mARS Standard Operating Procedure (SOP)

October 2018

# **Why use mARS?**

The microbial Antarctic Resource System (mARS) is an online platform that aims to support research in the Antarctic (or other cryo-environments) by making the data of past projects discoverable and exchangeable in the Antarctic scientific community and beyond. This is done through collecting and archiving metadata that can be searched, and linking this to the online data repositories of the microbial DNA or RNA sequences (e.g. European Nucleotide Archive ENA or NCBI’s Sequence read Archive (SRA), as well as the relevant publications about these data. More information on the vision of mARS can be found [here](http://share.biodiversity.aq/MARS/Vision/mARSVision.pdf).

**About this SOP**

This SOP details how you can upload your project meta- and environmental data to the microbial Antarctic Resource System (mARS). We distinguish between three aspects of microbial datasets: 1) metadata, or information about the data, such as who took the samples, who initiated the project, what is the project about, etc. 2) environmental data: this are ancillary measurements that were taken from the environment of the microbial isolates, such as pH, measurements of ions, etc. Finally, 3) the sequence data refers to the actual sequences of DNA fragments that were obtained from the organisms of interest. To be clear the from the outset, **mARS will NOT store sequence data**; it is intended that the sequence data produced by your project is submitted to a public repository such as the European Molecular Biology Laboratory (EMBL), the National Center for Biotechnology Information (NCBI) or the DNA Databank of Japan (DDBJ).

Instead, mARS aims to collect meta- and environmental data, which enables different projects to be searched and filtered more easily than on the public repositories. To ensure the procedure of gathering all your meta- and environmental data only has to been carried out once, the mARS team has devoted special care to following widely-used standards for biodiversity data, as promoted by the [Global Biodiversity Information Facility](http://www.gbif.org) (GBIF) and the [Genomics Standards Consortium](http://gensc.org). In this particular case, this SOP is built around two main types of standards, namely [DarwinCore](https://code.google.com/p/gbif-providertoolkit/wiki/DarwinCore), and the Minimum Information on any Sequence ([MIxS](http://wiki.gensc.org/index.php?title=MIxS)) standard (where the Minimum Information about a MARKer gene Sequence ([MiMARKS](http://wiki.gensc.org/index.php?title=MIMARKS)) has been designated). This ensures maximal interoperability with internationally-recognized data and metadata repositories.

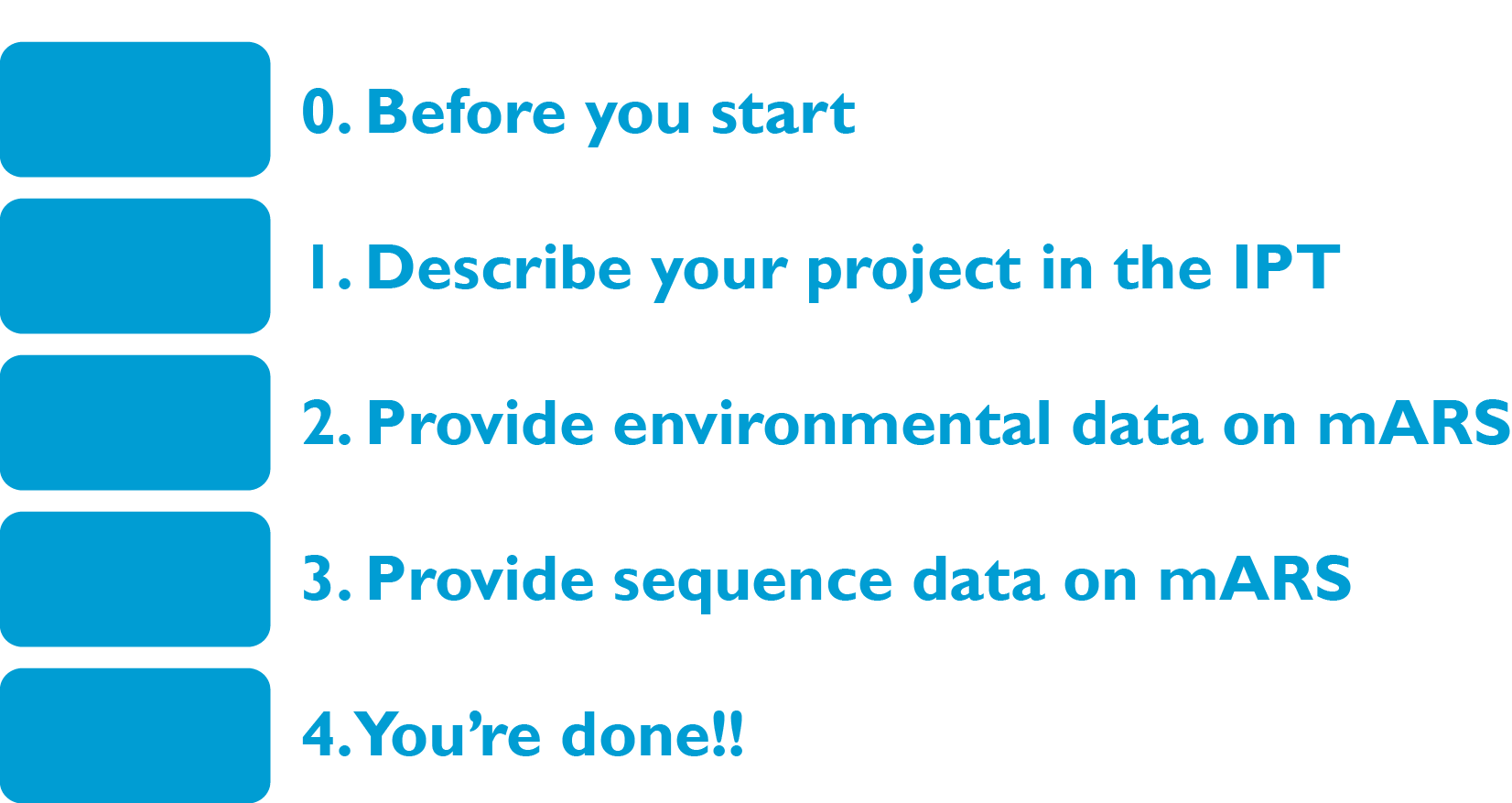
There are two different steps during which meta- and environmental data can be committed to mARS, and which also roughly correspond with different phases in the life-cycle of a research project.

In a first step, the project together with basic metadata can be registered through an interface with GBIF, using standardized DarwinCore terminology. At present, GBIF is mainly intended to accept occurrence data for macrorganism biodiversity data sets, and **can NOT accept molecular sequence-derived OTU occurrence data sets**. Therefore, mARS implements custom data system, in which the product (the project metadata) is discoverable across GBIF and mARS; while the latter also encloses the environmental data (see second step) as well as a link to the accompanying sequence data.

The second step typically accompanies the later phases of a research project, when sequence data has been generated and analyzed. For this, we are requiring that the sequence data is accessible in a public repository (ENA, Genbank, SRA, or other web-accessible sites). In most cases, like with high throughput sequencing datasets, the public repository will require a MiMARKS spreadsheet to be filled in ([Yilmaz et al. 2011](http://www.nature.com/nbt/journal/v29/n5/pdf/nbt.1823.pdf)). This file will contain valuable technical data on the samples and how they were sequenced, as well as all any environmental parameters that were gathered, using standardized ontologies and measurement units. This MiMARKS file is required to make your data discoverable on the mARS portal. If your sequence data has not yet been submitted to a public repository, an empty MiMARKS spreadsheet can be completed via mARS, and can later be used to save time when uploading your sequence data to a public repository.

Finally, MARS also requires a the “Microbial sequence set spreadsheet” to be filled in, which summarizes additional sequence-related information, and allows us to access and represent sequence data you have published in open repositories.

To upload your project information, environmental data, and sequence-related information simply follow the step-by-step procedure below. If you have any difficulties, don’t hesitate to [get in touch with us](http://mars.biodiversity.aq/site_pages/contact).



The diagram above provides an overview of the metadata publishing work flow, and steps that are further described in this SOP.

# **0. Before you start**

1. Send an email to request a username and password from the [IPT administrator](mailto:%20nyoudjou@gmail.com)
2. 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.
3. Make a copy or download the [MiMarks Googlesheet](https://docs.google.com/spreadsheet/ccc?key=0AjvjN1rFsR43dGhKU3doek43a05JTmozcUo3NmJXRnc&authkey=COanrdEC#gid=5) from the RDP MiMarks Googlesheet that we have a link to here (**click on “Make copy” or “from the “File” menu**). If you cannot access Google Documents, get in touch with us.
4. Make a copy of the[Microbial Sequence Set](https://docs.google.com/spreadsheet/pub?key=0AsB0418meERPdHpyaGZaMENmQXpVN0xXeVpPeFdIOWc&output=html) from the mARS Googlesheet (**click on “Make copy” from the “File” menu**). If you cannot access Google Documents, get in touch with us.

# **1.1.** **Describe your project in the IPT**

1. Login to the Integrated publishing toolkit ([IPT](http://ipt.biodiversity.aq/)) at biodiversity.aq using your credentials. This tool will guide you step by step through the procedure to publish your metadata, and uses the standardized terms and ontologies of DarwinCore to document your dataset.
2. After logging in, the “Manage Resources” tab will appear at the top of the page. Click on manage resources.
3. Use the form at the bottom of the “Manage Resource” page to create a new resource. Provide a unique "shortname" for your dataset.
4. Under “Type” select “Metadata only” or “Other”; at this point, the IPT, which is affiliated with GBIF is not designed to take environmental or molecular diversity information.
5. Click the “Create” button. You will arrive on the Resource Management page.
6. Click on the “Edit” button in the Metadata section on the left and fill in the details for the different metadata sections. A detailed instructions are available from [IPT quick reference guide](https://code.google.com/p/gbif-providertoolkit/wiki/IPT2ManualNotes#Quick_Reference_Guide). *Hint*: mention your grant number in the “Project Data” section, to allow us to link your resource to relevant projects in the GCMD/AMD.

***Hints for filling out the IPT forms:***

* “ \* ” signifies fields that need to be filled out
* There’s a navigation bar to the right with various forms that require information
* Basic Metadata sheet: under “Title” enter a descriptive title for your project.
* 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.
* Collection data: These terms don’t apply to the mARS data sets perfectly; though you can still enter relevant terms for your data. This data is particularly relevant if you do have archived samples either in your lab or in the DNA archive in New Zealand. Enter Not Applicable where appropriate.
* External Links: This is a good place to put the URL for a project website if you have one.
* Additional Metadata: If your data set is registered in other data bases (e.g. Antarctic Master Directory; GenBank or the Sequence Read Archive, BCO-DMO) then enter that information here.

# **1.2. Publish and register your metadata**

1. Back in the IPT webpage, from your Resource Management page, click on the “Publish” button in the “Published release” section on the left of the page. This action will upload your project metadata (and only the 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.
2. 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.
3. 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.

# **2.1. Prepare your MiMarks spreadsheet**

1. If your sequence data has been deposited on SRA or ENA, and you have the filled in MiMarks spreadsheat as a comma separated file (CSV) at hand, you can immediately go to step 2.2, uploading your MiMARKS CSV.
2. If your sequence data has not yet been deposited on SRA or ENA, then fill in your environmental data details in the [MiMarks Googlesheet](https://docs.google.com/spreadsheet/ccc?key=0AjvjN1rFsR43dGhKU3doek43a05JTmozcUo3NmJXRnc&authkey=COanrdEC#gid=5) you’ve created in step 0. The “Google Documents” interface can be used for this, and instructions are available from the MiMarks Googlesheet [documentation at RDP](http://rdp.cme.msu.edu/misc/googleSheetsHelp.jsp). Example files are available from the mARS website. By completing the MiMARKS template using standardized ontology terms, your environmental data can be coupled to the different sequence samples, and become comparable and interchangeable with other projects.
3. In the header for each column that will hold your unique identifier of your sequence data sample (i.e. sample name). Note that only one environmental package (i.e. spreadsheets) must be completed, in accordance with the origin of the samples.
4. Once you are finished, download your spreadsheet as a CSV (Comma-separated Values) file on your computer.

# **2.2. Upload your MiMarks CSV(s) in mARS [Still in DEV]**

For now just mail us your Mimarks and Microbial Sequence set as a comma separated file (CSV), which can, for instance, be obtained by using the ‘save as’ function under the file menu in Excell.

# **3.1. Prepare your Microbial Sequence Set spreadsheet**

1. In the [Microbial Sequence Set](https://docs.google.com/spreadsheet/pub?key=0AsB0418meERPdHpyaGZaMENmQXpVN0xXeVpPeFdIOWc&output=html) Googlesheet you’ve created in step 0, fill ***all*** the fields (replace the examples available from the Googlesheet)
2. Once you are finished, download your spreadsheet as a CSV file on your computer.

# **3.2. Upload your Microbial Sequence Set CSV(s) in mARS [Still in DEV]**

For now just mail us your Mimarks and Microbial Sequence set CSV as a spreadsheet (eg excel)

# **4. You’re done, go shopping or bar hopping!**

1. Get in touch with us when you’ve gone through the process.
2. Do you need a GCMD/AMD identifier? Send us an email, and we’ll sort that out for you.

**5. You realize you made a mistake, or want to update the IPT**

1. You can edit any field. Log into the IPT, go to manage resources overview, change the fields that you want to change. Go to the Published Release box on the left panel and hit “publish” again.

2. You must email the mARS administrator to notifiy them of your update.

# 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>