|  |  |
| --- | --- |
| Last Revised: | Feb 29, 2012 |
| Produced By: | CBIIT Development |
| Version: | 3.2 |

# Description

Initiated by a Data Owner, this use case enables the import of previously-uploaded data files and MAGE-TAB files into an experiment, resulting in association of data and annotations with samples.

# Preconditions

A Data Owner (typically a Lab Scientist) has logged into the System, and an experiment with uploaded files is available.

# Basic Flow: Import MAGE-TAB plus Data Files

The Data Owner chooses to import a subset of uploaded files including a MAGE-TAB IDF, a MAGE-TAB SDRF and all the data files referenced in the SDRF.

The System begins the validation and import process in the background, allowing the Data Owner to check status of the files. If validation fails (see Use Case: Validate Data Files), the entire import fails, otherwise the System proceeds with import.

On successful import, the System creates the appropriate biomaterials and associates them with the appropriate data files and annotations as specified in the MAGE-TAB files. In addition, for “parseable” data file types (see Use Case: Manage Data Files), the System saves individual data values and their corresponding design elements (probes) in the repository.

# Post Conditions

The files have a status of “Imported”, “Imported, not Parsed” or “Import Failed”.

# Alternative Flow: Import Data Files Only

The Data Owner chooses to import a subset of uploaded files including only data files and no MAGE-TAB. (This flow is available for only certain file types. See User’s Guide for a list of these.)

The System provides the following options to the Data Owner to decide how sample-data relationships will be allotted:

1. **System will auto-generate individual biomaterial chains**: If the user selects this option, then for every unique file name that is imported, the System will create a Source – Sample – Extract – LabeledExtract – Hybridization chain corresponding to each data file imported. These entities are identified based on the base name of the data file. E.g., importing mouse\_342.txt, mouse\_342.chp and mouse\_342.cel will result in one chain of biomaterials and Hybridization each named “mouse\_342.
2. **System will auto-generate a single biomaterial chain**: If the user selects this option, then the System will create a single Source, Sample, Extract, Labeled Extract and Hybridization, and associate all selected data files with this single chain.
3. **User will associate data files to a specific biomaterial or hybridization**: The System presents all available Sources, Samples, Extracts, Labeled Extracts and Hybridizations. The User selects one of these, and the data files are associated with that biomaterial or hybridization. Note that additional items in the chain (to the right of the selected biomaterial) may need to be generated by the System. E.g., let us say the User imports 4 data files: zebrafish\_6311.cel, zebrafish\_6311.chp, zebrafish\_6666.cel and zebrafish\_6666.chp. Let’s also say the User chooses to associate these data files with an existing Extract called zebrafish\_extract\_6. The System will auto-generate a LabeledExtract – Hybridization chain called zebrafish\_6311 and associate the first 2 data files with it. The System will also auto-generate a LabeledExtract-Hybridization chain called zebrafish\_6666 and associate the last 2 data files with it. Both these auto-generated chains will be associated with the zebrafish\_extract\_6 Extract that the User selected. The part of the chain to the left of the selected Extract (Sources, Samples) will remain unmodified.

The System generates biomaterials-data relationships as requested by the Data Owner.

# Alternative Flow: Partial Import of MAGE-TAB set

The Data Owner chooses to cancel a MAGE-TAB + Data Files import job that is in progress. (See Use Case: Manage Job Queue.) The System finishes importing the current subset of files that are being processed, and cancels the import of the remaining set of files. The Data Owner can choose to resume the import by selecting the remaining files and starting a new import job.

# Alternative Flow: Unsupported Files

The System determines during validation that one or more of the files cannot be imported (e.g., they are of unknown file type or they are an array design type). The System asks the Data Owner if the remaining files should be imported.

If the Data Owner confirms the import, the System proceeds with the partial import, otherwise the entire import process is cancelled.

# Exception Flow: Interrupted Import

If file IO fails or the database connection is broken during import, any changes must be rolled back.

If during the import process the Data Owner closes the browser, the import should continue even though the session is lost.

# Special Requirements

## Experiment Annotations Read-only During Import

While data is being imported into an experiment, all attributes and annotations of that experiment will be made read-only so that changes made through the UI don’t conflict with annotations being created as part of the import process.

## Missing Biomaterials in MAGE-TAB Auto-generated

In the MAGE-TAB SDRF being imported, if a biomaterial node is missing in the Source – Sample – Extract – LabeledExtract – Hybridization chain, appropriate intermediate nodes will be generated to complete the chain. The number of nodes generated will depend on the left side of the graph. E.g., if the SDRF describes 1 Source connected to 3 Extracts, 1 Sample will be auto-generated and inserted in the chain. On the other hand, if the SDRF describes 3 Sources combined to generate 1 Extract, 3 Samples will be auto-generated and inserted into the chain.

In another scenario, if the first (leftmost) biomaterial node in the SDRF is not a Source, then the chain to the left of the biomaterial node is auto-generated by the System. E.g., let us say the SDRF starts with the Extract column, and there are 2 unique Extracts in the SDRF called “liver\_2900” and “liver\_3200”. The System will auto-generate 2 Sources called “liver\_2900” and “liver\_3200”, and also 2 Samples called “liver\_2900” and “liver\_3200”. The Sources, Samples and Extracts will be linked in 2 separate chains.

## Protocols Associated Intelligently

If biomaterials missing in the SDRF are auto-generated (as described in Section ), the System will intelligently associate protocol applications with the most appropriate auto-generated node. Note that this applies only to protocol applications in the SDRF whose association to biomaterial nodes is ambiguous. E.g., if the SDRF contains a Sample followed by a Labeling Protocol followed by a Hybridization, the Labeling Protocol will be associated with the auto-generated Extract - LabeledExtract portion of the chain.

The specific rules are:

* A protocol of type “labeling” (MGED Ontology) will be associated with the Extract – LabeledExtract portion of the chain.
* A protocol of type “hybridization” (MGED Ontology) will be associated with the LabeledExtract – Hybridization portion of the chain.

## Biomaterial Annotations

A biomaterial (Source, Sample, Extract or Labeled Extract) can be followed by any number of Characteristics columns containing annotation information about that biomaterial. Some of these Characteristics are interpreted specially by caArray:

A characteristic with the OrganismPart category becomes the biomaterial’s tissue site.

A characteristic with the CellType category becomes the biomaterial’s cell type.

A characteristic with the DiseaseState category becomes the biomaterial’s disease state.

A characteristic with the Organism category becomes the biomaterial’s associated organism.

A characteristic with the ExternalId category becomes the corresponding biomaterial’s unique external identifier. This identifier is guaranteed to be unique for a biomaterial within an Experiment, but in future, will be unique for a biomaterial *across Experiments* as well. This external identifier can be used to uniquely identify a biomaterial across caArray as well as other applications.

## SDRF Decides Raw versus Derived Data File

If the user sets the file type of a data file to raw (using the Manage Data UI), but the SDRF designates that same data file as derived, then the designation from the SDRF overrides the one specified by the user in the Manage Data UI. This also applies in the reverse case where the user specifies a data file to be derived (using the Manage Data UI), but the SDRF specifies the data file to be raw. The designation in the SDRF is authoritative.

## MAGE-TAB References to Existing Biomaterials and Files

If the MAGE-TAB SDRF references a node (biomaterial, hybridization or data file) that has already been imported into the system previously, then that node will be reused. E.g., if the SDRF being imported has a Sample Name called “liver\_2900”, and the system already has a Sample with the name “liver\_2900”, the existing Sample will be reused, and any new characteristics, extracts, data files etc. will be associated with that Sample. This implies that Sample names are unique within an Experiment. This rule also applies to Source names, Extract names, Labeled Extract names and Hybridization names.

### Common usage scenarios

#### The user imports a set of MAGE-TAB plus data files, creating a Sample associated with 20 Extracts. The user later imports a new batch of MAGE-TAB plus data files, creating 20 new Extracts but associating them to the same Sample as in the first batch.

#### The user imports a set of MAGE-TAB plus raw data files. The user later modifies the SDRF to add derived data files (that were derived from the same raw data files that have already been imported), and imports this new SDRF with the derived data files.

#### The user imports a set of MAGE-TAB plus data files, creating a set of Samples and other annotations. At a later stage, clinical characteristics for those same Samples become available for the first time, or change from what was originally submitted. The user modifies the SDRF to add the new clinical characteristics (or to change values of already-existing characteristics), and re-imports the SDRF.

### Types of changes supported

#### **Changing values in already-existing characteristics**: This applies to Provider, Material Type and Label columns as well as to Characteristics[] columns in the SDRF. The values for these characteristics will be changed, added or deleted according to the values present in the new SDRF.

#### **Adding new characteristics**: Any Characteristics[] column with a new category will result in that characteristic being added to the biomaterials in the system.

#### **Changing biomaterial-data chains**: Only *additive* changes to linkages and nodes will be allowed. No deletion of linkages or nodes will be supported through MAGE-TAB import. Note that the same rules apply to auto-generated nodes (section 7.2) as well. Examples of different import scenarios and the expected results follow:

#### Old SDRF links Source A to Extract 1 and Extract 2. System auto-generates Sample A in between. New SDRF links Source A to Extract 3 and Extract 4. Expected results: System reuses generated Sample A, and resulting chain looks like this: Source A linked to Sample A linked to Extracts 1, 2, 3 and 4.

#### Old SDRF links Source A to Extract 1 and Extract 2. System auto-generates Sample A in between. New SDRF links the same Extracts 1 and 2 to a new Source B. Expected results: System generates Sample B, and resulting chains look like this: Source A linked to Sample A linked to Extracts 1 and 2; Source B linked to Sample B linked to Extracts 1 and 2.

#### Old SDRF links Source A to Extract 1 and Extract 2. System auto-generates Sample A in between. New SDRF links Source A and Source B to a new Extract 3. Expected results: System reuses previously-auto-generated Sample A, auto-generates Sample B, and the resulting chains look like this: Source A linked to Sample A linked to Extracts 1, 2 and 3; Source B linked to Sample B linked to Extract 3.

#### Old SDRF links Source A to Extract 1 and Extract 2. System auto-generates Sample A in between. New SDRF links Source A and Source B to existing Extract 2. Expected results: System auto-generates Sample B, and resulting chains look like this: Source A linked to Sample A linked to Extracts 1 and 2; Source B linked to Sample B linked to Extract 2.

#### Old SDRF links Hybridization A to raw/derived data files 1 and 2. New SDRF links the same Hybridization A to raw/derived data files 3 and 4. Expected results: System links Hybridization A to raw/derived data files 1, 2, 3 and 4.

#### **Changing units for existing characteristics or experimental factors**: Units for existing characteristics and experimental factors will be updated according to what is in the new SDRF.

#### **Changing/adding experimental factor values for existing hybridizations**: <Deferred> Note that factors following new hybridizations will be imported just as they normally are.

#### **Changing/adding protocols and parameters for existing biomaterials/hybridizations (including Image, Scan, Performer and Date)**: <Deferred> Note that protocols and parameters following new biomaterials/hybridizations will be imported just as they normally are.

#### **Changing array design associated with an existing hybridization**: Any new array design referenced in the SDRF will be added to the experiment’s set of array designs, and the array design associated with the existing hybridization will be updated to the new one.

#### **Changes/additions from the IDF**:

#### Experiment description and dates are updated according to the new IDF.

#### Attributes of existing experimental designs and factors cannot be changed. Note that new experimental designs and factors are added just as they normally are.

#### Attributes and roles of existing persons cannot be changed. Note that new persons are added just as they normally are.

#### Any new terms are added to the replicate types, normalization types and quality control types.

#### Attributes of existing publications cannot be changed. Note that new publications are added just as they normally are.