



Universidad
de Alcalá

BIBLIOTECA

Document downloaded from the institutional repository
of the University of Alcalá:

<https://ebuah.uah.es/dspace/>

This is a postprint version of the following published
document:

Kuhn, S. et al. (2021) 'NMReDATA: Tools and
applications', *Magnetic Resonance in Chemistry*, 2021, v.
59, n. 8, p. 792-803.

Available at <https://doi.org/10.1002/mrc.5146>

© 2021 Wiley

(Article begins on next page)



This work is licensed under a

Creative Commons Attribution-NonCommercial-
NoDerivatives
4.0 International License.

NMReDATA: Tools and applications

Stefan Kuhn¹ | Lianne H. E. Wieske² | Paul Trevorrow³ |
Daniel Schober⁴ | Nils E. Schlörer⁵ | Jean-Marc Nuzillard⁶ |
Pavel Kessler¹⁰ | Jochen Junker⁷ | Angel Herráez¹¹ |
Christophe Farés⁸ | Mate Erdelyi² | Damien Jeannerat⁹

¹School of Computer Science and Informatics, De Montfort University, The Gateway, Leicester LE1 9BH, UK

⁸Max-Planck-Institut für Kohlenforschung, Abteilung NMR, Kaiser-Wilhelm-Platz 1, 45470 Mülheim an der Ruhr, Germany

³Wiley, The Atrium, Chichester PO19 8SQ, UK

⁵Department of Chemistry, University of Cologne, Greinstr. 4, 50939 Köln, Germany

⁹NMRprocess, Switzerland

²Department of Chemistry - BMC, Uppsala Universitet, Husargatan 3, 752 37 Uppsala, Sweden

⁴MatterWaveSemantics, Ontology Development, Südharz, Germany, and Leibniz Institute of Plant Biochemistry, Stress and Developmental Biology, Weinberg 3, 06120 Halle (Saale), Germany

⁶Université de Reims Champagne Ardenne, CNRS, ICMR UMR 7312, 51097 Reims, France

⁷Fundação Oswaldo Cruz - CDTS, Rio de Janeiro - RJ, Brazil

¹⁰Bruker BioSpin GmbH, Silberstreifen, 76287 Rheinstetten, Germany

¹¹Department of Systems Biology, Universidad de Alcalá. Alcalá de Henares, Spain

The NMReDATA format has been proposed as a way to store, exchange, and to disseminate 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

Nuclear Magnetic Resonance (NMR), NMReDATA, chemical information, data standard, peak assignment

This version is the accepted version of the manuscript, after undergoing full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as it was published: *Magnetic Resonance in Chemistry* 2021;1–12. <https://doi.org/10.1002/mrc.5146>

1 | INTRODUCTION

The NMReDATA format [1] was introduced recently for reporting and exchanging 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 PDF documents. The NMReDATA was 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 supports 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, 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

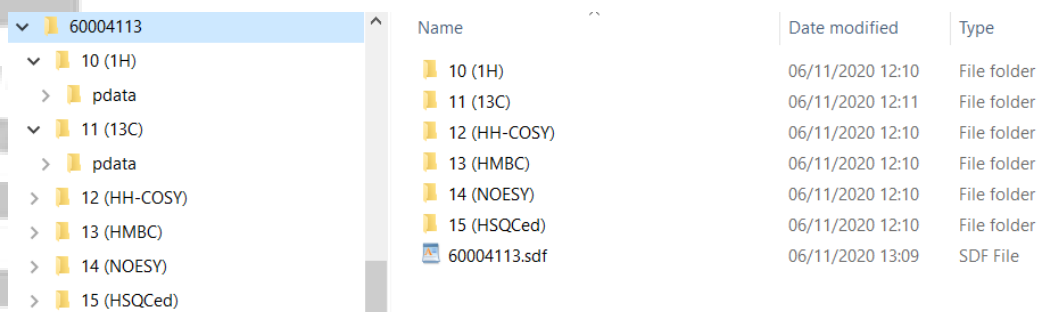
In this section we show how the NMReDATA format improves the structure verification workflow for NMR-based investigations of small molecules. The outcome of this workflow can ultimately be used for direct deposition of the resulting standardised data associated with a scientific journal article as well as registration and deposition of the data to relevant repositories. As example for the workflow we will use the published NMR data from 5 α -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 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/Spectrum 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, some are open source.

Name	Vendor	Language/ Operating system	Functions	URL	License/ Availability
Topsin/CMCse	Bruker	Windows, Linux, macOS	Export: NMRRecord+ NMRData Import/Visualize/ Validate: NMRData Export/Import: NMR Record/ NMRReDATA Export NMR Record /NMRReDATA Read/Export/	https://www.bruker.com/service/support-upgrades/software-downloads/nmr.html https://mestrelab.com/software/mnova/ https://www.acdlabs.com/products/adh/ spectrusprocessor/index.php	Commercial, academic licenses available Commercial Commercial
Mnova	MestreLab	Windows, Linux, macOS	NMR Record/ NMRReDATA Export NMR Record /NMRReDATA Read/Export/	https://mestrelab.com/software/mnova/ https://www.acdlabs.com/products/adh/ spectrusprocessor/index.php	Commercial Commercial
Spectrus Processor	ACD/Labs	Microsoft Windows 10	Record /NMRReDATA Read/Export/	spectrusprocessor/index.php	Commercial
nmrshiftdb2	nmrshiftdb2.org	Online	Validate NMRReDATA Read/Write/ Vizualize NMRReDATA Read/Export/	https://www.nmrshiftdb.org http://www3.uah.es/nmr_e_data/reader/reader.htm	Free online Free online
NMRReDATA_J_reader	Angel Herráez	Online	Vizualize NMRReDATA Read/Export/	http://www3.uah.es/nmr_e_data/reader/reader.htm	Free online
nmrdata.com	cheminfo.org	Online	NMR Record/ NMRReDATA Read/Write/ Vizualize NMRReDATA Export to LSD	http://nmrdata.com/ https://github.com/NMRReDATAInitiative/javatools	Free online Free online
NMRReDATA javatools	NMRReDATA.org	Java	Vizualize NMRReDATA Export to LSD	https://github.com/NMRReDATAInitiative/javatools	Free online

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



Name	Date modified	Type
10 (1H)	06/11/2020 12:10	File folder
11 (13C)	06/11/2020 12:11	File folder
12 (HH-COSY)	06/11/2020 12:10	File folder
13 (HMBC)	06/11/2020 12:10	File folder
14 (NOESY)	06/11/2020 12:10	File folder
15 (HSQCed)	06/11/2020 12:10	File folder
60004113.sdf	06/11/2020 13:09	SDF File

FIGURE 1 A screenshot of the NMR record of 5α -Cyprinol. The NMReDATA file (60004113.sdf) sits in the root directory of the record, the Bruker data are contained in their original form. Each of the directories 10 to 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.

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 1D and 2D NMR spectra acquired for the compound is reported in [3], comprising ^{13}C and ^1H 1D and ^1H - ^1H COSY, ^1H - ^{13}C HMBC, ^1H - ^1H 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 has been processed and the initial NMReDATA file has been created, the outcomes of numerous NMR post-processing 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 [8, 9, 5], 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 1D and 2D 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 NMReDATA J_reader, an HTML-based tool [11]. It is multiplatform and can be operated either online or offline. 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 3D 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

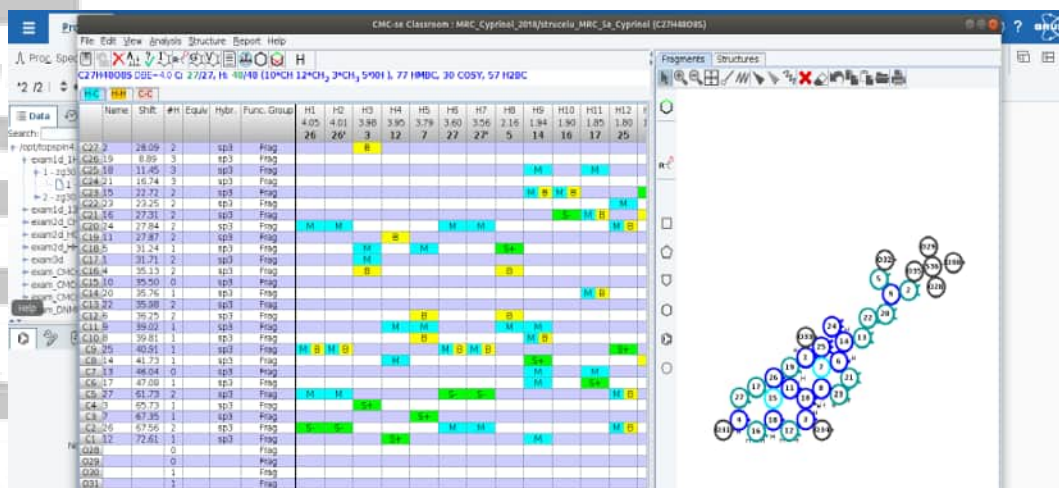


FIGURE 2 A screenshot of the data of 5 α -Cyprinol opened in Bruker CMC-se [4]. The data had been acquired using Bruker equipment, the data were processed using Topspin, then opened in CMC-se and saved as NMRDATA from there.

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

Finally, the website <http://nmrdata.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 NMRDATA 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 NMRDATA file, two tools were developed. One is a Java-based software, called *NMRDATA javatools*, which is available as a library and also as a standalone Java program from <https://github.com/NMRDATAInitiative/javatools>. Figure 5 shows the NMRDATA file for 5 α -Cyprinol sulfate opened with the standalone-version of the *javatools*. The second tool is a JavaScript-based software available from <https://github.com/cheminfo/nmrdata>. 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, e. g. whether atoms used in the assignment exist in the structure. There is no check for the chemical validity, e. g. 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



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

online on the “QuickCheck” tab of the www.nmrshiftdb.org website. This module accepts an NMRdata file as its input and generates a validation report as shown in Fig. 6. Of course, a validation report is also directly generated along with the NMRdata file when shifts are entered manually as mentioned in Section 2.2.1. The report for each ^{13}C and ^1H 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 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 ^1H , giving the confidence that the suggested structure is correct.

2.2.3 | Validation of 2D spectra correlations

Further verification of the results can be conducted by using the LSD (Logic for Structure Determination) software [14] to compare the archived molecular structure in an NMRdata file to those suggested by LSD. The NMRdata editor, included in NMRdata javatools, allows exporting an NMRdata file in the LSD input format. Since LSD requires 1D ^1H and ^{13}C as well as 2D COSY, HSQC and HMBC spectra, those must be contained in the NMRdata 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 $-\text{OH}$ 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 [15].

An alternative validation approach is implemented in CMC-se. The NMRdata file may be imported and the built-in structure verification procedure executed. The coupling path length related to all available correlations is assessed,

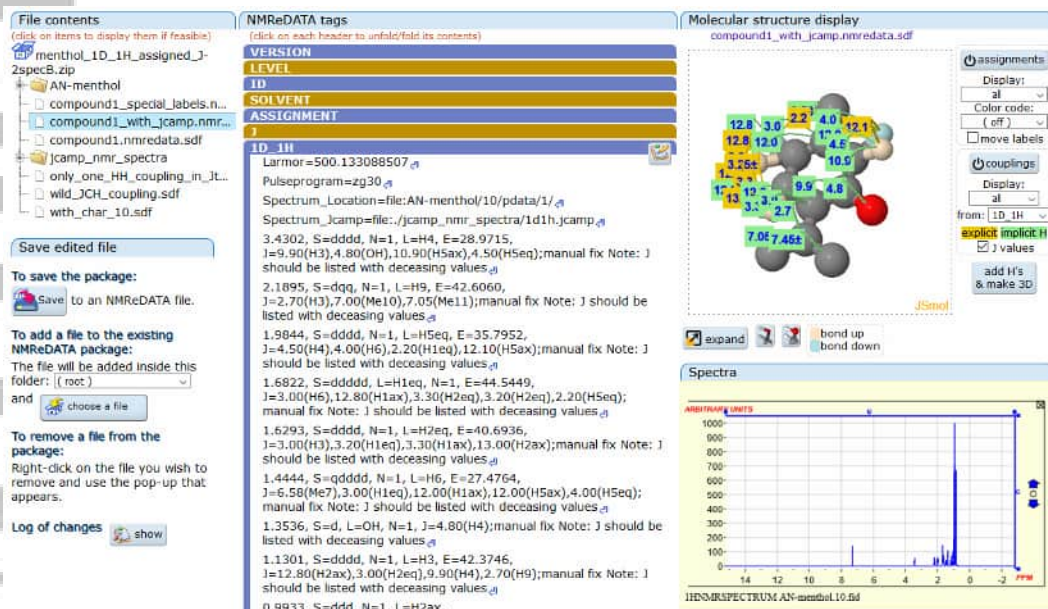


FIGURE 4 The NMRReDATA J_reader interface. Left panel: contents of the NMRReDATA file (top) and package editing tools (bottom). Middle panel: contents of each selected file in the package. Top right: display of the structure, shifts, assignments and couplings. Bottom right: Display of spectra.

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. 2J and 3J HMBC), 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 NMRReDATA record.

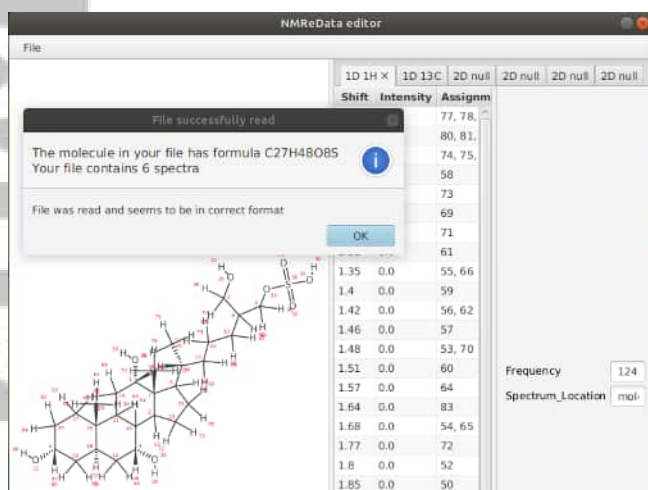
2.3 | Publishing and deposition

Since NMRReDATA is a 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 FAIR-compliant data handling (see 3).

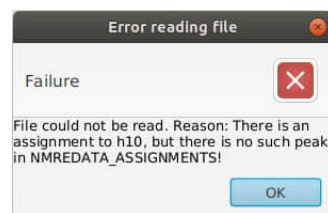
A requirement for deposition is that the proposed molecule and assignments pass all described validations. The data, together with the reports and the original NMRReDATA file, will be saved in a repository and proposed for review. Which repository is used will be decided by the data producer. It could 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 NMRReDATA uploads. The data for 5α -Cyprinol sulfate have been uploaded to nmrshiftdb2 and are available at http://nmrshiftdb.nmr.uni-koeln.de/molecule/60004113/dataset/MRC_Methano1-D4+%28CD3OD%29. 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



(a) A screenshot of the data of 5 α -Cyprinol opened in with NMRData 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 NMRData javatools viewer.

will submit their spectral assignment article, together with their raw data or NMRData 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 Reusable (FAIR). In order to achieve an acceptable degree of Data Fairness we discuss how NMRData supports the FAIR principles along the published FAIR metrics criteria [17]. We demonstrate that the format ensures that data, if available as NMRData 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: Since NMRData 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, e.g. for format conversions by means of parameter mapping tables. We have decided to base NMRData 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 NMRData file and display the structure). This advantage outweighs that of an explicit KR language, but will consider an XML or linked-data

If you [login](#), your Quick Check results will be available later!

1D spectra 2D spectra

Atom No. #	¹³ C Shift	¹ H Shift	² D shift for diastereotopic atoms
1	72.61 Keep unchanged	73.20, diff: 0.59 H41 3.953 Keep unchanged	3.94, diff: 0.01
2	67.56 Keep unchanged	67.56, diff: 0.00 H52 4.007 Keep unchanged	4.03, diff: 0.02
3	67.347 Keep unchanged	67.35, diff: 0.00 H45 3.793 Keep unchanged	3.79, diff: 0.00
4	65.73 Keep unchanged	71.05, diff: 5.32 H51 3.983 Keep unchanged	3.65, diff: 0.33
5	61.73 Keep unchanged	61.73, diff: 0.00 H54 3.559 Keep unchanged	3.58, diff: 0.02
6	47.08 Keep unchanged	47.08, diff: 0.00 H40 1.846 Keep unchanged	1.83, diff: 0.00
7	46.04 Keep unchanged	46.04, diff: 0.00	
8	41.73 Keep unchanged	42.38, diff: 0.65 H37 1.944 Keep unchanged	1.94, diff: 0.00
9	40.91 Keep unchanged	40.91, diff: 0.00 H56 1.796 Keep unchanged	1.80, diff: 0.00
10	39.81 Keep unchanged	39.81, diff: 0.00 H38 1.477 Keep unchanged	1.48, diff: 0.00
11	39.02 Keep unchanged	39.02, diff: 0.00 H39 1.679 Keep unchanged	1.68, diff: 0.00
12	36.25 Keep unchanged	36.25, diff: 0.00 H46 1.346 Keep unchanged	1.38, diff: 0.04
26	6.89 Keep unchanged	6.89, diff: 0.00 H75 0.82 Keep unchanged	0.85, diff: 0.03
27	28.09 Keep unchanged	30.50, diff: 2.41 H78 1.635 Keep unchanged	1.64, diff: 0.00

Submit ¹³C Input list: [Input format](#)

Submit ¹H Input list: [Input format](#)

Quality report for carbon spectrum: 7: [Show full report!](#) Quality report for hydrogen spectrum: 10: [Show full report!](#)

FIGURE 6 Key elements of the display of an evaluation of the NMR assignment of 5 α -Cyprinol sulfate from [3] in nmrshiftdb2. The list of shifts has been shortened.

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 provides 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 http://edamontology.org/format_3824. Alignment with the nmrML controlled vocabulary (<http://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 since they are specifically for raw data.

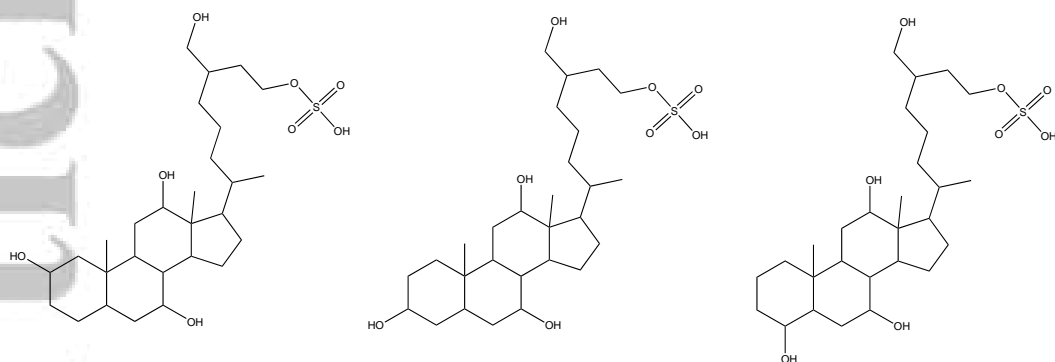


FIGURE 7 The three candidate structures generated by LSD for the example data. The structure in the centre is the correct one for 5 α -Cyprinol.

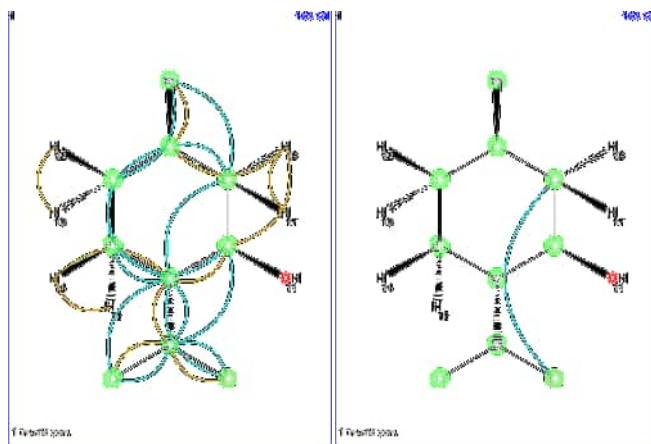


FIGURE 8 NMRData record verification in CMC-se. All available standard and long-range correlations are displayed. The difference between experimental and predicted ^{13}C chemical shifts is color-coded.

FM-R1.1-Accessible Usage License: NMRData files can carry any license, which is specified in NMRData_LICENSE. 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.

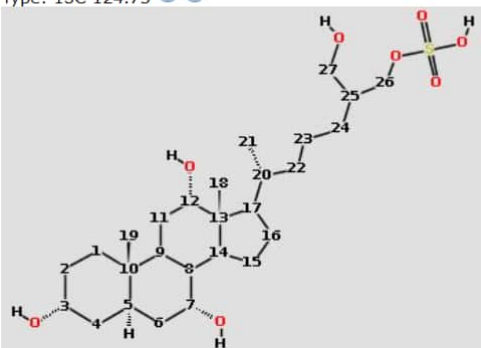
FM-R1.2-Detailed Provenance: For the standard, this is handled by having a clear versioning system for NMRData (currently versions 1.0, 1.1, and 2.0 have been defined). For data using the standard, this is handled by repositories and outside the scope of the format.

FM-R1.3-Meets Community Standards: NMRData 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, i.e. via developers from the Metabolomics Standards Initiative (MSI), <http://cosmos-fp7.eu/msi.html>, who sanctioned the nmrML standard.

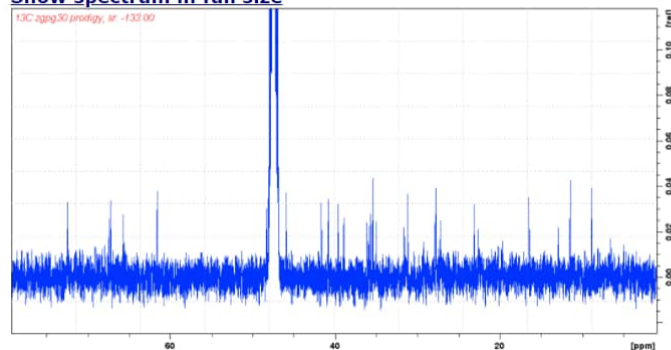
In summary, it is clear that NMRData, being a data format, cannot provide a full data management solution complying with FAIR principles. It lays the foundations, mainly in the area of interoperability, standards, and tool

NMRReDATA_MRC/Methanol-D4 (CD3OD)

Spectral Data Additional Data Download

Type: ¹³C 124.75

Show spectrum in full size



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 5 α -Cyprinol sulfate from [3] with nmshiftdb2. Further spectra are found by scrolling down and not shown here.

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

4.1 | Use of NMReDATA java tools for checking compliance

The NMReDATA java tools contain a class `de.unikoeIn.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 which 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 java tools, 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 Fig. 10. For example, `Asunaprevir.java` contains specific tests for `Asunaprevir.nmredata.sdf` data set. The test method is as follows:

```
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());
}
```

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 H_{1a} , H_{1b} , 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 `readme.txt` files in the directories explaining the key issues with the files. Some areas covered are:

- Different data sources/generators: `asunaprevir`, [19] 1,2-bis(pyridylethynyl)benzene, [20] cyclic-decapeptide, [21] 8-prenylmilloidrone [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

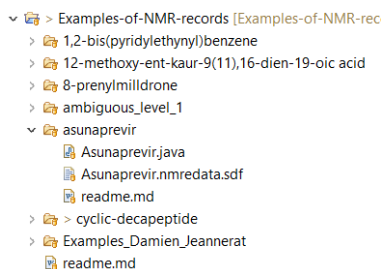


FIGURE 10 An NMReDATA file and the associated test cases in the NMReDATA sample data directory. The file `Asunaprevir.nmredata.sdf` is accompanied by the test file `Asunaprevir.java` and a readme file that describes the scope of the example.

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-prenylmiltidrone` 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 1D spectra, additional information to chemical shifts can be given. For example, for `8-prenylmiltidrone`, 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.
- 2D spectra: 2D spectra of different types can be specified alongside the 1D spectra, referring to the same set of shifts. For example, for `8-prenylmiltidrone` a TOCSY spectrum is defined by:

```
> <NMReDATA_2D_1H_TJ_1H>
Larmor=799.873759389\
CorrType=TOCSY\
Pulseprogram=dipsi2gpshzs ;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 facilitates the fulfillment of the FAIR principles, and together with appropriate repositories and journal publication policies, ultimately contributes 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 RDF). NMReDATA also forms the core of a wider initiative for chemical data, called CHEMeDATA [24].

ACKNOWLEDGMENT

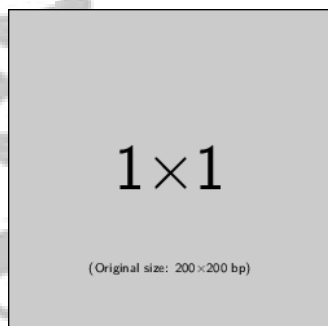
DS work was financed by Phenomenal (H2020 654241) at the initiation-phase of this effort, current work in kind contribution. HD is supported by National Heart Lung and Blood Institute grant T32 HL007575. This project made use of the NMR Uppsala infrastructure, which is funded by the Department of Chemistry – 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).

REFERENCES

- [1] Pupier M, Nuzillard JM, Wist J, Schlörner NE, Kuhn S, Erdelyi M, et al. NMReDATA, a standard to report the NMR assignment and parameters of organic compounds. *Magnetic Resonance in Chemistry* 2018;56(8):703–715. <https://onlinelibrary.wiley.com/doi/abs/10.1002/mrc.4737>.
- [2] Schober D, Jacob D, Wilson M, Cruz JA, Marcu A, Grant JR, et al. nmrML: A Community Supported Open Data Standard for the Description, Storage, and Exchange of NMR Data. *Analytical Chemistry* 2018;90(1):649–656. <https://www.mcponline.org/content/10/1/R110.000133>.
- [3] Hahn M, von Elert E, Bigler L, Díaz Hernández MD, Schloerer NE. 5 α -Cyprinol sulfate: Complete NMR assignment and revision of earlier published data, including the submission of a computer-readable assignment in NMReDATA format. *Magnetic Resonance in Chemistry* 2018;56(12):1201–1207. <https://onlinelibrary.wiley.com/doi/abs/10.1002/mrc.4782>.
- [4] Kessler P, Godejohann M. Identification of tentative marker in Corvina and Primitivo wines with CMC-se. *Magnetic Resonance in Chemistry* 2018;56(6):480–492. <https://onlinelibrary.wiley.com/doi/abs/10.1002/mrc.4712>.
- [5] Mnova - Mestrelab; Accessed: 2020-2-27. <http://mestrelab.com/software/mnova/>.
- [6] Macros & Scripts for ACD/Labs Software and Solutions; Accessed: 2020-10-11. <https://www.acdlabs.com/resources/knowledgebase/macros/index.php>.
- [7] Dashti H, Westler WM, Tonelli M, Wedell JR, Markley JL, Eghbalnia HR. Spin System Modeling of Nuclear Magnetic Resonance Spectra for Applications in Metabolomics and Small Molecule Screening. *Analytical Chemistry* 2017;89(22):12201–12208. <https://doi.org/10.1021/acs.analchem.7b02884>, PMID: 29058410.
- [8] Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *Journal of Biomolecular NMR* 1995;6(3):277–293. <https://link.springer.com/article/10.1007/BF00197809>.

- [9] Norris M, Fetler B, Marchant J, Johnson BA. NMRfX Processor: a cross-platform NMR data processing program. *Journal of Biomolecular NMR* 2016;65(3-4):205–216. <https://link.springer.com/article/10.1007/s10858-016-0049-6>.
- [10] Steinbeck C, Krause S, Kuhn S. NMRShiftDB - Constructing a Free Chemical Information System with Open-Source Components. *Journal of Chemical Information and Computer Sciences* 2003;43(6):1733–1739. <https://doi.org/10.1021/ci0341363>, PMID: 14632418.
- [11] Herraes A, NMReDATA J_reader: an HTML interface for displaying the contents of NMReDATA files, molecular structure, NMR data and spectra; Accessed: 2020-11-12. http://www3.uah.es/nmr_e_data/.
- [12] Jmol: an open-source Java viewer for chemical structures in 3D; Accessed: 2020-11-12. <http://jmol.sourceforge.net/>.
- [13] Kuhn S, Schlorer NE. Facilitating quality control for spectra assignments of small organic molecules: nmrshiftdb2—a free in-house NMR database with integrated LIMS for academic service laboratories. *Magnetic Resonance in Chemistry* 2015;53(8):582–589. <https://onlinelibrary.wiley.com/doi/full/10.1002/mrc.4263>.
- [14] Plainchont B, de Paulo Emerenciano V, Nuzillard JM. Recent advances in the structure elucidation of small organic molecules by the LSD software. *Magnetic Resonance in Chemistry* 2013;51(8):447–453.
- [15] Nuzillard JM, Plainchont B. Tutorial for the structure elucidation of small molecules by means of the LSD software. *Magnetic Resonance in Chemistry* 2018;56(6):458–468. <https://onlinelibrary.wiley.com/doi/abs/10.1002/mrc.4612>.
- [16] FAIR principles; Accessed: 2020-2-27. <https://www.go-fair.org/fair-principles/>.
- [17] Wilkinson MD, Sansone SA, Schultes E, Doorn P, Bonino da Silva Santos LO, Dumontier M. A design framework and exemplar metrics for FAIRness. *Scientific Data* 2018;5:180118.
- [18] EDAM Ontology; Accessed: 2020-9-27. <http://edamontology.org/EDAM.owl>.
- [19] Reviriego C. Asunaprevir. HCV serine protein NS3 inhibitor, Treatment of hepatitis C virus. *Drugs of the Future* 2012;37:247–254.
- [20] Lindblad S, Mehmeti K, Veiga AX, Nekoueshahraki B, Grafenstein J, Erdelyi M. Halogen Bond Asymmetry in Solution. *Journal of the American Chemical Society* 2018;140:135037–13513.
- [21] Danelius E, Andersson H, Jarvoll P, Lood K, Grafenstein J, Erdelyi M. Halogen bonding: a powerful tool for modulation of peptide conformation. *Biochemistry* 2017;56:3265–3272.
- [22] Deyou T, Makungu M, Heydenreich M, Pan F, Gruhonjic A, Fitzpatrick P, et al. Isofalvones and Rotenoides from the leaves of *Millettia oblata* ssp *teitensis*. *Journal of Natural Products* 2017;80:2060–2066.
- [23] Yaouba S, Valkonen A, Coghi P, Gao J, Guantai EM, Derese S, et al. Crystal structures and cytotoxicity of ent-Kaurane-Type Diterpenoids from Two *Aspilia* Species. *Molecules* 2018;23:3199–3272.
- [24] CHEMeDATA Initiative; Accessed: 2020-11-12. <https://chemedata.github.io/>.

GRAPHICAL ABSTRACT



Please check the journal's author guideline for whether a graphical abstract, key points, new findings, or other items are required for display in the Table of Contents.

This article is protected by copyright. All rights reserved.