# MIRRI Information System (MIRRI-IS) How to compile the sheets

(MIRRI-IS Dataset for Microorganisms Version 5.1, delivered on March 24th, 2023)

## Introduction

This is a short guide on how to compile the MS Excel file for the submission of catalogue information to MIRRI for inclusion in the MIRRI Information System (MIRRI-IS).

It complements the documents

- “MIRRI-IS dataset_Specifications” (Version 5.1, delivered on March 24th, 2023)
- “Preface to the MIRRI-IS dataset” (Version 5.1, delivered on March 24th, 2023)

It is meant as a guide to the compilation of the Excel template file

- MIRRI-IS_dataset_template_v5.1_20230324.xlsx

The template file is not intended to be filled in manually. Although it could be compiled manually, instead, it is foreseen that many culture collections (CCs) will create the file programmatically, i.e. they will use some software to create it. However, we **strongly suggest to compile the template file for some strains** of the collection in order to achieve a better understanding of the desired format for submission of data, of allowed values for each data field and of all interrelations between data fields.

In the following paragraph, we suggest how to perform this exercise.

## Sheets of the template

The template file includes various sheets having different aims and use. CCs must understand the differences in order to appropriately fill them. Note that a pink background highlights all cells referring to mandatory data. Specifically, the template file includes:

- Fixed reference sheets: these sheets include the complete list of allowed values for some of the data fields. They are fixed, i.e. cannot be changed by the CC, and usually hidden. As an example, consider the sheets for organism types and for forms of supply.

- Extendable reference sheets: these sheets include partial lists of allowed values for some of the data fields and can be updated by the CC in order to include some specific values for their needs. As an example, consider the sheets on sexual states and on markers that are not exhaustive and can be extended by the CCs according to their data.

- Collection specific reference sheets: these are aimed to include reference lists for each CC and must be filled in by the CC before inserting data in the main sheets. As an example, consider the sheets on growth media used by the CC, geographic origin of strains held by the CC and literature related to the CC strains.

- Catalogue sheets: these sheets are aimed to include the catalogue data, i.e. specific information on the strains included in the catalogue of the CC. These namely are the 'Strains' and 'Genomic information' sheets.

The following table report all sheets, their type and the relation to other sheets.

| Sheet | Data | Sheet type | Needed by |
|---|---|---|---|
| Resource types | List of allowed values for organismType | Fixed reference, hidden | Strains |
| Forms of supply | List of allowed values for supplyForms | Fixed reference, hidden | Strains |
| Ploidy | List of allowed values for ploidy | Fixed reference, hidden | Strains |
| Axenic | List of allowed values for axenicCulture | Fixed reference, hidden | Strains |
| Growth media | List of growth media used by the CC | Collection specific reference | Strains |
| Geographic origin | List of origin localities for strains in the CC | Collection specific reference | Strains |
| Literature | List of publications for the strains in the CC. | Collection specific reference | Strains |
| Sexual state | List of allowed values for the 'sexualState' field in the 'Strains' sheet | Extendable reference | Strains |
| Strains | Data related to strains in the CC | Catalogue | Genomic information |
| Ontobiotope | List of Ontobiotope terms for the relative field in the 'Strains' sheet | Fixed reference | Strains |
| Markers | List of marker names for the related field in the 'Genomic information' sheet | Extendable reference | Genomic information |
| Genomic information | Data related to sequences for the strains in the CC | Catalogue | - |

# Compilation steps

Due to the interrelation between sheets, for a proper compilation of the CC specific file the following steps should be performed in order.

1. Make a copy

   This is needed to preserve the original template file and keep it as a reference.

   There is not a standard naming convention, but we suggest that the name of the new file includes both the acronym of the collection and the date. E.g., LMG_2023-02-24.xlxs.

2. Complete the collection specific reference sheets

   Collection specific reference sheets must be compiled before inserting data in the 'Strains' sheet because they are used as a list of allowed values for some of the 'Strains' data fields. Such sheets are 'Growth media', 'Geographic origin' and 'Literature'. Read carefully the guidelines on how to insert your data into them.

   In the template, these sheets include some fictitious data meant to support their appropriate compilation. Fictious data must be removed before starting the compilation of the 'Strains' sheet.

   Note that all rows in these sheets has a progressive number in the first column. These numbers must be included in the appropriate cells of the 'Strains' sheet in order to link it to the reference sheets.

   Note that in the collection specific reference sheets a pink background highlights all mandatory data.

   A special case is represented by the Literature sheet where information can be submitted according to the following possibilities:

   - Only reference ID and PMID. The Pubmed ID of a publication is able to identify it uniquely. As a consequence, when it is included, all other reference data may be omitted. The ID column must be compiled.

   - Only reference ID and DOI. The DOI of a publication is able to identify it uniquely. As a consequence, when it is included, all other reference data may be omitted. The ID column must be compiled.

   - The complete reference divided in its usual main components, i.e. list of authors, title, journal, year, volume, issue, starting page, ending page. This format should normally be used only when neither the PMID nor the DOI are available. In this case, each component must be included in the relative column. and leave all other columns, but the ID column, empty.

   - The complete reference as free text. Since many CCs may have the bibliographic references stored as a unique text, without any distinction between its usual components (list of authors, title, journal, year, volume, issue, etc…), this format is accepted, although it should be considered only when none of the previous options is possible. In this case, include the reference in the 'Full reference' column and leave all other columns, but the ID column, empty.

3. <u>Check and complete the extendable reference sheets</u>

   The extendable reference sheets include some of the possible values for the related data fields in the 'Strains' or 'Genomic information' sheets. CCs must check their contents and add all missing values that are needed for their strain descriptions. There are only two such sheets, 'Sexual state' and 'Markers'.

   Note that a pink background highlights all mandatory data.

4. <u>Complete the 'Strains' sheet</u>

   Once all reference sheets have been compiled, the 'Strains' sheet can be filled in. This is because some of its data fields take the list of allowed values from the reference sheets.

   On its turn, the 'Strains' sheet must be filled in before the 'Genomic information' sheet because the latter includes a data field for strain accession numbers that takes as allowed values the list of accession numbers of the 'Strains' sheet.

   Note that a pink background highlights all mandatory data.

5. <u>Complete the 'Genomic information' sheet</u>

   Note that a pink background highlights all mandatory data. In this case, the INSDC accession number should usually be submitted, but when it is not available, the sequence can be submitted instead. In any case, at least one of these two fields must be filled in.