Data Import in GSRS 3.1

A User’s Guide

Document History

|  |  |  |
| --- | --- | --- |
| Version | Date | Author |
| 0.1 | 30 June 2023 | MM |
| 0.2 | 5 July 2023 | MM |
| 0.3 | 6 November 2023 | MM |
| 1.0 | 12 December 2023 | MM |

# Background

GSRS has supported data exchange since at least version 2.0. The ability was limited to files of GSRS JSON that can be exported from one installation of GSRS after a search or while browsing the entire database and imported in their entirety in another installation of GSRS.

Any selection or transformation of the data required custom tools that were used on the exported file before it was loaded into the next installation.

A diagram of data exchange features

Description automatically generated with low confidence

With GSRS 3.1, the team developed a vision of building data selection and transformation tools *inside* facGSRS itself to reduce the need for custom software and to provide standardization in the handling of data.

The initial vision of the team was fairly broad: support import of a variety of file types:

* GSRS compressed JSON files [available in version 3.1]
* SD (structure data) files [available in version 3.1]
* Delimited text files [coming in a future release]
* Excel spreadsheets [coming in a future release]

Because data import functionality is so powerful – giving a single user the ability to add thousands of records to GSRS in a brief period – we have limited access to import functionality (at least for now) to GSRS administrators.

For the 3.1 release, we concentrated our efforts on 2 points:

* Support robust facilities for importing gsrs file (with capabilities that were not available earlier) as well as SD files (including comprehensive field mapping selections).
* Create the architecture that serves as a base for further file types and operations.

### Overview of the import process

**Initial View:**

Adjust data interpretation

**Staging Area:**

(Temporary area within GSRS)

**File upload**

**Load**

Start:

User has a data file

**View, select, process…**

Data now in GSRS

# How to import an SD file

## Background

SD files are a means of exchanging chemical information. They consist of a molfile (representation of a chemical structure), optionally followed by text/numeric data fields in a sequential format.

More information is available here:

<https://en.wikipedia.org/wiki/Chemical_table_file>

<https://discover.3ds.com/ctfile-documentation-request-form>

## Steps

SD files are very commonly used by chemists and chemically savvy organizations to send information from one place to another. For example, PubChem (https://pubchem.ncbi.nlm.nih.gov/), ChemBL (<https://www.ebi.ac.uk/chembl/>) provide data in SD file format for download. Organizations within the GSRS community often receive data in SD file format and wish to load into GSRS. Therefore, we prioritized SD file import for GSRS 3.1

Make sure you’re logged into GSRS as an administrative user.

Select the Admin Panel from the drop-down menu at the top of the screen, on the right side.

A screenshot of a phone

Description automatically generated with medium confidence

* A picture containing text, screenshot, font, algebra

  Description automatically generatedSelect the ‘Data Import’ tab on the Admin Panel.

Use the ‘Select File to Import’ button to bring up an open file dialog box:

A screenshot of a computer

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated

Select an SD file.

The system now recognizes that you have selected an SD file and selects the appropriate adapter. (Note: if, for some reason, the system selects an incorrect adapter, you can change the selection of adapter by selecting the appropriate radio button.)

A screenshot of a computer

Description automatically generated with medium confidence

Click the Upload button.

The system now attempts to guess the appropriate processing action for each field in your file:

A screenshot of a computer

Description automatically generated with medium confidence

Each field is listed along with the selected action and a hotlink to change the action.

For example, the file field PUBCHEM\_CACTVS\_ROTATABLE\_BOND field is recognized as a GSRS property.



The preview panel on the right side of the screen shows how the data would look if the current set of processing instructions were followed.

A screenshot of a cell phone

Description automatically generated with medium confidence

Clicking on the Settings link brings up a dialog box where you can change the way the field is mapped to GSRS fields:

A screenshot of a computer settings

Description automatically generated with low confidence

Figure Mapping dialog as it initially appears.

A screenshot of a computer

Description automatically generated with medium confidence

Figure Mapping Dialog after we have changed the processing of this field to 'Create Note Action'

You can click the ‘Ignore Field’ checkbox to have GSRS skip processing of a field.

After making changes to field processing, you can, optionally, click the ‘Reload Preview’ button to reinitialize the preview of the data.

Once the data look the way you intend them to look, click ‘Use Setting and Continue.’

At this point, GSRS reads each record in your file, applies the mapping actions you selected and creates GSRS substances. These substances are validated, compared to the records already in your GSRS database, stored within a temporary database (the ‘staging area’ as explained below) and indexed for searching.

Depending on the size of your file and the provisioning of your server, this process can take a while.

A picture containing text, font, screenshot

Description automatically generated

# How to import a GSRS file

A GSRS file is a formatted text file where each line of the file contains a JSON representation of one GSRS record from another GSRS instance, partially compressed by removing newline, carriage return, and space characters. There are two tab characters at the start of each line so the JSON starts at the third column. Also, the file is gzipped. The extension may be .gsrs or .gz. The file format has been used by GSRS for data exchange for several years.

The first few steps for importing a GSRS file are identical to importing an SD file.

Make sure you’re logged into GSRS as an administrative user.

Select the Admin Panel from the drop-down menu at the top of the screen, on the right side.

A screenshot of a phone

Description automatically generated with medium confidence

* A picture containing text, screenshot, font, algebra

  Description automatically generatedSelect the ‘Data Import’ tab on the Admin Panel.

Use the ‘Select File to Import’ button to bring up an open file dialog box:

A screenshot of a computer

Description automatically generated with low confidence

The next step is somewhat different.

After you select your data file and click the Upload button, notice that the table of import actions is blank:

A white background with black text

Description automatically generatedThat is because, unlike SD files, GSRS file contain data already assigned to fields in the GSRS data model. (There is no mapping to select.)

You can examine your data in the Preview panel.

Once you click ‘Use Settings and Continue,’ the data will be loaded into the Staging Area and the process looks like it does for SD files, as described above.

# Staging Area

The temporary database for just-loaded data is called the ‘Staging Area.’ Data in the Staging Area can be dissected using facets, searched, browsed, changed and then loaded into GSRS and or deleted.

[how to get to the staging area after some other operation with screenshot]

A screenshot of a computer

Description automatically generated

Description of action items within staging area record view:

**New Data Import** – start importing a new file.

**Data Cleaning Options** – select parts of the data record to be removed before inserting into GSRS.

**Select All on Page**– add records on the page to the set of action-ready records.

**Select All Records** –add all records in the current list to the set of action-ready records.

**Clear Selections** – empty the set of action-ready records.

**Create** – add one record to GSRS.

**Edit** – make changes to one record within the Staging Area.

**Merge** – copy items (such as name or codes) from one record into a match record in GSRS. **Note: this action is not yet available**.

**Reject** – mark a record within the Staging Area as unwanted.

When a new record goes into the Staging Area, we calculate hashes of the structure of chemicals as well as the sequences for proteins and nucleic acids, and use this information to determine potential matches against records within GSRS. This allows you to see duplicates *before* they are introduced into the database.

Browsing data within the staging area looks a lot like browsing substances in GSRS.

* You can view individual records by clicking the name at the top of the individual record card. For example, click on ‘(9*R*,10*S*,11*S*,13*S*,16*S*,17*R*)-9-chloro-11,17-dihydroxy-17-(2-hydroxyacetyl)-10,13,16-trimethyl-6,7,8,11,12,14,15,16-octahydrocyclopenta[a]phenanthren-3-one’ in blue to bring up a detail page for the record.

A screenshot of a computer

Description automatically generated with medium confidence

Figure One record within the staging area. Except for the 'Staging Area Record' label at the top of the screen, you might think you're looking at a record within the main repository!

* You can edit (make changes to) records within the staging area, keeping the changed record within the staging area. To access the edit page for a record, either click the Edit button while browsing or click the pencil icon while viewing a single staging area record.
  + When you are done with the edits, click the ‘Validate and Submit’ button, as you do for edits to substances within GSRS.
  + Address any errors or warnings.
  + Click the Submit button.

## Facets

A screenshot of a phone

Description automatically generated

Figure Example view of facets within the Staging Area

You can use facets to zero in on data of interest within the staging area. (There are some facets that are unique to the staging area, such as ‘source’ that contains the name of the file from which each record was taken, as well as standard substance facets, such as those for code systems and molecular weights.)

2 common uses of facets within the Staging Area:

1. Quickly view all records that have no errors and have no matches. (This may be inserted directly into GSRS.)
2. Quickly view all records that have exactly 1 match (this may serve as the basis of a data merge, once the option is available.)

Immediately after you load a new file and browse the data in the staging area, GSRS applies a ‘source’ facet so you see just the newest records.

Additional facets that you may find useful in the staging area:

* Loaded by – the name of the user who loaded the file.
* Date Loaded – when the file was loaded into the staging area.
* Validation Type – the high-level result of validation rules applied to the substance when it was loaded into the staging area. Typical values: Error, Warning, Info.
  + You may want to select all records with an error so that you can determine the specific error and address it.
* Match Count – the number of records in the permanent repository that match your record in one or more ways.

One possible usage scenario: select facets as follows:

* Source = recently loaded file
* Match Count = 0 (no duplicates in the main database)
* Validation Type != error (exclude records with a data error)

And then move all matching record (as described below) into GSRS.

## Moving Data from the Staging Area into GSRS

The endpoint of the data import process is to add records into GSRS.

There are 2 ways to accomplish:

* Adding individual records by clicking on the ‘Create’ button next to the record.
* Selecting a set of records (using the ‘Select All Records’ link, for example), then using the ‘Bulk Actions’ button to push the records into GSRS.

You’ll probably want to delete records from the staging area once the data have been pushed into GSRS.

Note that deleting a record from the staging area after it has been pushed into GSRS does not delete the new data from GSRS; the data are *copied* into GSRS. After the copy, you don’t need the version within the staging area.

## Data Cleaning

Before pushing new records to GSRS, you can elect to have a few clean-ups performed on the data:

A screenshot of a computer program

Description automatically generated with medium confidence

* Removing codes by system allows you to select one or more code systems to remove from the input records. (You can also select code systems to *keep* and force the other codes to be deleted.)
* You can remove Approval IDs from the input data and/or copy the value of the approval ID to a code with a code system of your choice.
  + If such a code already exists, the copy step will be skipped.
  + This step only applies to data imported from a .gsrs file; data from SD files will not have approval IDs.
* You can remove UUIDs from imported data so the UUID of imported records will be different from the source data.
  + This step only applies to data imported from a .gsrs file.
  + You can also copy the value of the top-level record UUID into a code with a designated code system. (If such a code already exists, the copy step will be skipped.)
* You can clear out the value of all standardized names in the input data so that standardized names can be regenerated automatically from the main name.
* You can elect to clear out audit information (the name of the user who created the record and the user who last edited the record, as well as the date of creation and date of last editing) when importing.