NMReDATA, a standard to report the NMR assignment and parameters of organic compounds

Marion Pupier[a], Jean-Marc Nuzillard[b], Julien Wist[c], Nils E. Schlörer[d], Stefan Kuhn[e], Mate Erdelyi[e], Christoph Steinbeck[f], Antony J. Williams[g], Craig Butts[h], Tim Claridge[i], Bozhana Mikhova[j], Wolfgang Robien[k], Hesam Dashti[l], Hamid R. Eghbalnia[l], Christophe Farès[m], Kessler Pavel[n], Fabrice Moriaud[o], Mikhail Elyashberg[p], Dimitris Argyropoulos[q], Manuel Pérez[r], Patrick Giraudeau[s,t], Roberto R. Gil[u], Paul Trevorrow[v], Damien Jeannerat[a,]*

To be submitted to Magnetic Resonance in Chemistry as Research article

[a]Department of Organic Chemistry, University of Geneva, 30 Quai E. Ansermet, 1211 Geneva 4, Switzerland, Fax: (+41 22 379 32 15), E-mail: damien.jeannerat@unige.ch

[b]Institut de Chimie Moléculaire de Reims, UMR CNRS 7312, BP 1039, 51687 Reims Cedex 2, France, ORCID : 0000-0002-5120-2556

[c]Chemistry Department, Universidad del Valle, 76001, Cali, Colombia

[d]Department of Chemistry, University of Cologne, Greinstr. 4, 50939 Köln, Germany

[e]Department of Chemistry - BMC, Uppsala University, Husargatan 3, 752 37 Uppsala, Sweden, *ORCID :0000-0003-0359-5970*

[f]Institute for Inorganic and Analytical Chemistry, Friedrich-Schiller-University, Lessingstr. 8, 07743 Jena, Germany, ORCID :0000-0001-6966-0814

[g]National Center for Computational Toxicology, Environmental Protection Agency, 109 T.W. Alexander Drive, Room D131I, Mail Drop D143-02, Research Triangle Park, NC 27711, ORCID :0000-0002-2668-4821

[h]School of Chemistry, Bristol University, BS8 1TS, Bristol United Kingdom ORCID :0000-0001-6678-8839

[i]Oxford University, Chemistry research laboratory, Mansfield Road, Oxford, OX1 3TA, United Kingdom

[j]Institute of Organic Chemistry with Centre of Phytochemistry, Bulgarian Academy of Sciences, Akad. G. Bonchev Str. Bl.9, Sofia 1113, Bulgaria

[k]University of Vienna, Department of Organic Chemistry, Währingerstr. 38, 1090 Vienna, Austria

[l]Department of Biochemistry, National Magnetic Resonance Facility at Madison (NMRFAM), 433 Babcock Drive, Madison, USA

[m]Max-Planck-Institut für Kohlenforschung, Abteilung NMR, Kaiser-Wilhelm-Platz 1, 45470 Mülheim-an-der-Ruhr, Germany

[n]Bruker BioSpin GmbH, Silberstreifen, 76287 Rheinstetten, Germany

[o]Bruker BioSpin AG, Industriestrasse 26, 8117 Fällanden, Switzerland

[p]Advanced Chemistry Development, Moscow Department, 6 Akademik Bakulev Street, Moscow 117513, Russian Federation

[q]Advanced Chemistry Development, Inc. (ACD/Labs), Venture House, Arlington Square, Downshire Way, Bracknell, Berkshire, RG12 1WA, United Kingdom

[r]Mestrelab Research, S.L. Feliciano Barrera 9B – Bajo, ES-15706 Santiago de Compostela, Spain

[s]EBSI Team, Chimie et Interdisciplinarité : Synthèse, Analyse, Modélisation (CEISAM) Université de Nantes, CNRS, UMR 6230, BP 92208, 2 rue de la Houssinière, 44322 Nantes, France.

[t]Institut Universitaire de France, 1 rue Descartes, 75005, Paris Cedex 05, France

[u]Department of Chemistry, Carnegie Mellon University, 4400 Fifth Ave., Pittsburgh, PA 15213, USA, ORCID :0000-0002-8810-5047

[v]Wiley, The Atrium, Chichester, PO19 8SQ, United Kingdom

ABSTRACT

Even though NMR has found countless applications in the field of small molecule characterization, there is no standard file for the NMR data relevant to structure characterization of small molecules. A file format is introduced to associate the NMR parameters extracted from 1D and 2D spectra of organic compounds to the assigned chemical structure. These NMR parameters, which we shall call NMReDATA, include chemical shift values, signal integrals, intensities, multiplicities, scalar coupling constants, lists of 2D correlations, relaxation times and diffusion rates. The file format is an extension of the existing SDF (Structure Data Format), which is compatible with the commonly used MOL format. The association of an NMReDATA file with the raw and spectral data from which it originates constitutes an NMR record. This format is easily readable by humans and computers and provides a simple and efficient way for disseminating results of structural chemistry investigations, automating the verification of published result, and for assisting the constitution of highly needed open-source structural databases.

Keywords

NMR

computer-assisted structure elucidation (CASE)

data format

database

NMR record

NMReDATA

Extracted data

1 Introduction

Exploiting the NMR spectra of compounds consists of extracting the NMR parameters (which we shall call NMReDATA) including chemical shift values, scalar coupling constants, 2D correlations, *etc*. and assigning them to a chemical structure. The reliability of the structure determination depends heavily on the accuracy and extent of the NMReDATA which need to be reported in a format compatible with a maximum number of software packages. Here, we introduce a file format to effectively associate these NMReDATA with the assigned structure.[1] These electronic reports will be generated by software assisting chemists in the interpretation of spectra, or by using tools to create them in the case where the spectra were interpreted using "paper and pencil". The commercial software providers Bruker, Mestrelab Research, Advanced Chemistry Development (ACD/Labs), and the non-commercial software platform Nmrshiftdb2[2] and C6H6.org[3], have all committed to use the format in future releases of their products. These formatted NMReDATA will facilitate the accessibility and exchange of the partial or full assignment within the chemistry community. More importantly, it will constitute an easier way for the inspection and verification by reviewers when assigned data are submitted to scientific journals. Simply opening the NMReDATA file will permit a user to interrogate and confirm the assignments associated with the signals in the spectra. This contrasts with the current unsatisfactory situation where NMR parameters are dispersed in peak lists, tables of correlations and low-quality image-based spectra typically located in supplementary material, commonly in PDF format only and thereby making it difficult for both humans and computers to access and assess the data.[4]

Because of the need to keep both the NMReDATA and the associated spectra together to allow for verification of their consistency, we have also introduced the concept of "NMR record". This refers to a compressed folder including the NMReDATA file and all related 1D and 2D spectra (see Figure 1) including the raw data, that is, the FIDs and all the acquisition and processing parameters. Such NMR records will also be optionally exported and imported by NMR software packages such as CASE (Computer-Assisted Structure Elucidation) software[5] and could be submitted with manuscripts for publication. For *Magnetic Resonance in Chemistry*, the requirement that all "spectral assignment" manuscripts submissions be accompanied by full NMR records was announced recently and will be gradually enforced during 2018.[6] Other journals have demonstrated interest in this initiative and may follow suit. The intention of the NMReDATA initiative is the following. When journal articles are accepted for publication, the associated NMR records are deposited with keywords and metadata about the publication in open databases thereby making them permanently available to the community. The database should provide a DOI (Digital Object Identifier) to insure stability of the data. Making spectral data available should also satisfy the "Open Data" requirements of a growing number of funding bodies.

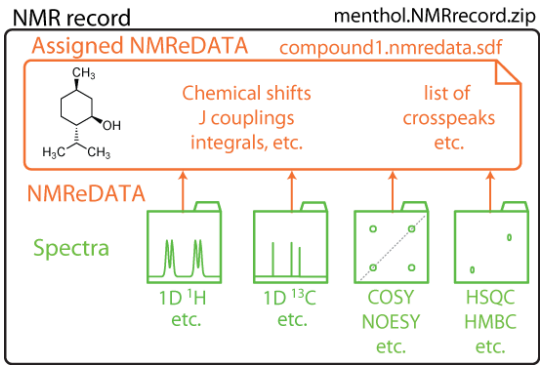


Figure 1. An "NMR record" consists of a compressed folder in .zip format and includes a set of NMR spectra with the time domain FIDs, the acquisition and processing parameters (in green), and the NMReDATA file (.nmredata.sdf file in orange) containing a chemical structure and the extracted NMR data .

Note that related efforts, e.g. NMR-STAR[7-10] and nmrML[11] intend to cover the maximum number of data types. In contrast, the NMReDATA initiative intends to restrict to a core set of parameters relevant to support structure determination. This should facilitate the interoperability of the experimental spectra and their associated meta-data.

2 Structure Data Format (SDF)

NMR-extracted data (NMReDATA) are associated with a chemical structure in a Structure Data Format (SDF) file. This file format was introduced by Molecular Design Limited (MDL) to associate one or more chemical structures with diverse types of data.[12] The structures are encoded in the commonly used MOL (.mol) format, also introduced by MDL, dealing with 2D and 3D structures with and without implicit hydrogen atoms. The additional data are included in the file in the form of property value pairs called "tags" (see Figure 2). SDF files can be considered as MOL files with additional metadata. This file format is the primary import and export format for chemical structure drawing software packages. It is also used by the vast majority of publicly available chemistry databases (e.g. PubChem,[13] ChemSpider,[14] ChEBI[15]) and across the field of cheminformatics to associate chemical structures with properties such as solubility, affinity to a drug target, or more trivial information such as storage locations, *etc*.

The .nmredata.sdf files will therefore include the assigned chemical structure (black in Figure 2) and all the NMR data (and possibly also including non-NMR data) in a compact and convenient manner (see tags in Figure 2). Structure visualization software suites (such as ChemDraw,[16] Jmol,[17] etc.) will open the SDF files and only display the chemical structure encoded in MOL file format while NMR software packages will be able to access the NMReDATA tags and allow for their visualization, manipulation, validation, etc.

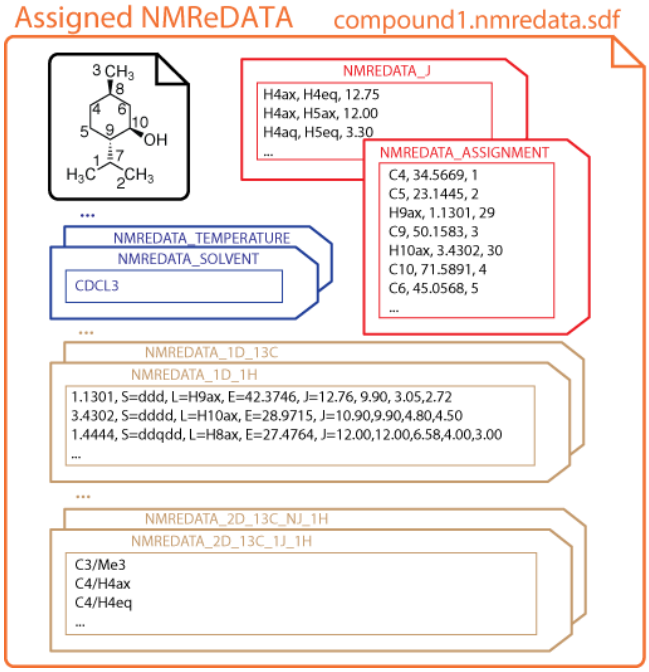


Figure 2. Content of an SDF file included in NMR records. It consists of a structure (in black) and the NMR extracted data (NMReDATA tags). The property value pairs comprise information on the sample (blue tags), compilations of the chemical shift and scalar coupling information (red tags) and descriptions of 1D and 2D spectra (brown tags).

3 Chemical structure

A key requirement of an NMR assignment is to report the structure of the compound appropriately. The soundness of a molecular structure can be judged, if not proven, by the quality of the pairing of its structural features with its extracted NMR parameters. Indeed, depending on the data quality and requirements, an NMR characterization may be reported with different levels of structural precision. In the simplest case, one may only be able (or wish) to determine the connectivity of atoms (sometimes called "flat structure") with no additional information about the three-dimensional structure (see Figure 3A). In some cases, the exploitation of specific NMR parameters (e.g. a scalar coupling) may provide a local structural feature (e.g. the relative configurations of a pair of carbons as shown in Figure 3B). Finally, in the most exhaustive cases, the full analysis of the coupling constants, NOE data, anisotropic parameters or the knowledge of the absolute configuration obtained by non-NMR means, allows the unambiguous configuration of all chirality elements of the molecule to be specified as in Figure 3C. In all cases, it is important to use a representation of the structure reflecting the available knowledge in order to avoid under- or, more problematically, over-interpretation. Note that given the inability of NMR to determine absolute configuration with classical methods, one should always consider the absolute configuration of the reported structures as not being ascertained.

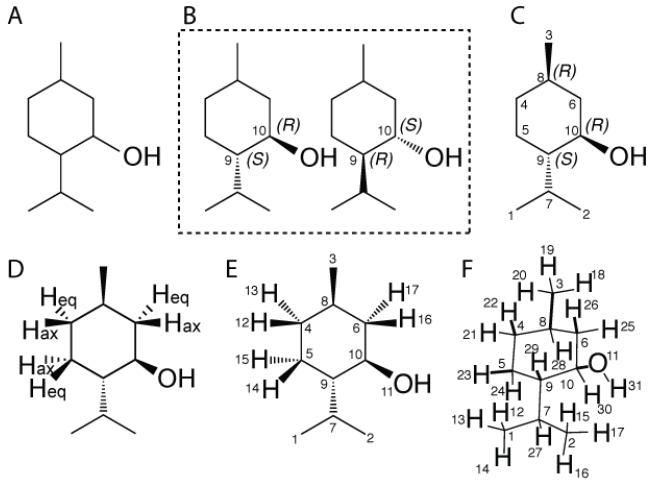


Figure 3. Different levels of interpretation of the NMR spectra of menthol. (A) The absolute configuration of the carbons is unknown. (B) The relative configuration of carbons 9 and 10 is defined based on the observation of a characteristic large *trans*-diaxial coupling (J > 10 Hz). (C) The chirality is fully determined. (D) Structure with explicit hydrogen atoms depicted for unambiguous assignment of the pairs of non-equivalent protons of the methylene groups. (E) Representation and ALATIS numbering of non-equivalent atoms. (F) True 3D model structure with all hydrogen atoms, including equivalent ones, with ALATIS numbering. InChI=1S/C10H20O/c1-7(2)9-5-4-8(3)6-10(9)11/h7-11H,4-6H2,1-3H3/t8-,9+,10-/m1/s1 as reported by ALATIS. Note that even if the chirality of menthol is fully determined, the assignments of carbon 1 and 2 are usually not unambiguously assigned because it has no consequence on the determination of the structure.

The choice of which structure to include in an NMReDATA file also depends on the completeness of the assignment. When the protons of the methylene groups of menthol are only assigned to their directly attached carbons, a structure with implicit hydrogen atoms is appropriate (Figure 3C). But when axial and/or equatorial protons are discriminated (typically through the observation of a large *trans*-diaxial coupling constant), a structure with explicit protons is necessary to unambiguously assign the two hydrogen atoms (see Figure 3D).

In general, chemical structures can be expressed with InChI String codes. But the fact that one of the roles of the structure is to provide atom indices for the assignment requires us to make sure that the interconversion always produces the same atom numbering. This can be done using the ALATIS code [18] (where ALATIS stands for Atom Label Assignment Tool using InChI String). This ensures that structures generated by diverse sources result in a defined and unique atom numbering.

In the context of NMReDATA, we therefore recommend the structure part of NMReDATA to use the ordering of the atoms of ALATIS-generated structures. This allows one to safely replace the structure with an ALATIS code making it possible to generate pure-text assignment reports where the assignment is performed using atom indices.

4 NMReDATA tags in SDF files

NMReDATA are encoded using a set of tags (see Figure 2 and Table 1). The tag names start with the prefix "NMREDATA_" in order to avoid confusion in those cases where the SDF files also include property types other than the NMR related ones. The *NMReDATA Initiative* defined, for its version 1, the set of "tags" listed in Table 1.

Table 1. NMReDATA tags names and content (Version 1.0)[1]

NMREDATA_VERSION

1.0

NMREDATA_SOLVENT

Solvent, or mixture of solvent with their proportions, *etc*.

NMREDATA_ID

Doi=...	DOI of the NMR record
Record=...	URL pointing to the directly accessible ziped record
Path=...	Location of the nmredata.sdf file relative to the root of the NMR record
...	Other references can be added by the software packages generating the files and the databases storing them.

NMREDATA_TEMPERATURE

Temperature of the sample during the NMR acquisition. (in Kelvin)

NMREDATA_CONCENTRATION *(optional)*

Concentration in mM

NMREDATA_LEVEL

Specify the complexity of the content of the NMReDATA files[2]

NMREDATA_FORMULA

Chemical formula of the compound

NMREDATA_INCHI

InChI string of the chemical structure

NMREDATA_ALATIS

Standard InChI string of the chemical structure as generated by ALATIS

NMREDATA_ASSIGNMENT

-

Associate the labels used for the assignment to a chemical shift and the

atom numbers in the chemical structure

NMREDATA_J

Compile assigned scalar couplings (only present when couplings were assigned)

NMREDATA_1D_*Isotope*

List the peaks extracted from 1D spectra (see Table 4)

NMREDATA_2D_ *Isotope (F1)_Code of the mixing_ Isotope (F2)*

List peaks extracted from 2D spectra (see Table 5)

[1] Refer to the website of the Initiative for more details. [1]

[2] A non-zero value indicates the presence of ambiguous assignment (see section "Ambiguities").

4.1 General information about the sample and the NMR experiment

A minimal set of information about the sample and the performed NMR experiments is part of the NMReDATA format. It includes the solvent composition, the temperature and, when known, the concentration of the compound. The spectrometer magnetic field is not listed here, but can be inferred from the Larmor frequency of the detected nuclei for each NMR spectrum (see below).

Other data, such as the equipment used, pulse sequences, pulse durations, referencing, etc. are not part of the format, but can be searched in the acquisition and processing parameters that are present in the associated NMR record.

We strongly recommend to include the chemical formula, the SMILES and InChI code, but only if they truly reflect the chemical structure. These can facilitate searches in online databases. If they are not present in NMReDATA generated by the NMR software, they should be added before they are made available on databases.

4.2 The Assignment Tag

The assignment tag provides, for each atom label used in the assignment of the spectra, the chemical shift (with adequate number of decimal places[19]) and the list of atom indices they refer to (see Figure 2). Symmetry properties, such as the degeneracy resulting from the fast rotation of methyl groups, can be expressed by associating to a label, the list of indices of the atoms it refers to (*i.e.* the ones of the three hydrogen atoms). Similarly, the non-equivalent protons of a CH_2 will be usually designated with a single label. If necessary, for example in cases of magnetic non-equivalence (see Figure 4) and when different coupling constants need to be reported, the two hydrogen atoms should have different labels allowing the coupling constants to be assigned to a given coupling partner. In this case, the identical chemical shift will indicate the chemical equivalence.

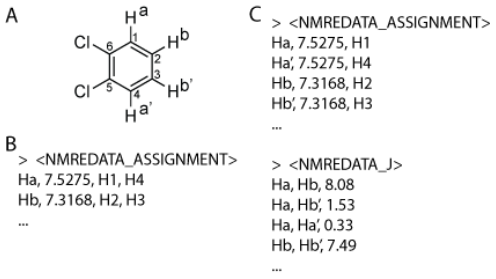


Figure 4. Examples of NMREDATA_ASSIGNMENT tags for the description of the ^{1}H spectrum of orthodichlorobenzene (A). For simplicity only hydrogen atoms were listed in the figure. In general, the list includes reference to carbon, fluorine, phosphorus, etc. atoms when their spectra are part of the NMR record. (B) When ignoring the complex structures of the AA'BB' system, the protons can be described using a single label ("a" and "b"). (C) When the coupling constants were extracted using optimal fitting with spectral simulations, distinct labels for the equivalent protons (a, a', b, b') are necessary to report all coupling constants using a NMREDATA_J tag (see next paragraph). Note that because hydrogen atoms were implicit in structure (A), "H" was added and the indices of the atoms are the ones of the directly attached heavy atoms.

The redundancy of the reporting of the chemical shifts and coupling constants with respect to the description of the 1D spectra (see below) is deliberate. The compilation of the data found in a set of spectra facilitates access to these essential NMR parameters, avoiding the need to search and analyze the values found in the individual 1D and 2D spectra. Another reason is that in some spectra, the chemical shifts may be reported in an ambiguous manner reflecting the limits of the specific experiment. For example, a set of protons may overlap in a 1D ^{1}H spectrum and the chemical shift reported as a broad chemical shift range, while an HSQC or a *J*-resolved spectrum may provide clear-cut values for each individual chemical shift. The chemical shifts in the assignment tag can be seen as the aggregation of the chemical shift measured through the set of NMR spectra of the record.

4.3 The J-coupling tag

In some situations, the analysis of the scalar coupling constants is essential for the determination of chemical structures. By analogy to the compilation of the chemical shift, the NMReDATA format defines a tag to list all of the assigned coupling constants. Typically, the J_{HH} and J_{CH} are used to determine the 3D structure of natural and synthetic products (see the example of menthol in Fig. 5). In other cases, such as when measuring long-range J_{CH} constants or when reporting $^{1}J_{CH}$ including residual-dipolar coupling, the list will contain only (or mostly) $^{2,3}J_{CH}$ or $^{1}D_{CH}$ respectively. Obviously, other couplings such as J_{CF}, J_{HP}, *etc*. can also be included.

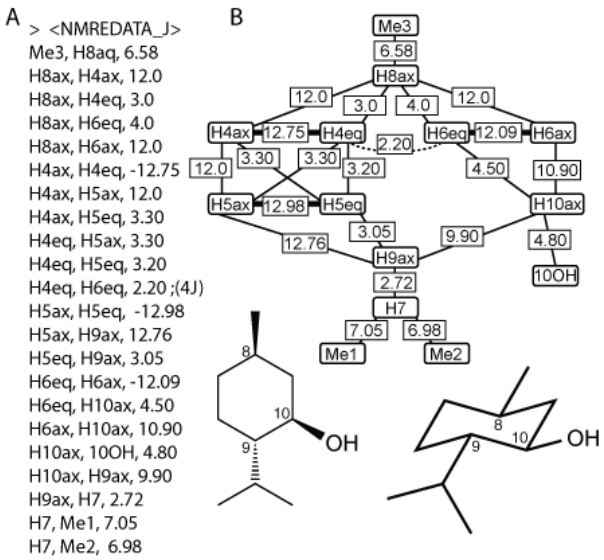


Figure 5. (A) Example of a *J*-coupling tag. (B) Corresponding ^{1}H-^{1}H coupling network. The all-equatorial substitution of the cyclohexane ring of menthol can be deduced from the six characteristically large coupling constants (> 8 Hz) of pairs of axial protons. Thick, normal and dotted lines represent ^{3}J, ^{2}J and ^{4}J respectively originating from the spectral tag (see following sections). When known, the signs of the coupling constants should be specified, otherwise the absolute values are listed. Optionally, the number of bonds between the pair of spins can be specified after "nb=". (eg. "Me3, H8aq, 6.58, nb=3").

4.4 Tags describing spectra

For each spectrum included in NMR records, a tag will describe the extracted information. The name of the tag (see Table 2) encodes the type of spectrum (1D v/s 2D), and the detected isotope. For 2D correlation spectra, the isotope evolving during t_1 and the type of mixing is specified before the detected isotope.

Table 2. Nomenclature of spectral tag names [1]

Tag name	Type of spectrum
NMREDATA_1D_1H	1D ^{1}H spectrum
NMREDATA_1D_13C	1D ^{13}C spectrum
NMREDATA_1D_13C#2	Second 1D ^{13}C spectrum (for example, for DEPT spectrum)
NMREDATA_2D_13C_1J_1H	2D spectrum with ^{13}C evolution during t_1, ^{1}H evolution during detection and ^{1}J mixing (HSQC, or HMQC experiments)
NMREDATA_2D_1H_NJ_1H	2D COSY spectrum

[1] More details can be found on the website of the NMReDATA initiative.

Before listing the peaks extracted from the spectra, a header provides important characteristics of the spectrum using the keywords (see Table 3). The Larmor frequency of the detected isotope follows "Larmor=". Decoupling or absence of decoupling (when it is assumed by default - as for HSQC, HMBC spectra) is specified using "Decoupled=*isotope*" and "NonDecoupled=*isotope*". Specifying the type of experiment ("CorType=", COSY, HSQC, etc.) is particularly important to allow the verification of the compatibility of the reported NMR parameters with the chemical

structure (see Validation of NMReDATA). The format of the peak list differs for 1D and 2D spectra, and is discussed in the next two sections.

A pointer to the original spectrum (with FID, acquisition and processing parameters in the crude original format of the manufacturer) is mandatory to permit verification (or refinement, correction, etc.) of the NMReDATA. The "Spectrum_Location=" is a directory path, relative to the root of the zip file, pointing to the spectrum. When the NMR record is available in a web-accessible database, the HTML link having as prefix a DOI (to insure reliability of data access) and pointing directly to the readily downloadable zip file of the record is also specified.

Table 3. Spectral properties (1D and 2D spectra)

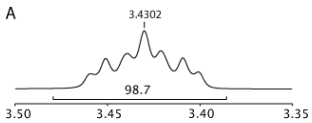
Keyword	Spectrum property
Larmor=*value in MHz*	Larmor frequency of the detected isotope
CorType=*type*	Type (1H, 13C, DEPT135, COSY, HSQC, *etc*.)
Decoupled=*isotope*	List if decoupled isotopes[1]
Nondecoupled=*isotope*	Denies implicit decoupling[1]
Spectrum_Location =	*Path to the spectrum relative to the root of the NMR record*
Pulseprogram =	*Name of the pulse program*

[1] For HSQC and HMQC experiments proton and carbon decoupling is implicit in F1 and F2 respectively. For HMBC experiments proton decoupling is implicit in F1.

4.5 Peak list of 1D spectra

For each signal extracted from 1D spectra, one line lists its parameters. The first field is the chemical shift or chemical shift range. The others, called "attributes", provide additional information such as multiplicity, integral, assignment, *etc*. (See the complete list in Table 4). An example of a multiplet described using various level of details is given in Figure 6.

Note that depending on the ambition of the user, the quality of the spectra and the capabilities of the software package used to analyze the spectra, not all NMR parameters will be reported. The type and the number of peak attributes reported in the NMReDATA will therefore reflect the level of detail of the analysis and may be quite sparce.



B > <NMREDATA_1D_1H>
Larmor=500.13
.....
C 3.475-3.385, E=98.7
D 3.475-3.385, E=98.7, N=1
E 3.4302, M=m, E=98.7, N=1
F 3.4302, I=1.705E5
G 3.4302, M=dddd, L=H10ax
H 3.4302, M=dddd, L=H10ax, J=10.90, 9.90, 4.50, 4.80
I 3.4302, M=dddd, L=H10ax, J=10.90 (H6ax), 9.90(H9ax), 4.50(H6eq), 4.80(10OH)

Figure 6. Examples of descriptions of the same 1D ^{1}H signal. (A) Multiplet of proton H10 of menthol. (B) Tag name and spectral characteristics of the 1D ^{1}H spectrum. (C) Signal defined as a chemical shift range with the crude

integral. (D) Normalization of the integral to one proton. (E) Signal identified with a single chemical shift with an "m" multiplicity. (F) Signal with only its peak intensity, typical for 1D ^{1}H decoupled ^{13}C spectra. (G) Signal identified as a "dddd" and assigned to the proton H10ax. (H) The coupling constants measured in the multiplet are listed and assigned in (I).

Table 4. Peak attributes of 1D spectra

Key letter	Peak attributes	*examples*
-	Chemical shift	4.1238
-		4.1312-4.1278
S=	Multiplicity	s, m, d, t, dd, etc.
J=	Scalar coupling constants	9.90
		9.90 (H9ax)
N=	Number of nuclei	1, 2, 3, etc.
L=	Assignment (label)	H10ax
E=	Integral (crude)	99.87
I=	Signal intensity (relative)	1.67E5
W=	Width at half height (Hz)	1.3
T1=	T_1 relaxation time (s)	1.54
T2=	T_2 relaxation time (s)	1.24
Diff=	Diffusion rate (m^2/s)	3.1E-9

Note that if the software generating the NMReDATA does not allow for the assignment of the coupling partners (Figure 6H), a program observing that couplings with (nearly) identical values are found for protons that are known to be two or three bonds away could be used to assign coupling constants (Figure 6I) and construct the coupling network (Figure 5). As mentioned earlier, more sophisticated tools using spectral simulations[20-23] could further increase the reliability of the assignment and the precision of the coupling constants, and also provide reliable spectral parameters (chemical shifts and scalar coupling constants) in strongly coupled systems.

4.6 Peak list of 2D spectra

For each 2D spectrum, the pairs of correlated signals are listed on separate lines. Other signal properties such as the intensity (either at the exact coordinates of the chemical shifts of signals or slightly off to avoid minimal values in the middle of doublet, or antiphase patterns) are optional (see Table 5).

Table 5. Peak attributes of 2D spectra

Key letter	Peak attributes	*examples*

-	Assignment of the peak F1/F2	C10/H10
E=	Volume (crude)	7.34E6
I=	Signal intensity (relative)	1.67E5
W1 =	Width at half height in F1 (Hz)	2.3
W2=	Width at half height in F2 (Hz)	1.3
Ja=	Active scalar coupling (Hz)	10.9
J1 =	Passive scalar coupling in F1 (Hz)	5.60
J2=	Passive scalar coupling in F2 (Hz)	7.34
		9.90(H9ax)
T1=	T_1 relaxation time (s)	1.54
T2=	T_2 relaxation time (s)	1.24
Diff=	Diffusion rate (m^2/s)	3.1E-9

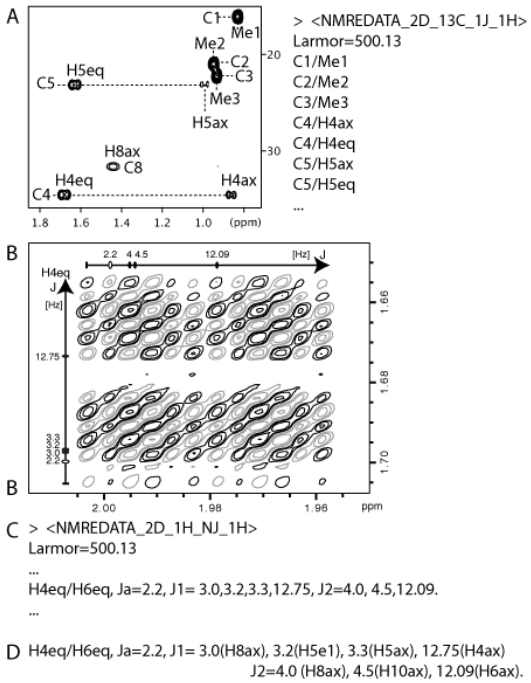


Figure 7. Examples of NMReDATA tag of 2D spectra of menthol. (A) Assignment of region of interest of a 2D HSQC spectrum. (B) Cross peak between the protons H4eq and H6eq of a high-resolution DQF-COSY spectrum illustrating how coupling constants can be included in the NMReDATA without (C) or with assignment (D).

Even if it is still uncommon to measure coupling constants in the multiplet structure of high-resolution 2D COSY spectra,[24] NMReDATA has the capability of incorporating the results of such an analysis (see Figure 7B for an example).

Note that even if this is normally not used, the NMReDATA format allows the inclusion of unassigned cross peaks; in these cases, to define a signal, one can use the corresponding chemical shift values of the signals in place of the atom labels. This allows simple peak-picking software to

generate NMReDATA files, which can be updated by a subsequent assignment process. Note as well that when the NMReDATA files are used to report assigned data, unassigned signals may be ignored.

5 Validation of NMReDATA

Providing electronically readable NMR spectral parameter sets, NMReDATA files allows the manual or automatic verifications of the reported data[25-27] and the evaluation of the soundness of the determined structure. In the following paragraph, we shall only present a broad overview of how this is in principle possible, without discussing the known difficulties associated with automated analysis. We will consider a typical set of spectra used for a structural analysis of an organic compound (often called "full analysis") that is usually 1D ^{1}H and ^{13}C/DEPT-135 and 2D COSY, HSQC, and HMBC spectra. Note that some of the possibilities listed below are already being introduced on commercial software like ACD/Spectrus,[28] CMC-se,[29] Mnova[30] and other non-commercial platforms such as C6H6.[3]

Having the associated chemical structure and the description of 1D ^{1}H and ^{13}C spectra allows for the determination of whether all protons and carbons have been documented. One of the difficulties in this process is the possibility of degenerate signals resulting from symmetry, this is overcome by the assignment of a tag listing the equivalent atoms or the presence of multiple signals with the exact same chemical shifts. In such a way, a program could take into account the nature of the solvent to consider whether the proton from OH or NH are expected to exchange with the ones of the solvent. Note that we encourage that exchanging protons be named “Ex”. Note that in the case of menthol, the hydroxyl proton is not exchanging (as demonstrates the coupling with H10ax) otherwised it would have been called “Ex”. This name will indicate to an automatic verification program that a group of protons assigned to a single signal is not due to symmetry, a property which can be assessed using cheminformatic tools such as CDK.[31]

A verification of the reported multiplicity of signals can be performed regarding the spectrum and the chemical structure. Simply considering that multiplicity should include one “d” for each spin 1/2 located 1 - 3 bonds from the reference nucleus and keeping in mind that “t” counts for “dd”, “q” counts as “ddd”, etc. should be enough in most cases. More sophisticated methods could take into account nearby spins >1/2 (such as ^{2}D, ^{11}B, ^{17}O) or elements with multiple isotopes (resulting to the small doublet due to the minor ^{29}Si isotope or the two doublets due to the ^{117}Sn and ^{119}Sn isotopes). The values of the coupling constants could also be taken into account. A simulation of multiplets could be compared to the experimental spectrum using the reported values of the couplings and optionally using predicted values of the coupling constants.[32] Note that second-order effects could be anticipated by observing, from the NMReDATA, that coupling partners present a small difference in chemical shifts. Determining the coupling constants, even in the case of strong coupling and magnetic equivalence, should be possible by spectral simulation and fitting procedures found in SPINACH[20, 21] (all kinds of spectra), ANATOLIA[22] (currently only 1D ^{1}H spectra) or GISSMO.[23]

The 2D correlations reported in NMReDATA can also be verified by checking the presence of signals at the coordinates of the chemical shifts of the correlated spins. The consistency of the reported data with the structure can also be tested by simply verifying that the number of bonds between the correlated spins is compatible with experiments based on scalar coupling (1 HSQC, 2, 3, (4) for COSY and HMBC data). More sophisticated tools such as Logic for Structure Determination (LSD)[33] can confirm that the structure is compatible with the NMReDATA, and when the structure is not uniquely determined, LSD provides the user with the list of the alternative structures fully compatible with the spectral data.

Quality factors based on the comparison of the experimental chemical shifts and coupling constants with predicted ones[2, 32, 34, 35] could be easily calculated from the NMReDATA.

6 Ambiguities

Ambiguities are commonplace in NMR assignment. In some cases, ambiguous assignment has no consequence on the determined structure. For example, the inability to assign two methoxy groups does not mean that the structure does not have two OMe, but only that their assignments could not be determined. At a different level of the assignment, the lack of resolution of an HSQC spectrum may not allow unambiguous assignment of a pair of carbons (resulting to the quite common "these carbons are interchangeable") but still permit the determination of a structure. These two types of ambiguities were introduced into the NMReDATA format. Because dealing with ambiguities complicates the visualization of the data, we introduced a tag called “NMREDATA_LEVEL” to indicate when ambiguities are reported in the NMReDATA. This allows software tools that are not able to deal with ambiguities to reject such data and facilitate the quick development of programs dealing with the most common non-problematic cases.

6.1 Ambiguities in the assignment of spin

The first type of ambiguities occurs at the level of the assignment - that is when a single chemical shift is associated with a list of atoms in the structure. It consists of listing a set of signals/labels for which a set of different assignments is possible (see Table 6). Another assignment ambiguity case occurs, for example when a methoxy group (with ^{1}H and ^{13}C labeled "CH3a" and "Ca", respectively) is not distinguished from a second methoxy group because of a lack of signal in the HMBC spectrum. One way to produce NMReDATA with these type of ambiguities consists of duplicating (or multiplicating when ambiguities include more than two possibilities) the table of references to the atom indices. These types of ambiguities can be dealt with by considering a set of possible atom numbering tables (instead of a unique one).

Table 6. Ambiguities in the assignment tag

	Example	Meaning
1	Interchangeable=Ha, Hb	The spins Ha and Hb may be interchanged
2	Interchangeable=(CH3a, Ca), (CH3b, Cb)	The spins of CH3a together with Ca, may be interchanged with those of the (CH3b, Cb)

6.2 Ambiguities in the assignment of a peak

Ambiguities can also be specified at the level of the assignment of a peak to a 1D or 2D spectrum, for example if an HMBC signal is too broad in the indirect dimension to point to only one carbon (see the fourth entry in Table 7). In such a case, possible candidate signals are listed in parentheses. One consequence of the presence of such ambiguities is that the validation tool should use a logical "or" on the list of possible assignments.

Table 7. Ambiguities in spectral tags

	Example	Meaning
1	1.2324, L=Ha	Unambiguous assignment of 1D signal
2	1.2324, L=(Ha|Hb)	The signal is assigned to either Ha or Hb.
3	Ha/C1	Unambiguous assignment of a 2D signal
4	Ha/(C1|C2)	The signal is assigned to C1 and/or C2 in the F1 dimension

5	(Ha|Hb)/(C1|C2)	The signal is assigned to one or more of the four possible combinations Ha/C1, Ha/C2, Hb/C1 and Hb/C2

7 Pure text descriptions

In many cases such as reports, PhD theses, etc. a text document is used to report spectral assignments and including the content of NMReDATA files may not be possible or desirable. We therefore support a textual form of the core information of nmredata.sdf files for cases without ambiguities. It consists in a single line of text composed of the content of the second column of the Table 8. (See the Supplementary material for an example).

Table 8. Pure text translation of nmredata.sdf files

	nmredata.sdf file	Pure text equivalent
		NMReDATA(V1)
1		Doi=*doi* Record=*url* [5] Path=*path of sdf file*
2	MOL block	ALATIS InChI code of the molecule
3	ASSIGNMENT tag	ChemShifts:
4	Ha, 1.3453, 10	"10"=1.3453[1]
5	Ha, 1.3453, H10	"H10"=1.3453 (for implicit hydrogen atoms)[1]
6	Hb, 1.5442, 4, 5, 6	"4-6"=1. 5442[1]
7	J tag	J:
8	Ha, Hb, 7.50	("10","4-6")=7.50[1]
9	1D_*isotope* tag	1D(Type[6], Larmor frequency[2], solvent[3], temperature[4]):
10	1.3453, S=*text1*, N=*text2*, J=*text3*, L=*text4*, ...	
11		1.3453 (*text1*, <*text4*>[7], *text2* H, J= *text3*)[1]
12		Path=*path* [5]
13	2D_*iso1_mix_iso2* tag	2D(CorType[8], *iso1*, *iso2*, Larmor frequency[2], solvent[3], temperature[4]):
14	text1/text2	"text1"/"text2"[1]
15		Path=*path*[5]

[1] A comma followed by a space is added as separator between entries. A period followed by a space is added after the last entry.

[2] From the "Larmor=" property of 1D and 2D spectra.

[3] From the NMREDATA_SOLVENT tag.

[4] From the NMREDATA_TEMPERTURE tag.

[5] From the "Spectrum_Location" properties of 1D and 2D tags.

[6] Either the isotope of the spectrum (1H, 13C) of the name of the pulse sequence (DEPT135, DEPT90, *etc*.).

[7] In the absence of comma to list multiple atoms, the "<" and ">" delimiters are not mandatory.

[8] From the "CorType=" of property of 2D tags.

After a header announcing "NMReDATA" and its version number, the chemical structure is encoded using the ALATIS format. This ensures that all information about the structure is retained and that the sequence of atoms is unambiguous. This allows the replacement of assignment tag labels with indices of the atom in the ALATIS format within double quotes. Then follows the compilation of chemical shifts and (when present) the assignment of the couplings. Afterwards, for each 1D spectrum, the chemical shifts are listed with the most relevant attributes in parentheses and the DOI and path pointing to the specific spectrum within the record. Finally, for each 2D spectrum, the list of correlated signals is given by direct reference to the atoms indices.

Note the absence of italics, Greek letters, or sub- or superscripts, facilitating the extraction of the text and making it possible to compress the text as QR codes if desired.

8 Analysis of mixtures

Because the NMR record includes the spectra of a given sample it may, by nature, contain multiple components (mixtures, slowly exchanging isomers, etc.). We therefore had to consider those cases where more than one compound is associated with an NMR record. When more than one compound is assigned to a set of spectra, one should simply generate one NMReDATA file per assigned compound (see Figure 8).

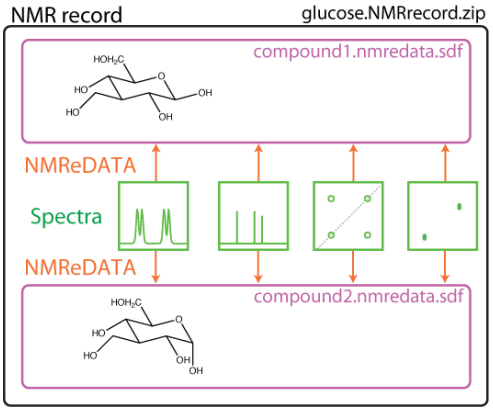


Figure 8. Structure of the NMR record of a sample containing the two slowly exchanging isomers of glucose. Two NMReDATA files (compound1.nmredata.sdf and compound2.nmredata.sdf) corresponding to the two anomers of glucose are included in the NMR records.

9 Conclusion

Despite the wide application of NMR for the study of small molecules, there is no standardized procedure for the exchange of structure related NMR parameters between software suites. As a consequence, retrieving different types of experimental information from reports and publications has been a challenge in the NMR community. Here, we address this challenge by presenting the data structure of NMReDATA records (Version 1.0) that constitutes a concise human- and machine-readable format. It results from the collective work by a wide range of stakeholders including experts in structure elucidation, software developers and journal editors. The format will allow the generation and the exchange of NMR data and facilitate their validation along the way going from the chemists, their supervisors, NMR experts, authors of publications, reviewers, and, finally, all users of the published data. The urgent need of an electronic data format in combination with an

open-access repository of NMR spectral data is further supported by the large number of wrong structures found in the literature[26, 27] and the difficulty in revising them.

After integration of the new format into NMR software, the generation of NMR records including the NMReDATA files will take no additional time. A "save as..." option or a simple copy/paste is all that will be necessary to save or transfer NMR records to dedicated NMR databases such as cheminfo.org. When deposited for review, they will initially be embargoed and only accessible to the reviewers using a coded URL included in the manuscript. When the article is published, the NMR records will become open to the public and relevant records could be merged into databases such as C6H6.org,[3] nmrshiftdb2,[2, 36] (which contents will become part of C6H6.org) BMRB,[10] ChemSpider[14, 37] and Metabolights.[38]

The NMReDATA initiative collaborates with the Biological Magnetic Resonance Bank[10] to develop interfaces between the NMReDATA and the NMR-STAR[7-10] data formats. These interfaces are designed to guarantee seamless access to reference metabolites from BMRB in NMReDATA data format, and also to readily deposit published NMReDATA data into BMRB.

Version 1 of the format should be seen as a compromise, which may evolve in order to integrate new features as motivated developers join the initiative. A future release could address the question of the format of the raw spectra, and additional requirements for the Open Data initiative. Currently, NMR records can include any format generated by spectrometer manufacturers - only the NMReDATA *per se* are standard - but accepting a single standard for the spectra would further simplify the utilization of NMR records.

We believe that in the future, all users of NMR data will have access to software packages allowing them to easily check chemical structures for consistency with all experimental data used for structure verification or elucidation.[5] This shall significantly reduce the number of wrong structures published in chemical journals.

Acknowledgements

DJ and MP thank the State of Geneva and the Swiss NSF (grant no. 200021_147069 and 206021_128746) for funding. H.D. and H.R.E are supported by the U.S. National Institutes of Health (NIH) grant P41GM103399 and P41GM111135 (NIGMS). NES gratefully acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, IDNMR project, Grant SCHL 580/3-1). John Markley is thanked for useful discussions and for comments on the manuscript.

Disclaimer: The views expressed in this paper are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

References

[1] http://www.nmredata.org.

[2] S. Kuhn, N. E. Schlorer, *Magn. Reson. Chem.* **2015**, *53*, 582-589. doi: 10.1002/mrc.4263

[3] L. Patiny, M. Zasso, D. Kostro, A. Bernal, A. M. Castillo, A. Bolaños, M. A. Asencio, N. Pellet, M. Todd, N. Schlörer, S. Kuhn, E. Holmes, S. Javor, J. Wist, *Magn. Reson. Chem.* **2017**. doi: 10.1002/mrc.4669

[4] D. Jeannerat, *Magn. Reson. Chem.* **2017**, *55*, 7-14. doi: 10.1002/mrc.4527

[5] M. E. Elyashberg, A. J. Williams, K. Blinov, *Contemporary Computer-Assisted Approaches to Molecular Structure Elucidation*, RSC Publishing, Cambridge, **2012**.

[6] R. Gil, G. Martin, *Magn. Reson. Chem.* **2017**, *55*, 1057-1058. doi: 10.1002/mrc.4631

[7] N. Spadaccini, S. R. Hall, *J. Chem. Inf. Model.* **2012**, *52*, 1901-1906. doi: 10.1021/ci300074v

[8] S. R. Hall, N. Spadaccini, *J. Chem. Inf. Model.* **1994**, *34*, 505-508. doi: 10.1021/ci00019a005

[9] S. R. Hall, *J. Chem. Inf. Model.* **1991**, *31*, 326-333. doi: 10.1021/ci00002a020

[10] E. L. Ulrich, H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C. F. Schulte, D. E. Tolmie, R. Kent Wenger, H. Yao, J. L. Markley, *Nucleic. Acids Res.* **2008**, *36*, D402-408. doi: 10.1093/nar/gkm957

[11] D. Schober, D. Jacob, M. Wilson, J. A. Cruz, A. Marcu, J. R. Grant, A. Moing, C. Deborde, L. F. de Figueiredo, K. Haug, P. Rocca-Serra, J. Easton, T. M. D. Ebbels, J. Hao, C. Ludwig, U. L. Gunther, A. Rosato, M. S. Klein, I. A. Lewis, C. Luchinat, A. R. Jones, A. Grauslys, M. Larralde, M. Yokochi, N. Kobayashi, A. Porzel, J. L. Griffin, M. R. Viant, D. S. Wishart, C. Steinbeck, R. M. Salek, S. Neumann, *Anal. Chem.* **2018**, *90*, 649-656. doi: 10.1021/acs.analchem.7b02795

[12] A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. Gushurst, D. L. Grier, B. A. Leland, J. Laufer, *J. Chem. Inf. Model.* **1992**, *32*, 244-255. doi: 10.1021/ci00007a012

[13] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, S. H. Bryant, *Nucleic Acids Res.* **2016**, *44*, D1202-1213. doi: 10.1093/nar/gkv951

[14] http://www.chemspider.com.

[15] J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, C. Steinbeck, *Nucleic Acids Res.* **2013**, *41*, D456-463. doi: 10.1093/nar/gks1146

[16] http://www.cambridgesoft.com/software/overview.aspx.

[17] http://wiki.jmol.org/index.php/Main_Page.

[18] H. Dashti, W. M. Westler, J. L. Markley, H. R. Eghbalnia, *Sci. Data* **2017**, *4*, 170073. doi: 10.1038/sdata.2017.73

[19] G. F. Pauli, S. N. Chen, D. C. Lankin, J. Bisson, R. J. Case, L. R. Chadwick, T. Godecke, T. Inui, A. Krunic, B. U. Jaki, J. B. McAlpine, S. Mo, J. G. Napolitano, J. Orjala, J. Lehtivarjo, S. P. Korhonen, M. Niemitz, *J. Nat. Prod.* **2014**, *77*, 1473-1487. doi: 10.1021/np5002384

[20] H. J. Hogben, M. Krzystyniak, G. T. Charnock, P. J. Hore, I. Kuprov, *J. Magn. Reson., Ser. A* **2011**, *208*, 179-194. doi: 10.1016/j.jmr.2010.11.008

[21] I. Kuprov, *Magn. Reson. Chem.* **2017**. doi: 10.1002/mrc.4660

[22] D. A. Cheshkov, K. F. Sheberstov, D. O. Sinitsyn, V. A. Chertkov, *Magn. Reson. Chem.* **2017**. doi: 10.1002/mrc.4689

[23] H. Dashti, W. M. Westler, M. Tonelli, J. R. Wedell, J. L. Markley, H. R. Eghbalnia, *Anal Chem* **2017**, *89*, 12201-12208. doi: 10.1021/acs.analchem.7b02884

[24] D. Jeannerat, *Magn. Reson. Chem.* **2000**, *38*, 156-164. doi: 10.1002/(SICI)1097-458X(200003)38:3<156::AID-MRC610>3.0.CO;2-R

[25] M. Elyashberg, A. J. Williams, K. Blinov, *Nat. Prod. Rep.* **2010**, *27*, 1296-1328. doi: 10.1039/c002332a

[26] W. Robien, in *Progress in the Chemistry of Organic Natural Products, Vol. 105*, 2017 ed. (Eds.: A. D. Kinghorn, H. Falk, S. Gibbons, J. I. Kobayashi), Springer, **2017**, pp. 137-215.

[27] K. C. Nicolaou, S. A. Snyder, *Angew. Chem. Int. Ed.* **2005**, *44*, 1012-1044. doi: 10.1002/anie.200460864

[28] Advanced Chemistry Development, Inc, Toronto.

[29] Bruker Spectrospin.

[30] Mestrelab.

[31] E. L. Willighagen, J. W. Mayfield, J. Alvarsson, A. Berg, L. Carlsson, N. Jeliazkova, S. Kuhn, T. Pluskal, M. Rojas-Cherto, O. Spjuth, G. Torrance, C. T. Evelo, R. Guha, C. Steinbeck, *J. Cheminform.* **2017**, *9*, 33. doi: 10.1186/s13321-017-0220-4

[32] A. Navarro-Vázquez, R. Santamaría-Fernández, F. J. Sardina, *Magn. Reson. Chem.* **2017**. doi: 10.1002/mrc.4667

[33] J. M. Nuzillard, B. Plainchont, *Magn. Reson. Chem.* **2017**. doi: 10.1002/mrc.4612

[34] S. G. Smith, J. M. Goodman, *J. Am. Chem. Soc.* **2010**, *132*, 12946-12959. doi: 10.1021/ja105035r

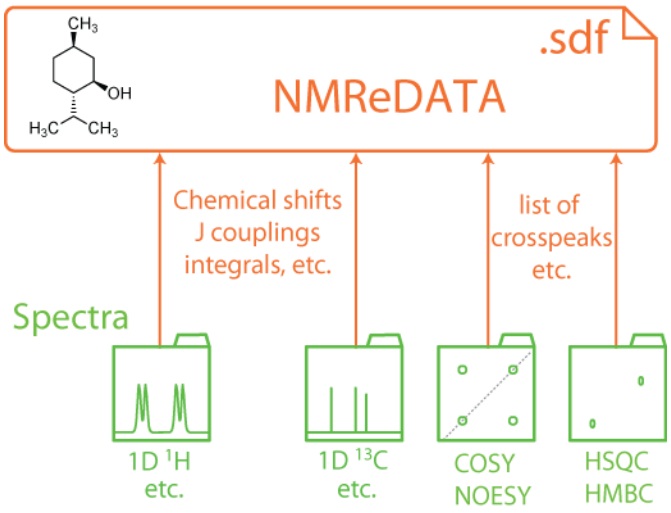
[35] N. Grimblat, M. M. Zanardi, A. M. Sarotti, *J. Org. Chem.* **2015**, *80*, 12526-12534. doi: 10.1021/acs.joc.5b02396

[36] http://www.nmrshiftdb.org.

[37] H. E. Pence, A. Williams, *J. Chem. Education* **2010**, *87*, 1123-1124. doi: 10.1021/ed100697w

[38] http://www.ebi.ac.uk/metabolights.

Graphical abstract



Text of the graphical abstract:

A format for the data extracted from a set of NMR spectra (chemical shifts, coupling constants, 2D correlations, etc.) will make easier to report, compare, verify, validate, share and archive NMR data relevant to structure determination.