MIRRI Information System (MIRRI-IS)
Guidelines for catalog uploading

MIRRI ICT Task Force

(Version 2, approved on June 17th, 2020, after teleconferences with CC curators)

Index of contents

PREMISE

This document is meant to provide guidelines to culture collections (CC) curators for the provision of their catalogs to MIRRI for inclusion in the MIRRI Information System (MIRRI-IS). These guidelines were prepared during various meetings, both face-to-face and virtual, of the ICT Task Force of MIRRI (MIRRI-ICT), and discussed with curators of MIRRI CCs during 2 teleconferences, held on June 10th and 17th, 2020.

During the meetings of the ICT Task Force, it was decided that:

- MIRRI-IS will be implemented using BioloMICS on the basis of an agreement that MIRRI must sign with BioAware. BioAware provided a first version to MIRRI in March 2020 already. MIRRI board must review it.
- Only catalogs from collections of MIRRI partner countries will be initially included in MIRRI-IS.
- The content of MIRRI-IS will be initially limited to an agreed list of information items (MIRRI-IS dataset), whose data format have been defined by the task force.
- All information included in the dataset is recommended, according to the resource type, and should therefore be submitted when available, because MIRRI aims at gathering all interesting information into the MIRRI-IS without limiting to the essential data.
- The MIRRI dataset will be progressively extended in order to improve its usefulness towards user needs and expected applications.
- A detailed description of the MIRRI dataset format will be provided to CC curators.
- For catalogs already available in a BioloMICS implementation, CBS/BioAware will extract the MIRRI dataset and include it in the MIRRI-IS, provided that the involved CC agrees.
- For the other catalogs, the extraction of data and its submission to CBS/BioAware for inclusion in the MIRRI-IS is under responsibility of the CC curators, which can be supported at the national level (CECT (ES), HSM (IT), VKM (RU) already expressed their willingness to support other collections at a national level).
- CBS/BioAware will validate some essential data of the catalogs and return to CC curators eventual lists of inconsistencies and errors found, so that they can carry out all appropriate corrections.

At the end of its work, the ICT Task Force delivered the MIRRI-ID dataset version 2020.04.1 as well as the initial version of these guidelines. During the teleconferences held with CC curators, clarifications were provided and the dataset was redefined according to the results of the discussion, until the definition of the current version 2020.06.1.

In the following part of this document, we report the MIRRI-IS dataset version 2020.06.1.

Annexes to this report are an MS Excel file describing the MIRRI-IS dataset version 2020.06.1, with examples, and an MS Excel file to be used as a template for uploading catalogs.

METHODOLOGY AND REFERENCES

This work was carried out starting from the existing standards and comparing them with the information effectively available in the CC catalogs, by taking into account some of the most recent and relevant international regulations and European rules as well:

- Annex A "From CABRI data fields to MIRRI data objects: the evolution of a standard" of the Deliverable D8.3 of MIRRI preparatory phase. This document includes a simple table that examines CABRI data sets for bacteria and, for each field, specify how it could be better implemented for MIRRI and the possible data validation processes.
- Mapping between data fields of catalogs submitted by CCs and the CABRI expanded datasets, compiled by Alexander Vasilenko, VKM.
- The ISO/AWI 21710 document on “Specification on data management and publication in microbial resource centres” (TC 276, WG 5).

Starting from the above information, the task force analysed each data field with the aim of defining:

- whether it had to be included in the MIRRI-IS dataset,
- if the CABRI and ISO/AWI 21710 definitions were in agreement,
- the format that it should have in the MIRRI-IS dataset and its related possible values.

A special attention was devoted to data fields related to regulations. A few additional data fields were included in the MIRRI-IS dataset, although not explicitly mentioned in CABRI and ISO standard definitions. Examples of these are Nagoya related information, dual use of strains and quarantine. For some of these fields advice was obtained from experts outside the task force.
A special effort was put on the definition of codes and enumerations to be associated to data fields, instead of free text, whenever possible.

GENERAL CONSIDERATIONS

Resource type focus

The ICT Task Force focused its effort on bacteria, archaea, filamentous fungi and yeasts, mainly because of the specific competence of its members. Other resource types require other information. The Task force is willing to incorporate all needed information in the dataset in future releases. This applies, e.g., to plasmids, phages, viruses, algae and cyanobacteria.

Mandatory data fields

All data fields are strongly recommended: the CC should make all possible efforts to provide it in the requested format. There presently are only a few mandatory fields that are listed here below and highlighted in the description of the dataset. MIRRI-IS will check the provided data and may eventually discard records missing mandatory data fields and inform the CC.
Presently, mandatory data fields are the following:

- Accession number
- Restrictions on use
- Nagoya protocol compliance conditions
- Risk Group
- Organism type
- Taxon name

- Recommended growth temperature
- Recommended medium for growth
- Form of supply
- Geographic origin

Taxonomy

With reference to the taxonomic identity, it was decided that the value for this data field has to be compiled by retrieving the taxonomic exact and complete definition from authoritative sources. These include MycoBank (see http://www.mycobank.com/) for fungi and yeasts, the Prokaryotic Nomenclature Up-to-Date (PNU see https://www.dsmz.de/services/online-tools/prokaryotic-nomenclature-up-to-date), that recently joined with the List of Prokaryotic names with Standing in Nomenclature (LPSN see https://www.bacterio.net/) to conform a new site (see https://lpsn.dsmz.de/), for bacteria and archaea, and AlgaeBase (see http://www.algaebase.org/) for algae. The reference nomenclature for viruses is the taxonomy provided by the International Committee on Taxonomy of Viruses (ICTV see https://talk.ictvonline.org/taxonomy/).

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. A list of examples is provided in the annex. When a resource is not completely identified yet, or the species is not described yet, the genus name only (no "sp."), or even a higher rank name, must be specified.

It should be noted that a free text field for supplementary information on the taxon name, 'Comment on taxonomy', is available. This field should be used anytime 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 the taxon name that is present in the catalog in the 'Taxon name' field AND include a comment (free text) in the 'Comment on taxonomy' data field.

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

MIRRI-IS will implement the BioloMICS method for synonyms. All strains of a given taxon and of its synonym taxa will be retrieved together, when querying by any of those names. This implies that synonyms are kept once defined by the CCs as taxon name: if CCs are using a synonym instead of the name in the catalog, this information can be kept.

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.

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.
When just the accession number is provided, MIRRI-IS will retrieve the relative sequence and will return it to the CC for inclusion in the catalog.

Growth media

Growth medium information must be provided as detailed as possible, in a separate table. A list of common growth media will be compiled and made available in future versions of the MIRRI-IS dataset.

Geographic origin

As to geographic origin, information for country, region, city, locality must be submitted in a separate table. This will improve precision of user queries and possibly allow the gathering of longitude and latitude values from GeoNames (see https://www.geonames.org/), when available with a sufficient precision. However, when coordinate information is available from CCs, these data fields must also be submitted and will be used. Locality names can be expressed either in the local language or in English.

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 in a separate table. In this case, as a validation task and in order to improve catalogs, MIRRI-IS will try to extract the missing Pubmed ID and DOI on the basis of the available information and return them to the CC for inclusion in the catalog.

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

The vast majority of 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 compliance to Nagoya protocol. Logical values (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 should be left empty.

Data validation

MIRRI-IS will implement a tool for checking coherence of taxonomic identity among those strains that are available by more than one CC. In case of discrepancies, all relevant CCs will be informed and requested to verify their data. A similar check will be implemented for genomic sequences. Further checks will be carried out for the majority of data fields, according to the description of validation tasks in the dataset description.

Other CC numbers associated with each strain will be inter-checked for coherence. Strains will then be connected within the MIRRI-IS. Those strains that do not explicitly mention some or all of the other CCc names will be linked too, but in a different field because this association is not explicitly provided by CCs and it is solely responsibility of MIRRI-IS.

MIRRI-IS DATASET (version 2020.06.1)

Name	**MIRRI Accession number**
Description	Unique identifier of the strain in the MIRRI-IS. It will be created on the first submission of a strain in the MIRRI-IS in a one-to-one connection with the Accession number of the strain in the CC. It is meant as a reference within the MIRRI-IS and as a unique reference for interoperability with other Life Science tools. It will include a version extension and be used as a reference for provenance issues as well. This information will be returned to CCs in association with the relative strain accession number. CCs are invited to include it in their catalogue and return it to MIRRI-IS at every following submission.
Syntax	The MIRRI Accession number will be composed by the 'MIRRI' prefix followed by a space character and an alphanumeric code made of two letters followed by four digits (this allow for more than 6.5 million codes).
Values	On first submission, this field should not be compiled. At following submissions, the accession number returned by MIRRI-IS should be included.
Validation	When missing, check whether the strain accession number was already included. If not, a new value will be created and returned to the collection. If the strain accession number was already submitted in the past, MIRRI-IS will retrieve the related values and assign it to the strain again. When a value is submitted by the collection, MIRRI-IS will check that the correct syntax is used and that the MIRRI and the strain accession numbers are properly associated. If not, inform the collection.
Examples	MIRRI AA0234

Name	**Accession number (MANDATORY FIELD)**
Description	Unique identifier of the strain in the CC.
Syntax	CC acronym followed by a space character and a number or code. If a code is used, it cannot include spaces. If the current accession number is not compliant with the new rule, it must be redefined. In this case, the previous number must be included in the "Other culture collection numbers".
Values	Free text, according to defined syntax.
Validation	Check that the correct syntax is used.
Examples	LMG 25 DSM 790 CBS 1546.1

Name	**Other culture collection numbers**
Description	Accession numbers of the same strain in other CCs, when known.

Syntax	Accession numbers formatted as above specified and separated by a semicolon character. Should not include accession numbers that do not follow the relative syntax. As an exception, Herbarium numbers can be included here.
Values	Free text, according to defined syntax.
Validation	For accession numbers of strains of CCs available in MIRRI-IS: • control that taxon names or synonyms are identical and if the assigned name is not the current name warn the original CC; • if the name is incorrect warn all the CCs having the strain.
Examples	CBS 316.51; NRRL 1944; QM 191; MUCL 9645

Name	**Restrictions on use (MANDATORY FIELD)**
Description	Report if the strain can be used for commercial development or not.
Syntax	One of the allowed values.
Values	One of the following values: 1 (no restrictions apply), 2 (for research use only), 3 (for commercial development a special agreement is requested).
Validation	Check that one and only one of the allowed values is used. Report errors to the CC.
Examples	1

Name	**Nagoya protocol restrictions and compliance conditions (MANDATORY FIELD)**
Description	Situation of the strain in relation to the Nagoya protocol.
Syntax	One of the allowed values.
Values	One of the following: 1 ("No known restrictions under the Nagoya protocol"), 2 ("Documents providing proof of legal access and terms of use available at the collection"), 3 ("Strain probably in scope, please contact the culture collection").
Validation	Check that one and only one of the allowed values is used. Report errors to the CC.
Examples	1

Name	**ABS related files**
Description	Uniform Resource Locator (URL) of the Internationally Recognized Certificates of Compliance (IRCC) providing evidence that the strain was accessed in accordance with Prior Informed Consent (PIC) and Mutually Agreed Terms (MAT).
Syntax	The field must include a properly formatted URL, including a scheme (such as http, https of ftp), a hostname, possibly a path, and a file name. See https://url.spec.whatwg.org/#urls for details.
Values	A valid and complete URL.
Validation	Check URLs. Report errors to the CC.
Examples	http://www.domain.tld/path/abs_file.pdf

Name	**MTA file**
Description	Strain specific Material Transfer Agreement (MTA) document

Syntax	Working URL or file name of the strain specific MTA document, if any. The MTA document must be made available and accessible on-line. To this aim, CCs must provide the URL. In case, all such files are made available by the CC in the same folder of its web site, the file name is sufficient, but the CC must separately provide the URL of the folder to MIRRI-IS. Links to MTA files will then be created by MIRRI-IS according to the user requests.
Values	Free text including filename or URL.
Validation	Check that the path of the folder is accessible and that a file with the given name is available on-line. Alternatively, check URLs. Report errors to the CC.
Examples	mta_strainA.pdf http://www.domain.tld/path/mta_strainA.pdf

Name	**Strain from a Registered Collection**
Description	Strain included in the registered CC according to the EU Regulation 511/2014. Unregistered CCs can omit this information.
Syntax	One of the allowed values.
Values	One of the following values: 1 (for No), 2 (for Yes)
Validation	Check that one and only one of the allowed values is used. Report errors to the CC.
Examples	2

Name	**Risk Group (MANDATORY FIELD)**
Description	Risk group according to EU Directive 2000/54/EC and its amendments and corrections.
Syntax	One of the allowed values.
Values	Allowed values: 1, 2, 3, 4.
Validation	Check that one and only one of the allowed values is used. Report errors to the CC.
Examples	3

Name	**Dual use**
Description	Specify whether the strain has the potential for a harmful use according to EU Council Regulation 2000/1334/CE and its amendments and corrections.
Syntax	One of the allowed values.
Values	One of the following values: 1 (for No), 2 (for Yes).
Validation	Check that one and only one of the allowed values is used. Report errors to the CC.
Examples	2

Name	**Quarantine in Europe**
Description	Specify whether the strain is subject to quarantine according to European Directive 2000/29/CE and its amendments and corrections. The list of

	quarantine organisms is available in the Commission Implementing Regulation (EU) 2019/2072.
Syntax	One of the allowed values.
Values	One of the following values: 1 (for No), 2 (for Yes)
Validation	Check that one and only one of the allowed values is used. Report errors to the CC.
Examples	2

Name	**Organism type (MANDATORY FIELD)**
Description	The type of the resource.
Syntax	One of the allowed values. Alternatively, in special cases, both Filamentous Fungi and Yeast can be specified, separated by a ";".
Values	One of the following terms: Algae, Archaea, Bacteria, Cyanobacteria, Filamentous Fungi, Phage, Plasmid, for Virus, Yeast.
Validation	Check that one and only one of the allowed values is used, but for the special case above. Report errors to the CC.
Examples	Archaea

Name	**Taxon name (MANDATORY FIELD)**
Description	Taxon name including genus, species and variant names, as taken from an authoritative nomenclature reference, including Mycobank for fungi and yeasts, the Prokaryotic Nomenclature Up-to-date for bacteria and archaea, AlgaeBase for algae and cyanobacteria, and ICVT for viruses.
Syntax	According to the appropriate nomenclature. For Archaea, Bacteria, Filamentous Fungi and Yeasts, genus name followed by species name and by the subspecies and variant names, when appropriate. The subspecies name must be preceded by "subsp.". The variant name must be preceded by "var.". When the species name is not available, do not include "sp.". When the genus name is not available, specify the family name instead. In order to cope with delays in nomenclature updates, the most updated taxon name can be used, even when it is missing from the current version of the reference. In this case, a remark must be included in the 'Comment on taxonomy' data field. For hybrid strains, more than one taxon name can be specified. The semicolon ";" must be used as a separation character.
Values	All taxon names included in the authoritative nomenclature references, reported according to the given syntax.
Validation	Check for the correct syntax and the existence of the taxon name(s) in the reference nomenclatures. Report errors to the CC.
Examples	Candidaceae Candida Candida albicans Candida albicans var. claussenii

	Actinomyces globisporus subsp. Flaveolus

Name	**Infrasubspecific names**
Description	Infrasubspecific names including biovar, chemovar, cultivar, morphovar, pathovar, phagovar, serovar, forma specialis, phase.
Syntax	The infrasubspecific name, usually preceded by a short specification of its type, e.g. “pv.” for pathovar and “sv.” for serovar.
Values	Free text
Validation	None
Examples	pv. lachrymans sv. Typhi

Name	**Comment on taxonomy**
Description	Any comment and/or note on the taxonomy of the strain. It may be used, e.g., for information on new species or revised nomenclatures. It must be used when the Taxon name data field includes a name that is not present in the nomenclature reference.
Syntax	None
Values	Free text
Validation	None

Name	**Status**
Description	For type strains, specify their type (type, neotype, holotype, epitype, etc). A list of allowed values is not defined and this information can be provided as free text. Future improvements of the dataset will likely foresee a list of values.
Syntax	None
Values	Free text
Validation	None
Examples	Holotype

Name	**History of deposit**
Description	Transfers of the strain between isolation and deposit in the CC.
Syntax	The field includes entries separated by "<" meaning "received from". Entries may include persons or CCs. The name of the CC should be followed by the month, when available, and year of the acquisition. Between parentheses, the strain designation or CC numbers and/or a name can also be entered when a name change has occurred.
Values	Free text, according to above syntax
Validation	Check for the validity of the format. Report errors to the CC.
Examples	CECT, 1995 < CBS, 1990 < ATCC, 1989 NCTC, Nov. 1973 (Bacillus loehnisii) < T. Gibson, 1935 < Kral Collection (Bacillus probatus)

Name	**Depositor**
Description	Name, institute and town / country of the depositor.
Syntax	None
Values	Free text
Validation	None
Examples	M. Sebald, Inst. Pasteur, Paris, France P. Hirsch, Inst. Allg. Mikrobiol. Univ. Kiel, Germany

Name	**Date of deposit**
Description	Date when the strain was deposited at the CC
Syntax	May include a full date in the ISO 8601 format. YYYY-MM-DD or YYYYMMDD for full dates, YYYY-MM for year and month only, YYYY for year only. See https://en.wikipedia.org/wiki/ISO_8601 for a quick introduction.
Values	A valid date in one of the above formats
Validation	Check for the validity of the format. Report errors to the CC.
Examples	1999-02-20

Name	**Collected by**
Description	Name, institute and town / country of the collector.
Syntax	None
Values	Free text
Validation	None
Examples	J. Fraser, Moredun Res. Inst., Edinburgh, UK

Name	**Date of collection**
Description	Date when the sample was collected.
Syntax	May include a full date in the ISO 8601 format. YYYY-MM-DD or YYYYMMDD for full dates, YYYY-MM for year and month only, YYYY for year only. See https://en.wikipedia.org/wiki/ISO_8601 for a quick introduction.
Values	A valid date in one of the above formats
Validation	Check for the validity of the format. Report errors to the CC.
Examples	1999-11-27

Name	**Isolated by**
Description	Name, institute and town / country of the isolator.
Syntax	None
Values	Free text
Validation	None
Examples	I. Orskov, Ser. Inst., Copenhagen, Denmark

	D. Haas, Inst. Pasteur, Paris, France

Name	**Date of isolation**
Description	Date when the strain was isolated from the sample.
Syntax	May include a full date in the ISO 8601 format. YYYY-MM-DD or YYYYMMDD for full dates, YYYY-MM for year and month only, YYYY for year only. See https://en.wikipedia.org/wiki/ISO_8601 for a quick introduction.
Values	A valid date in one of the above formats
Validation	Check for the validity of the format. Report errors to the CC.
Examples	2019-08-17

Name	**Date of inclusion in the catalogue**
Description	Date when the strain was included in the catalog and/or an accession number was assigned to it.
Syntax	May include a full date in the ISO 8601 format. YYYY-MM-DD or YYYYMMDD for full dates, YYYY-MM for year and month only, YYYY for year only. See https://en.wikipedia.org/wiki/ISO_8601 for a quick introduction.
Values	A valid date in one of the above formats
Validation	Check for the validity of the format. Report errors to the CC.
Examples	1996-12-13

Name	**Tested temperature growth range**
Description	The lowest and the highest temperature at which the strain was tested for growing.
Syntax	Temperatures are expressed as decimal numbers in Celsius degrees and must be separated by a semicolon. The symbol ° and the letter C should not be included.
Values	Decimal numbers
Validation	Check for the validity of the format. Report errors to the CC.
Examples	15;35

Name	**Recommended growth temperature (MANDATORY FIELD)**
Description	The recommended growing temperature for the strain.
Syntax	The temperature is expressed as decimal number in Celsius degrees. The symbol ° and the letter C should not be included.
Values	Decimal number
Validation	Check for the validity of the format. Report errors to the CC.
Examples	24

Name	**Recommended medium for growth (MANDATORY FIELD)**
Description	The medium that is recommend for growing the strain.

Syntax	A textual reference, usually an acronym, to the appropriate growth medium in a table provided by the CC.
Values	CCs are invited to submit a table including a list of the growth media they use. The table should include at least an acronym and a description for each growth medium. A full description of the recipe is also welcome. All descriptions should be in English. In future versions of the MIRRI-IS dataset, a table of shared descriptions with acronyms will be provided.
Validation	Check for the presence of the textual reference in the provided table of growth media. Report errors to the CC.
Examples	AGA GYA

Name	**Form of supply (MANDATORY FIELD)**
Description	The forms of supply of the strain to users.
Syntax	One or several of the allowed values, separated by a ";".
Values	Allowed values: Agar, Cryo, Dry Ice, Liquid Culture Medium, Lyo, Oil, Water.
Validation	Check for the validity of the format. Report errors to the CC.
Examples	Cryo Agar; Lyo

Name	**Other denomination**
Description	Unofficial names that are often used for the strain, e.g. in publications, or a name given to the strain by the isolator before its deposit at the collection.
Syntax	None
Values	Free text
Validation	None
Examples	S288c; AB1157

Name	**Coordinates of geographic origin**
Description	The geographic coordinates of the location where the sample was collected.
Syntax	Latitude, longitude and precision. Latitude and longitude are expressed in decimal degrees. Cardinal directions North and West are implicit and must not be reported. Precision can be omitted. When included, it must be expressed in kilometers. Values are separated by semicolons. Conversion of latitude and longitude values from the sexagesimal format (as in 40° 26′ 46″) to the decimal format (40.446) can easily be achieved as follows: decimal degree = sexagesimal degree + (sexagesimal minutes/60) + (sexagesimal seconds/3600)
Values	Decimal numbers from -180 to 180 for longitude and -90 to 90 latitude. Decimal numbers for precision.
Validation	Check for the validity of the format and values. Report errors to the CC.

	When the information is missing, MIRRI-IS will try to determine it through the geographical database GeoNames (see http://geonames.org/) using the geographic origin information and will return it to the CC.
Examples	51.3456;15.46456 44.4111;8.89552;0.2

Name	**Altitude of geographic origin**
Description	The altitude of the location where the sample was collected.
Syntax	None
Values	Decimal number.
Validation	Check for the validity of the format and values. Report errors to the CC.
Examples	1286 -20

Name	**Geographic origin (MANDATORY FIELD)**
Description	The locality where the sample was collected, defined with the highest possible precision.
Syntax	Reference to a separate table, which includes all localities where at least one strain was collected. For organisms constructed in a lab, use the address of the depositor.
Values	The geographic location should be defined with the highest possible precision, but unambiguously. It should include locality, city, province, region, country. For old strains for which the geographic origin is not known, make reference to the special locality 'Unknown'. Avoid specifying countries and continents only. The table can include either separate fields for the geographic details or one single text including all details. The first format is preferred over the second. In order to improve the description of the location, you can check if it is described in GeoNames (see http://geonames.org/) and use its 'Administrative hierarchy' to include further rows with information missing in the table, e.g. administrative commune and region, until you find the country. NB! While querying GeoNames, you may also recover geographic coordinates and altitude of the locality.
Validation	Check for the presence of the reference in the table of localities. Report errors to the CC.
Examples	In order to insert Altafjorden, look at GeoNames. You will find it associated to the record n. 780944 whose administrative hierarchy reports Norway as country, Troms og Finnmark as adm1 and Alta as adm2. You will also retrieve 70.05765, 23.08293 for geographic coordinates. Altitude is not specified since this is a fiord. In your table you should include, either in separate cells or in a unique description, Altafjorden, Alta, Troms og Finnmark, Norway.

Name	**GMO**

Description	Specify whether the strain is a Genetically Modified Organism (GMO).
Syntax	One of the allowed values.
Values	One of the following values: 1 (for No), 2 (for Yes).
Validation	Check that one and only one of the allowed values is used. Report errors to the CC.
Examples	1

Name	**GMO construction information**
Description	Information on the construction of the GMO. By now, this information can be provided as free text. Future improvements of the dataset will likely foresee some syntactical rules and/or list of values.
Syntax	None
Values	Free text
Validation	None

Name	**Mutant information**
Description	Information on mutant strains. By now, this information can be provided as free text. Future improvements of the dataset will likely foresee some syntactical rules and/or list of values.
Syntax	None
Values	Free text
Validation	None
Examples	X-ray mutant of NRRL 1951.B25 Glutamine auxotroph of strain 74-A

Name	**Genotype**
Description	Information on the genotype of the strain. By now, this information can be provided as free text. Some syntactical rules and/or list of values are foreseen in the next version of the MIRRI-IS dataset.
Syntax	None
Values	Free text
Validation	None
Examples	leu2-3 leu2-112 his4-519 can1 gln-1b

Name	**Literature**
Description	Information on literature linked to the identification and properties of the strain. Does not include literature related to the sequence of the strain, which should be included in the field “Literature linked to the sequence/genome”. For publications indexed by Pubmed or having an official DOI number, collections should provide the relative identifiers, respectively PMIDs and DOIs.

	In this case, MIRRI-IS will retrieve additional information and complete the bibliographic data. When neither a PMID nor a DOI are available, all usual bibliographic fields used for citing a paper, a book, a patent, or a document available on-line, including, e.g., authors, title, journal, volume, issue, pages, editors, publishers, etc… must be submitted as separate fields in a distinct table. In this case, identifiers linking to the separate literature sheet must be included here. Multiple papers can be included for a single strain just by reporting more PMIDs, DOIs, and table identifiers separated by ";".
Syntax	PMIDs, DOIs, and numbers separated by a semicolon ";"
Values	Valid PMIDs and DOIs or reference numbers of the literature sheet.
Validation	MIRRI-IS will try to extract any missing PMID and DOI on the basis of the provided information and return it to the collection. Any errors and inconsistencies will also be reported to the CCs.
Examples	120; 26492633; 10.1371/journal.pcbi.1004525

Name	**Sexual state**
Description	Information on strain sexual state / mating type, for relevant resource types.
Syntax	One of the allowed values. More can be added by CCs.
Values	Mata Matalpha Mata/Matalpha Mata Matb Mata/Matb MTLa MTLalpha MTLa/MTLalpha MAT1-1 MAT1-2 MAT1 MAT2 MT+ MT-
Validation	Check that one and only one of the allowed values is used. Report errors to the CC.
Examples	Mata MTLa/MTLalpha

Name	**Ploidy**
Description	Information on the ploidy level of the strain.
Syntax	One of the allowed values.

Values	One of the following values: 0 (Aneuploid), 1 (for Haploid), 2 (for Diploid), 3 (for Triploid), 4 (for Tetraploid), 9 (for Polyploid (over 4n)).
Validation	Check that one and only one of the allowed values is used. Report errors to the CC.
Examples	2

Name	**Interspecific hybrid**
Description	This field reports whether the strain is an interspecific hybrid.
Syntax	The value ”Yes” should be included, when the strain is an interspecific hybrid.
Values	One of the following values: 1 (for No), 2 (for Yes).
Validation	Check that one and only one of the allowed values is used. Report errors to the CC.
Examples	2

Name	**Pathogenicity**
Description	Information about pathogenicity of the strain for plants, humans and animals. Can include specification for the Belgian plant pathogenicity code.
Syntax	None
Values	Free text
Validation	None
Examples	Pathogenic to Agaricus bisporus. Transmissible murine colonic hyperplasmia.

Name	**Enzyme production**
Description	Information about enzyme production by the strain. By now, this information should be provided as free text. Future improvements of the dataset will likely foresee some syntactical rules and/or list of values.
Syntax	None
Values	Free text
Validation	None
Examples	Decarboxylase, Isomerase, Pectinase.

Name	**Production of metabolites**
Description	Information about metabolite production by the strain. By now, this information should be provided as free text. Future improvements of the dataset will likely foresee some syntactical rules and/or list of values.
Syntax	None
Values	Free text
Validation	None
Examples	Degradation of beta-phenylpropionic acid Capreomycin; oxytetracyclin.

Name	**Applications**

Description	Information about applications of the strain. By now, this information should be provided as free text. Future improvements of the dataset will likely foresee some syntactical rules and/or list of values.
Syntax	None
Values	Free text
Validation	None
Examples	Biomass electricity generation Studies of pathway of beta-phenylpropionic acid metabolism Environmental restoration

Name	**Remarks**
Description	Any further note that is not present in the other fields.
Syntax	None
Values	Free text
Validation	None
Examples	Two stable colony types giving identical gel electrophoretic protein profiles. Strain was preserved after several local lesion passages with Nicotiana tabacum cv. Java as host plant.

Name	**Plasmids**
Description	Information about plasmids in the strain. It may include plasmid name and type (original plasmid, cloning vehicle, recombinant plasmid), restriction sites, relevant genes (e.g., origin of replication, transposons, promoters, terminators, structural genes). By now, this information should be provided as free text. Future improvements of the dataset will likely foresee some syntactical rules and/or list of values.
Syntax	None
Values	Free text
Validation	None
Examples	pUZ8 PO100 of HfrR4 Unknown Plasmid free

Name	**Plasmids collections fields**
Description	Information about availability of strain plasmids in CCs of plasmids.
Syntax	It should include the name of the plasmid followed by the CC number in parentheses. More than one plasmid can be reported, separated by ";".
Values	Plasmid names should be provided as free text. CC numbers should be composed by the CC acronym followed by a number separated by a space.
Validation	Check the syntax of the information. Report errors to the CC.

Examples	pUZ8 (LMBP 8011)

Name	**Substrate/host of isolation**
Description	Information about the substrate and the host of isolation of the strain. It may include the detailed substrate from which the strain was isolated and the name of host plant/animal. By now, this information should be provided as free text. Future improvements of the dataset will likely foresee some syntactical rules and/or list of values.
Syntax	None
Values	Free text
Validation	None
Examples	Soil under Pinus sylvestris. Flowering plant of Helleborus foetidus. Arachis hypogaea.

Name	**Isolation habitat**
Description	Information about the biotope where the species was found. It should include environmental physical factors, such as humidity, range of temperature, pH and light intensity, as well as biotic factors, such as the availability of food and the presence or absence of predators. It may also include information already specified in the related fields Geographic origin, Geographic origin coordinates and Altitude. By now, this information should be provided as free text. Future improvements of the dataset will likely foresee some syntactical rules and/or list of values.
Syntax	None
Values	Free text
Validation	None
Examples	Tropical rain forest Salt marsh, Salicorna habitat. Forest litter, radioactivity 1.5x10(4) Bq/kg.

Name	**Ontobiotope term for the isolation habitat**
Description	Information about the habitat where the species was found provided by using the most specific term(s) of the Ontobiotope ontology of microorganism habitats. Note that this ontology is mainly for bacteria.
Syntax	The id(s) of the term(s) should be provided. An id includes the prefix “OBT:” followed by an integer of six digits, as in “OBT:001119” for forest. Alternatively, the preferred name(s) for the term(s) can also be provided. When submitting more id(s) or preferred name(s), separate them by a semicolon “;”. Id(s) and preferred names(s) must not be mixed.
Values	Any valid term id or preferred name from the Ontobiotope.

	See the Ontobiotope browser.
Validation	Check for validity of ids and preferred names. Check for the syntax.
Examples	OBT:001119; OBT:002941 Forest

Name	**Genomic sequences and accession numbers**
Description	Known genomic sequences and related INSDC accession numbers of the strain. According to the resource type, these include, but are not limited to, the nuclear ribosomal Internal Transcribed Spacer (ITS), the nuclear ribosomal Large SubUnit (LSU) and the 16S rRNA gene. Any further gene or marker that is considered of relevance by the CC, such as Calmodulin (CaM) and β-Tubulin, can be included. These data must be submitted in separate table including, in distinct fields, the following information: accession number of the strain in the CC, marker name, INSDC accession number of the marker sequence, sequence.
Syntax	Fields in the table follow different syntaxes: Strain accession number: as defined in the related field of the MIRRI-IS dataset. Marker name: the short name of the marker. INSDC accession number: An INSDC accession number is an alphanumeric code made by a fixed number of letters followed by a fixed number of digits, without any separation. For sequences, the code is currently made of two letters followed by six numbers. Sequence: Any valid genomic sequence.
Values	Values of fields in the table are as follows: Strain accession number: any accession number in the CC. Marker name: any common marker designation. INSDC accession number: Any valid INSDC accession number. Sequence: Genomic sequence, any format, any length.
Validation	Check for the validity of the syntaxes, formats and values. Check that the sequence in INSDC actually relates to the named gene sequence of the given strain. Report errors and discrepancies to the CC.
Examples	See attached table.

Name	**Literature linked to the sequence/genome**
Description	Information on literature linked to the sequences or genome of the strain. Do not include here literature linked to the identification and properties of the strain. Include identifiers linking to a separate literature sheet in the same file. For publications indexed by Pubmed or having an official DOI number, collections should provide the relative identifiers, respectively PMIDs and DOIs. In this case, MIRRI-IS will retrieve additional information and complete the bibliographic data.

	When neither a PMID nor a DOI are available, all usual bibliographic fields used for citing a paper, a book, a patent, or a document available on-line, including, e.g., authors, title, journal, volume, issue, pages, editors, publishers, etc… must be submitted as separate fields in a distinct table. In this case, identifiers linking to the separate literature sheet must be included here. Multiple papers can be included for a single strain just by reporting more PMIDs, DOIs, and table identifiers separated by ";".
Syntax	PMIDs, DOIs, and numbers separated by a semicolon ";"
Values	Valid PMIDs and DOIs or reference numbers of the literature sheet.
Validation	MIRRI-IS will try to extract any missing PMID and DOI on the basis of the provided information and return it to the collection. Any errors and inconsistencies will also be reported to the CCs.
Examples	120; 26492633; 10.1371/journal.pcbi.1004525