GSRS Find ChemIDplus

MITCH MILLER
SCIENTIFIC THINKING, LLC
NOVEMBER 16, 2018

Overview

- Introduction to GSRS Find
- Update on progress
- Areas of investigation
- ChemIDplus

Name clarification: GSRS Find

- The utility formerly known as ‘g-srs Excel Tools’

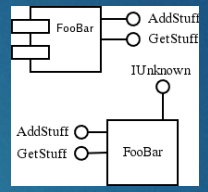
Whys and wherefores

- g-srs provides interactive data entry in the web application
  - Fine for a handful of records or data that requires a lot user attention
- Suppose you have
  - 1000 records to which you want to add a new synonym?
  - 2000 records whose code URLs have changed formats?
  - 100 records from a legacy system to load?
- However, when you have 1000s of records to load, contact the GSRS team for an alternative solutin
- Bulk data loading in Excel provides a solution

Top-level categories of operations

- Retrieval
  - Adding columns to an existing sheet
  - Add a new sheet with data that matches one key column of an existing sheet.
- Sheet creation
  - Creates a worksheet with columns ready for data loading
    - Builds drop-downs for fields with controlled vocabularies
  - Optional but easier than creating a sheet by hand
- Data entry
  - Add/modify/delete information

Top-level components of the client system



- Configuration data
  - Stored in an external file
  - Which server to use, etc.
- C# code
  - Manages top-level operations
  - User interface (most of it)
  - Communicates with the worksheet(s) and with the g-srs API via JavaScript
  - Add processing results back to the sheet
- JavaScript
  - Process data and send to server
- Sheet with data to load

How these tools work for data retrieval

- Open a worksheet with one key field already populated
  - Name, CAS number, etc.
- Select the fields you want to retrieve
- Tool fills the new data fields into the current sheet or a new sheet

How these tools work for data loading

- Worksheets within an Excel workbook are configured for a specific operation (such as adding a name to an existing record)
  - One type of operation per worksheet.
- Each record represents one transaction of that operation.
  - For example, one new name for one existing record
  - Each record is processed atomically.

Recent Progress

Progress

- Half-time developer assigned to the project
  - me!
- Major restructuring of the application
- Script improvements
- Miscellaneous

Excel Tool Structure

- The Excel Tools are now a Visual Studio solution
  - Written in C#
    - More ‘modern’ language
  - Debugging is easier
  - Unit tests
  - Maintenance is easier
  - JavaScript editable in the same editor as C# code
  - Easier to add and modify dialog, buttons, ribbon commands, icons
- Tool installed via .MSI (Microsoft Installer) file
  - Installed centrally rather than in specific workbooks
  - Configuration stored centrally/automatically
  - Standard for software on Windows
  - Possible downside: requires administrative privilege on the client

Controlled Vocabularies

- Many fields within the g-srs software are limited to a list of values
  - For example, Substance Class can be on of
    - Chemical, Concept, Mixture, Nucleic Acid, Protein, Specified Substance
- These fields have internal representations that are sometimes very different from what users see in the g-srs (web) application.
  - For example ‘cn’ for ‘Common Name’
- We now create drop-downs for these fields
  - Sorted alphabetically
- Optional: when you build your own worksheets for loading data, you can use the internal data representation

Sheet Creation easy to access

- There was a feature within the add-in that created worksheets with the correct format for data loading
- There is now a ribbon ‘Create Loading Sheet’ button

Greater stability

- Used to fail after 1-2 thousand records
- Currently handles 10-20 thousand records
- Confirmed by a test user

Logging

- Aids troubleshooting
- Uses standard log4net library
- Default: c:\temp but changeable
- Can be reduced/suppressed by editing file C:\Program Files (x86)\NCATS\g-srs Excel Tools\log4net.config
- In case of trouble, send us the log file from your computer!
- Turn on ‘debug mode’ within configuration dialog for JavaScript logging

Script improvements

- Several scripts that create data (names, codes, relationships) allow you to specify substances by **either** UUID or Preferred Term (main name)
  - In some cases, you can also use BDNUM
- Added top-level validators
  - A validator checks that parameters are correct for data loading
  - We have had validators on individual parameters
  - Now one validator can check the full set
    - For example, if parameter A has a value, parameter B should not have a value

Fetcher improvements

- All fetchers will run without crashing even when no compatible data found.
  - For example, it is safe to run the protein sequence fetcher on a set of records that include chemicals

Areas of Experimentation

Properties

- Background: g-srs allows you to store a variety of properties for each substance
  - Numeric, text
  - Data format is flexible.
    - You can have a maximum value, a minimum value and/or an average
    - Assign text values
    - Assign units
  - Can add many of these to each substance
- Experimented with creation of Properties in Scripts
  - Volume of Distribution
- Experimented with retrieval in a Fetcher

SD Files

- Strategy: create Substances (class = ‘Chemical’) for records in an SD File
- Steps:
  - Read data, including molfile, into sheet
    - Generate depictions of molfiles and perform duplicate checking
  - Arrange the column headers so a script can perceive fields
  - Use regular Load Data to create substances using a script
  - [Include some flexible handling of the fields within an SD file]
- This week, we turned a prototype over to selected users

Your feedback and participation in GSRS Find are welcome!

- On priorities for the project
- Scripts that you would like to use to populate your database
- Fetchers you would like to see for retrieving data

ChemIDplus

- ChemIDplus is a database of "over 400,000 chemical records. More than 300,000 of those record include chemical structures. ChemIDplus is searchable by Name, Synonym, CAS Registry Number, Molecular Formula, Classification Code, Locator Code, Structure, and/or Physical properties. Enhanced structure display is available in ChemIDplus Advanced."
- Chemicals of significance to human health
  - Pharmaceuticals
  - Environmental chemicals
- Provides a **curated** gateway to information about on the worldwide web.
- Part of the Toxnet system at National Library of Medicine
  - https://toxnet.nlm.nih.gov/

ChemIDplus

- Available since the web in 1998
  - One of the first databases searchable by chemical structure on the web
  - Replaced older, mainframe-based systems
- Several flavors:
  - 'Lite' (without structures) https://chem.nlm.nih.gov/chemidplus/chemidlite.jsp
  - 'Advanced' https://chem.nlm.nih.gov/chemidplus/
  - https://druginfo.nlm.nih.gov/drugportal/drugportal.jsp
- Technical underpinnings
  - Lots of custom Java
  - Oracle database with PL/SQL
  - BIOVIA Direct for chemical registration and searching
  - ChemAxon Marvin components
    - JavaScript components for chemical drawing
    - Server library for rendering
  - Corina for 3D structures
    - Graciously provided by Molecular Networks/Altamira

ChemIDplus News

- New team leader: Ying Sun
  - Ying has been a computer scientist at NLM for fifteen years, majorly doing design, development and research on web, mobile applications of Toxicology, bioinformatics, and chemistry. She holds degrees on computer science and chemistry.

ChemIDplus RESTful API

- Provides the ability to search and retrieve the data available on the ChemIDplus web site
- RESTful URLs for ChemIDplus records have been available for 3 years
  - https://chem.nlm.nih.gov/chemidplus/rn/58-85-5
  - https://chem.nlm.nih.gov/chemidplus/id/0000058855 (equivalent)
- Some searches have been available via RESTful URLs for a while
  - https://chem.nlm.nih.gov/chemidplus/name/contains/caffeine
- Power user search:
  - https://chem.nlm.nih.gov/chemidsearch
  - Allows you to build queries as complex as you like
  - Create individual clauses for field + operator + value
- Programmer access:
  - https://chem.nlm.nih.gov/api/data/<field>/<operator>/<predicate>?data=<selection>
  - Documentation: https://chem.nlm.nih.gov/chemidsearch/api

Thank you for your attention!

- mitch.miller@thinkscience.us