CHEMeDATA: Driving FAIRness in Chemistry Through Community-Based Structured Data and Tools

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

[1] NMRprocess.ch, Geneva, Switzerland

[2] Mestrelab Research S.L.U., Avenida de Barcelona 7, 15706, Santiago de Compostela, Spain

Mestrelab Research
chemistry software solutions

The NMReDATA initiative

The CHEMeDATA is a community-based standard for chemistry information that supports the FAIR data principles aiming to improve the exchange and accessibility of chemical data. Its primary goals include enabling straightforward search and re-use of chemical information stored in scientific repositories, eliminating the need for proprietary formats and the limitations of traditional import/export features (Figure 1). It also facilitates the seamless transfer of specific data elements—such as assignment data, lists of peaks, or multiplet information—between different software environments using simple copy-and-paste or drag-and-drop interactions.

We distinguish five categories of CHEMeDATA objects:

CHEMeDATA object type	Code	Relations to other types
- samples* (describe things that exist)	S	S' : {P, S}
- processes* (describe transformations of things)	P	
- measurements* (describe measurements made on samples*)	M	M : {S, instrument}
- claims* (describe facts established through research activities)	C	C: {A, M, S/P}
- abstraction* (abstraction of what claims* are about)	A	

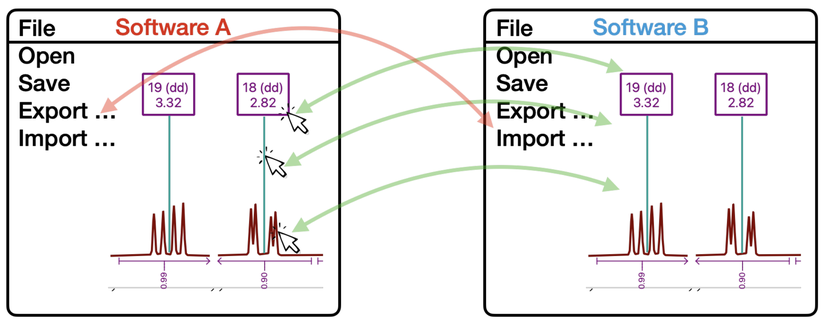

Figure 1: *Usually,* data exchange between software applications is performed using export and import tools (red arrows). Import/export steps may be necessary to find satisfactory solutions. The CHEMeDATA object definitions for specific types of chemical data (spectra, multiplet analysis, chemical shifts, regions of interest in spectra, *etc.*) could facilitate data exchange between commercial software.

CHEMeDATA objects

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

- an NMR sample (sample*)
- a 3D molecular structure (object*)
- a NMR spectrum (measurement*)
- a description of purification, synthesis, *etc*. (processes*)

… or much more specific chemical information:

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

The objects follow a hierarchy allowing anybody to derive existing CHEMeDATA objects so that they better fit his/her specific 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 the multiple phases of their NMR samples.

Viewer of CHEMeDATA objects

We recommend providing a web browser viewer for every CHEMeDATA object. We wrote JavaScript classes that create SVG graphics in web browsers using D3.js for the following objects:

-**NMR spectra** (bottom of Figure 2). Note the options to select regions of interest - see the scale near 2.5 and 2.3 ppm.

-**NMR assignment of chemical shifts** (Vertical gray lines running from multiplet to the labels of the assigned protons at the very top)

-**Scalar compling constants** (J-graph at the top of the Figure 2 where horizontal lines connect pairs of assigned coupling constants and open circles unassigned couplings)

-**3D molecular structures** using the JSmol visualizer

A mecanism of signal transmission was introduced in the base class of viewers allowing interactions between graphical objects (see Figures 2 and 3).

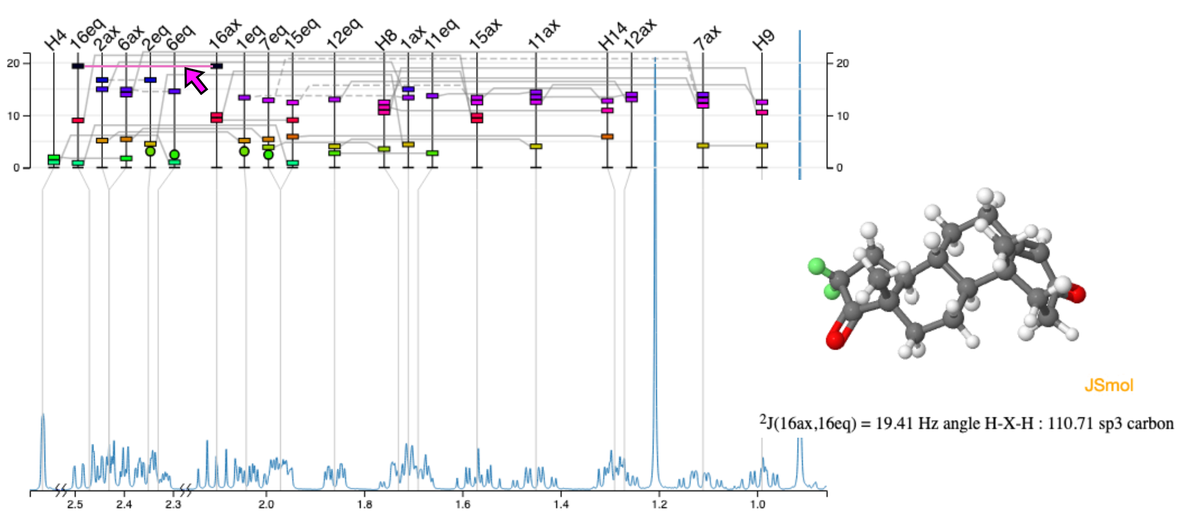

Figure 2: Screen captures of a web browser illustrating the integration of viewers for the spectrum, the assignment, and the J-graph. Zooming in on the spectrum adjusts the assignment and J-graph accordingly. When the mouse pointer rolls over an assigned coupling (pink arrow), the corresponding hydrogen atoms are highlighted (in green) in the JSmol visualizer.

As an alternative to coding a viewer, a contributor to the CHEMeDATA initiative introducing, say, a peak list, could write javaScript code to transform the peak list into an NMR spectrum and use that viewer instead.

Compatibility with Mnova

Starting from the version 14.1 of Mnova, spectra, peak-lists and signal assignment can be exported and imported in the json format. An API's can be used to interconvert the content of Mnova json files into CHEMeDATA objects.

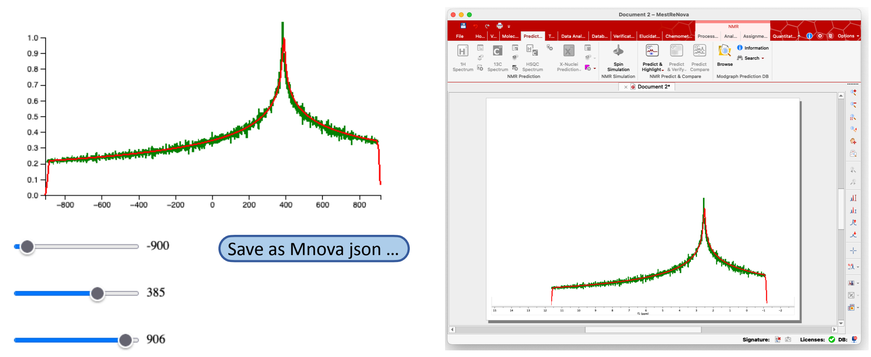

Figure 3: Schematic illustration of the generation of solid-state powder-pattern spectra for visualization in Mnova. The main CHEMeDATA page defining a CSA tensor (left) can generate the corresponding NMR spectrum object and automatically take advantage of the existing exporter to a Mnova JSON file.

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 (Figure 4) .

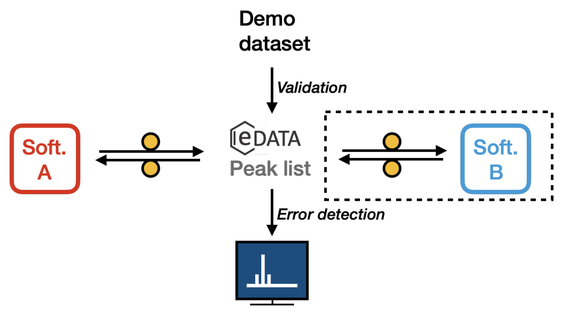

Figure 4: CHEMeDATA objects have demonstration web pages with editable examples, schema validation, and visualization (bottom). The interconversion of data into CHEMeDATA objects involves small units of code, for which demonstration examples, schema validation, and error detection can help newcomers (Software B in the dotted frame) ensure that the data they produce conform to the standard.

The output of (say) a Gaussian geometry optimization could consist of a set of JSON objects representing the electronic energy, geometry, chemical properties, and more. These objects could be generated by small pieces of code, making maintenance easier and allowing unit tests to detect any compatibility issues. Writing a converter from Gaussian files to CHEMeDATA atomic property objects would benefit all users of this software by enabling compatibility with all existing viewers and tools. These could include the generation of solid- and liquid-state NMR spectra from DFT/GIAO calculations, J-graph visualizations (see Figure 2), shielding tensor visualization, *etc*.

Chemistry electronic notebooks

CHEMeDATA aims to become the standard export format for chemistry electronic notebooks. Its object model—covering entities such as *samples*, *transformations*, and more—is designed to adapt to diverse formats, making it a common language of choice for enhancing the generation of FAIR data in chemistry research.

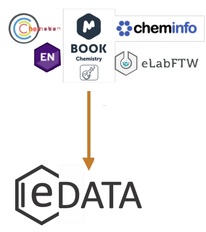

Figure 5: CHEMeDATA can be seen as the Rosetta Stone for electronic lab notebooks, enabling interoperability between diverse data formats and facilitating the standardized exchange of chemical information.

Technical details

CHEMeDATA instances are schema-based and include linked data to support multiple references to a single object.

Each CHEMeDATA object is associated with a main web page and a JavaScript class, which enables:

- Schema validation of object instances — particularly useful during development
- Import from common chemical file formats
- Version management, ensuring that changes to the format or the addition of new fields do not render repository data obsolete or incompatible with legacy viewers
- Generation of related objects (*e.g.* spectra from peak lists)
- Call of user-defined CHEMeDATA viewer or third-party software solutions.

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.

Acknowlegements

Damien Jeannerat thanks Jean-Marc Nuzillard for many years of stimulating discussions, Bob Hanson and other members of the IUPAC working group for the *Development of a Standard for FAIR Data Management of Spectroscopic Data*, members of NFDI4CHEM, Stefan Kuhn, Luc Patiny, Julien Wist and other contributors to the NMReDATA initiative and Pierre-Yves Burgi, Hugues Cazeaux, Mathieu Vonlanthen, Lamia Friha and their colleagues at the eResearch of the University of Geneva, for creating the conditions for this project to arise. We also thank Yair Rodríguez de la Peña for his contribution during his internship at Mestrelab.

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

PANACEA -nmr.eu