FAIR data management using pISA-tree: Standard project directory tree

A. Blejec, Ž.Ramšak, M.Petek, Š. Baebler, A. Coll, K. Gruden

2018-01-08

This protocol describes a system for organisation of your experiment in the findable accessible interoperable and reproducible (FAIR) manner, thus allowing integrative multiscale/multilevel analysis. It is set in accordance with ISA-tab standard and is compatible with FAIRdom, using SEEK and JERM systems as a basis.

To properly manage and annotate the data within the project, one needs to design pISA project **before** the samples are eventually collected. Thus data management setup should be done in parallel with the Wet-lab Experimental Setup. So far this was not done in a systematic way as experiments were in general less complex and included mainly only one type of variables measured per one sample (e.g. only microarray analysis was done combined with limited qPCR analysis...). When dealing with experiments with data collected with multiple technologies (NGS, microarrays, qPCR...), for different molecular levels (mRNA, miRNA, proteins, metabolites), together with structural information (macroscopic, SM and EM data) the data can only be properly analysed if organised in the way described below.

When designing the experiments it would be best if also data management plan would be prepared. This would allow proper management of resources for storage and analysis data, as well define the vocabularies used for data annotation and expose other problems related with interopearability. And example of **data management plan** is stored in the same folder as this Instructions (DM\_indie\_v3.doc).

pISA is organised in concordance with PrimerDB, quantGenius (formerly known as qPCR\_Calculator) and GeneCloud applications and will allow your experiment to be stored and analysed as smoothly as possible with the above mentioned tools as well as with applications developed in R-environment (currently all by AndrejB).

# What is the pISA-tree?

pISA-tree provides a set of batch files that is used to create a standard directory tree for research projects.

## Layers

At the top layer, you have to create a directory (here called pISA\_projects root directory) which will contain future projects. The root directory is the place for initial batch files. Step one is to copy/download the batch files from the <https://github.com/ablejec/pISA/> and install it into the folder[**..\..\pISA\_Projects**](file:///\\srvljfs\fito\DEJAVNOSTI\OMIKE\pISA-Projects) into the root directory. Detailed instructions for installations are given in README.MD file (generated when unzipping the downloaded application).

**Project** is organized as a collection of one or more **investigations**. An **investigation** is similarly organized as a collection of one or more **studies**. Each **study** has it's own collection of one or more **assays**. Assays can be of specific type (e.g. connected to MicroArray, NGS, Modeling, Statistical analysis, ...) and are structured accordingly.

## Batch files

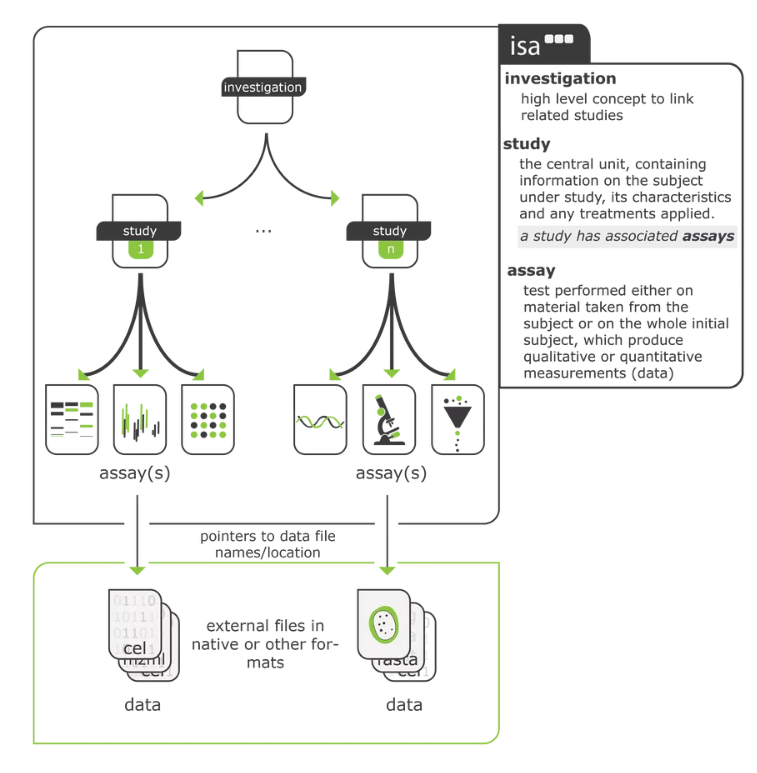
* makeproject.bat - makes a new **project** directory tree
* makeInvestigation.bat - makes a new **investigation** directory tree (subdirectory tree within the **project**)
* makeStudy.bat - makes a new **study** (subdirectory tree within the **investigation**)
* makeAssay.bat - makes a new **assay** (subdirectory tree within the **study**)

# Creation of the directory tree

The directory tree is a way to enforce the subordination of pISA layers. To emphasize the layer type, directory names are constructed automatically using the standard prefix and short layer ID. Standard prefixes are:

* \_p\_for project
* \_I\_for investigation
* \_S\_for study
* \_A\_for assay

Schematic overview of the directory tree, according to ISA-tab standard:



Here are the examples of each level of the directory tree:

|  |  |  |
| --- | --- | --- |
| pISA-tree | Describtion | Example |
| Program | International collaborations | PVY |
| Project |  | Lesions |
| Investigation | *The high level concept to bring together related studies.*  Has the Master sample table & Features summary table | Hormonal treatments |
| Study | *The central unit, containing information on the subject under study, its characteristics and any treatments applied.*  One batch of plants in the growth chamber. A study is related to sampling date and time – e.g. samples taken to be put for qPCR, transcriptomics and proteomics assays.  In the dry lab, a new study should be defined when we are integrating between studies. If the integration is within a study itself, then several assays. | HT\_series1\_sampling1\_stu  HT\_series2\_sampling1\_stu  HT\_series2\_sampling1\_ath |
| Assay | *Test performed either on material taken from the subject or on the whole initial subject, which produce qualitative or quantitative measurements (data).*  One test, it can be a batch of chips, plates if applicable. This can be organised according to researcher preference. Wetlab and drylab assays have different features to assist the researcher and consequently also different structure. | RNA\_Isolation,  qPCR  product\_model-stat  annotation |

## project

To create a new project, run (double click) the file makeProject.bat and enter the project short name (ID). This will make a direcory tree, metadata files and a local copy of makeInvestigation.bat. Short project name (ID), automatically prefixed with \_p\_, is used as the name of the directory (no spaces or special characters, file name conventions apply!!). If you set the project ID as bla the project directory name will automatically become \_p\_bla.

## Investigation

To create a new investigation, run (double click) the file makeInvestigation.bat and enter the study short name (ID). This will make a direcory tree, metadata files and a local copy of makeStudy.bat. Short investigation name (ID), automatically prefixed with \_I\_, is used as the name of the directory (no spaces or special characters, file name conventions apply!!). The investigation directory name for the investigation ble will be \_I\_ble.

## Study

To create a new study, run the makeStudy.bat and enter the study ID (Short name). This will make a direcory tree with several standard folders, metadata and auxiliary files and a local copy of makeAssay.bat. The study folder name will be \_S\_blu for a study with short name (ID) blu.

*Note: New study should be initiated which each new bacth of samples collected!*

## Assay

Analyses for each study are collected in the folder of that study. To make a new assay, run the makeAssay.bat file.

You will be asked to enter assay **Class** [Wet/Dry]:

* Wet: measurements on the biological material (MicroArray, NGS, PCR, ...)
* Dry: process data (Statistcs, Modelling, ... )

Then you will enter the assay **Type** (see documentation for covered types: NGS, MA, STat ...) and assay ID (Short name, for example ASSAY1). Short assay name and type (separated by '-' and prefixed by \_A\_) are used as the name of the assay directory tree (for example: \_A\_ASSAY1-NGS).

Folders in assay directory trees for different Classes slightly differ, according to the need of the specific Class.

Assay classes and types are defined as subfolders of the Templates subfolder. An example is ../Templates/Wet/RT. The folder names define assay Class “Wet” and assay Type “RT”. To add another type of *Wet* assay (named *mytype*) , one can create a new subfolder with the name *mytype*: ../Templates/Wet/mytype. To add another class, create subfolder *myclass*: ../Templates/myclass.

Note: currently only a limited number of assays are available, some by SEEK (modeling and simulations); some mainly related to transcriptomics (RNA isolation, DNAse treatment and reverse transcription, chech the folder for any updates) and statistics by NIB

When creating either of these levels a certain folder structure is created. Descriptions of generated subfolders and needed files is given here:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Folder | INVESTIGATION | STUDY | ASSAY wetL | ASSAY dryL |
| Additional files | metadata file  master sample table = phenodata.txt  FeatureSummary table | metadata file | metadata file following specific minimal standards  Feature table = featuredata.txt | Description file following specific minimal standards  Feature table = featuredata.txt |
| /input/ | NG | NG | NG | input for the analysis |
| /reports | project files (e.g. applications) which make it easier to understand the investigation or to report to the outside world | protocols that hold true for the sample handling (treatments) prior to any assay | assay specific protocols for the analysis of images, etc.  to do: align the structure with the workflows in SciNote | documented procedure of the analysis and tools used in the assay |
| /scripts | NG | NG | NG | scripts used in the assay |
| /output/ | NG | NG | Output files/data of the procedures used | output files/data of the procedures used |
| /output/raw | NG | NG | original data from the machine | NG |
| /presentations/ | All presentation prepared summarising results and experimental design within investgation | NG | NG | NG |
| /other/ | NG | NG | For anything not included elsewhere | For anything not included elsewhere |
| readme.md | Can be used to make notes (LOG file) | Can be used to make notes (LOG file) | Can be used to make notes (LOG file) | Can be used to make notes (LOG file) |

NG – not generated on this level

# Additional files that need to be prepared by users or are automatically generated by pISA-tree

## Metadata files

Each level has a *metadata* file (a file with additional information needed to describe the experiment with enough information to be repeatable), listing the informative items for that level. These metadata files are tab delimited files with two columns:

1. item name (ended by a colon)
2. item metadata or value

Item metadata can be some text (for example investigator's name or a longer metadata of the study, study description and analysis description etc.) or a value (for example the path to *phenodata* file). Each item metadata should be typed in one line. Special escaped characters for a line break (\n) and tab (\t) are allowed. Be careful if the metadata contains prime symbol (' ,as in 5'), it is safer to spell it, like 5-prime.

Two examples of the metadata entry are given below (tab character is presented as right arrow, 🡪):

Investigator: 🡪Miha Mihav  
Phenodata: 🡪./data/phenodata.txt

When running batch file to create new project, investigation, study or assay the user is asked to enter basic metadata. Additionally, layer batch files will create a text file named common.ini in the particular level directory. This file contains fields and metadata that are fixed for the project/investigation/study/assay, for example the principal investigator name, contact address etc. Information in this file will be appended to the metadata files for all investigations, studies, and assays within a project. A dummy/template common.ini file is created automatically in each project.

Metadata files and common.ini file are plain text files. You can open and edit the files with any text editor (e-g- Notepad or WordPad) or Excel also anytime later not only when starting new level in pISA-tree. In the text editors, tab character that is separating item name and item value might be invisible. You can detect the tab character presence by using left/right arrow and notice that the cursor jumps to the right of the colon. If you will use Excel, the file will be presented in two columns and might be more readable and easier to edit. Do not forget to save Excel opened file as Text (Tab delimited) file and do not change the name and extension (.TXT or .INI) of the saved file.

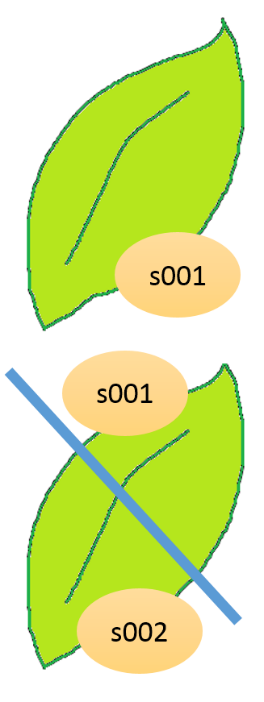
## Phenodata files

Phenodata files (=sample definition tables) are tab delimited txt files that describe your **samples**. Phenodata file should be stored in the Investigation folder. Start of new Study is related to collection of new samples. Already before starting the real experiment, e.g. growing plants, one should create a phenodata file together with the basic pISA-tree investigation structure.

Now the definition of a sample is a complex issue:

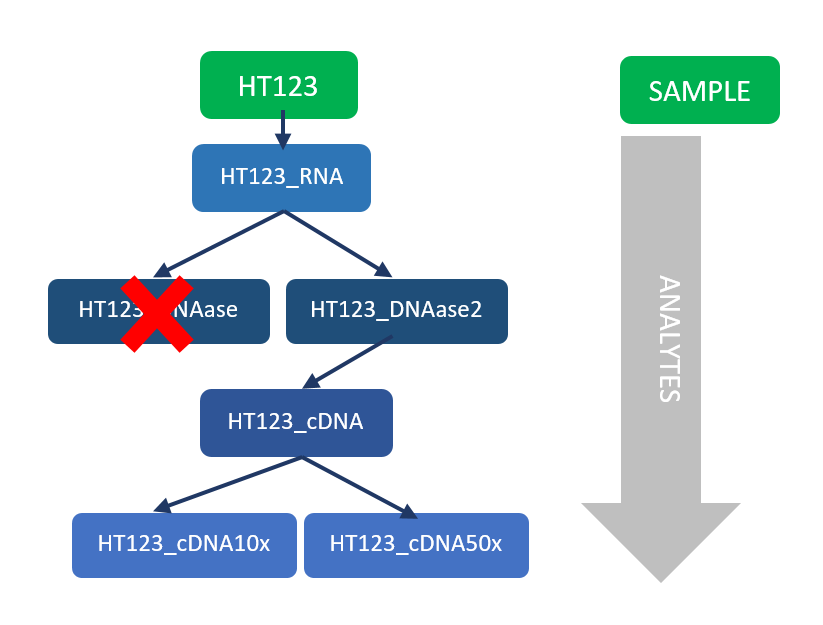
1. It starts with the sample collection itself (image bottom, left):

* the image below displays, how a full leaf which was then homogenized (left) and used for two techniques (transcriptomics and metabolomics) would get the same ID
* however, if the leaf is first cut in half and then one part is used for transcriptomics and other for metabolomics, they would get different identifiers
  + of course, the columns of the phenodata file should be created to allow us to still extract the information, that these two samples are from the same experiment, same plant and same leaf



1. Now other issues arise during the analyses – if a protocol for sample preparation for another technology (microarrays etc.) contains several steps, then these unique sample ID’s are repeated several times (image bottom, right).

This is why introduce the term analyte, which we can create additional unique IDs for, that combine both the sample ID and the substance that was produced during analysis, e.g. s001\_substance, s001\_RNA, s001\_DNA, s001\_labeling is ok, when the researcher consistently and persistently throws away the ones that were repeated, DON’T WORRY! This additionaly analyte IDs are created automatically within the specific assays. See Chapter Analytes files for more details on that.



The headers of columns in phenodata.txt are partially prescribed, but also any additionally columns that help better describe the samples collected can be added:

Firstly, all samples used in an investigation must have unique sample IDs, which are a combination of the two-letter study acronym (e.g. HT for hormonal treatments) and a three-digit number, e.g. HT001–HT999.

By definition unique means that:

* it will not be repeated twice in any other experiment within the same investgation
* it really directly relates to the ‘real life’ sample collected

Besides SampleID in the first column, phenodata file should contain sampleName (a bit longer version allowing biologists to related biology to the samples), all further columns should contain sample descriptions e.g. time after inoculation (dpi; e.g. 1, 2, 3, ...), treatment data (e.g. mock, PVY), genotype (e.g. NT, coi, nahG...), position of the sample on the plant (e.g. upper leaf...) and any further info you consider relevant. When creating this descriptions you should again should NOT(!!!) use any spaces (use \_ instead) NOR č,š,ž characters (see also FitoBase Manual).

The templates for these files are automatically generated when you start a new investigation of pISA-tree.

*Note1:* for further analysis (especially dry lab assays) it is very practical to add info in which assays which samples were used. The simplest was=add columns for Assays (column name = AssayID) -> mark used samples in each assays with “1”. In the next version of pISA-tree this will be added automatically.

*Note2:* to allow machine readability of the files (=easy automatic integration of results) standard vocabulary should be used when filling in the phenodata file, e.g. always the same word for the same describtion (including the same use of spaces, hyphens, etc), Information on minimal information to be entered can be found in various standards, see Annex 1 for this

## Analytes.ini files

In addition to the basic items, one can also use assay specific items, depending on the assay type. The assay specific items are pre-specified in the Analytes.ini files, placed in the appropriate *Class/Type* subfolder of subfolder Templates in the pISA projets root (main) directory. The makeAssay batch file will generate questions (if any) to add information to assay metadata file. The Analytes.ini file is specific for each used assay Type in your system.

The Analytes.ini files are plain text files. You can open and edit them by Notepad, WordPad or Excel (Use 'Open with ...'). Each line represents one Item-Value pair, separated by the tab character (marked below as right arrow 🡪). The first part (Item name) will appear as the assay specific question during the assay creation. The second part (Item value) will be offered as the default (or the first) value for information entry.

An example of the Analytes.ini file:

Item name 🡪 Item value  
Isolation Protocol 🡪Rneasy\_Plant  
Operator 🡪John/Bob  
Date Homogenisation🡪%today%  
RNA ID 🡪 RNA\_$  
ng/ml 🡪 Blank

The Analytes.ini will not need to be tackled with by standard user of pISA-tree application.

The user will be asked about the assay specific items (defined by assay specific Analytes.ini file) when running makeAssay.bat and those will be included in the ASSAY\_METADATA file. In addition, they are used as the assay specific description of samples used in an assay and are automatically added as the assay specific extension to the phenodata file. Assay specific metadata will be copied into columns of the Analytes.txt file, which contains information about the samples used in the assay.

Syntax rules in item value part are used for support of choices in menu--like data entry. This reduces errors in spelling, spacing, and use of the character case.

**Fields with one or more choices**

Item value choices, if more than one, are separated by the slash (/) character. See the example above for the items named Isolation Protocol and Operator. To select the operator name, a simple menu will be presented to the user:

1 John  
2 Bob  
3 Other

User will use numbers (1 to 3) to select name to use. The last line (“Other”) is automatically added and enables ad--hoc addition of any new choice. If the choice is likely to occur in future, it can be added into the Analytes.ini file.

**Date field**

Date fields are considered in the same way as ordinary choice fiels. Special bookmark %today% will be replaced by current date in a data entry menu.

**Sample ID replacement**

New sample related identification codes are sometimes needed. Sample ID can be automatically inserted in the place of a dollar character ($) to form new IDs. In example above for the field RNA ID and Sample names SMP001, SMP003, SMP007 one would get new IDs: RNA\_SMP001, RNA\_SMP003 and RNA\_SMP007.

**Blank fields**

The word Blank as item value signals the column that has to be left blank in the Analytes.txt file

## Analytes.txt files

Prior to the creation of the assay level (using makeAssay.bat) the user creates a text file with a list of samples that will be used in the assay. The list of samples is in a (tab delimited) text file and placed at the Study level of the future assay and should be named Analytes.txt. A small example of the Analytes.txt three samples could be:

ID  
SMP\_001  
SMP\_003  
SMP\_007

When makeAssay.bat file is executed, user will be asked to enter the standard and assay specific items (metadata). The Analytes.txt file with the following content will be generated at the Assay level combining the information about the samples in Analytes.txt file on the study level and information entered as metadata:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | Isolation protocol | Operator | Date Homogenisation | RNA\_ID | ng/ml |  |
| SMP\_001 | Rneasy\_Plant | Bob | 24. 04. 2018 | RNA\_SMP\_001 |  |  |
| SMP\_003 | Rneasy\_Plant | Bob | 24. 04. 2018 | RNA\_SMP\_003 |  |  |
| SMP\_007 | Rneasy\_Plant | Bob | 24. 04. 2018 | RNA\_SMP\_007 |  |  |

## Featuredata files

Featuredata file (=annotation file) lists and describes the features (=variables) measured in a particular assay (=experiment). Besides the unique IDs (e.g. gene ID, metabolite ID, etc) the file that describes the features provide additional information about the features, e.g. short name, description, GO, EC, MapManBin, any technical issue (e.g. specificity problem, quantification problems)...

The file should be created or downloaded (gal file in the case of microarrays, gff file in case of RNASeq). The file should be prepared in a tab delimited format (can be also \*.xls or \*.xlsx) where the first column contains list of all features and is named FeatureName (see also Note 3), followed by any number of columns that give improved knowledge and understanding of the feature.

The templates for these files are automatically generated when you start a new level of pISA-tree.

Note 1: In case of transcriptomics, proteomics or metabolomics the annotation files can be quite complex, but it has to contain at least two columns: FeatureName and Description.

Note 2: For microarray analysis this file normally includes also information on feature positions on the microarray which are provided by the manufacturer of the microarrays.

Note 3: For all transcriptomics (microarrays, NGS, qPCR) and proteomics experiments we will link the features to Genes. Consequently, first column in the Annotation file should list GeneIDs and be named “geneID”.

Note 4: For qPCR experiments Annotation file is actually the export from PrimerDB and does not need to be prepared de novo

# Auxiliary batch files

#### showMetadata.bat

Collects all metadata files in a tree below the current level. Descriptions are typeset in the file **METADATA.MD**. (plain text file in a markdown format; all text files can be edited by any text editor, e.g. Notepad, Wordpad or Excel and Word as long as they are saved as the text files. Use Open with option to select the non-default program to open such data).

#### xcheckMetadata.bat

Checks all metadata files for missing required information (\*) in a tree below the current level. Produces the file named **xCheckMetadata.md** which is similar to the one produced by showMetadata.bat but lists only lines with asterisks (\*).

#### showTree.bat

List a directory tree below the current level in the file **TREE.TXT**.

#### update.bat

Replaces batch files in existing tree with the updated versions in the project directory. Use after update from [GitHub](https://github.com/ablejec/pISA/)

# Future plans:

* FeaturesSummary tables: which feature is measured in which assay, ne damo cisto vseh ampak le bolj podrobne ananlize
* AnalytesSummary tables: ali je isto? Dodaj v phenodata kot stolpec oziroma se popravlja samo, z R-jem k obo v zda-super
* Archiving within the assay

# Annex 1: standards helping in setting up appropriate phenodata files

Plenty of various platform dependent standards exist for the description of experimental data; consequently all these standards are assay dependent (e.g. qPCR assay that involves sample preparation for it).

* ISA-TAB creator allows us to modify existing templates to suit our purposes or create new ones
* some of the existing templates: default ISA-TAB templates, MIAPPE template that improves on the phenotyping, for metabolomics: CIMR-MTBLS, MetaboLights; ScientificData templates
* pISA tree templates are stored in folder /Templates/WET/
* wet lab related minimum information standards ([MIBBI, now on BioSharing](https://biosharing.org/standards/?selected_facets=isMIBBI:true)):

|  |  |  |  |
| --- | --- | --- | --- |
| Standard | Type | ExperimentType | ImportantNotes |
| MINSEQE | - | - | minimum information about a high throughput sequencing experiment |
| MIPFE | - | - | minimum information for protein functional evaluation |
| MIARE | wetL | ? | minimum information about a RNAi experiment |
| CIMR | wetL | metabolomics | core information for metabolomics; see MIAPE-MS too I suppose |
| MIAPE | wetL | proteomics | <http://www.psidev.info/>  proteomic experiments covering all MS-workflow steps   * gel electrophoresis; sample processing; mass spectrometry   proteomic data repositories:   * MIAPEGelDB (gel electrophoresis) * PRIDE (mass spectrometry) * ProteoRed MIAPE Generator Tool (GE + MS) |
| MIMIx  MIAPAR  MIABE | wetL | proteomics | <http://www.psidev.info/>  minimum information about molecular interactions, protein affinity reagents, bioactive entities |
| MIAME | wetL | transcriptomics | gene expression microarray  repository: GEO/ArrayExpress |
| MIQE | wetL | transcriptomics | minimum information for qPCR experiments |
| MIGS  MIMS | dryL or wetL? | (meta)genomics | Genomics Standards Consortium  minimum information about a (meta)genome sequence |
| MINIMESS | dryL | (meta)genomics | minimal metagenome sequence analysis standard |
| MIASE | dryL | models | minimum information about a simulation experiment |
| MIRIAM | dryL | models | minimum information required in the annotation of models |
| MIAPA | dryL | sequences | minimum information about a phylogenetic analysis |
| MIACA | - | - | minimum information about a cellular assay; high-throughput cell biological analyses (cells in culture); extension of minimum information captured by primary nucleotide sequence archives |
| MIAPepAE | - | - | minimum information about a peptide array experiment |
| MIASPPE | - | - | minimum information about sample preparation for a phosphoproteomics experiment |
| MIDE | - | - | minimum information required for a DMET experiment |
| MIFlowCyt | - | - | minimum information about a flow cytometry experiment |
| MIGen | - | - | minimum information about a genotyping experiment |
| MIMARKS | - | - | minimum information about a MARKer gene sequence |
| MINI | - | - | minimum information about a neuroscience investigation; electrophysiology |
| MIQAS | - | - | minimum information for QTL and association studies |
| MIRAGE MS | - | - | minimum information required for a glycomics experiment – MS analysis |
| MIRING | - | - | minimum information for reporting NGS sequencing genotyping |
| MISFISHIE | - | - | minimum information about in situ hybridization and immunohistochemistry experiments |
| STRENDA | - | - | standards for reporting enzymology data guidelines |