MIRRI-IS PLASMID DATASET

(version 2022.03.1)

Benyahia Z, Figge M, Olbrechts A, Romano P.

Introduction 3
Correspondences among MIRRI-IS Guidelines 3
Mandatory and non-mandatory fields 3
Fields' relevance for sorting of results 3
Grouping of information 3
Literature 5
Dates 5
Data fields description 6
Identity group 6
MIRRI Accession number (MANDATORY FIELD) 6
Collection number (MANDATORY FIELD) 6
Plasmid name (MANDATORY FIELD) 7
Other culture collection numbers 7
Other denomination 8
Parental clone 8
Distribution group 8
Restrictions on use (MANDATORY FIELD) 8
Additional restrictions 9
Nagoya protocol compliance conditions (MANDATORY FIELD) 9
ABS related files 10
MTA file 10
Strain from a Registered Collection 11
Form of supply (MANDATORY FIELD) 11
Host for distribution (MANDATORY FIELD) 11
Genotype of the host for distribution 12
Recommended growth temperature (MANDATORY FIELD) 12
Recommended medium for growth (MANDATORY FIELD) 12
Cultivation remarks 13
Biological information group 13
Organism type (MANDATORY FIELD) 13

Class 14
Risk group (MANDATORY FIELD) 14
Risk group remarks 14
Dual use 15
Plasmid status 15
GMO 16
GMO construction information 16
Origin information group 16
Geographic origin 16
Country (MANDATORY FIELD) 17
Deposit information group 18
History of deposit 18
Depositor 18
Date of deposit 18
Constructed / isolated by 19
Date of inclusion in the catalogue 19
Properties group 19
Cloned DNA 20
Promoter 20
Terminator 20
Replicon (MANDATORY FIELD) 21
Ribosome binding site 23
Applications 23
Selection marker 24
Host range 24
Remarks 25
Sequence information group 25
Plasmid sequences accession numbers 25
Plasmid sequencing results 25
Sequence file 26
Circular map 26
Sequence detail 27
Plasmid length 27
Literature information group 28
Literature 28

MIRRI-IS PLASMID DATASET

(version 2022.03.1)

Benyahia Z, Figge M, Olbrechts A, Romano P.

Introduction

This document specifies guidelines for uploading data of **catalogs of plasmids** from culture collections of MIRRI partners to the Information System of MIRRI (MIRRI-IS).
They were required because the information that is needed for the description of plasmids differs significantly from the information required for microbial strains (archaea, bacteria, yeasts, filamentous fungi, microalgae), for which the 'Guidelines for catalog uploading' were already delivered by the ICT Task Force of MIRRI.

Correspondences among MIRRI-IS Guidelines

Although these guidelines include data fields which are specific for plasmids, there also are data fields which are the same or similar to fields described for microbial strains.
In order to highlight them, existing correspondences between these two guidelines are annotated for each data field in the new field feature labeled "MIRRI-IS DATASET (version 2020.06.1)", where the corresponding field of the guidelines for microbial strains is reported, when available. If the field is not exactly the same, a short note on differences is also reported. New fields, i.e. fields which are specific for plasmid data, are annotated with the term "N/A", meaning "not available".

Mandatory and non-mandatory fields

Mandatory fields are highlighted. MIRRI-IS will discard records missing mandatory data fields and inform the CC. All other data fields are strongly recommended and collections are invited to submit all data they have. When no information is available for a given data field, it should be left empty.

Fields' relevance for sorting of results

Results of queries carried out on the MIRRI-IS integrated catalog will be sorted by relevance first and then by completeness of data. This means that plasmids having the same relevance to the query will be sorted on the basis of the quantity of filled in data fields.
To this aim, a list of non-mandatory fields which are considered more important than others, and should therefore be used for sorting results, have been defined as follows:
Sequence file, Circular map, Cloned DNA, Promoter, Terminator, Replicon, Selection marker, Applications, Host range, Ribosome binding site, Literature.

Grouping of information

In this document, data fields were grouped for clarity sake on the basis of their relatedness to the following information categories: Identity, Distribution, Taxonomy, Origin, Deposit, Properties, Sequence, Literature. Such grouping can support a better understanding of the required information,

but it does not have any immediate consequence. It could however be considered for data visualization, for the definition of a format for data exchange and for targeted web services.

Group	Fields
Identity	MIRRI Accession number (MANDATORY FIELD) Collection number (MANDATORY FIELD) Name Plasmid (MANDATORY FIELD) Other culture collection numbers Other denomination Parental clone
Distribution	Restrictions on use (MANDATORY FIELD) Additional restrictions Nagoya protocol compliance conditions (MANDATORY FIELD) ABS related files MTA file Strain from a Registered Collection Recommended growth temperature (MANDATORY FIELD) Recommended medium for growth (MANDATORY FIELD) Form of supply (MANDATORY FIELD) Host for distribution (MANDATORY FIELD) Genotype of the host for distribution
(Taxonomy) Biological information	Organism type (MANDATORY FIELD) Class Risk group (MANDATORY FIELD) Risk group remarks Dual use Plasmid status GMO GMO construction information
Origin	Geographic origin Country (MANDATORY FIELD)
Deposit	History of deposit Depositor Date of deposit Constructed / isolated by Date of inclusion in the catalogue
Properties	Cloned DNA Promoter Terminator Replicon (MANDATORY FIELD) Ribosome binding site Applications Selection marker Host range

	Remarks
Sequence information	Plasmid sequences accession numbers Plasmid sequencing results Sequence file Circular map Sequence detail Plasmid length
Literature	Literature

Literature

For publications indexed by Pubmed, CCs should provide the relative Pubmed IDs. When available, DOIs should also be provided. When either the Pubmed ID or the DOI are available, there is no need to submit further information. On the contrary, when neither Pubmed ID nor DOI are available, complete information on authors, title, journal, volume, issue, pages, must be submitted as separate data fields.

Dates

Some dates are requested by the MIRRI-IS dataset and it can be useful to clarify their meaning in this context.

- Collection: when the sample is collected, usually in some habitat.
- Isolation: when the strain is isolated, usually in a laboratory.
- Identification: when the strain is identified with the current taxon.
- Deposit: when the strain is deposited at the collection.
- Accession: when the strain is included in the catalog or an accession number is assigned to it.

MIRRI-IS PLASMID DATASET (version 2022.03.1)

Data fields description

Identity group

Name	**MIRRI Accession number** (MANDATORY FIELD)
Short name	mirriAccessionNumber
Description	Unique identifier of the plasmid 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 seven digits.
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 plasmid collection number was already included. If not, a new value will be created and returned to the collection. If the plasmid collection 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 plasmid collection number are properly associated. If not, inform the collection.
Examples	MIRRI1234567
MIRRI-IS DATASET (version 2020.06.1)	Same as "MIRRI Accession number".

Name	**Collection number** (MANDATORY FIELD)
Short name	collectionNumber
Description	Unique identifier of the plasmid in the collection.
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 collection 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	LMBP 4227
MIRRI-IS DATASET (version 2020.06.1)	Same as “Accession number”.

Name	**Plasmid name (MANDATORY FIELD)**
Short name	plasmidName
Description	Name of the plasmid, as given by the depositor or as published.
Syntax	None
Values	Free text
Validation	None
Examples	pUC18 pCAGGS pLenti6-V5 puro
MIRRI-IS DATASET (version 2020.06.1)	N/A.

Name	**Other culture collection numbers**
Short name	otherCollectionNumbers
Description	Collection numbers of the same plasmid in other CCs, when known.
Syntax	Accession numbers formatted as above specified and separated by a semicolon character. Should not include collection numbers that do not follow the relative syntax.
Values	Free text, according to defined syntax.
Validation	For collection numbers of plasmids of CCs available in MIRRI-IS: o control that taxon names or synonyms are identical and if the assigned name is not the current name warn the original CC; o if the name is incorrect warn all the CCs having the plasmid.
Examples	DSM 4439; NCIB 12392; NCCB 3011; PC-V 3010
MIRRI-IS DATASET (version 2020.06.1)	Same as “Other culture collection numbers”.

Name	Other denomination
Short name	otherDenomination
Description	Unofficial names that are often used for the plasmid, 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	
MIRRI-IS DATASET (version 2020.06.1)	Same as “Other denomination”.

Name	Parental clone
Short name	parentalClone
Description	Names of the plasmids it originates from. The same information of the related plasmid name field must be specified so that a link can be established.
Syntax	When more than one name, they should be separated by a semicolon character ';'
Values	Free text
Validation	Check for the existence of the target plasmid (parent) and add an internal link to its record.
Examples	pUC18 pCAGGS pcDNA3.1 pLenti6-V5 puro
MIRRI-IS DATASET (version 2020.06.1)	N/A.

Distribution group

Name	Restrictions on use (MANDATORY FIELD)
Short name	useRestrictions
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 known restrictions apply), 2 (only for non-commercial use), 3 (for commercial development a special permission is

	requested), 4 (permission from depositor is requested for all uses)
Validation	Check that one and only one of the allowed values is used. Report errors to the CC.
Examples	1
MIRRI-IS DATASET (version 2020.06.1)	Same as “Restrictions on use”.

Name	**Additional restrictions**
Short name	additionalRestrictions
Description	Special restrictions imposed by the depositor at the time of deposit, with request to be mentioned somewhere.
Syntax	None
Values	Free text
Validation	None
Examples	“Restricted to academic users” “The depositor will be informed of the customer’s identity upon release of a sample” "Acknowledgment of the depositor in any publication resulting from the use of the MATERIAL” "RECIPIENT agrees to refer to the ’Plasmid reference’ in the first publication making use of the MATERIAL.” "When appropriate in accordance with academic customs, RECIPIENT agrees to include the depositor(s) as co-author(s) in the first publication making use of the MATERIAL” "When making use of the MATERIAL, RECIPIENT agrees to refer to ‘OPENPichia’ in the Materials and Methods section and in the acknowledgments of all publications, as well as when launching commercial products” "Only for research and teaching purposes”
MIRRI-IS DATASET (version 2020.06.1)	N/A.

Name	**Nagoya protocol compliance conditions (MANDATORY FIELD)**
Short name	nagoyaConditions
Description	Situation of the plasmid 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
MIRRI-IS DATASET (version 2020.06.1)	Same as "Nagoya protocol compliance conditions".

Name	**ABS related files**
Short name	absFiles
Description	URL of the IRCC document permitting the holding of the plasmid at the CC.
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
MIRRI-IS DATASET (version 2020.06.1)	Same as "ABS related files".

Name	**MTA file**
Short name	mtaFile
Description	Plasmid specific Material Transfer Agreement (MTA) document
Syntax	List of file names, separated by a semicolon character. Files must be made available all in the same folder, accessible on-line and the URL of the folder must be provided by the CC to MIRRI-IS. Links to MTA files will then be created by MIRRI-IS according to the user requests. Alternatively, CCs may provide working URLs for the files.
Values	Free text including filenames or URLs.
Validation	Check that the path of the folder is accessible and that files with the given names are available on-line. Alternatively, check URLs. Report errors to the CC.
Examples	mta_file1.pdf ; mta_file2.pdf http://www.domain.tld/path/mta_file.pdf

MIRRI-IS DATASET (version 2020.06.1)	Same as “MTA file”.

Name	**Strain from a Registered Collection**
Short name	registeredCollection
Description	Plasmid included in the registered CC according to the EU Regulation 511/2014. Unregistered CCc 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
MIRRI-IS DATASET (version 2020.06.1)	Same as “Strain from a Registered Collection”.

Name	**Form of supply (MANDATORY FIELD)**
Short name	supplyForms
Description	The forms of supply of the strain to users.
Syntax	One or several of the allowed values, separated by a “;”.
Values	Allowed values: 1 (for AGAR), 2 (for CRYO), 3 (for Dry Ice), 4 (for Liquid Culture Medium), 5 (for LYO), 6 (for OIL), 7 (for WATER), 8 (plasmid DNA).
Validation	Check for the validity of the format. Report errors to the CC.
Examples	4; 8
MIRRI-IS DATASET (version 2020.06.1)	Same as “Form of supply”, with new values on purpose.

Name	**Host for distribution (MANDATORY FIELD)**
Short name	distributionHost
Description	Strain designation of the host organism in which the plasmid is distributed. This field should consist of three parts: taxonomic designation, group (when applicable), strain designation. In case a helper plasmid is present, put it here between squared brackets [ ].
Syntax	None
Values	Free text.

	http://cabri.org/guidelines/catalogue/CPdishostplasm.html
Validation	None
Examples	Escherichia coli K12 DH5α Escherichia coli K12 MC1061(λ) Escherichia coli K12 SM10(λpir) Escherichia coli K12xB DB3.1
MIRRI-IS DATASET (version 2020.06.1)	N/A.

Name	**Genotype of the host for distribution**
Short 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	gyrA462 endA1 Δ(sr1-recA) mcrB mrr hsdS20 glnV44 (=supE44) ara14 galK2 lacY1 proA2 rpsL20 xyl5 leuB6 mtl1
MIRRI-IS DATASET (version 2020.06.1)	Same as “Genotype”, with reference to the host for distribution.

Name	**Recommended growth temperature (MANDATORY FIELD)**
Short name	recommendedTemperature
Description	The recommended growing temperature for the plasmid carrying host.
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
MIRRI-IS DATASET (version 2020.06.1)	Same as “Recommended growth temperature”.

Name	**Recommended medium for growth (MANDATORY FIELD)**

Short name	recommendedMedium
Description	The medium that is recommend for growing the plasmid carrying host.
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	LB-Lennox + ampicillin (100 µg/ml) + kanamycin (50 µg/ml) + streptomycin (25 µg/ml) M17 + 0.5% glu (GM17) + chloramphenicol (25 µg/ml)
MIRRI-IS DATASET (version 2020.06.1)	Same as “Recommended medium for growth”.

Name	**Cultivation remarks**
Short name	cultivationRemarks
Description	Remarks on cultivation
Syntax	None
Values	Free text
Validation	None
Examples	"Selection of transformants on chloramphenicol; subsequent cultivation of a single colony in liquid medium with ampicillin.” "This culture grows very slow: take 24 hours into account for saturated growth!"
MIRRI-IS DATASET (version 2020.06.1)	N/A

Biological information group

Name	**Organism type** (MANDATORY FIELD)
Short name	organismType
Description	The type of the resource

Syntax	One of the allowed values.
Values	7 (for Plasmids)
Validation	Check that the allowed value is used. Report errors to the CC.
Examples	7
MIRRI-IS DATASET (version 2020.06.1)	Same as “Organism type”, but only the plasmid specific value is allowed.

Name	**Class**
Short name	Class
Description	Origin of the element.
Syntax	One of the allowed values.
Values	One of the following values: "Recombinant plasmid" or "Natural plasmid"
Validation	Check for the validity of the format. Report errors to the CC
Examples	Recombinant plasmid
MIRRI-IS DATASET (version 2020.06.1)	N/A.

Name	**Risk group (MANDATORY FIELD)**
Short name	riskGroup
Description	Risk group in the host for distribution, usually E. coli, according to regulations (EU Directive 2000/54/EC and its amendments and corrections, EU directive 2009/41/EC and its updates).
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
MIRRI-IS DATASET (version 2020.06.1)	Same as “Risk Group”.

Name	**Risk group remarks**
Short name	riskGroupRemarks

Description	Specify when risk group is different in other hosts.
Syntax	None
Values	Free text
Validation	None
Examples	L1 in E. coli; L2 in mammalian cells
MIRRI-IS DATASET (version 2020.06.1)	N/A

Name	**Dual use**
Short name	dualUse
Description	Specify whether the strain has the potential for a harmful use according to EU regulations (EU Council Regulation 2000/1334/CE and its amendments and corrections, EU Reg 2021-821).
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
MIRRI-IS DATASET (version 2020.06.1)	Same as “Dual use”.

Name	**Plasmid status**
Short name	plasmidStatus
Description	Based on the status of data processing, incl. quality control tests that have already been performed, plasmids can be classified as preliminary, non-core and core.
Syntax	One of the allowed values
Values	Allowed values: “Preliminary”, “Non-core”, “Core”
Validation	None
Examples	Check that one and only one of the allowed values is used. Report errors to the CC.
MIRRI-IS DATASET (version 2020.06.1)	N/A

Name	**GMO**
Short name	gmo
Description	Specify whether the plasmid 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
MIRRI-IS DATASET (version 2020.06.1)	Same as “GMO”.

Name	**GMO construction information**
Short name	gmoConstruction
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
MIRRI-IS DATASET (version 2020.06.1)	Same as “GMO construction information”.

Origin information group

Name	**Geographic origin**
Short name	geographicOrigin
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.
MIRRI-IS DATASET (version 2020.06.1)	Same as "Geographic origin".

Name	**Country (MANDATORY FIELD)**
Short name	Country
Description	The country where the sample was collected or constructed. This information should be provided with reference to the name of the country where the locality is located now, if possible. Former country names should be specified only when the locality or its current country are unknown.
Syntax	This information must be expressed by using the ISO-3166 standard for country codes. The preferred set is ISO 3166-1 alpha-2 (two letters code), but ISO 3166-1 alpha-3 (three letters code) is also accepted. Former country codes must follow standard's part three ISO 3166-3 (four letters code). Only one code can be included.
Values	All codes included in ISO 3166-1 alpha-2 (two letters code), ISO 3166-1 alpha-3 (three letters code) and ISO 3166-3 (four letters code). In the examples, IT is the two letters code for Italy, GBR is the three letters code for United Kingdom, and CSHH is the four letters code for the former country Czechoslovakia.
Validation	Check for the validity of codes. Return errors to CCs.
Examples	IT GBR

	CSHH
MIRRI-IS DATASET (version 2020.06.1)	N/A

Deposit information group

Name	**History of deposit**
Short name	depositHistory
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, possibly according to above syntax
Validation	Check for the validity of the format. Report errors to the CC.
Examples	
MIRRI-IS DATASET (version 2020.06.1)	Same as “History of deposit”.

Name	**Depositor**
Short name	Depositor
Description	Name, institute and town / country of the depositor.
Syntax	None
Values	Free text
Validation	None
Examples	M. Mergeay and A. Provoost, SCK.CEN, Mol, Belgium
MIRRI-IS DATASET (version 2020.06.1)	Same as “Depositor”.

Name	**Date of deposit**
Short name	depositDate
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
MIRRI-IS DATASET (version 2020.06.1)	Same as “Date of deposit”.

Name	**Constructed / isolated by**
Short name	constructor
Description	Name, institute and town / country of the constructor or isolator.
Syntax	None
Values	Free text
Validation	None
Examples	J. Fraser, Moredun Res. Inst., Edinburgh, UK
MIRRI-IS DATASET (version 2020.06.1)	N/A

Name	**Date of inclusion in the catalogue**
Short name	accessionDate
Description	Date when the plasmid 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
MIRRI-IS DATASET (version 2020.06.1)	Same as “Date of inclusion in the catalogue”.

Properties group

Name	**Cloned DNA**
Short name	clonedDNA
Description	Gene or other DNA fragment cloned into the MCS (Multiple Cloning Site) For genes, it should include, when appropriate, the following information: species name, gene name, gene symbols and gene ID (NCBI Gene database). For tags/epitopes/recognition sites there is no gene ID available.
Syntax	None
Values	Free text
Validation	When a geneID is available, the visualization should be completed by a link to NCBI Gene. See e.g. https://www.ncbi.nlm.nih.gov/gene/?term=948538
Examples	Yersinia enterocolitica outer membrane adhesin protein gene (yadA, P1, yopA) Histidine tag (His-tag); C-terminal Escherichia coli maltose ABC transporter periplasmic binding protein cDNA (malE, GeneID 948538); mature sequence Dermatophagoides farinae allergen Der f 2 cDNA (Der f 2, Der f2)
MIRRI-IS DATASET (version 2020.06.1)	N/A.

Name	**Promoter**
Short name	Promoter
Description	Specific DNA region required for initiation of transcription For promoters, it should include, when appropriate, the following information: species name, gene name, and gene symbols.
Syntax	None
Values	Free text
Validation	None
Examples	Escherichia coli lac operon promoter Human cytomegalovirus immediate early promoter (CMV-IE) and enhancer Phage SP6 promoter Simian virus 40 early promoter (SV40 early)
MIRRI-IS DATASET (version 2020.06.1)	N/A.

Name	**Terminator**
Short name	Terminator

Description	Specific DNA region causing termination of transcription For terminators, it should include, when appropriate, the following information: species name, gene name, and gene symbols.
Syntax	None
Values	Free text
Validation	None
Examples	Aspergillus nidulans tryptophan C terminator (trpC) Bovine growth hormone polyadenylation signal (BGH polyA) Simian virus 40 polyadenylation signal (SV40 polyA)
MIRRI-IS DATASET (version 2020.06.1)	N/A.

Name	**Replicon (MANDATORY FIELD)**
Short name	Replicon
Description	Names of replicon elements (sequences allowing initiation of replication, enabling plasmids to reproduce themselves and survive within their host).
Syntax	Check the CABRI agreed list at http://www.cabri.org/guidelines/catalogue/CPrepliconplasm.html and use those terms, if possible. Also check the values listed here below.
Values	Free text. Examples of possible texts are: Acinetobacter calcoaceticus plasmid pWH1277 origin Bacillus thuringiensis plasmid pHT1030 origin Bacteriophage M13 origin Bacteriophage M13 origin; with deletion Bifidobacterium catenulatum L48 plasmid pBC1 origin BK virus origin (BKV) Bovine papilloma virus origin (BPV) Broad-host-range Gram-negative Bordetella bronchiseptica S87 plasmid pBBR1 replicon Broad-host-range plasmid pSa replication origin Broad-host-range plasmid pWV01 origin; temperature sensitive variant (pVE6007) Broad-host-range plasmid RK2 origin of transfer (oriT) Broad-host-range plasmid RK2/RP4 origin of transfer (oriT) Broad-host-range plasmid RK2 vegetative replication origin Broad-host-range plasmid RSF1010 origin of transfer (oriT) Broad-host-range plasmid RSF1010 vegetative replication origin; IncQ Broad-host-range plasmid RSF1030 origin of replication Candida albicans autonomously replicating sequence ARS

	Candida albicans autonomously replicating sequence ARS2 Candida albicans autonomously replicating sequence ARS3 Chloroplast replication origin (oriA) Enterobacter agglomerans plasmid pPIGDM1 origin Enterobacter cloacae plasmid ClDF13 origin Enterobacter cloacae plasmid CloDF13 origin; derivative (CDF) Enterococcus faecalis plasmid pAD1 origin Enterococcus faecalis plasmid pAMß1 origin, incl. repD and repE Enterococcus faecalis plasmid pAMß1 origin, modified sequence incl. repD and repE Epstein-Barr virus origin of replication (oriP) Escherichia coli incompatibility group P plasmid origin (IncP) Escherichia coli plasmid ColE1 origin Escherichia coli plasmid p15A origin Escherichia coli plasmid pMB1/ColE1 origin Escherichia coli plasmid pMB1 origin Escherichia coli plasmid pMB1 origin; fragment Escherichia coli plasmid pMB1 origin; truncated Escherichia coli plasmid pSa origin Escherichia coli plasmid R6K origin; defective pir gene Human adenovirus 5 origin (Ad5) IncA Lactobacillus plantarum A112 plasmid pA1 origin, incl. repA and repB Mycobacterium fortuitum plasmid pAL5000 origin (OriM) Mycobacterium fortuitum plasmid pAL5000 temperature-sensitive origin (tsOriM) Phage f1 origin Pichia pastoris autonomously replicating sequence (ARS1) Plasmid pSC101 origin derived from R6-5 Plasmid pSC101 origin derived from R6-5; temperature sensitive Polyoma virus origin Pseudomonas plasmid pVS1 origin Runaway replication mutant of plasmid R1 (R1drd19) Saccharomyces cerevisiae 2 micron plasmid origin (2µ) Saccharomyces cerevisiae 2 micron plasmid origin (2µ); incl. region conferring stability (STB) Saccharomyces cerevisiae autonomously replicating sequence (ARS1) Saccharomyces cerevisiae centromeric region 4 (CEN4) Saccharomyces cerevisiae centromeric region 6 (CEN6)/ S. cerevisiae autonomously replicating sequence ARSH4 Simian virus 40 bidirectional origin (SV40) Staphylococcal plasmid pC194 origin Staphylococcal plasmid pC194 origin; mutant Streptococcus parasanguinis plasmid pFW213 origin

	Yersinia enterocolitica plasmid pYV227 ori Yersinia enterocolitica plasmid pYV40 ori
Validation	None
Examples	ColE1; pMB1 Escherichia coli plasmid pMB1 origin Escherichia coli plasmid R6K origin; defective pir gene Simian virus 40 bidirectional origin (SV40) Saccharomyces cerevisiae 2 micron plasmid origin (2µ), incl. region conferring stability (STB)
MIRRI-IS DATASET (version 2020.06.1)	N/A.

Name	**Ribosome binding site**
Short name	ribosomeBindingSite
Description	Sequence upstream of a gene, which, after transcription, will bind to ribosomes in order to initiate translation
Syntax	None
Values	Free text
Validation	None
Examples	Ribosome binding site (RBS) of the Moraxella strain TA144 lipase 2 gene (lip2) Ribosome binding site (RBS) of the Escherichia coli lac Z gene (lacZ) Internal ribosome entry site (IRES) of the encephalomyocarditis virus (EMCV) polyprotein gene
MIRRI-IS DATASET (version 2020.06.1)	N/A.

Name	**Applications**
Short 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. If possible, logically valid combinations of the following values: Empty cloning/expression vector, Reporter vector, Gateway vector, Retroviral vector, RNAi vector, Crispr vector, Suicide vector, Promoter analysis vector, Yeast two-hybrid vector, Bicistronic vector, Glycosylation vector, Human insert vector,

	Mouse insert vector, Bacterial insert vector, Fungal insert vector, Yeast insert vector, Plant insert vector, Viral insert vector, Other insert vector, epitope mapping, export signal probe, general cloning vector, insertion vector, insertion/deletion mutagenesis, positive/negative screening vector, promoter probe vector, protein production plasmid, replacement vector, sequencing vector, tool plasmid, transcomplementation of vectors for mobilization, transcription termination probe, translation initiation probe, YAC
Validation	None
Examples	Insertion vector
MIRRI-IS DATASET (version 2020.06.1)	Same as “Applications”.

Name	**Selection marker**
Short name	selectionMarker
Description	Feature(s) encoded by the plasmid enforcing its host strain to maintain the plasmid in order to survive under appropriate cultivation conditions, allowing selection for plasmid-containing host cells
Syntax	None
Values	Free text
Validation	None
Examples	Ampicillin (amp) Blasticidin (bsd) Neomycin (neo; G418; kanamycin (kan)) LEU2; auxotrophic
MIRRI-IS DATASET (version 2020.06.1)	N/A.

Name	**Host range**
Short name	hostRange
Description	Host(s) in which the element can be propagated and/or used and, when applicable, the peculiar features that are necessary for making use of the special characteristics of the plasmid.
Syntax	None
Values	Free text
Validation	None
Examples	Escherichia coli; use a ccdB-resistant strain for propagation

	Candida albicans; leu2-, integrative Gram-negative bacterial strains Mammalian cells; SV40 permissive cells
MIRRI-IS DATASET (version 2020.06.1)	N/A.

Name	**Remarks**
Short name	Remarks
Description	Any further note that is not present in the other fields.
Syntax	None
Values	Free text
Validation	None
Examples	
MIRRI-IS DATASET (version 2020.06.1)	Same as “Remarks”.

Sequence information group

Name	**Plasmid sequences accession numbers**
Short name	plasmidSequencesAN
Description	Known INSDC accession numbers of the plasmid.
Syntax	List of INSDC accession numbers, separated by semi-colon “;”. The INSDC accession number is an alphanumeric code made by a fixed number of letters followed by a fixed number of digits, without any separation sometimes with underscore, plus version .1, .2, etc…
Values	List of any valid INSDC accession number.
Validation	Check for the validity of the syntaxes, formats and values (INSDC must exist). Add link when showing the data. Report errors and discrepancies to the CC.
Examples	BC110618.1; AF292100.2; LVXZ01000065.1; NC_001147.5; U79115
MIRRI-IS DATASET (version 2020.06.1)	N/A

Name	**Plasmid sequencing results**
Short name	plasmidSequencingResults

Description	The sequences must be provided on-line in separate files whose URLs must be included in this field.
Syntax	URLs separated by space.
Values	Valid URLs
Validation	Check for the existence of the files at the URLs.
Examples	https://bccm.belspo.be/catalogues/files/lmbp-plasmids/SangerSequences/ef32243390.fasta https://bccm.belspo.be/catalogues/files/lmbp-plasmids/NextGenSequences/c02-gc-dec2018.fasta
MIRRI-IS DATASET (version 2020.06.1)	N/A

Name	**Sequence file**
Short name	sequenceFileName
Description	GenBank formatted file with nucleotide sequence of the plasmid
Syntax	Files must be made available in the same folder, accessible on-line and the URL of the folder must be provided by the CC to MIRRI-IS. Link to the sequence file will then be created by MIRRI-IS. Alternatively, CCs may provide working URLs for the files.
Values	Free text including filename or URL
Validation	Check that the path of the folder is accessible and that the file with the given name is available on-line. Alternatively, check URL. Report errors to the CC.
Examples	
MIRRI-IS DATASET (version 2020.06.1)	N/A.

Name	**Circular map**
Short name	Map
Description	pdf file with circular representation of the different features of the plasmid
Syntax	Files must be made available in the same folder, accessible on-line and the URL of the folder must be provided by the CC to MIRRI-IS. Link to the circular map file will then be created by MIRRI-IS. Alternatively, CCs may provide working URLs for the files.
Values	Free text including filename or URL
Validation	Check that the path of the folder is accessible and that the file with the given

	name is available on-line. Alternatively, check URL. Report errors to the CC.
Examples	
MIRRI-IS DATASET (version 2020.06.1)	N/A.

Name	**Sequence detail**
Short name	sequenceDetail
Description	Detailed nucleotide sequence of specific regions for clarifying reasons, incl. reading frame punctuation, designation of restriction enzymes, start & termination codon and/or amino acid number & 3 letter code: e.g. for: - mutated versus wild-type sequence - used primers for construction or … - junction between two fragments cloned together - at the start or the end of fusion genes (with tags in frame or not) - etc…
Syntax	monospaced font required for display otherwise restriction enzymes and amino acid codes are not put on the correct place below the nucleotide sequence
Values	
Validation	
Examples	See https://bccm.belspo.be/catalogues/lmbp-plasmids-plasmid-details?NUM=4123
MIRRI-IS DATASET (version 2020.06.1)	N/A.

Name	**Plasmid length**
Short name	plasmidLength
Description	number of nucleotides
Syntax	An integer or decimal positive number, followed by ‘bp’ for ‘base pairs’ or ‘kb’ for ‘kilobase pairs’.
Values	Positive numbers followed by ‘bp’ or ‘kb’.
Validation	Check syntax. Report errors.
Examples	4,5 kb 4589 bp
MIRRI-IS DATASET (version 2020.06.1)	N/A.

Literature information group

Name	Literature
Short name	Literature
Description	Information on literature, only for the plasmids not for the host. This information must be provided in a distinct table that includes all bibliographic references for plasmids of the collection, each of which must be identified by a unique id, e.g. a progressive number. 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 the literature data field, identifiers that are relevant to the described plasmid must be included separated by a semi-colon ";".
Syntax	Identifiers of the relevant bibliographic references, separated by a semicolon ";"
Values	Identifiers present in the separate 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; 130; 210
MIRRI-IS DATASET (version 2020.06.1)	Same semantics as "Literature", but with revised contents (all bibliographic references are included in this field) and format (PMIDs and DOIs must be included in the separate table).