MIRRI Information System (MIRRI-IS)

Preface to the MIRRI-IS dataset

(Version 5.1, delivered on March 24th, 2023)

Index of contents

PREFACE

This document introduces the **guidelines** to culture collections (CCs) curators for the provision of their catalogs to MIRRI for inclusion in the MIRRI Information System (MIRRI-IS).

It is essential for them to keep in mind that:

- Only catalogs from collections of MIRRI partner countries are being included in MIRRI-IS.
- The content of MIRRI-IS is currently limited to an agreed list of information items (MIRRI-IS dataset), whose up-to-date data format is included in the MIRRI-IS dataset document.
- All information included in the dataset is **recommended** and **should be submitted** when available.
- For catalogs available in a BioloMICS implementation, BioAware will extract the required data and include them in the MIRRI-IS, provided that the involved CC agrees.
- For the other catalogs, the extraction of data and its submission to BioAware is under responsibility of the CC curators, which can be supported at the national level or by CECT.
- MIRRI has implemented a web tool for the validation of catalogues before submission. In order to use the validation tool, the catalogue must be submitted as an Excel file, according to the template provided by MIRRI.
- BioAware will validate some essential catalog data and return to CC curators lists of inconsistencies and errors found, so that they can carry out all appropriate corrections before the following submission.

The **MIRRI-IS dataset** version 5, delivered on September 9, 2022 and revised as version 5.1 on February 24, 2023, is reported in an associated document. Annex to the dataset is an MS Excel file describing the MIRRI-IS dataset version 5, with examples, to be used as a template for uploading catalogs.

GENERAL CONSIDERATIONS

Resource type focus

This document relates to bacteria, archaea, filamentous fungi, yeasts, algae and cyanobacteria only. Specifications for other resources, such as plasmids, phages, and viruses, are included in separate guidelines.

Mandatory data fields

All data fields are strongly recommended and should be provided in the requested format. There presently are only a few mandatory fields that are highlighted in the description of the dataset. Submitted data will be checked: records missing mandatory data may be discarded.

Taxonomy

The taxonomic identity of resources must be compiled according to the exact and complete taxonomy from authoritative sources.

These include:

- Fungi and yeasts: MycoBank (see https://www.mycobank.org/),
- Bacteria and archaea: Prokaryotic Nomenclature Up-to-Date (PNU), that recently joined with the List of Prokaryotic names with Standing in Nomenclature (LPSN) to conform a new site (see https://www.dsmz.de/services/online-tools/prokaryotic-nomenclature-up-to-date, https://www.bacterio.net/ and https://lpsn.dsmz.de/),
- Algae and cyanobacteria: AlgaeBase (see http://www.algaebase.org/).

CCs are required to submit only genus and species names, plus possible subspecies and variant names. Terms like 'sp.', 'spp.', 'aff.' or other attributes that would not permit finding the proper species name are not allowed. Only recognized and validly published Latin names will be accepted. When a resource is not completely identified yet, or the species is not described yet, the genus name only (without "sp."), or even a higher rank name, must be specified.

A field for supplementary information on the taxon name, 'Comment on taxonomy', is available. This field should always be used when some extra notes can be provided on the taxon name of the strain. This includes, e.g., when the strain refers to a new species and when the reference taxonomical database has not yet been updated with reference to some changes in the classification of a genus or species. In such cases, include both the taxon name that is actually used in the catalog in the 'Taxon name' field and a free text comment in the 'Comment on taxonomy' data field.

The comment field can also include remarks, e.g. related to morphology, that may suggest a different classification. A comparison with a second taxon with which there could be a confusion, along with the reasons why it is not that, may also be included here, as well as the taxonomic name which was used by the depositor. It should not be used for information of exclusive interest for the collection and therefore not useful/of interest for the users.

In case of hybrid strains, more than one taxon name can be specified. In this case, names must be separated by a semicolon character ";".

Genomic data

Genomic reference data is of extreme importance in view of the development of MIRRI-IS and of the tools that will exploit its data, some examples of which are already available at https://catalog.mirri.org/.

CCs must provide the INSDC accession numbers of the sequence for all genes and markers that they consider of relevance for identification and other applications. Among them, are included the Internal Transcribed Spacer (ITS) regions, the Large Subunit (LSU), the 16S of the nuclear ribosomal RNA (rRNA), betatubulin (BenA), calmodulin (CaM), Actin (ACT), elongation factor

1-alpha (EF-1α), Ribosomal RNA-coding genes (RPB1 and RPB2). This list however is not exhaustive. The sequence too can be submitted, if the collection agrees, when it is known, even if it has not been submitted to any sequence databank.

Growth media

Growth medium information must be provided as detailed as possible, in a separate table. Resources must then be linked to that table by using a unique identifier for the growth medium.

Geographic origin

As to geographic origin, information for country, region, city, locality must be submitted in a separate table. Names should be expressed in English, when the English name exists, e.g. for countries, regions, main cities. Resources must then be linked to that table by using a unique identifier for the locality.

Literature

MIRRI-IS will include exact bibliographic references whenever possible, so that strains can be directly linked to their respective publications. To this aim, Pubmed IDs and DOIs should be provided, when available. When neither the Pubmed ID nor the DOI are available, information on authors, title, journal, volume, issue and pages are requested as distinct data fields, if possible. Literature information must be provided in a separate table. Resources must then be linked to the appropriate references in that table by using the relative unique row IDs.

Dates

Some dates are included in the MIRRI-IS dataset. Here is their intended meaning.

- Collection date: when the sample was collected, usually in in situ condition.
- Isolation date: when the strain was isolated, usually in a laboratory.
- Identification date: when the strain was identified with the current taxon.
- Deposit: when the strain was deposited at the collection.
- Inclusion date: when the strain was included in the catalog and/or received its accession number.

The “access date” as defined by the Nagoya protocol is not included in the dataset

Codes vs numeric values and empty data fields

Many enumerations (i.e. short lists of allowed values for a given data field) have been converted into numbers. This is the case, e.g., for the field “Nagoya protocol restrictions and compliance conditions” (shortName: nagoyaConditions). Logical values (True/False, Yes/No) have been converted to numbers too, for sake of uniformity.

These conversions can be done by the CC just before submitting the catalogue: the catalogue does not need to be changed for this.

When no information is available for a given data field, it must be left empty.

Data validation

MIRRI has implemented a tool for validating the catalogs before uploading to MIRRI-IS.