Improved Interoperability of Chemistry Data with CHEMeDATA

Damien Jeannerat,[1,2] Felipe Seoane,[1] Carlos Cobas[1]

[1] Mestrelab Research SL, Feliciano Barrera 9B-Bajo, 15706, Santiago de Compostela, Spain

[2] NMR Process, Geneva, Switzerland

Goal

Develop a community-based standard for chemistry information realizing the objective of FAIR data.

The CHEMeDATA initiative aims to facilitate the exchange of data between software, increase the findability of chemical data, make them accessible to AI driven tools, *etc*.

Aspiration

More specifically, we wish to facilitate:

- The search and re-use of chemical information stored on science repositories in a straightforward manner.

- Bypassing proprietary formats and the limitations of import/export features.

- Copy/paste any specific element of a document from one software to another (assignment data, list of peaks, multiplet data, *etc*.)

- The comparison of GIAO NMR shielding data with experimental spectra with simple drag-and-drop.

- The development of tools to generate any specific type of chemistry data that can be visualized, and use in different software environments.

The CHEMeDATA Initiative aims at standardizing small unit of chemical information (CHEMeDATA objects) to facilitate the exchange and visualization of chemical information.

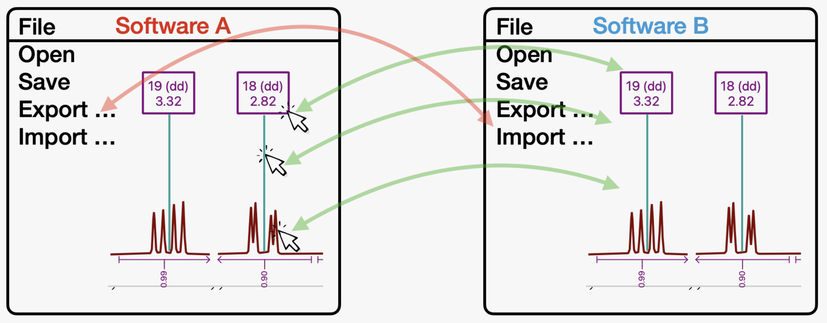


Figure 1: Usually, data exchange between software applications is made using export and import tools (red arrows). The content and compatibility of diverse formats are difficult to assess and documentation often insufficient. Tedious import/export may be necessary to find satisfactory solutions. In many cases the files need to be modified between the export and the import steps. A standard for specific type of chemical data (multiplet analysis, chemical shifts, regions of interest of spectra, *etc*.) would facilitate data exchange.

CHEMeDATA objects

A CHEMeDATA object can represent any chemical information. Considering the field of magnetic resonance, it could be:

- a 3D molecular structure (.mol, *etc*.)
- a NMR spectrum
- a sample description
- a description of a chemical transformation (purification, synthesis, *etc*.)

… or much more specific chemical information:

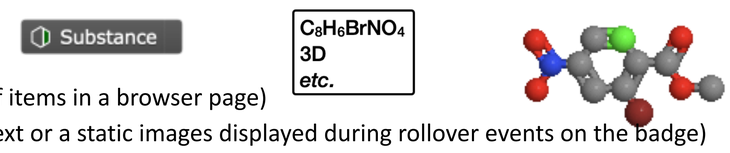
- a list of peaks extracted from an experimental spectrum or generated by quantum calculation
- the NMR shielding from a GIAO calculations
- the NMR assignment data of a set of NMR spectra (NMReDATA)
- the description of a biphasic system in a 5 mm NMR tube
- *etc*.

The objects follow a hierarchy allowing anybody to derive existing CHEMeDATA object types into types that are tailored to his/her specific and specialized needs.

For example, biphasic NMR samples may not be quite common, but still very important for the activity of some research groups. This group may want to benefit from the visualizer and data types of the usual CHEMeDATA NMR spectra object but add some specific fields to account for multiple phase of their NMR sample. They can simply derive the existing objects into a variant that fits their uses and set the schema for the data and metadata.

Visualization of CHEMeDATA objects

An important feature of CHEMeDATA is the visualization of the objects and the exploration of their hierarchy (parent-child relationship).



CHEMeDATA objects have:

- a Badge (typically in a list of items in a browser page)
- a preview (typically some text or a static images displayed during rollover events on the badge)
- a default visualizer (typically opening when clicking on the badge or display in a table cell)

Visualization of 3D structures

We are using JSmol as the default visualizer for CHEMeDATA 3D structures but developers can introduce alternative visualizer.

JSmol proposes very diverse options for the visualization of atoms, bonds, labels, and allows user to control interactively many properties such as view angle, change of color, *etc*. Multiple compounds can be visualized in a single HTML table, allowing other chemical objects on the same page to interact with it.

Visualization of assigned coupling constants and other chemical properties will be introduced in the future. Features allowing to display the results of the comparison of pairs (or series) of data set directly on the 3D model are also considered.

A simple visualization of NMR spectra is under development. The possibility to create quick previews of regions of interest will be introduced as well as the ability to compare pairs or series of spectra. For more complex interactions links will open the default NMR processing software of the user provided it recognizes CHEMeDATA objects.

Visualization of NMR spectra

A simple visualization of NMR spectra is under development. The possibility to create quick previews of regions of interest will be introduced as well as the ability to compare pairs or series of spectra. For more complex interactions links will open the default NMR processing software of the user provided it recognizes CHEMeDATA objects.

Visualization of other graphical objects

Anybody can introduce and share customized visualizer using D3/svg, Fabric/canvas, *etc*. We recommend to use self-containing classes in order to limit the JavaScript needed to include interactive CHEMeDATA objects In HTML documents. This will make the HTML document containing CHEMeDATA objects almost completely code free which will facilitate their generation.

A parallel coordinate D3 object has been introduced for the analysis of multiple-variable dataset and other simple visualizer will be introduced as demonstration examples.

Compatibility of Mnova

Starting from the version 14.1 of Mnova, spectra, peak-lists and signal assignment can be exported and imported in the json format. The API and mapping of the json serialization into CHEMeDATA is ongoing and will be released independently from the release of Mnova to take advantage of the flexibility of GitHub release generation.

Technical details

For each CHEMeDATA objects a default visualizer uses a JavaScript class object displaying text or graphics in a HTML canvas or svg.

Interaction methods allow for:

- the selection of an atom on a 3D structure and highlight a relevant part of a spectrum and vice-versa.
- the user to open the dataset in a default desktop application, produce reports, *etc*.
- the object to point to other related objects or along its hierarchy tree.

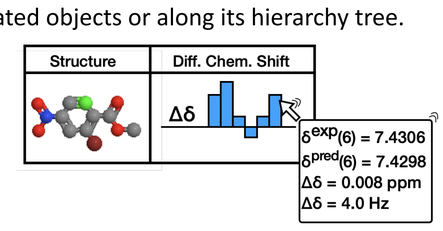


Figure 2: Illustration of an interaction: When the mouse hover on a graphical element (one of the blue column, for example) it triggers the temporary appearance of a text box with details about the hovered column and highlights the relevant atoms on the chemical structure from the model located in a separate column.

Complex objects combining other objects, such as the one of the assignment data of an NMR spectrum to a compound, should include (or link to) the structure of the compound and the relevant NMR spectra, both readily accessible electronically for visualization, verification, validation, reuse, *etc*.

Chemistry files and CHEMeDATA

Desktop applications often combine different chemical information (spectra, extracted NMR properties, *etc.*) into a single file. The philosophy of CHEMeDATA is to work at a finer level: the individual element and their relationships. For example, the output of Gaussian geometry optimization would consist in a set of json objects for the electronic energy, the geometry, chemical properties, *etc*. Each fulfilling criteria of descriptively, units, documentation, *etc*. Objects could be generated by small pieces of code making maintenance easy and allowing unit tests to detect any compatibility issues.

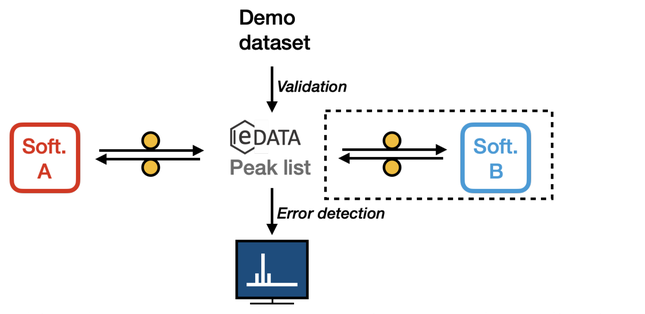


Figure 3: Each CHEMeDATA object has a schema, demonstration examples, validation processes and error detection of the default viewer (bottom). The interconversion of data into CHEMeDATA objects are small units of code for which demonstration examples, schema validation and error detection can be used by newcomers (Software B in the dotted frame) to insure the data they produce are conforming to the standard.

CHEMeDATA and NMReDATA

CHEMeDATA aims at generalizing the principle developed for the NMReDATA initiative. The idea of introducing high-level format for the assignment data of small molecule NMR is extended to other spectroscopies and more broadly to any chemical information. A major difference is that while NMReDATA was relying on tags in .sdf files, CHEMeDATA is based on json serialization supported by a schema and an object hierarchy. A json version of the NMReDATA will be introduced when the generic chemistry concepts and the derivations rule will be stable enough to allow for the construction of complex CHEMeDATA objects such as the one needed for assignment information.

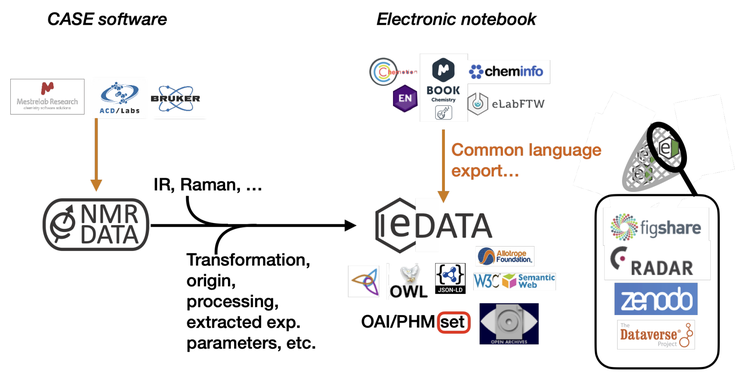


Figure 4: The NMReDATA initiative succeeded as making Computer-Assisted Structure Elucidation (CASE) software use a common format for NMR assignment (left). The CHEMeDATA can be seen as an equivalent project applied to electronic notebooks and other general-chemistry data organizers (right).

Acknowlegements

This project has received funding from the European Union's Horizon 2020 research and innovation program PANACEA under grant agreement No. 101008500.