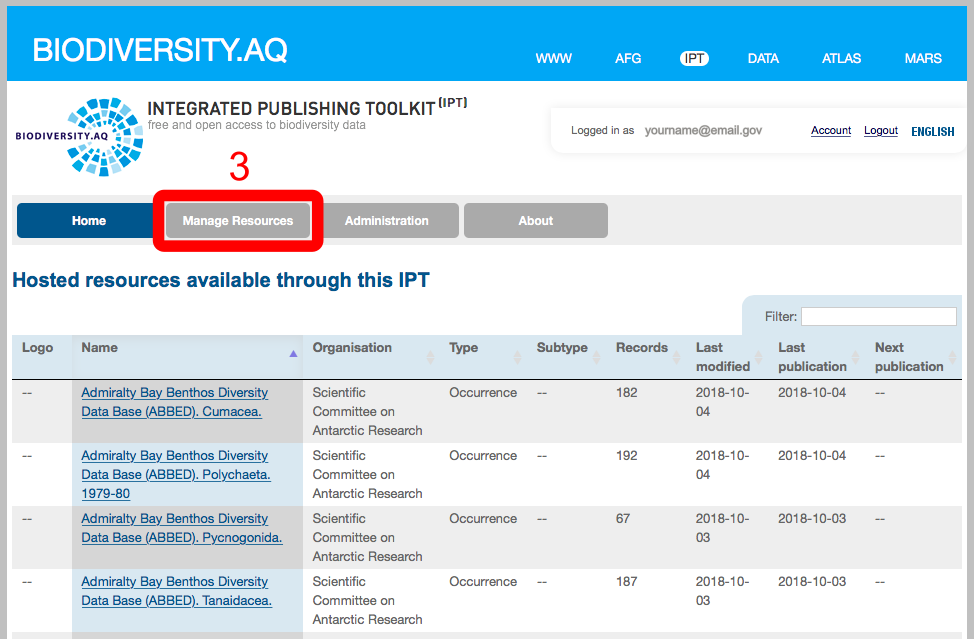
Metadata refers to any information about the data (e.g. who took the samples, when, for what project, what phylogenetic group was targeted, etc.). Strictly speaking, metadata, does not include the actual sequence data or environmental parameters that were measured.

To fill in your metadata using DarwinCore terms, we will use the Integrated Publishing toolkit (IPT) that was developed by GBIF.

To go there, surf to ipt.biodiversity.aq or click on the IPT icon at the top right of the mARS webpage. Next, login with your email address and your credentials. If you are not yet registered with mARS, please send an email to request a login and a password to one of the mARS administrators.

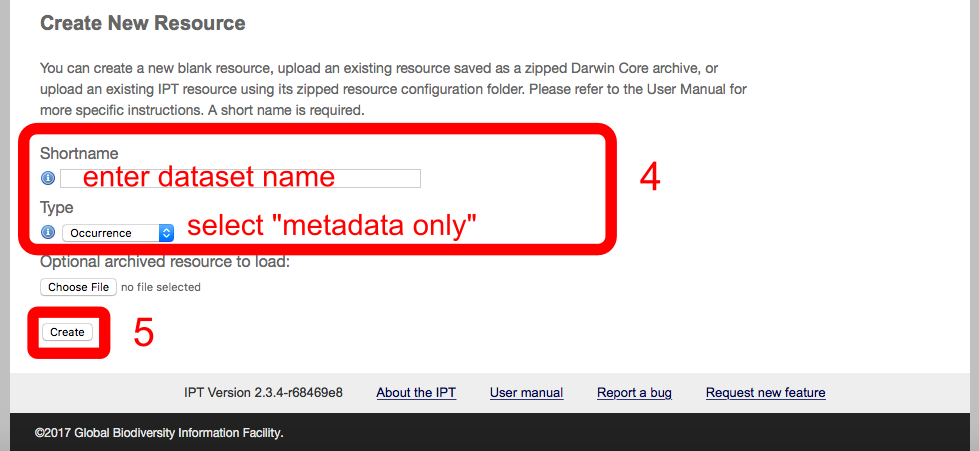


After logging in, the “Manage Resources” tab will appear at the top of the page. Click on manage resources.

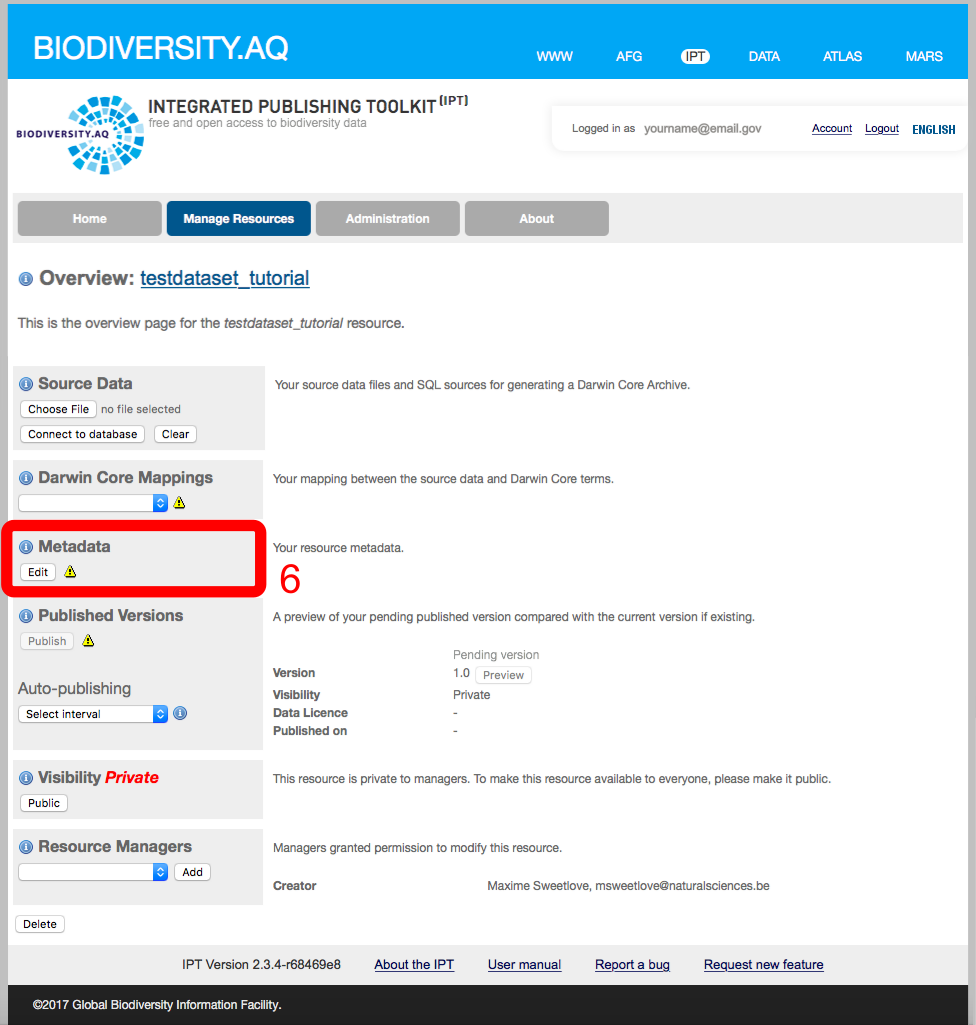


A new dataset entry (also referred to as a ‘resource’) can be created by scrolling down to the bottom of the “Manage Resource” page. Provide a unique "shortname" for your dataset. Because GBIF and DarwinCore archives are at this point only designed to take occurrence data of species, and not environmental or molecular diversity information that is commonly at the core of microbial datasets, “Metadata only” or “Other” has to be selected under the “Type” header.

Click on the “Create” button (step 5). This will direct you to the “Resource Management” page of your dataset.

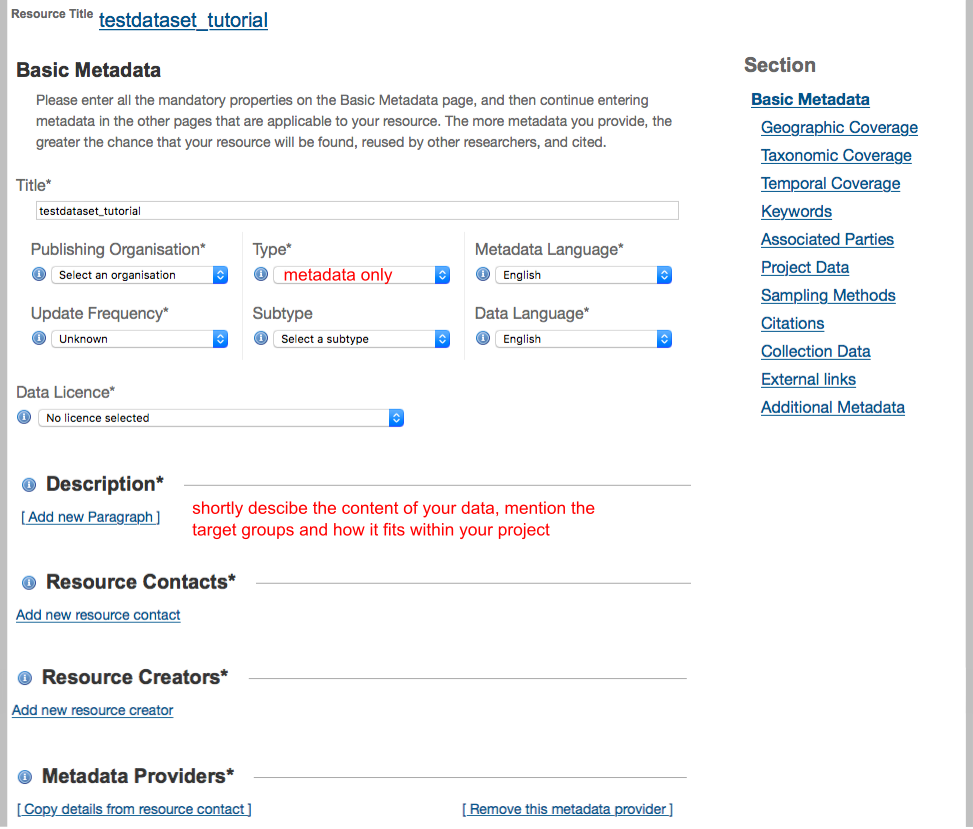


On the “Resource Management” page, click on the “Edit” button in the Metadata section (step 6). There, twelve sections need to be filled in to provide all the details for the different metadata sections can be filled in to document your project.



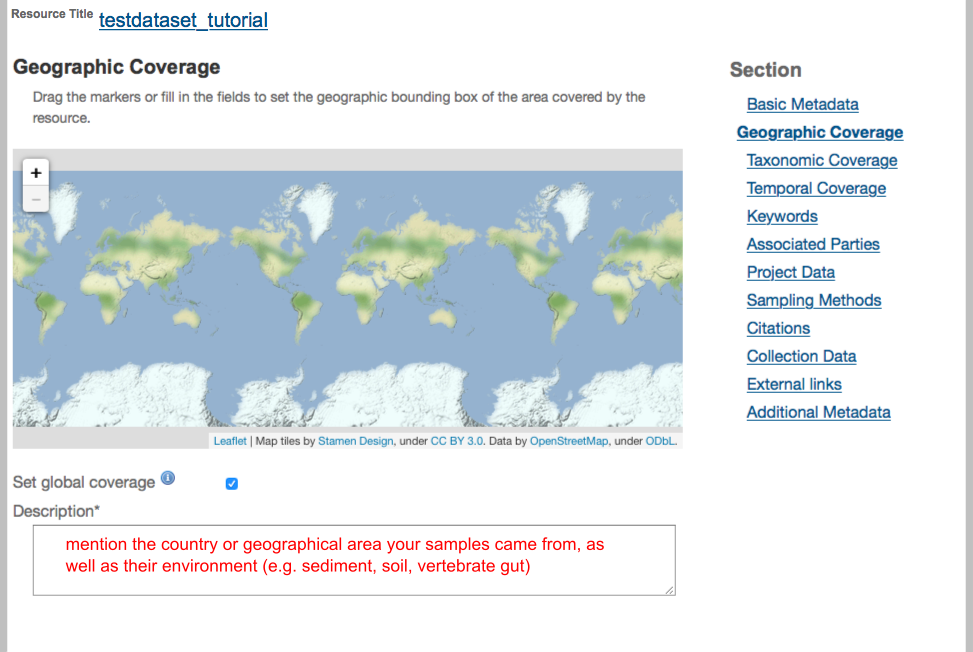
Next, the IPT website will guide you step by step through the procedure of documenting and publishing your metadata. The section below runs through the main pages, en provides additional information. Highly detailed instructions are also available from [IPT quick reference guide](https://code.google.com/p/gbif-providertoolkit/wiki/IPT2ManualNotes#Quick_Reference_Guide). “ \* ” signifies fields that need to be filled out. Don’t forget to save the queries regularly, which can be done at the bottom of each page.

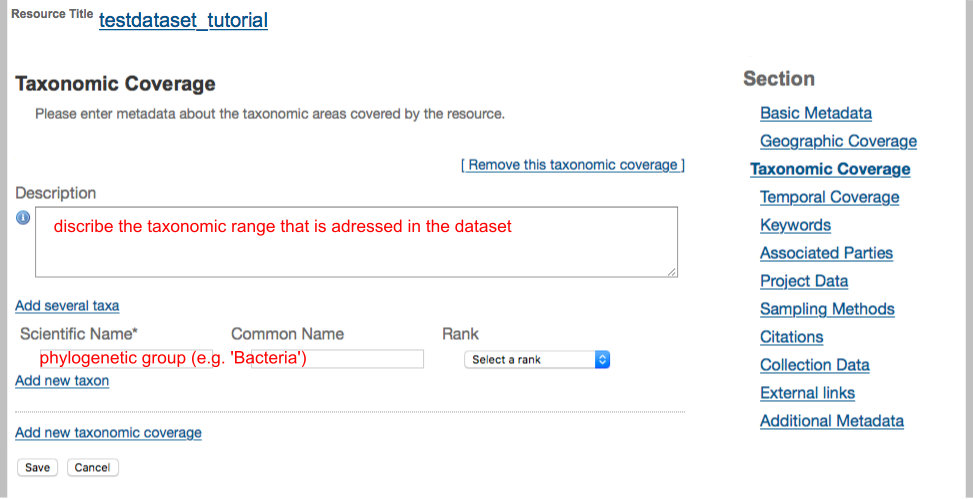
EXTRA TUTORIAL: filling in the IPT



=> getting help

Also see “Section” menu on the right

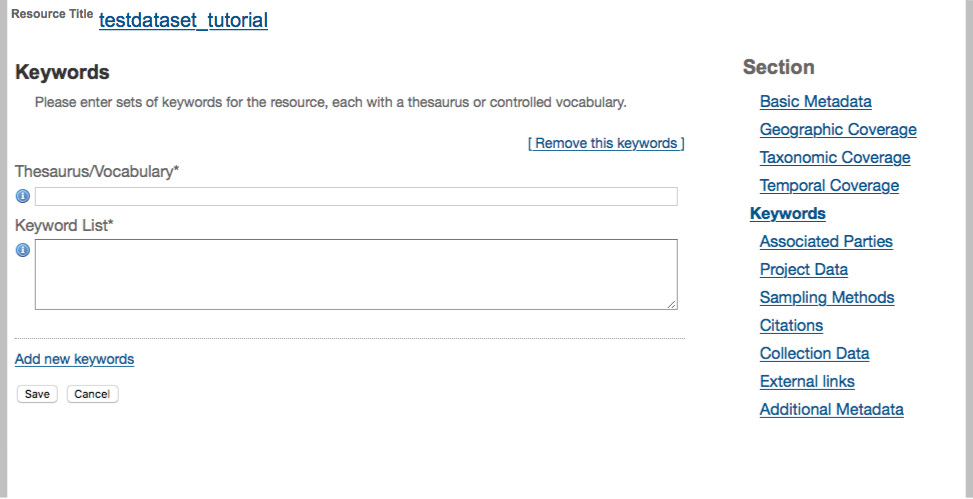


**

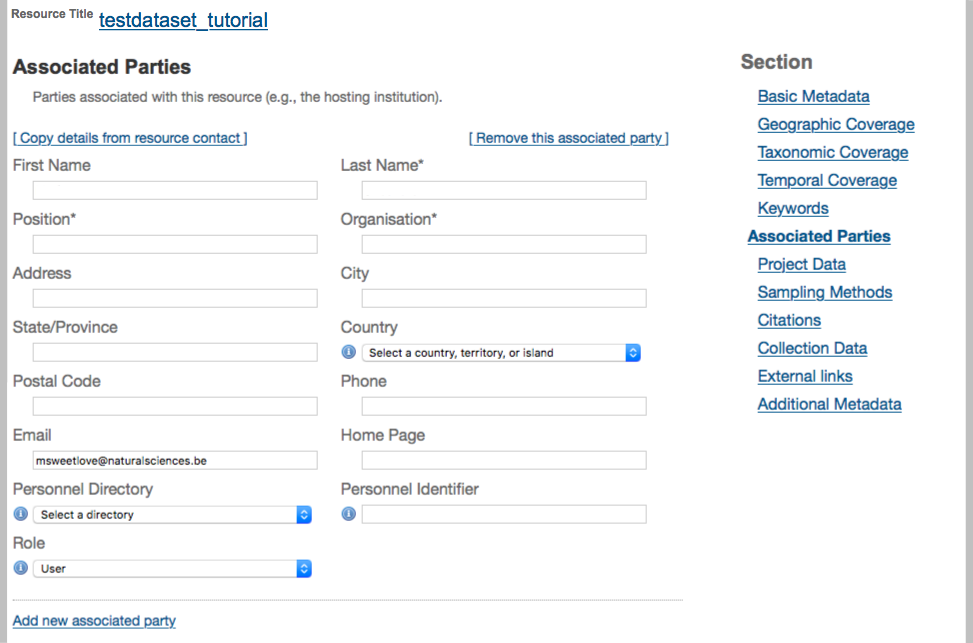
On the Temporal Coverage sheet, you can either enter specific sample seasons or dates if you have few samples, or if a long term project, then enter the complete date range.

******

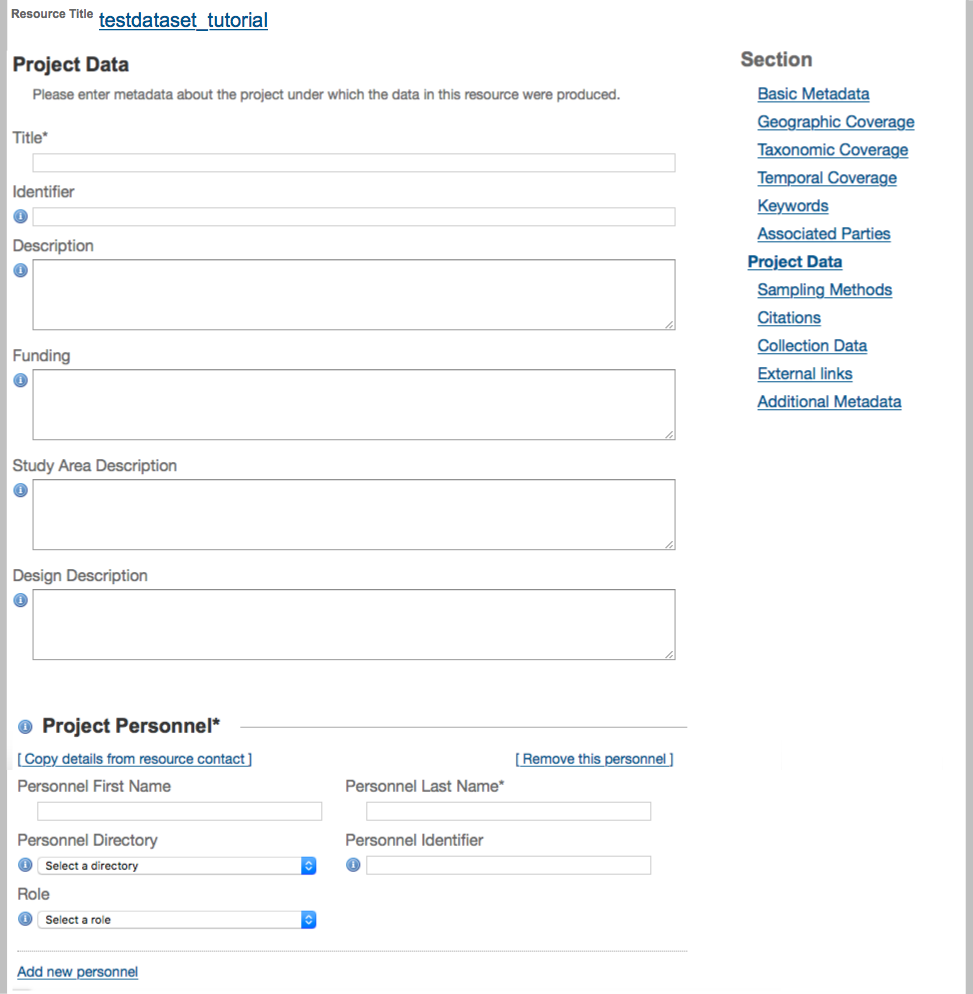
Keywords: You can enter “n/a” in the Thesaurus Vocabulary box. Enter Keywords that can be used to find your project. This will be searchable in GBIF among other search engines.



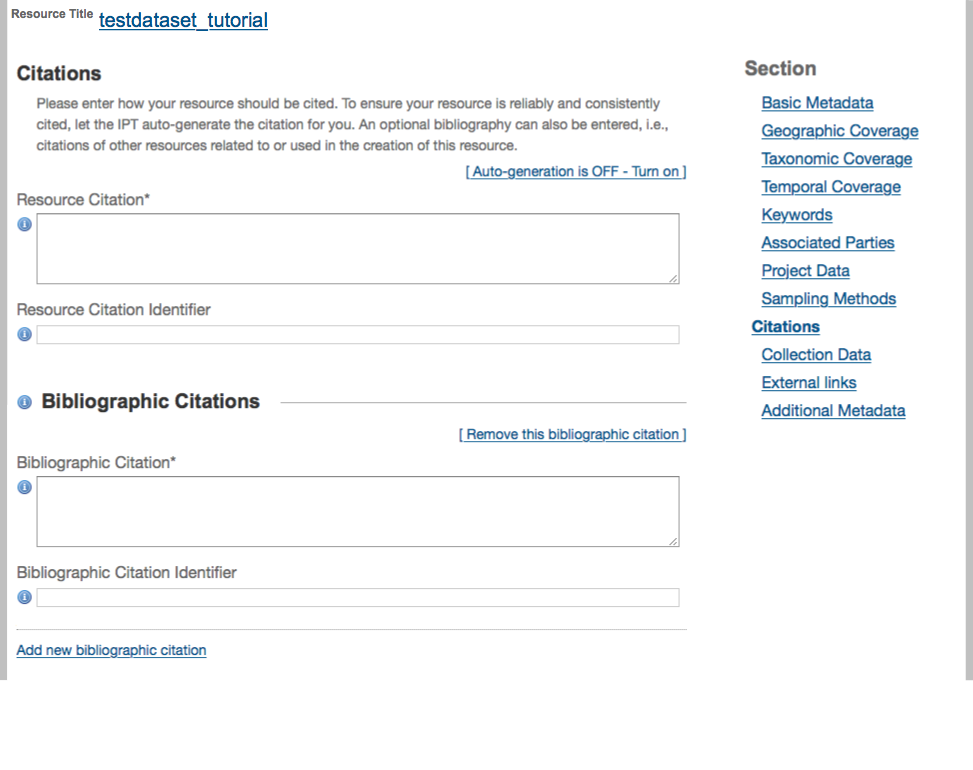
In the “Associated Parties” sheet, persons or institutions can be mentioned that originally gathered the data, authored publications in which that dataset featured, or owns the dataset (see drop menu under “Role”).



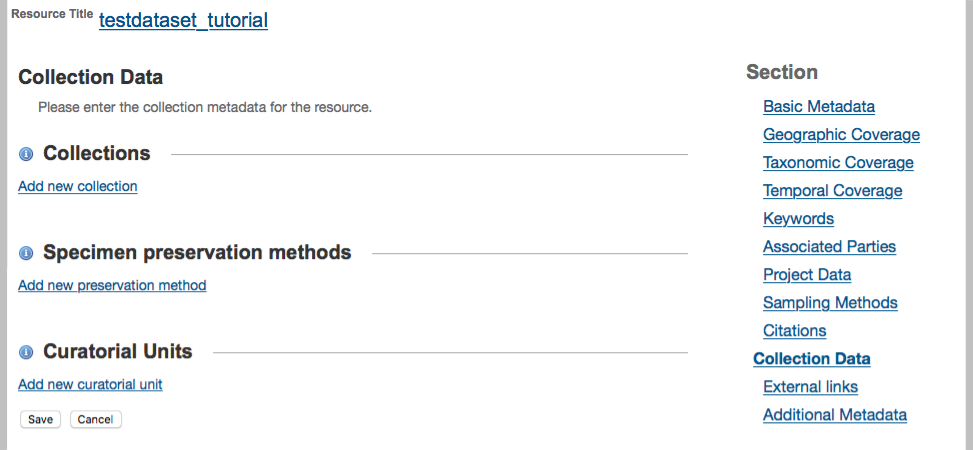
Grant numbers and other important project information can be mentioned in the “Project Data” section. This also allow mARS to link the resource to relevant projects in the GCMD/AMD. For projects that are in their early stages, this information can help increase the visibility and discoverability of a project, and can be interesting for further collaboration with other research groups.



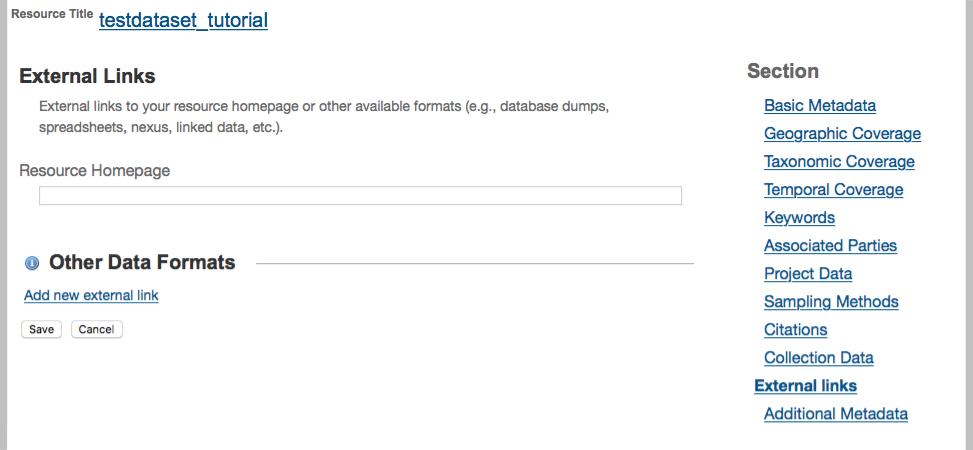
Under “Citations”, the correct reference to the dataset can be provided. This can also be auto-generated using the information provided in the previous sections. The “Bibliographic Citations” section can hold references to any publication where the dataset was used.



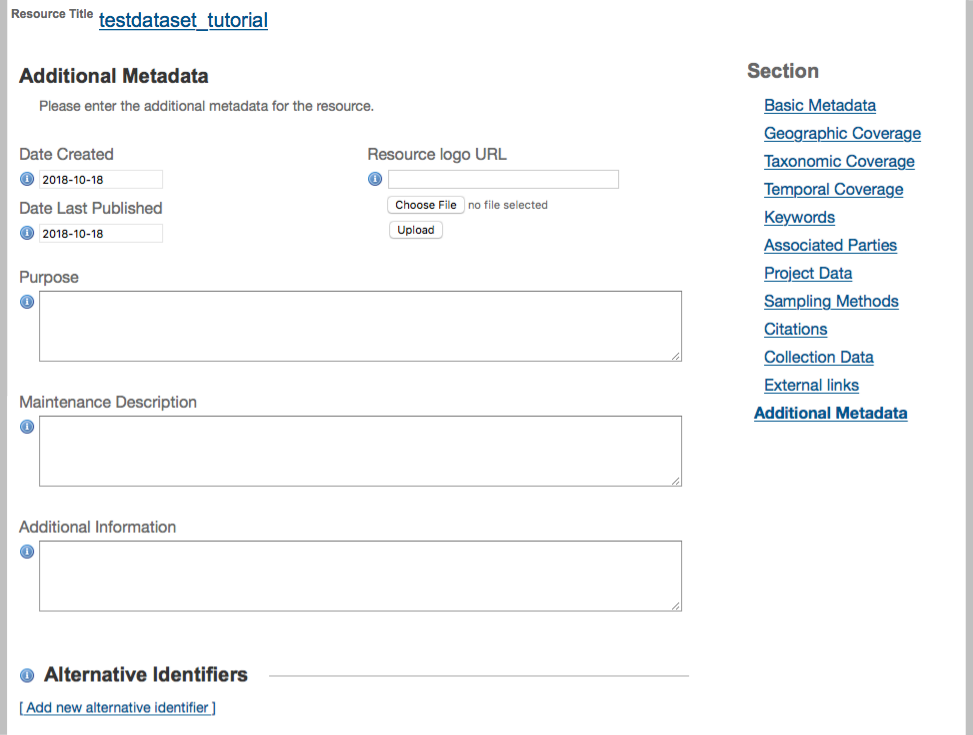
The terms under the section “Collection data” don’t apply to the microbial data sets perfectly; though you can still enter relevant terms for your data. This can, for instance be relevant if (sub-) samples have been archived in your lab or in the DNA archive in New Zealand. Enter Not Applicable where appropriate.



“External Links” is a good place to put the URL for a project website if you have one.

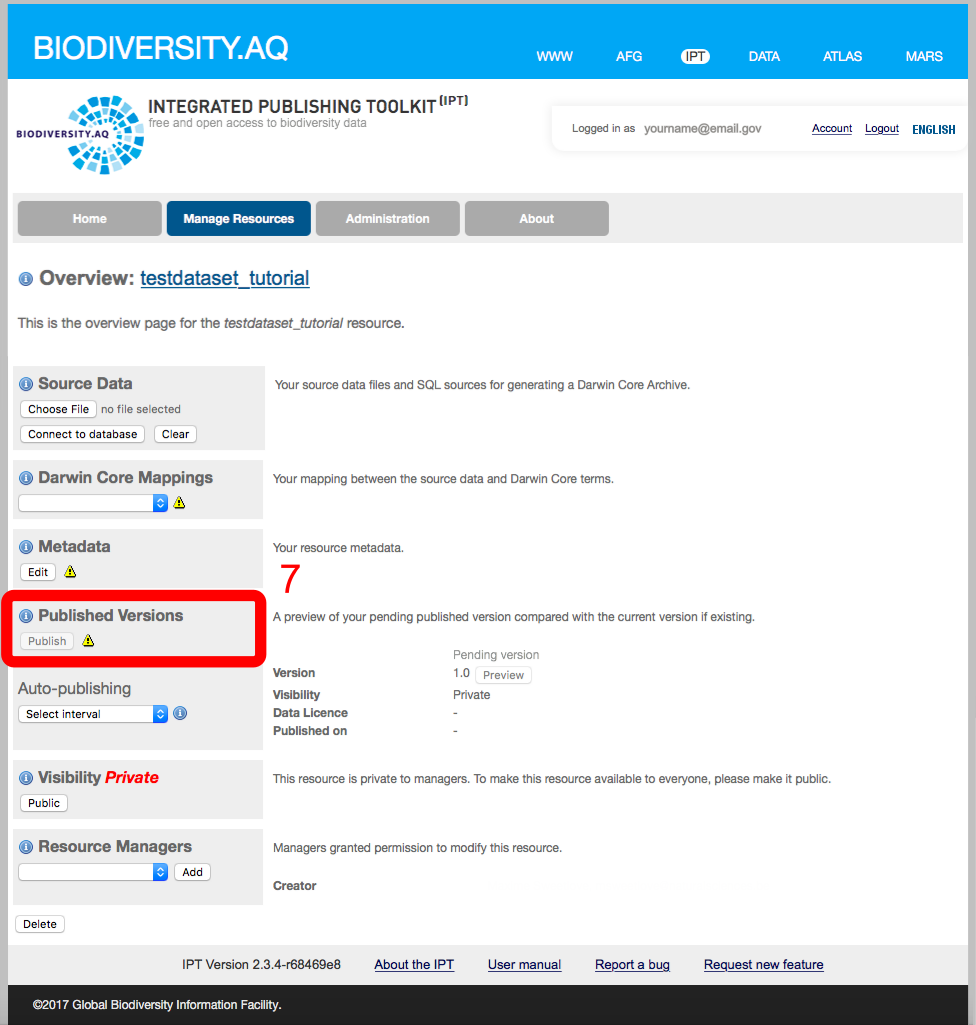


If your data set is registered in other data bases (e.g. Antarctic Master Directory; GenBank or the Sequence Read Archive, BCO-DMO) then that information can be entered in the “Additional Metadata” section.

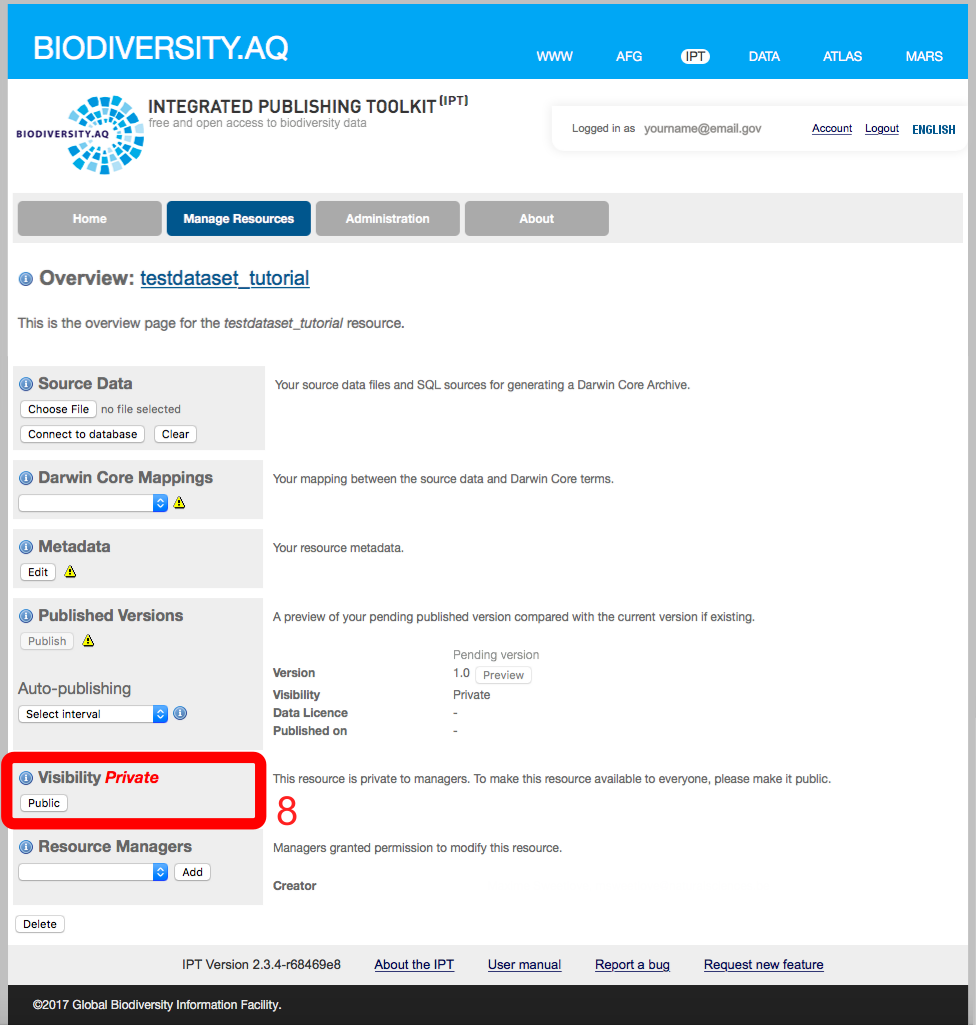


After filling in all the metadata about your dataset, go back to the IPT website by clicking on your resource title.

Click on the “Publish” button in the “Published release” section on the left of the page. This action will upload your project metadata to the MARS portal. Do not worry when you see a warning message “Source data or Darwin Core mappings missing. No data archive generated”. Please note that this is perfectly normal if you are publishing a metadata-only resource. Generating a DarwinCore archive requires also to upload file with occurrence data that is mapped to DarwinCore terms, and is combined with the metadata. This, however, is currently not yet possible for microbial data.



By default, your resource’s visibility is set to “*Private*”. To allow your resource to become visible on the IPT for all users, click on the “*Public*” button in the “Visibility” section.



After entering all the metadata, send an email to one of the mARS data administrators to request final “registration” of your data set. Registration differs from publishing in that this will allow your metadata to be discoverable through online biodiversity information networks, including GBIF. Your data set receives a stable, unique identifier on the web.

**Tutorial 2: providing the environmental data in the MiMARKS template (for high throughput sequencing of marker sequences)**

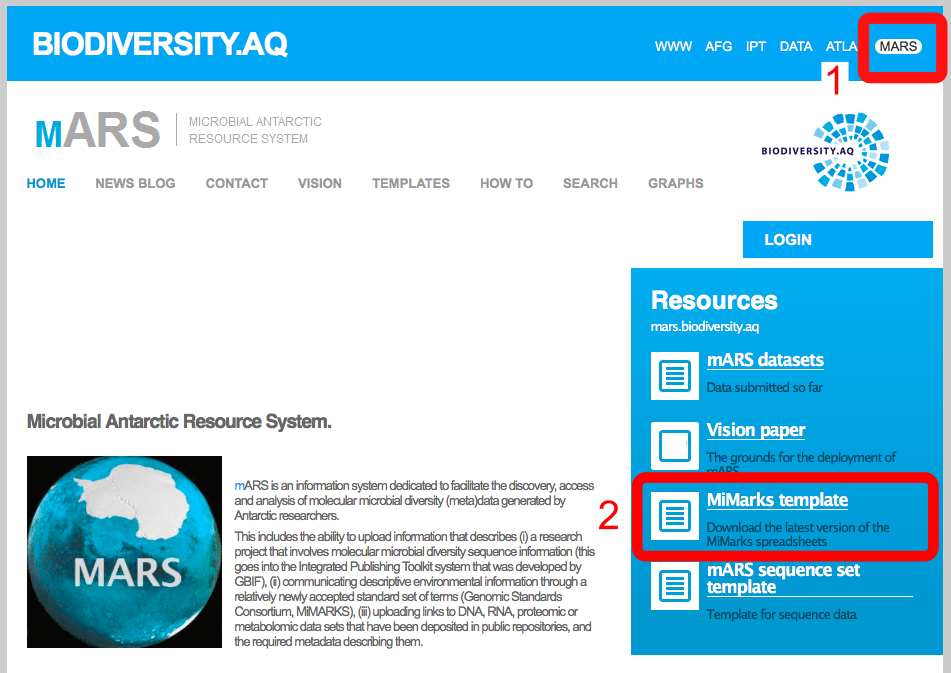
The aim of this tutorial is to document your environmental data and ancillery measurements using the Minimum Information about any Sequence (MixS) framework, and more specifically, the Minimum information on MARKer Sequences (MiMARKS) specifications.

The MiMARKS specifications were designed by the Genomic Standards Consortium (GSC) to capture the most fundamental (hence ‘minimum’) information that should accompany a set of sequences (Yilmaz et al., 2011). MiMARKS also allows to document any other ancillary information (e.g. environmental parameters) that are available using standardized units and therminology.

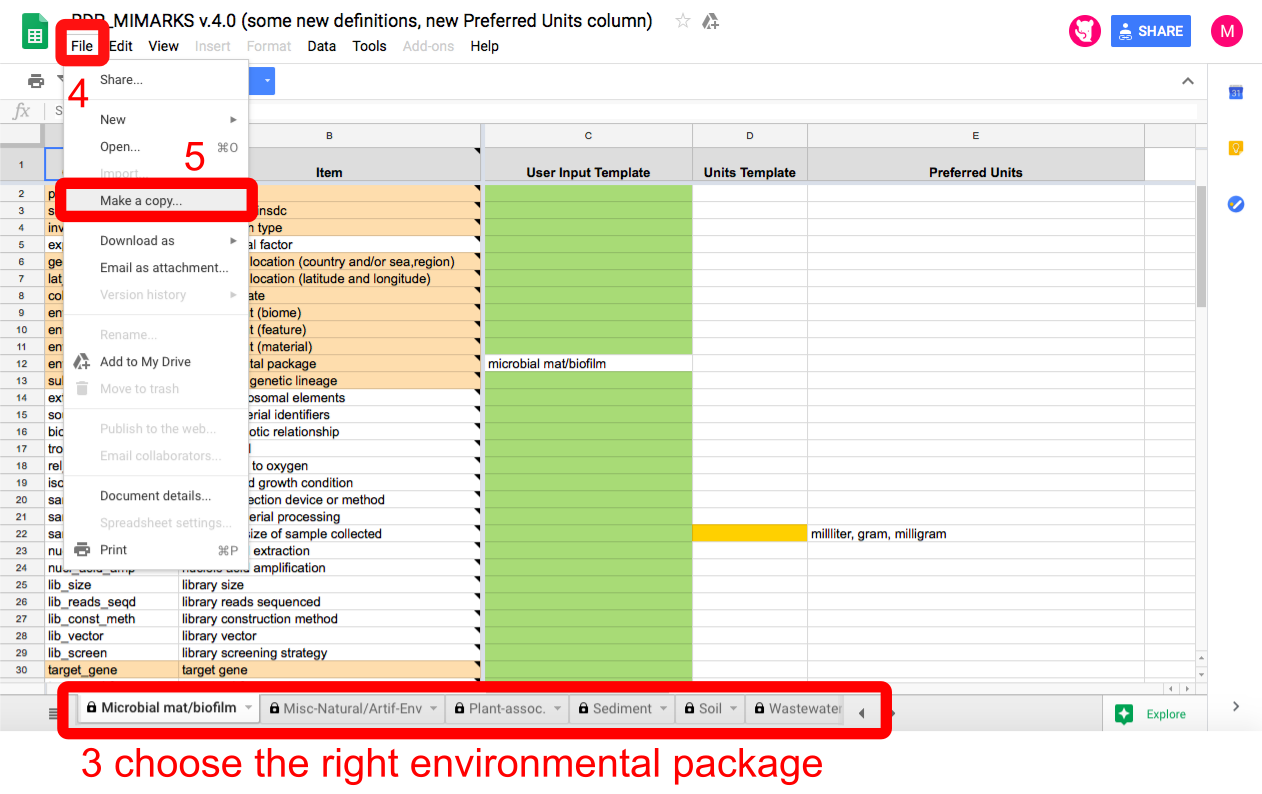
There are several slightly different variants of the MiMARKS template, which represent different environments from which your samples can be taken (e.g. soil, lake water, human skin, etc.). These variants are referred to as ‘packages’, and differ in some terms that are unique to the environment it represents. They can be found as individual spreadsheets in the MiMARKS file. Only the spreadsheet of the most appropriate package for your dataset needs to be completed.

Note that most public sequence repositories also require a completed MiMARKS file to accompany the sequences

To access the MiMARKS template at biodiversity.aq, click on the MARS icon at the top right of the webpage, and click on “MiMARKS template” on the bottom right. For this, a google account is required.



To fill in the MiMARKS template, **click** on “Make copy” from the “File” menu to create a local copy, or download it as a Comma Separated File (CSV) by clicking on “Download as” in the “File” menu. Make sure to do this with the correct environmental package for your data (step 3). If you cannot access Google Documents, get in touch with us.



Some help filling in the MiMARKS form:

MiMARKS uses standardized onthologies, which means that for some fields, only a limited number of stringently defined terms can be provided as user input. For every field, the desired or allowed input can be checked by clicking on the explanation in the “item” column. User content should be provided in the green column (“User Input Template”), while data from different samples can be filled in different columns, one next to the other. At the end, the “Units Template” can be used to specify the measurements units where necessary (i.e. yellow fields).

Fields that are colored in a light-orange denote mandatory terms, and represent the minimum information requirements. White fields can be used to provide additional information whenever applicable or available (e.g. environmental parameters), but do not need to be filled in.

Fields can be filled in by copy-pasting values or columns from your original environmental data file. Note that the format of MiMARKS at this point seems somewhat counter-intuitive, with samples as columns, and parameters as rows. Therefore, feel free to transpose the MiMARKS template to better fit the way your environmental data is stored (which is commonly a matrix with samples as rows, and parameters as columns).

Concerning the data formats for each field:

* use “.” as decimal separator
* avoid non-numeric and non-alphabethic characters (e.g. /,-,+,\*,\ etc.) if not explicitly specified in the description of the field.
* An underscore (“\_”) is preferred to a space (“ “) if not explicitly specified.

Some info on the mandatory fields:



Provide a short name of acronym of the project for which the data was acquired. This is not a unique identifier, and can be the same for different samples or replicates.



Specify (TRUE/FALSE) whether the sequence data has been submitted to any of the public databases ( insdc = International Nucleotide Sequence Database collaboration, e.g. Genbank, SRA, ENA,…)



Specify the main type of organisms or organelles hat was targeted. The allowed terms for this field are restricted, and can only be eukaryote, bacteria, virus, plasmid, organelle, metagenome, mimarks-survey, or mimarks-specimen.



Name of the geographic location (ocean, country, island), as accepted by INSDC (International Nucleotide Sequence Database collaboration) or the GAZ (Gazetteer) onthology. The INSDC terms list can be consulted at <http://insdc.org/country.html>. To access the GAZ onthology, go to <http://purl.bioontology.org/ontology/GAZ>, click on the “Ontologies” tab in the blue top menu, and search for the ontologies that are related to geography (e.g. the ontology of geographical regions)

Note: not all locations may be present in the list, in that case, there are three options: 1) take the closest match, 2) use the name of the mother-country, 3) propose a new ontology (e.g. by pressing the “Submit New Ontology” button in GAZ).



Latitude and longitude values that define the sample location (only numeric), separated by one space. The values should be reported in decimal degrees following the WGS84 system. Use the fines scale resolution as possible.

Note that: latitude values in the Southern Hemisphere and longitude values in the Western Hemisphere are negative in this notation system. Also note that this differs from how latitude and longitude is noted in the NCBI MiMARKS specifications.



The date, noted in ISO 8601 standard (YYYY-MM-DD or YYYY-MM-DDThh:mm:ss, e.g. 1990-11-05T11:32:55). Use the fines scale resolution as possible.



The environmental biome, describing the broad ecological context of a sample, as defined by the EnvO ontology. These terms are based on biological factors such as plant structures, leaf types, plant spacing, and physical factors like climate. Examples include: desert, taiga, deciduous woodland, or coral reef. The EnvO biome terms can be found at https://www.ebi.ac.uk/ols/index, by searching for the term “biome”.

Alternatively, at the EnvO site, you can also go to the page of the “biome” term (ENVO:00000428), and browse through the children of the “biome” term by clicking on the “+” signs in the interactive tree on the left side of the page.



Specify more small-scale environmental features that characterized the sampling site. Here too, only terms from the EnvO ontology are accepted. Examples include: harbor, cliff, or lake.

EnvO terms can be found at <https://www.ebi.ac.uk/ols/index>. Appropriate terms for this field may be found under “environmental feature” (ENVO:00002297), “hydrogeographic feature” (ENVO:00000012), “glacial feature” (ENVO:00000131), or “construction” (ENVO:00000070). You can browse through the sub-terms (children) of each term by clicking on the “+” signs in the interactive tree on the left side of the page.



Environmental material is the material that (partially) make up the sample, or in which the sample was embedded prior to the sampling event.

Only terms from the EnvO ontology are accepted. Examples include: air, soil, or water.

EnvO terms can be found at <https://www.ebi.ac.uk/ols/index>. Appropriate terms for this field may be found under “environmental material” (ENVO:00010483) or “material” (NCIT:C48187). You can browse through the sub-terms (children) of each term by clicking on the “+” signs in the interactive tree on the left side of the page.



The environmental packages field should be automatically filled in by selected the approprate package in the MiMARKS template. This field should have the same value for all samples within a package.



In the subspecific genetic lineage field, you can specify information about the most fine scale taxonomic resolution of the target organisms in the sample. When sequencing was done on cultures, this can be biovar (i.e. physiological differences of the strain compared to other strains of the genus/species) or serovar (antigenetic properties that differ between strains) information. When multiple species/taxa were targeted in the sample, then the finest scale taxonomic information should be provided, such as the kingdom, phylum or genus of that was targeted. Examples include: ‘Eukarya’, or ‘Bacillariophyta’, or ‘Aulacoseira’.



The name of the gene (or gene region /locus) that was targeted. Examples include ITS, 16S V3-V4, rbcL, etc.



Sequencing method used. This should be as complete as possible, and can also include the sequencing platform and sequencing chemistry used. Examples are: Sanger, pyrosequencing, Illumina MiSeq 300bp PE, Illumina HiSeq2500 125bp PE v4 chemistry, etc.

Once you are finished, download your spreadsheet as a CSV (Comma-separated Values) file on your computer.

Next, mail the Mimarks and Microbial Sequence set CSV files to the mARS team. We will perform a brief quality check and archive the data in our database. Please also indicate whether the data can be made public or not (default is not public).

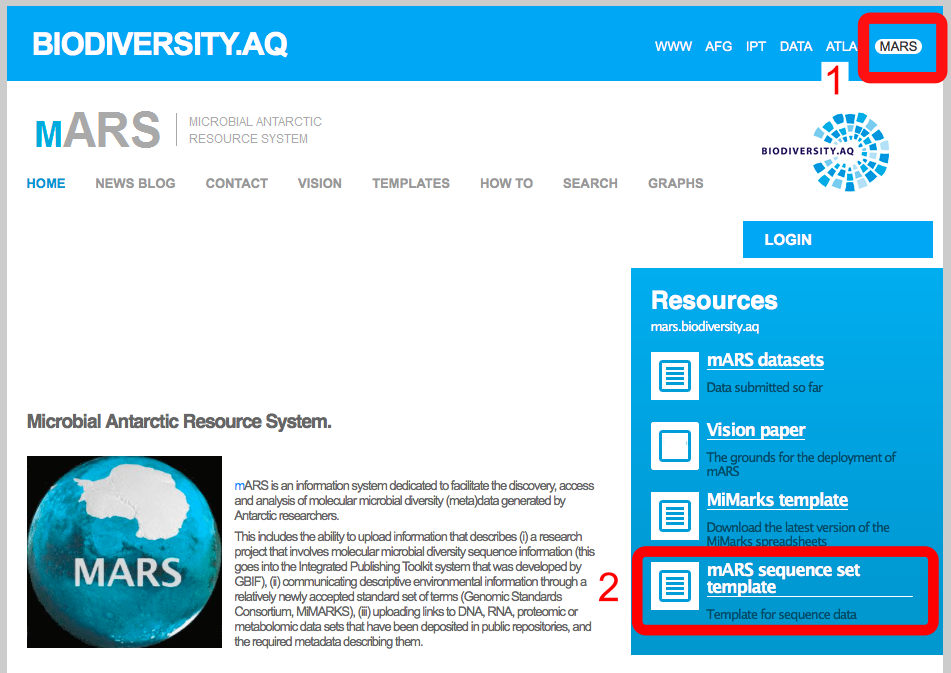
**Tutorial 3: filling in the sequence set template**

The aim of this tutorial is to provide additional information on the sequencing data and its location on public repositories. This way, the sequencing data and environmental data can be linked, and users of the mARS website can find the sequences by querying the metadata or the environmental data.

As this information is currently not presented in the MiMARKS format, we designed a separate file (‘sequence set template’) in which this information can be provided.

We define a sequence set as the digital library of sequences that originated from a single environmental sample. So, a sequence set corresponds to the sequences obtained from one DNA extract of a sample.

The mARS sequence set template can be found at biodiversity.aq, by clicking on the MARS icon at the top right of the webpage, and following the link to “mARS sequence set template” on the bottom right. For this, a google account is required.

****

In the sequence set template should be filled in analogous to the MiMARKS template, with entries per sample (i.e. sequence set) in each column. In total, 11 fields are mandatory and are highlighted in orange.



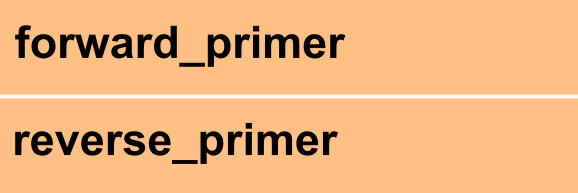
A unique identifier for your sample. For amplicon sequencing, this can be your SRA sample identifier.



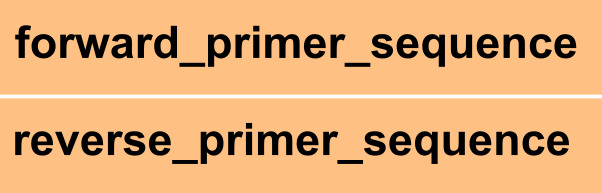
Fill in one of the following terms: marker gene, genome, metagenome, metatranscriptome, metaproteome, metabolome.



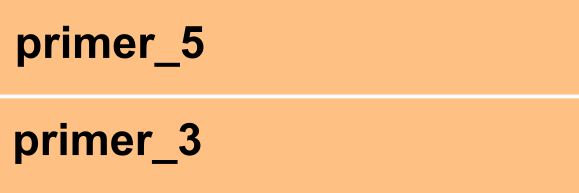
Same as Investigation\_type in the MiMARKS sheet. The expected input is one or more names of taxonomic groups (e.g. Bacteria, Nematoda,…)



Official name of the forward and reverse primers



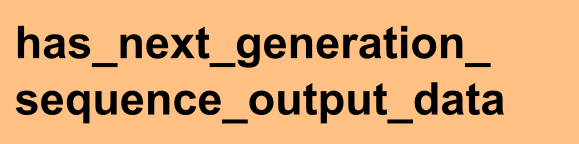
Forward and reverse primer sequence (written in 5`-3`format)



If available: sequence position of the the base that follows the last base (3' end) of the 5' primer or the last base (3' end) of the 3' primer.



The sequencing technology or platform used. (same as seq\_methin the MiMARKS sheet)



In addition to these mandatory fields, please also specify the ENA or SRA accession numbers

**Tutorial 4: providing the environmental data in the MIMS template (for high throughput shotgun metagenomes)**

**Tutorial 5: providing the environmental data in the MIGS template (for single organism genomes)**

**Tutorial 6: uploading your sequence data to SRA or ENA**

# References/links:

Yilmaz P., R. Kottman, D. Field, R. Knight, J. R. Cole, L. Amaral-Zettler et al. (2011). The “Minimum Information about a MARKer gene Sequence” (MIMARKS) checklist: Capturing contextual data about marker gene sequences and introducing MIxS, a unified standard for sequence checklist development including environmental data. *Nat Biotechnol.* [*http://www.nature.com/nbt/journal/v29/n5/pdf/nbt.1823.pdf*](http://www.nature.com/nbt/journal/v29/n5/pdf/nbt.1823.pdf)

DarwinCore: <https://code.google.com/p/gbif-providertoolkit/wiki/DarwinCore>

MiMarks Wiki: <http://wiki.gensc.org/index.php?title=MIMARKS>

MiMarks Googlesheet documentation at RDP: <http://rdp.cme.msu.edu/misc/googleSheetsHelp.jsp>

IPT quick reference guide: <https://code.google.com/p/gbif-providertoolkit/wiki/IPT2ManualNotes#Quick_Reference_Guide>

mARS white paper: <http://marswhitepaper.blogspot.be>

mARS FAQs: <http://mars.biodiversity.aq/howto>

Global Change Master Directory: <http://gcmd.nasa.gov>

Antarctic Master Directory: <http://gcmd.gsfc.nasa.gov/KeywordSearch/Home.do?Portal=amd&MetadataType=0>