

NMReDATA: Tools and applications

Stefan Kuhn, Lianne H. E. Wieske, Paul Trevorrow, Daniel Schober, Nils E. Schlörer, Jean-Marc Nuzillard, Pavel Kessler, Jochen Junker, Angel Herráez, Christophe Farés, et al.

▶ **To cite this version:**

Stefan Kuhn, Lianne H. E. Wieske, Paul Trevorrow, Daniel Schober, Nils E. Schlörer, et al.. NMReDATA: Tools and applications. Magnetic Resonance in Chemistry, 2021, 10.1002/mrc.5146 . hal-03191016

HAL Id: hal-03191016
https://hal.univ-reims.fr/hal-03191016

Submitted on 13 Apr 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Check for updates

Received: 6 December 2020 | Accepted: 5 March 2021

DOI: 10.1002/mrc.5146

TUTORIAL

WILEY

NMReDATA: Tools and applications

Stefan Kuhn[1] | **Lianne H. E. Wieske**[2] | **Paul Trevorrow**[3] | **Daniel Schober**[4,5] | **Nils E. Schlörer**[6] | **Jean-Marc Nuzillard**[7] | **Pavel Kessler**[8] | **Jochen Junker**[9] | **Angel Herráez**[10] | **Christophe Farès**[11] | **Mate Erdélyi**[2] | **Damien Jeannerat**[12]

[1]School of Computer Science and Informatics, De Montfort University, Leicester, UK

[2]Department of Chemistry - BMC, Uppsala Universitet, Uppsala, Sweden

[3]Wiley, The Atrium, Chichester, UK

[4]Ontology Development, MatterWaveSemantics, Südharz, Germany

[5]Leibniz Institute of Plant Biochemistry, Stress and Developmental Biology, Halle (Saale), Germany

[6]Department of Chemistry, University of Cologne, Köln, Germany

[7]CNRS, ICMR UMR 7312, Université de Reims Champagne Ardenne, Reims, France

[8]Bruker BioSpin GmbH, Rheinstetten, Germany

[9]Center for Technological Development in Public Health, Fundação Oswaldo Cruz - CDTS, Rio de Janeiro - RJ, Brazil

[10]Department of Systems Biology, Universidad de Alcalá, Alcalá de Henares, Spain

[11]Abteilung NMR, Max-Planck-Institut für Kohlenforschung, Mülheim an der Ruhr, Germany

[12]NMRprocess, Geneva, Switzerland

Correspondence
Stefan Kuhn, School of Computer Science and Informatics, De Montfort University, The Gateway, Leicester LE1 9BH, UK.
Email: stefan.kuhn@dmu.ac.uk

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Numbers: SCHL 580/3-2, SCHL 580/3-1; National Heart, Lung, and Blood Institute, Grant/Award Number: T32 HL007575; Phenomenal, Grant/Award Number: H2020 654241

Abstract

The nuclear magnetic resonance extracted data (NMReDATA) format has been proposed as a way to store, exchange, and disseminate nuclear magnetic resonance (NMR) data and physical and chemical metadata of chemical compounds. In this paper, we report on analytical workflows that take advantage of the uniform and standardized NMReDATA format. We also give access to a repository of sample data, which can serve for validating software packages that encode or decode files in NMReDATA format.

KEYWORDS

chemical information, data standard, peak assignment, NMReDATA, nuclear magnetic resonance (NMR)

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2021 The Authors. *Magnetic Resonance in Chemistry* published by John Wiley & Sons Ltd.

1 | INTRODUCTION

The nuclear magnetic resonance extracted data (NMReDATA) format[1] was introduced recently for reporting and exchanging nuclear magnetic resonance (NMR) data of small molecules. This text-based format maintains a good readability for humans and can be easily interpreted by computers, contrarily to chemical drawings and the associated NMR data tables published by scientific journals as portable document format (PDF) documents. The NMReDATA were thus designed to facilitate the communication between producers and users of scientific findings in the field of structural organic chemistry. In this paper, we demonstrate how this is done in practice, showing how NMReDATA support the NMR-based discussion of proposed molecular structures using a diverse set of tools. We also provide a free access to a set of test data. These files are given to illustrate the features of the format and to serve as a didactically sound reference point for future users eager to understand its fine details, as a complement to https://nmredata.org/wiki/NMReDATA_tag_format. They can also serve as a test set for software handling NMReDATA files.

It is important to emphasize that the NMReDATA format is not limited to a specific vendor, even though the example uses the Bruker software suite. The format can capture one- and two-dimensional spectra and contains a set of NMR features (i.e., assignment, chemical shifts, and couplings) with a chemical structure representation and thus is independent of the instrument used to generate the data. The NMReDATA file can be combined with raw data in both the time and frequency domains for any type of NMR spectrometer in the NMR record. The raw data can be included in a vendor format as well as in the open nmrML[2] raw data format.

2 | DATA ANALYSIS PROCESS

In this section, we show how the NMReDATA format improves the structure verification workflow for NMR-based investigations of small molecules. The outcome of this workflow can ultimately be used for direct deposition of the resulting standardized data associated with a scientific journal article as well as registration and deposition of the data to relevant repositories. As example for the workflow, we will use the published NMR data from 5α-cyprinol sulfate (found in PubChem at https://pubchem.ncbi.nlm.nih.gov/compound/160665 with CID:160665).[3]

The NMR record and the NMReDATA file can be authored using either instrument vendor software or third-party data processing applications. So far, Bruker, Mestrelab and Advanced Chemistry Development

TABLE 1 An overview of the software mentioned in this paper (order as mentioned in the paper)

Name	Vendor	Language/operating system	Functions	URL	License/availability
Topspin/CMCse	Bruker	Windows, Linux, macOS	Export: NMRRecord+NMReData Import/Visualize/Validate: NMReData	https://www.bruker.com/service/support-upgrades/software-downloads/nmr.html	Commercial, academic licenses available
Mnova	MestreLab	Windows, Linux, macOS	Export/Import: NMR Record/NMReDATA	https://mestrelab.com/software/mnova/	Commercial
Spectrus Processor	ACD/Labs	Microsoft Windows 10	Export NMR Record/NMReDATA	https://www.acdlabs.com/products/adh/spectrusprocessor/index.php	Commercial
nrmshiftdb2	nmrshiftdb2.org	Online	Read/Export/Validate NMReDATA	https://www.nmrshiftdb.org	Free online
NMReDATA J_reader	Angel Herráez	Online	Read/Write/Visualize NMReDATA	https://www3.uah.es/nmr_e_data/reader/reader.htm	Free online
nmredata.com	cheminfo.org	Online	Read/Export/NMR Record NMReDATA	http://nmredata.com/	Free online
NMReDATA javatools	NMReDATA.org	Java	Read/Write/Visualize NMReDATA Export to LSD	https://github.com/NMReDATAInitiative/javatools	LGPL v3.0

(ACD)/Labs have included NMReDATA file creation into their software suites. Bruker allows the export of NMReDATA files from its Topspin software via the CMC-se Structure Elucidation Module.[4] The Mestrelab Mnova software suite[5] and the ACD/Spectrus Processor[6] are NMR processing software suites that support multiple instrument vendor formats, including Agilent, Bruker, and JEOL. Both feature tools for structure elucidation and spectra assignment, allowing for the export of NMReDATA files with the results.

Ultimately, we intend to gain a wide-ranging support for NMReDATA by vendors as well as third-party software suppliers and journals. The role of software suppliers is to ensure that the NMReDATA files can be generated, read, edited, and written, whereas journals will be requested to accept the format for supplemental materials, permanent data deposition, and also to promote format adoption. As the specification of the NMReDATA format is fully open, the different software suites can be used in any desired combination, thereby easing comparison of data sets generated with different tools. An overview of NMReDATA supporting software introduced can be found in Table 1. An up-to-date version is maintained at https://nmredata.org/wiki/Compatible_software. Some of the tools described here are freely available for use, and some are open source.

2.1 | Data preparation

The NMR spectra of the example compound were acquired using a 500 MHz Bruker Avance III HDX spectrometer and processed using Bruker TopSpin version 4.0 software. The complete list of 1-D and 2-D NMR spectra acquired for the compound is reported in Hahn et al.[3] comprising ^{13}C and ^{1}H 1D and ^{1}H–^{1}H correlation spectroscopy (COSY), ^{1}H–^{13}C heteronuclear multiple bond coherence (HMBC), ^{1}H–^{1}H nuclear Overhauser effect spectroscopy (NOESY), and other spectra. The initial NMR record was produced using the CMC-se module of the TopSpin software by Bruker. Figure 1 shows the files of the NMR record and the NMReDATA file produced by TopSpin.

Once the data have been processed and the initial NMReDATA file has been created, the outcomes of numerous NMR postprocessing software applications can be added to the NMReDATA file and saved in a single format. These outcomes, which include spin system matrix,[7] spectral peak lists,[5,8,9] and spectral peak assignments,[10] can be represented in different predefined tags in NMReDATA format. We are encouraging providers to expand the list of software programs that export their computational workspaces to NMReDATA format. For the example compound, the collected experimental data were processed using the Bruker CMC-se software for spectral analyses (initial evaluation, spectral peak picking, and assignment). CMC-se includes an option to export its results to NMReDATA format. After assigning peaks in CMC-se, the NMReDATA file will contain spectral peak lists of all 1-D and 2-D spectra and their associated assignments. Figure 2 shows the assignment as carried out in CMC-se.

If the spectrometer software available at an NMR facility does not export NMReDATA, or as an alternative option to perform the assignment and to export it as NMReDATA, three online tools are available. All are free to use. One is the "Quick Check" option of nmrshiftdb2, which does an automatic assignment (but allows editing this) and can export an NMReDATA file. The "Quick Check" module is available at the www.nmrshiftdb.org website. This needs a manual input of the structure and the shift lists (Figure 3 left) and produces an assignment from those (Figure 3 right). An NMReDATA file can be exported once the assignment is finalized, using manual correction if necessary (Figure 3 bottom).

Another option to explore the contents of an NMReDATA file in a visual and interactive way is

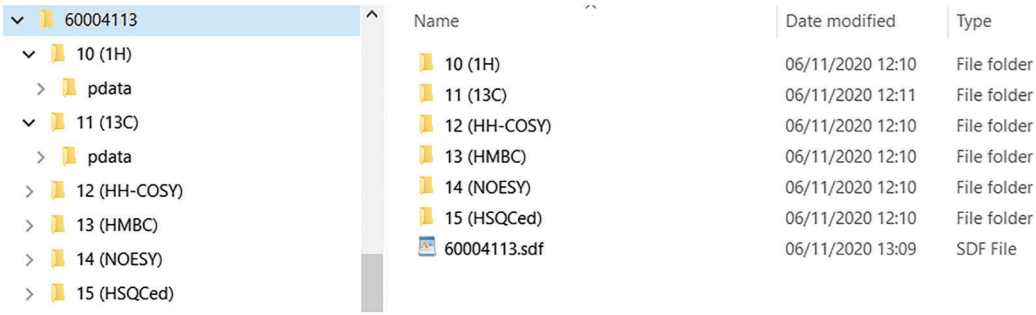


FIGURE 1 A screenshot of the nuclear magnetic resonance (NMR) record of 5α-cyprinol. The nuclear magnetic resonance extracted data (NMReDATA) file (60004113.sdf) sits in the root directory of the record, and the Bruker data are contained in their original form. Each of the directories 10–15 contains the raw data for one spectrum. The processed data (or pdata) directories are part of the Bruker output, alongside files not visible in the file browser

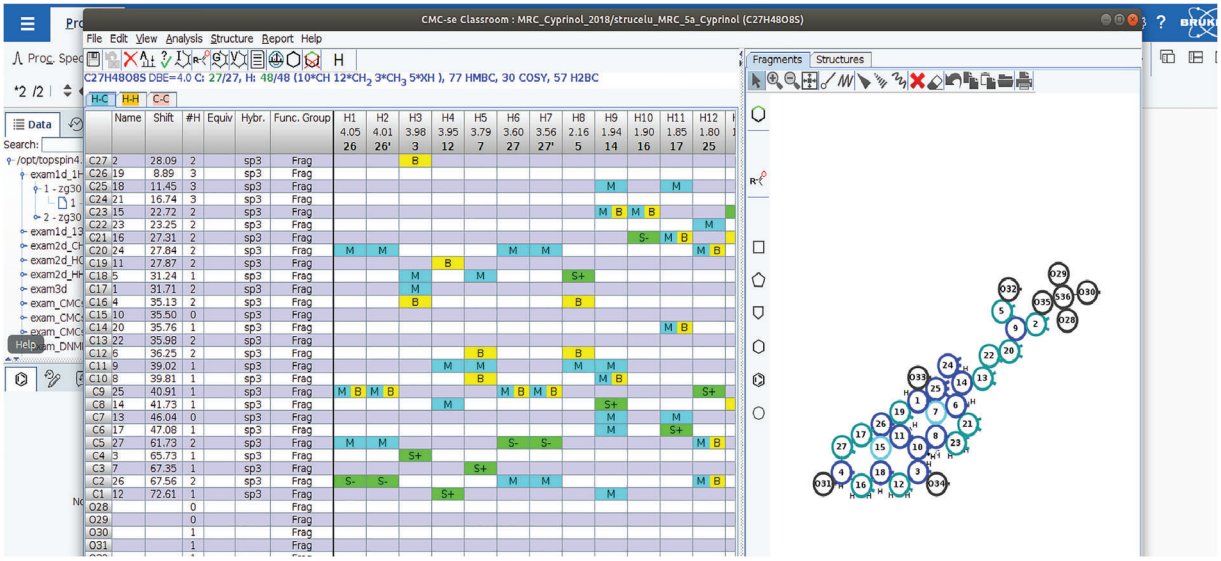


FIGURE 2 A screenshot of the data of 5α-cyprinol opened in Bruker CMC-se.[4] The data had been acquired using Bruker equipment, and the data were processed using Topspin, then opened in CMC-se and saved as NMReDATA from there

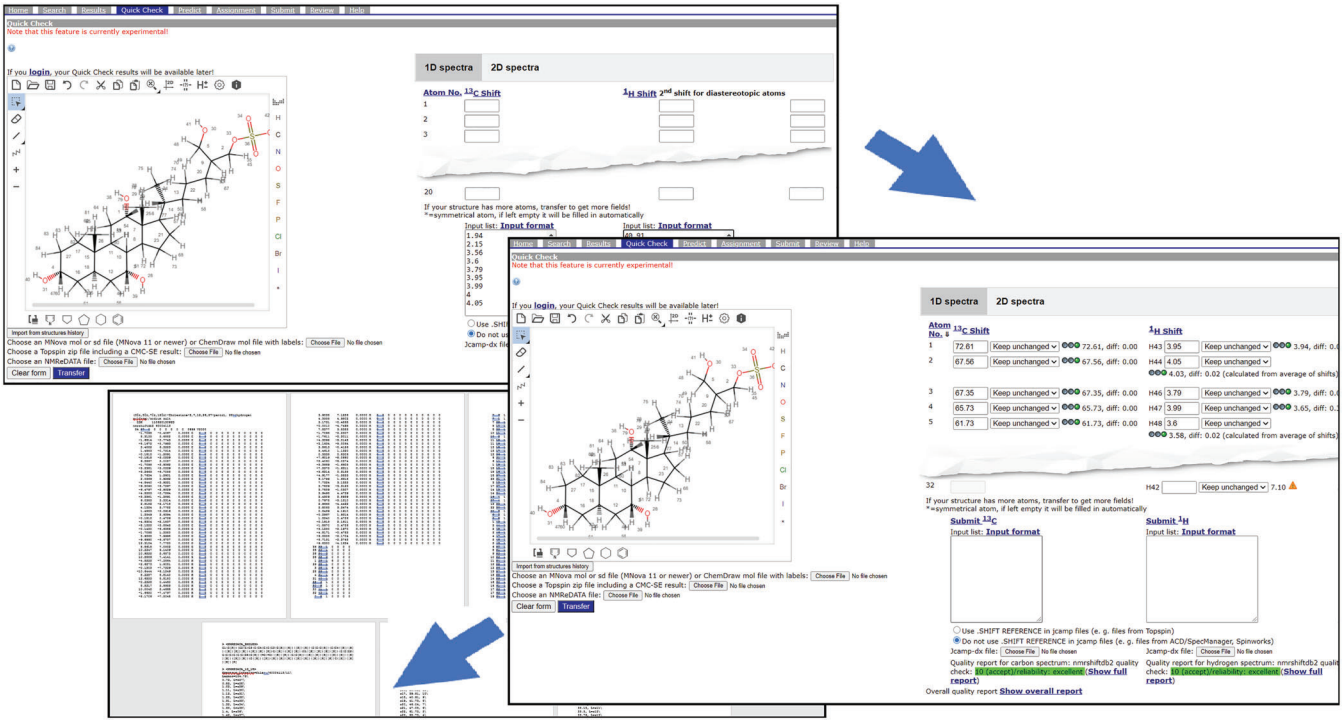


FIGURE 3 Outputs of the NMReDATA javatools viewer. Left upper: structure and shift list entered. Right: automatic assignment done and checked. Left lower: the NMReDATA file exported. The list of shifts has been shortened

NMReDATA J_reader, a Hypertext Markup Language (HTML)-based tool.[11] It is multiplatform and can be operated either online or off-line. All the contents in the file are exposed in a structured view, and their information is presented according to their format (Figure 4). The molecular structure is presented in an interactive 3-D display onto which chemical shifts, assignments, and couplings may be overlapped. JSmol, the JavaScript variant of Jmol,[12] is used for display of the structure as well as for data and file operations. JSpecView is used for display of spectra if they use the JCAMP format. Apart from its function as a viewer, NMReDATA J_reader can also be used to edit the NMReDATA tags composing the file. A special tool is included for adding implicit hydrogens

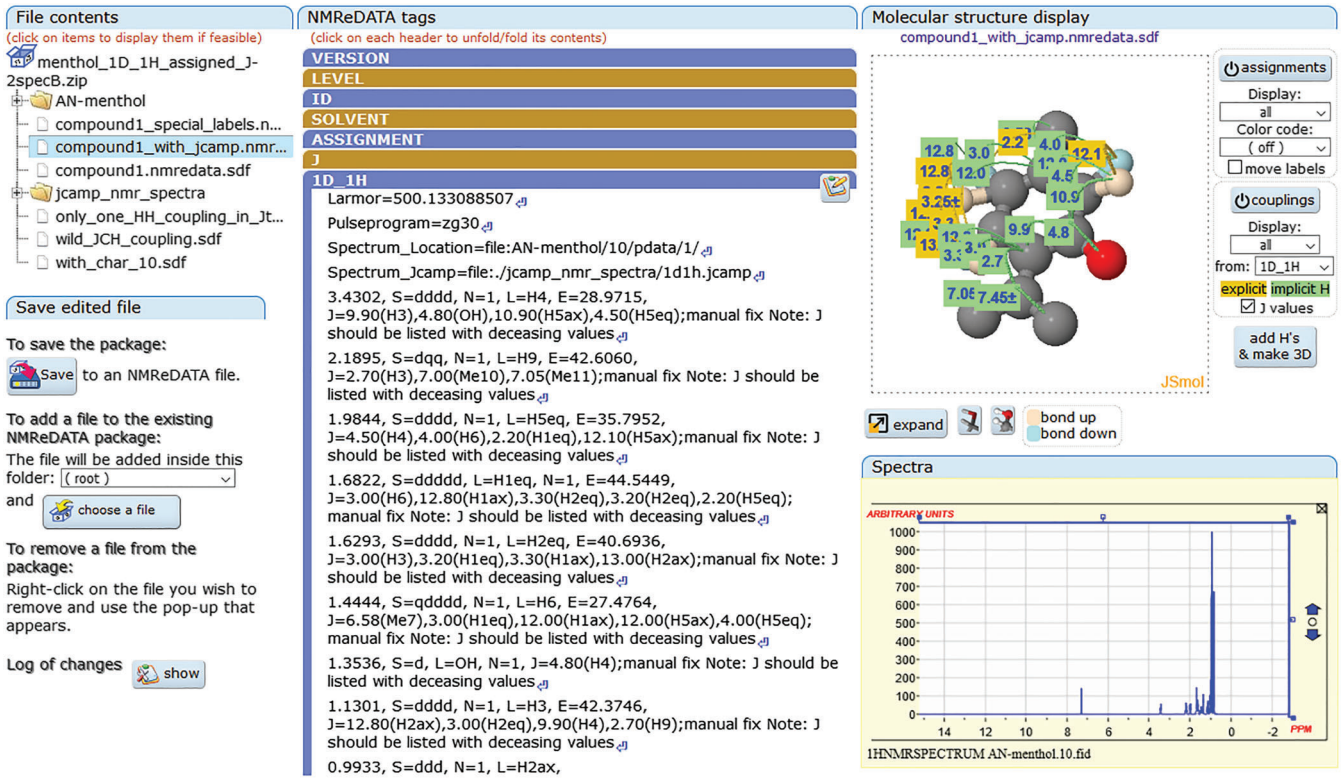


FIGURE 4 The nuclear magnetic resonance extracted data (NMReDATA) J_reader interface. Left panel: contents of the NMReDATA file (top) and package editing tools (bottom). Middle panel: contents of each selected file in the package. Top right: display of the structure, shifts, assignments, and couplings. Bottom right: display of spectra

and generating a 3-D structure that may be appended to the original 2-D structure. All will be saved back in the NMReDATA format, including a change log.

Finally, the website https://nmredata.com/ offers an online composer and viewer for NMR records. It allows import of raw data files and a structure and offers interactive peak picking and assignment.

In any case, the resulting NMReDATA file can be used for submission to journals or repositories where it can be validated in a two-step process described below.

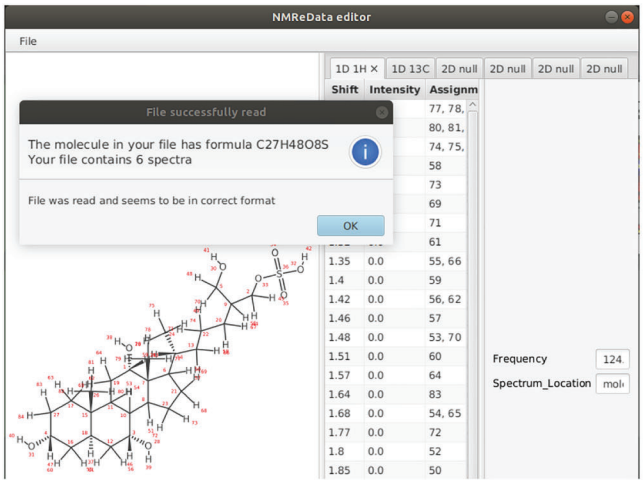
2.2 | Validation

2.2.1 | Formal validation

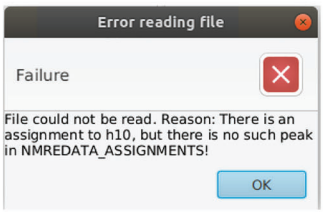
For a formal validation of an NMReDATA file, two tools were developed. One is a Java-based software, called *NMReDATA javatools*, which is available as a library and also as a stand-alone Java program from https://github.com/NMReDATAInitiative/javatools. Figure 5 shows the NMReDATA file for 5α-cyprinol sulfate opened with the stand-alone version of the javatools. The second tool is a JavaScript-based software available from https://github.com/cheminfo/nmredata. Both apply a syntactical validation of the file, ensuring that all required elements are contained and that the format of the tags is correct. The javatools also do some basic logic checking, for example, whether atoms used in the assignment exist in the structure. There is no check for the chemical validity, for example, whether the structure is compatible with the given NMR data.

2.2.2 | Validation of chemical shifts

The module for the validation of chemical shifts is conducted in collaboration with nmrshiftdb2[13] project. For the current example, the "Quick Check" module of the nmrshiftdb2 was used to verify the chemical shift lists and their assignments against the corresponding information calculated by nmrshiftdb2. The "Quick Check" module is available online on the "QuickCheck" tab of the https://www.nmrshiftdb.org website. This module accepts an NMReDATA file as its input and generates a validation report as shown in Figure 6. Of course, a validation report is also directly generated along with the NMReDATA file when shifts are entered manually as mentioned in Section 2.2.1. The report for each ^{13}C and ^{1}H shift gives a predicted value and calculates how close this is to the shift in the file. An overall quality score is generated from the



(a) A screenshot of the data of 5α-Cyprinol opened in with NMReDATA javatools. The file has found to be syntactically correct. The software displays an overview of the file content.



(b) A typical error message for an invalid file.

FIGURE 5 Outputs of the nuclear magnetic resonance extracted data (NMReDATA) javatools viewer

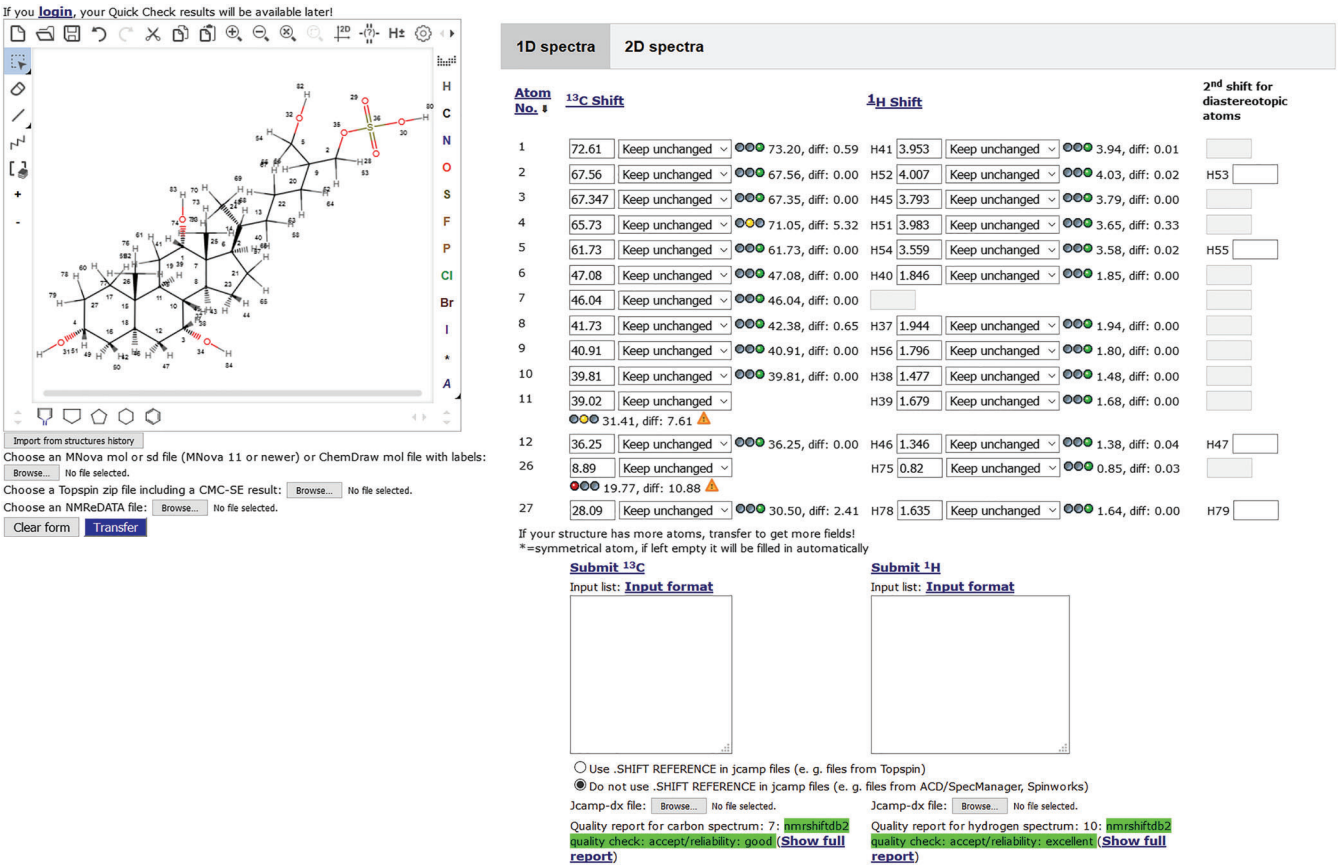


FIGURE 6 Key elements of the display of an evaluation of the nuclear magnetic resonance (NMR) assignment of 5α-cyprinol sulfate from Hahn et al.[3] in nmrshiftdb2. The list of shifts has been shortened

chemical shift deviations. In the example shown in Figure 6, there are two shifts for which nmrshiftdb2 identifies a larger deviation. On the other hand, the predicted shift is not fully reliable here (indicated by the orange triangles), so these have a low weight. The overall assignment is considered to be acceptable for ^{13}C and ^{1}H, giving the confidence that the suggested structure is correct.

2.2.3 | Validation of 2-D spectra correlations

Further verification of the results can be conducted by using the logic for structure determination (LSD) software[14] to compare the archived molecular structure in an NMReDATA file to those suggested by LSD. The NMReDATA editor, included in NMReDATA javatools, allows exporting an NMReDATA file in the LSD input format. Because LSD requires 1-D ^{1}H and ^{13}C as well as 2-D COSY, heteronuclear single quantum coherence spectroscopy (HSQC), and HMBC spectra, those must be contained in the NMReDATA file. It will then list all possible structures compatible with these spectra. Figure 7 shows the three structures LSD suggests as a fitting solution for the measured spectra of the example compound. The structures are very similar, differing only in the –OH group positions, with the middle one being the correct structure. This shows that the suggested structure is a good fit. If desired, the structures suggested by LSD can be ranked using the nmrshiftdb2 prediction. Details regarding the use of LSD can be found in the tutorial of Nuzillard and Plainchont.[15]

An alternative validation approach is implemented in CMC-se. The NMReDATA file may be imported and the built-in structure verification procedure executed. The coupling path length related to all available correlations is assessed, and the experimental ^{13}C chemical shifts are compared with the predicted ones. The verification protocol documents all correlations matching the standard coupling path length (e.g., ^{2}J and ^{3}J HMBC), and the optional long-range correlations are highlighted in a separate view. For the correlations, where the assignment is not unique, the shortest through-bond path is selected. Figure 8 shows an example. Additional interactive features are available if the spectra are available. The imported correlations are projected on the spectra. This allows for a detailed inspection or for a possible improvement of the NMReDATA record.

2.3 | Publishing and deposition

Because NMReDATA are data format, it cannot provide by itself a full solution for the problem of NMR data handling. For this, it needs to be integrated with repositories, databases, and search interfaces. We here sketch an ideal data deposition workflow to enable a full Findable, Accessible, Interoperable, and Reuseable (FAIR)-compliant data handling (see Section 3).

A requirement for deposition is that the proposed molecule and assignments pass all described validations. The data, together with the reports and the original NMReDATA file, will be saved in a repository and proposed for review. The selection of a suitable repository is left to the data producer. It could be a repository managed by a publisher, an institutional repository, or a third-party database. Besides data integrity, persistent unique identifiers (e.g., DOIs), versioning and query facilities for the data sets then improve findability and accessibility.

Nmrshiftdb2 is an example of a database accepting NMReDATA uploads. The data for 5α-cyprinol sulfate have been uploaded to nmrshiftdb2 and are available at

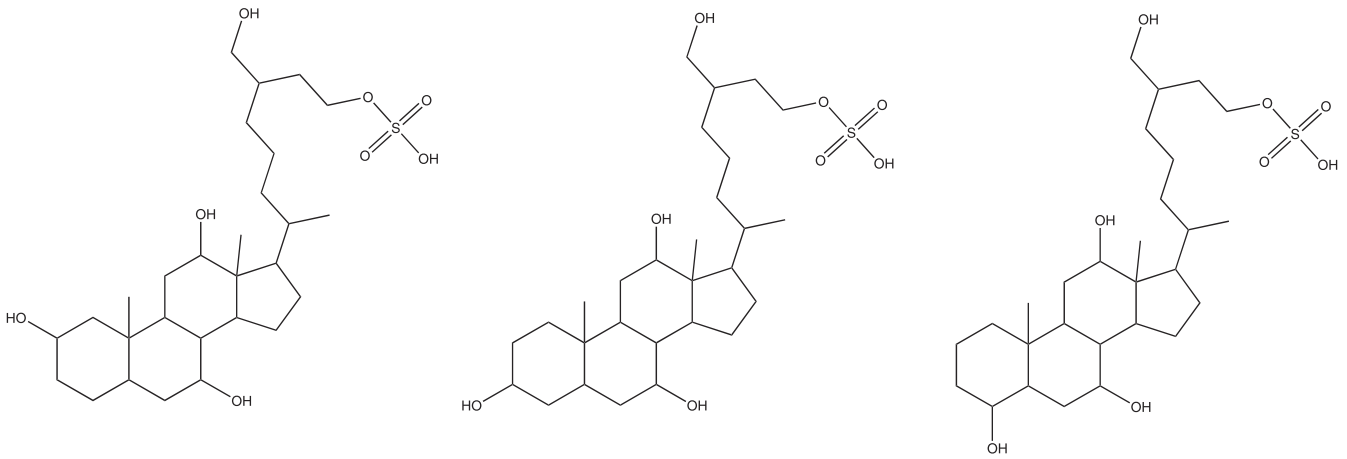


FIGURE 7 The three candidate structures generated by logic for structure determination (LSD) for the example data. The structure in the center is the correct one for 5α-cyprinol

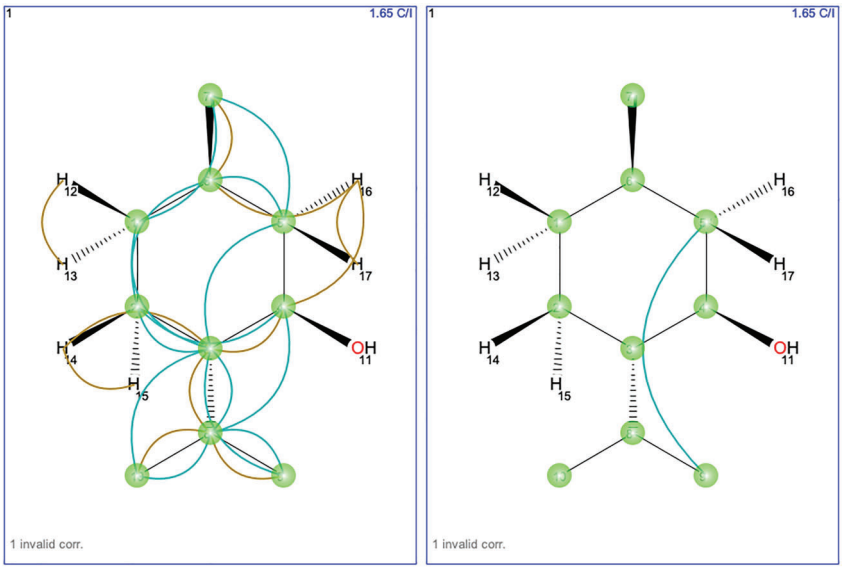
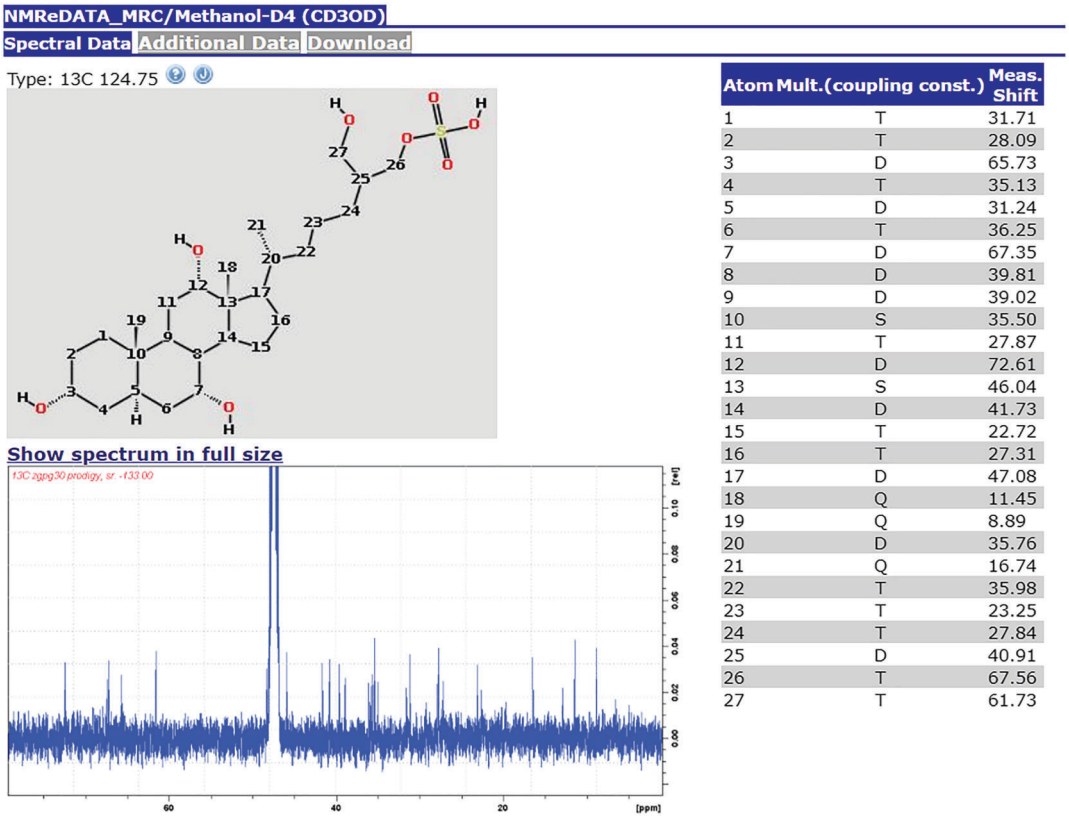


FIGURE 8 Nuclear magnetic resonance extracted data (NMReDATA) record verification in CMC-se. All available standard and long-range correlations are displayed. The difference between experimental and predicted ^{13}C chemical shifts is color-coded



Atom	Mult.	(coupling const.)	Meas. Shift
1	T		31.71
2	T		28.09
3	D		65.73
4	T		35.13
5	D		31.24
6	T		36.25
7	D		67.35
8	D		39.81
9	D		39.02
10	S		35.50
11	T		27.87
12	D		72.61
13	S		46.04
14	D		41.73
15	T		22.72
16	T		27.31
17	D		47.08
18	Q		11.45
19	Q		8.89
20	D		35.76
21	Q		16.74
22	T		35.98
23	T		23.25
24	T		27.84
25	D		40.91
26	T		67.56
27	T		61.73

FIGURE 9 The final deposition of 5α-cyprinol sulfate from Hahn et al.[3] with nmrshiftdb2. Further spectra are found by scrolling down and not shown here

https://nmrshiftdb.nmr.uni-koeln.de/molecule/60004113/dataset/MRC_Methanol-D4%2B%28CD3OD%29. Figure 9 shows the deposited data in nmrshiftdb2. The raw data for each spectrum are available on the "Download" tab.

Ideally, the submission of a spectral assignment article and of its associated data will be a seamless process. Authors will submit their spectral assignment article, together with their raw data or NMReDATA files. During peer review, both editors and reviewers can verify the data consistency by validating the assignment by themselves or to inspect existing reports. This validation step will help referees and editors to ascertain the assignment accuracy and likelihood of the submitted spectra.

Overall, the format, in conjunction with appropriate repositories, enables a full handling of NMR data from measurement to deposition and revision. In this respect, it forms the backbone of a FAIR-compliant NMR workflow.

3 | NMREDATA AND FAIR PRINCIPLES

The FAIR initiative[16] provides best practice guidelines to make data Findable, Accessible, Interoperable, and Reuseable (FAIR). In order to achieve an acceptable degree of Data Fairness, we discuss how NMReDATA support the FAIR principles along the published FAIR metrics criteria.[17] We demonstrate that the format ensures that data, if available as NMReDATA files, cover some of the metrics and that together with appropriate data repositories, a complete coverage can be achieved:

FM-F1A-identifier uniqueness and FM-F1B-identifier persistence: NMReDATA is a data format, it does not deal with these issues. Identifiers would be provided by repositories (e.g., nmrshiftdb2 IDs), which would also take care of persistency and versioning.

FM-F2-machine readable metadata: Although the format is not specified in an explicit knowledge representation (KR) language, its mol file inspired text format is semi-formal as parsers can read and write it, for example, for format conversions by means of parameter mapping tables. We have decided to base NMReDATA on an existing format to make adaption easier by use of existing tools (e.g., any molecular structure editor should be able to open an NMReDATA file and display the structure). This advantage outweighs that of an explicit KR language but will consider an Extensible Markup Language (XML) or linked-data serialization for the future. As our format leverages on the open nmrML raw data standard (XML with ontology support), this data section comes readily FM-F2 compliant.

FM-F3-resource identifier in metadata: The NMREDATA_ID tag allows inclusion of IDs generated by repositories in the metadata of a file.

FM-F4-indexed in searchable resource: This goal is achieved by the interplay of NMReData and repositories. Search functions are provided by the repositories (e.g., nmrshiftdb2 allows search by structure, spectrum, author, solvent, etc.).

FM-A1.1-access protocol, FM-A1.2-access authorization, and FM-A2-metadata longevity: These issues are mainly dealt with by the repositories. NMReDATA provide an important aspect of longevity, namely, a defined, vendor-independent format. Full metadata longevity has yet to be proven, but the community is building a rigid sustainability plan, which will contribute to NMReDATA metadata longevity. Longevity for the standard ensures, in turn, longevity for the data using the standard. Submission as NMR processed data standard to FAIRsharing is under discussion.

FM-I1-use a knowledge representation language: See FM-F2.

FM-I2-use FAIR vocabularies: Common terminology of the field has been used, and the format is published. The standard itself now has an EDAM term ID,[18] which can be found at https://edamontology.org/format_3824. Alignment with the nmrML controlled vocabulary (https://nmrml.org/cv/) is a task for a future release.

FM-I3-use qualified references: References are not used extensively in NMReDATA, but within an NMR record, there are links to the raw data in the NMReDATA file. Those links are fully qualified because they are specifically for raw data.

FM-R1.1-accessible usage license: NMReDATA files can carry any license, which is specified in NMReDATA_LICENCE. By default, the license is CC-BY to encourage data sharing. Other licenses, including closed licenses, are acceptable to enable adoption of the format. Due to having a default license, the user can always determine which license applies.

FM-R1.2-detailed provenance: For the standard, this is handled by having a clear versioning system for NMReDATA (currently, versions 1.0, 1.1, and 2.0 have been defined). For data using the standard, this is handled by repositories and outside the scope of the format.

FM-R1.3-meets community standards: NMReData was developed by practitioners and according to representative use cases in order to assure compliance with the NMR user communities requirements. We aligned our efforts with existing standardization bodies, that is, via developers from the Metabolomics Standards Initiative (MSI), who sanctioned the nmrML standard.

In summary, it is clear that NMReDATA, being a data format, cannot provide a full data management solution complying with FAIR principles. It lays the foundations, mainly in the area of interoperability, standards, and tool support. In conjunction with data repositories, full FAIR compliance can be achieved.

As part of the open development of the format, we provide tools under open-source licenses. The code of the javatools including the parsing and writing library and the Javascript library are available under open-source licenses.

4 | EXAMPLE AND TEST DATA

In order to enable testing of tools and to exemplify the format in practice, we have created a repository of

NMReDATA files at https://github.com/NMReDATA Initiative/Examples-of-NMR-records. This repository contains various examples, which cover a wide range of use cases. It comes in conjunction with the NMReDATA javatools, which can be used to check all NMReDATA files in the repository for their compliance to the standard. Any additional file can be checked as well.

4.1 | Use of NMReDATA javatools for checking compliance

The NMReDATA javatools contain a class `de.uni-koeln.chemie.nmr.ui.cl.CheckFormat`, which recursively checks a directory for any NMReDATA files and parses them. This directory can be a checkout of the sample data or any data by a user. By doing so, any syntactic problem in the files will be uncovered. Furthermore, the tool performs some semantic checks as well. For example, it will detect if there are labels used in the spectra that are not in NMREDATA_ASSIGNMENT, or it will complain if an atom number is used in NMREDATA_ASSIGNMENT, which is not in the structure. On the other hand, it does not check if the shifts match the structure (the tools in Section 2.2 would do so, though). This check can be used to validate future implementations of NMReDATA for compliance with the standard. If files produced by another tool can be read by the NMReDATA javatools, they can be assumed to be compliant with at least the basic requirements of the format.

This general parsing and testing can be supplemented by tests for individual files. This is achieved by adding a JUnit test case file to the directory where the NMReDATA file is located, with the same file name as the class but a different file extension. For some of the sample data, these test files can be found, as shown in Figure 10. For example, Asunaprevir.java contains specific tests for Asunaprevir.nmredata.sdf data set. The test method is as follows:

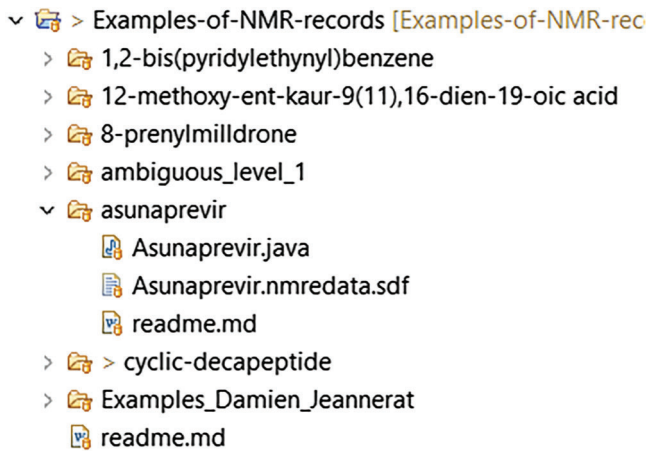


FIGURE 10 An nuclear magnetic resonance extracted data (NMReDATA) file and the associated test cases in the NMReDATA sample data directory. The file Asunaprevir.nmredata.sdf is accompanied by the test file Asunaprevir.java and a readme file that describes the scope of the example

It tests for specific number of atoms, bonds, spectra, and couplings. It then tests that the first coupling has a coupling constant of 11.8 Hz and that the atoms it refers to are both the first atom in the molecule. This coupling is `H1a, H1b, 11.8` in the NMReDATA file, whose connectivity table (CTAB) part does not contain explicit hydrogens. The NMReDATA reader does not add these, which is a deliberate decision, independent of the NMReDATA format design. The coupling, which is the geminal one between the hydrogen atoms attached to the first carbon, is assigned twice to the same carbon. These tests may seem trivial, but writing such ones has become a standard practice in software development to immediately identify problems when introducing new options or refactoring code.

4.2 | The sample data sets

The NMReDATA sample project directory contains samples of NMReDATA files and NMR records. The structure is shown in Figure 10. There are README.md or

```
public void test() {
    Assert.assertEquals(51, data.getMolecule().getAtomCount());
    Assert.assertEquals(55, data.getMolecule().getBondCount());
    Assert.assertEquals(8, data.getSpectra().size());
    Assert.assertEquals(6, couplings.size());
    Assert.assertEquals(11.8, couplings.get(0).getConstant());
    Assert.assertEquals(0, ((AtomReference)couplings.get(0)
    .getAssignments1()[0]).getAtomNumber());
    Assert.assertEquals(0, ((AtomReference)couplings.get(0)
    .getAssignments2()[0]).getAtomNumber());
    }
```

readme.txt files in the directories explaining the key issues with the files. Some areas covered are as follows:

- Different data sources/generators: Asunaprevir,[19] 1,2-bis(pyridylethynyl)benzene,[20]cyclic-decapeptide,[21] 8-prenylmilldrone,[22] and 12-methoxy-ent-kaur-9(11), 16-dien-19-oic acid[23] have been created using the export from MNova, whereas examples in ambiguous_level_1 have been exported from nmrshiftdb2, using the NMReDATA tools. This tests the compatibility of conversion outcomes.
- NMReDATA levels: Most files are NMREDATA_LEVEL 0, but in level_1, there are examples for ambiguous assignments. These are taken from the nmrshiftdb2 database. In line with many other repositories, nmrshiftdb2 can only hold unambiguous assignments, and text provides a hint that other assignments are possible. In contrast, the NMReDATA can hold it in a defined format. The files were manually edited to include the ambiguous assignment. The NMReDATA tools only read one assignment, which is checked in the java test files. A better handling of such assignments in processing software is encouraged by the NMReDATA project but not enforced.
- Explicit hydrogens: Asunaprevir, 1,2-bis(pyridylethynyl) benzene, cyclic-decapeptide, and 8-prenylmilldrone do not have explicit hydrogen atoms. Therefore, assignments of hydrogens are reported to the respective heavy atoms. In case of diastereotopic hydrogens, there are two shifts with different labels, but both assigned to the same atom. In contrast, the files in level_1 contain explicit hydrogen atoms and assignments to those hydrogens.
- Couplings and multiplicities: For 1-D spectra, additional information to chemical shifts can be given. For example, for 8-prenylmilldrone, multiplicities and integrals are given in the 1H spectrum, where shifts look like `7.5740, S=s, L=H3, E=34.8605`. Coupling constants are given for example in the line `H1a, H1b, 11.8` where NMREDATA_J indicates a coupling constant of 11.8 Hz between the atoms attached to the first atom.
- 2-D spectra: 2-D spectra of different types can be specified alongside the 1-D spectra, referring to the same set of shifts. For example, for 8-prenylmilldrone, a TOCSY spectrum is defined by the following:

```
> <NMREDATA_2D_1H_TJ_1H>
Larmor=799.873759389\
CorrType=TOCSY\
Pulseprogram=dipsi2gpphzs ;optional in V1\
Spectrum_Location=file:TSE_28F/14/pdata/1/\
zip_file_Location=https://www.dropbox.com/sh/ma8v25g15wylfj\
H17/H16\
```

The spectrum is defined as involving 1H resonances in the direct and indirect dimensions, with mixing over multiple bonds (TJ stands for total correlation spectroscopy through J couplings). After some additional attributes, the peaks are listed, the first being the one between H17 and H16, the reference of which are defined by the NMREDATA_ASSIGNMENT tag.

5 | CONCLUSION

We have shown how the NMReDATA format streamlines the process of NMR processing, data handling, verification, and archiving of the results. We also showed how the NMReDATA facilitate the fulfillment of the FAIR principles and, together with appropriate repositories and journal publication policies, ultimately contribute to a fully FAIR compliant NMR data handling process in the future. The NMReDATA format is readable for both humans and machines. This ensures that the format can be widely used, even if appropriate software is lacking, and will always be readable.

Apart from firmly establishing the format in the community, we plan to have a serialization of NMReDATA as linked data (for example, XML or Resource Description Framework [RDF]). NMReDATA also form the core of a wider initiative for chemical data, called CHEMeDATA.[24]

ACKNOWLEDGEMENTS

DS work was financed by Phenomenal (H2020 654241) at the initiation phase of this effort, current work in kind contribution. HD is supported by National Heart Lung and Blood Institute grant T32 HL007575. This project made use of the NMR Uppsala infrastructure, which is funded by the Department of Chemistry - BMC and the Disciplinary Domain of Medicine and Pharmacy. NES gratefully acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, IDNMR project, grants SCHL 580/3-1 and SCHL 580/3-2).

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/mrc.5146.

ORCID

Stefan Kuhn https://orcid.org/0000-0002-5990-4157
Lianne H. E. Wieske https://orcid.org/0000-0003-4617-7605
Paul Trevorrow https://orcid.org/0000-0002-1956-4135
Daniel Schober https://orcid.org/0000-0001-8014-6648
Nils E. Schlörer https://orcid.org/0000-0002-0990-9582
Jean-Marc Nuzillard https://orcid.org/0000-0002-5120-2556
Pavel Kessler https://orcid.org/0000-0003-0566-0545
Jochen Junker https://orcid.org/0000-0003-4002-4405
Angel Herráez https://orcid.org/0000-0002-9900-6845
Christophe Farès https://orcid.org/0000-0001-6709-5057
Mate Erdélyi https://orcid.org/0000-0003-0359-5970
Damien Jeannerat https://orcid.org/0000-0001-7018-4288

REFERENCES

[1] M. Pupier, J.-M. Nuzillard, J. Wist, N. E. Schlörer, S. Kuhn, M. Erdélyi, C. Steinbeck, A. J. Williams, C. Butts, T. D. W. Claridge, B. Mikhova, W. Robien, H. Dashti, H. R. Eghbalnia, C. Fars, C. Adam, P. Kessler, F. Moriaud, M. Elyashberg, D. Argyropoulos, M. Prez, P. Giraudeau, R. R. Gil, P. Trevorrow, D. Jeannerat, *Magn. Reson. Chem.* **2018**, *56*(8), 703. https://doi.org/10.1002/mrc.4737

[2] 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. Günther, 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*(1), 649. https://www.mcponline.org/content/10/1/R110.000133

[3] M. Hahn, E. von Elert, L. Bigler, M. D. Díaz Hernández, N. E. Schlörer, *Magn. Reson. Chem.* **2018**, *56*(12), 1201. https://doi.org/10.1002/mrc.4782

[4] P. Kessler, M. Godejohann, *Magn. Reson. Chem.* **2018**, *56*(6), 480. https://doi.org/10.1002/mrc.4712

[5] Mnova - mestrelab, http://mestrelab.com/software/mnova/, **2020**.

[6] Macros & scripts for acd/labs software and solutions, https://www.acdlabs.com/resources/knowledgebase/macros/index.php, **2020**.

[7] H. Dashti, W. M. Westler, M. Tonelli, J. R. Wedell, J. L. Markley, H. R. Eghbalnia, *Anal. Chem.* **2017**, *89*(22), 12201. https://doi.org/10.1021/acs.analchem.7b02884

[8] F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfeifer, A. Bax, *J. Biomol. NMR* **1995**, *6*(3), 277. https://doi.org/10.1007/BF00197809

[9] M. Norris, B. Fetler, J. Marchant, B. A. Johnson, *J. Biomol. NMR* **2016**, *65*(3-4), 205. https://doi.org/10.1007/s10858-016-0049-6

[10] C. Steinbeck, S. Krause, S. Kuhn, *J. Chem. Inf. Comput. Sci.* **2003**, *43*(6), 1733. https://doi.org/10.1021/ci0341363

[11] A. Herráez, Nmredata j_reader: an html interface for displaying the contents of nmredata files, molecular structure, nmr data and spectra, http://www3.uah.es/nmr_e_data/, **2020**.

[12] Jmol: an open-source java viewer for chemical structures in 3d, http://jmol.sourceforge.net/, **2020**.

[13] S. Kuhn, N. E. Schlörer, *Magn. Reson. Chem.* **2015**, *53*(8), 582. https://doi.org/10.1002/mrc.4263

[14] B. Plainchont, V. de Paulo Emerenciano, J.-M. Nuzillard, *Magn. Reson. Chem.* **2013**, *51*(8), 447. https://doi.org/10.1002/mrc.3908

[15] J.-M. Nuzillard, B. Plainchont, *Magn. Reson. Chem.* **2018**, *56*(6), 458. https://doi.org/10.1002/mrc.4612

[16] Fair principles, https://www.go-fair.org/fair-principles/, **2020**.

[17] M. D. Wilkinson, S. A. Sansone, E. Schultes, P. Doorn, L. O. Bonino da Silva Santos, M. Dumontier, *Sci. Data* **2018**, *5*, 180118. https://doi.org/10.1038/sdata.2018.118

[18] Edam ontology, http://edamontology.org/EDAM.owl, **2020**.

[19] C. Reviriego, *Drugs Future* **2012**, *37*, 247. https://doi.org/10.1358/dof.2012.37.4.1789350

[20] S. Lindblad, K. Mehmeti, A. X. Veiga, B. Nekoueishahraki, J. Grafenstein, M. Erdélyi, *J. Am. Chem. Soc.* **2018**, *140*, 135037. https://doi.org/10.1021/jacs.8b09467

[21] E. Danelius, H. Andersson, P. Jarvoll, K. Lood, J. Grafenstein, M. Erdélyi, *Biochem.* **2017**, *56*, 3265. https://doi.org/10.1021/acs.biochem.7b00429

[22] T. Deyou, M. Makungu, M. Heydenreich, F. Pan, A. Gruhonjic, P. Fitzpatrick, A. S. Koch, S. Derese, J. Pelletier, K. Rissanen, A. Yenesew, M. Erdélyi, *J. Nat. Prod.* **2017**, *80*, 2060. https://doi.org/10.1021/acs.jnatprod.7b00255

[23] S. Yaouba, A. Valkonen, P. Coghi, J. Gao, E. M. Guantai, S. Derese, V. C. K. W. Wong, M. Erdélyi, A. Yenesew, *Mol.* **2018**, *23*, 3199. https://doi.org/10.3390/molecules23123199

[24] Chemedata initiative, https://chemedata.github.io/, **2020**.

How to cite this article: Kuhn S, Wieske LHE, Trevorrow P, et al. NMReDATA: Tools and applications. *Magn Reson Chem*. 2021;1–12. https://doi.org/10.1002/mrc.5146