Dissemination of Original NMR Data Enhances Reproducibility and Integrity in Chemical Research

Journal:	*Natural Product Reports*
Manuscript ID	NP-PER-02-2016-000022.R1
Article Type:	Viewpoint
Date Submitted by the Author:	12-Apr-2016

Complete List of Authors:	Bisson, Jonathan; UIC Simmler, Charlotte; UIC Chen, Shao-Nong; UIC Friesen, J. Brent; Dominican University Lankin, David; University of Illinois at Chicago, College of Pharmacy McAlpine, Jim; UIC Pauli, Guido; UIC,

Natural Product Reports

VIEWPOINT

Dissemination of Original NMR Data Enhances Reproducibility and Integrity in Chemical Research

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Jonathan Bisson*, Charlotte Simmler*, Shao-Nong Chen, J. Brent Friesen, David C. Lankin, James B. McAlpine, and Guido F. Pauli

The notion of data transparency is gaining a strong awareness among the scientific community. The availability of raw data is actually regarded as a fundamental way to advance science by promoting both integrity and reproducibility of research outcomes. Particularly, in the field of natural product and chemical research, NMR spectroscopy is a fundamental tool for structural elucidation and quantification (qNMR). As such, the accessibility of original NMR data, i.e., Free Induction Decays (FIDs), fosters transparency in chemical research and optimizes both peer review and reproducibility of reports by offering the fundamental tools to perform efficient structural verification. Although original NMR data are known to contain a wealth of information, they are rarely accessible along with published data. This Viewpoint discusses the relevance of the availability of orginal NMR data as part of Good Research Practices not only to promote structural correctness, but also to enhance traceability and reproducibility of both chemical and biological results.

A. FIDs are Hidden Treasures

1. Originality of Nuclear Magnetic Resonance (NMR) Data

Natural product research heavily depends on valid structural information to ensure reliability and repeatability of experiments from the bench to the clinical trial. Today, the majority of structural evidence on nature's small molecules is gleaned from spectroscopic data. The vast majority of structures result from human and/or computer-assisted data interpretation that involves different levels of deductive reasoning, rather than from direct evidence. NMR spectroscopy has become one of the mainstays of this process. Today, NMR provides a rich tool set for the observation of C/H/N- and X-nuclear parameters (mainly δ, J, $\omega_{1/2}$, nOe), allowing the construction of molecular connectivities and 3D spatial relationships. Reflecting on these facts, it is important to recognize the crucial role of original NMR data. As knowledge increases and methodology advances, unfiltered (un-interpreted) data are the only truly objective reference. Moreover, such data preserves all chemical information that a given experiment encodes, the decoding of which depends on both the skills of the interpreters and the capability of available methodology. This in particular applies to the mother of all NMR experiments, the 1D ^{1}H NMR (HNMR) spectrum:[1] with the caveat of proton- or *J*-deficient molecules, HNMR spectra exhibit a very high information content. In fact, when mined appropriately, they encode virtually all the structural information pertaining to a natural product.[2]

In contemporary NMR, the Free Induction Decay (FID) represents the original data. Visualization of the spectrum requires Fourier Transformation (FT, Figure 1) and other processing steps, particularly in 2D NMR.[3] The underlying physical mechanisms make HNMR spectra highly information rich and uni-determinant for a given structure, as summarized in SI-1, Supporting Information. Its high sensitivity and information content makes HNMR *the* ubiquitous first pass experiment and ^{1}H FIDs the most valuable original data, as well as a natural focus of this Viewpoint. However, the following discussion applies equally to original (FID) data from other NMR experiments. Successful extraction of the vast NMR information greatly benefits from the use of post-acquisition processing tools (SI-2, Supporting Information). The wealth of chemical/structural information encoded in the FID is the primordial repository for any further processing, from which transformed spectra can be created according to specific needs. This makes the original FID *the* unique source for the transparency of experimental information, as it can be transformed into a spectrum at any time, for a specific purpose, and using any state-of-the-art methodology available at the time of re-processing (Figure 1).

2. FIDs Provide Unique Opportunities

One soul-searching question pertains to the reasoning, why a scientific community, which is well aware of the analytical facts behind NMR, has not yet adopted a public NMR dissemination model. The reasoning is likely as multi-facetted as a higher order ^{1}H multiplet resonance. The lack of FID inclusion requirements for publications, together with relatively slow-paced guideline revision and innovation cycles, are key contributing factors in classical dissemination. Moreover, public visibility generated by the need for sharing unvarnished NMR data is an invitation to scrutiny, which might be perceived as demanding – but could equally be considered a means of enhancing scientific accuracy, as with any other kind of analytical data. A third impact layer regards the general

Center for Natural Product Technologies (CENAPT), Department of Medicinal Chemistry and Pharmacognosy, College of Pharmacy, University of Illinois at Chicago, 833 South Wood Street, Chicago, Illinois, 60612, United States.

**CS and JB contributed equally.*

overburdening of peer reviewers and editors: flooded with manuscripts, they might consider FID addenda an unwarranted extra burden, at least initially. However, FIDs also open new opportunities for peer-review enhancement (e.g., integrity check) and acceleration, thus even overcompensating for this type of concern. Finally, NMR spectroscopic analysis including data processing might still be perceived as highly specialized ("skilled operator"), complicated ("vast parameter sets"), and non-transparent. Present-day NMR instruments, software, and educational resources clearly mitigate this apprehension.

B. FIDs Foster Reproducibility and Integrity

1. Transparency and Integrity

In the literature, the descriptive analysis of newly reported structures is supported by interpreted frequency domain NMR data. Predominant practice (Figure 1) is to convert this information into a tabulated summary of the chemical shifts, multiplicities, and coupling constants. However, the completeness of the latter two varies largely across reports. Most spectra, particularly HNMR spectra, are usually acquired and processed with standard methods, which often are not fully adequate to unravel the complex signal patterns that represent the wealth of encoded structural information. Furthermore, critical information is forcibly lost or misrepresented when the HNMR spectrum is converted into a table by means of a visual or software assisted analysis. Moreover, HNMR spectra are characterized by the variation in line widths due to the dynamic nature of the structure, signal overlap, coupling patterns, and higher order effects. This situation is further exacerbated when NMR signals are described as "multiplets" without more precision regarding these observables.

Making FIDs available, together with annotations, promotes the efficacious peer review (Figure 2) of proposed structures and, thus, encourages structural correctness. Even with the inclusion of figures containing annotated NMR spectra, typically relegated to the Supporting Information, it remains a challenge for editors, reviewers, and readers alike to validate a proposed structure. The result of this being one of the probable causes of the rather common publication of structural revisions.[1, 4] As the latter authors had proposed earlier: "However, one type of unfortunate error could potentially be avoided if chemists were to deposit all their spectral data into a universal database similar to that used for X-ray crystal structures: namely, the proposal of an incorrect structure for a natural product that has already been isolated and characterized."[4]

An equally undesirable outcome is that the accuracy and precision of data comparison for structural dereplication is compromised by the current practices of reporting. Particularly, when comparing data from different magnetic field strengths, the observable major variations in spectral manifestation are not reflected by the widely used means of reporting.[5] As such, the rapid assessment of structural identity (dereplication) via comparison of tabulated data is impeded by these imprecisions, as the available data often fail to accurately describe and correspond to the experimental (acquisition) parameters of the sample being dereplicated (Figure 1). There is clearly a need for more transparency in the structural description of compounds in chemistry in general.

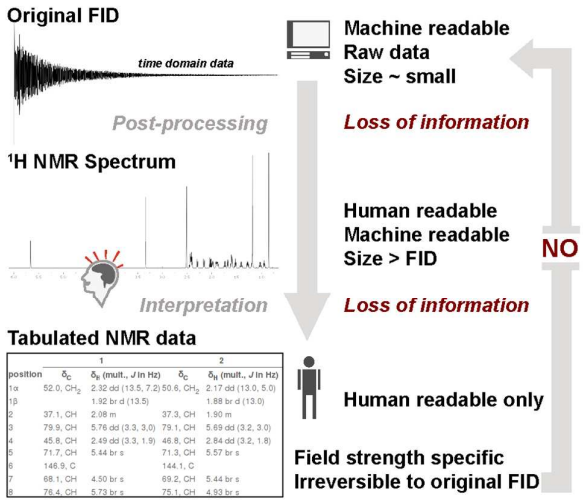

Figure 1: Characteristics of the workflow from the FID to the tabulated NMR data show the importance of the original FID as the fundamental raw data that encodes important structural information.

Notably, this true transparency can be approached by publication of the original FID data.

A database open to both publication and voluntary pipelines could offer reference codes and hyperlinks that make FIDs accessible to peers, become permanent bibliographic information, and are updated depending on review outcome and author choice. This would broadly foster opportunities for both reviewers and readers to validate and utilize structural information, and help diminish structural incorrectness (SI-3, Supporting Information). Whereas associated activities might initially be an unwelcome task, the utility and high information content of raw data will likely trigger rapid re-prioritization and adaptation of new methods to existing workflows.

In our experience, shared FIDs are essential to both detailed structural analysis and dereplication. We were intrigued by the recent reassignment of the aquatolide structure.[1] The structure itself offered some very interesting NMR correlations due to its bicyclo[2.1.1]hexane core structure, leading to the thought that additional insights could be gained from a detailed full-spin analysis. The original FID provided by the authors allowed us to reprocess the ^{1}H spectrum and analyze the previously reported "multiplets". In addition, the spectrum was analyzed with quantum mechanics (QM) based software to reveal the entire *J*-coupling information contained in complex signals.[1, 2] In the course of this work, the total synthesis of aquatolide was published.[6] Again, the authors provided the original FID, from which we could demonstrate the rapid dereplication of the isolated natural product vs. the synthesized compound. The resulting report[1] demonstrates the power of shared HNMR FIDs and adequate tabular description for the promotion of accurate structural elucidation and reliable dereplication.

2. Good Research Practices and Public Repository

WHAT? In order to promote the integrity of structural reports, and to establish Good Research Practices in natural product and chemistry research, all raw structural data (*e.g.*, FIDs) associated with each newly described and, ideally, all known reported structures have to be included as Supporting Information. To ensure compliance, academic publishers may

add a section in their author guidelines requiring the submission of FIDs, in addition to annotated spectra, and together with the tabulated NMR data including chemical shifts δ (precision 0.1-1 ppb) and coupling constants *J* (precision 10 mHz).[5]

WHERE? Dispersion of datasets throughout various journal websites in the form of Supporting Information still presents the problem of locating FIDs. A centralized repository, or at least a directory, are highly desirable. Several independent databases have been created in recent years.[7] The promising Open Spectral Database (www.osdb.info) is still in its infancy but has a great potential. The database at www.glycomics.ccrc.uga.edu contains structural information from various carbohydrates isolated from plants, fungi, and bacteria. The NMRshiftDB (www. nmrshiftdb.com) stores tabulated 1D NMR data, representations of spectra using simulated lines, and optionally FIDs converted to JCampDX format. The MetIDB (www.metidb.org), offering structural information for flavonoids through predicted and experimental HNMR spectra, along with the Human Metabolome (www.hmdb.ca) and the Madison-Qingdao Metabolomics Consortium databases at www.mmcd.nmrfam.wisc.edu are open repositories of original NMR data. However, these initiatives will benefit greatly from a harmonized referencing system, which would facilitate data retrieval from both literature and database sources.

HOW? Ideally, the dissemination of original NMR data will entail a unified public digital repository that accommodates the following: both time and frequency domain data; the tabulated and documented NMR data; putative or confirmed structures; and sample information (concentration, solvent, pH, composition). The time domain data typically comes in the form of manufacturer data, include the strength of the magnetic field and acquisition parameters, and can be best preserved in an archive format (ZIP, TAR). Quantum interaction and linkage tables (QuILTs)[1] offer a complete documentation of NMR data with the visualization of 1D *J*-coupling relationships, NOESY correlations, and heteronuclear experiments.

One of the obvious issues for building repositories of original scientific data, including NMR and other data from proprietary instrumentation, is the lack of a standardized format. The three most abundant NMR instruments (Agilent/Varian, Bruker, Jeol) each have their own format, with varying levels of standardization, change over time, and publicly available specifications. JCAMPDX was often thought of as universal format, but can contain different encoding schemes and implementations of 3rd party software,

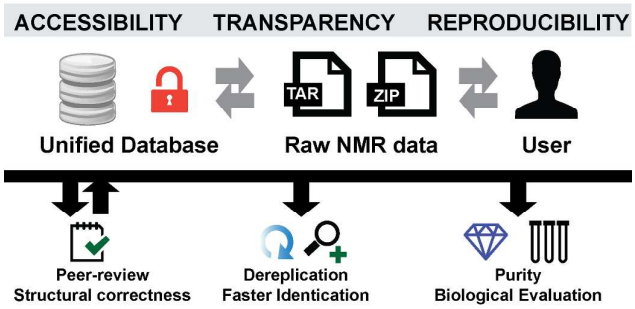

Figure 2: Freely accessible original NMR data advances chemical research by promoting transparency, accuracy, and reproducibility of results and downstream outcomes.

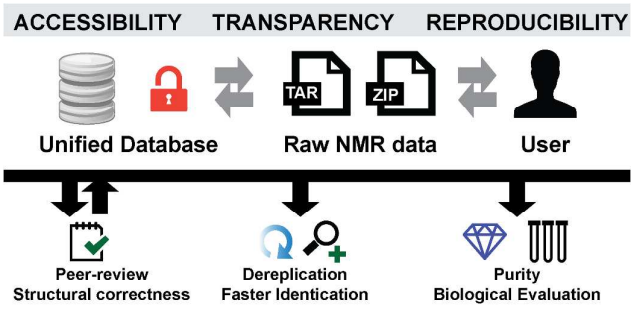

Figure 2: Freely accessible original NMR data advances chemical research by promoting transparency, accuracy, and reproducibility of results and downstream outcomes.

sometimes causing compatibility issues. However, most Free software (as in "Free speech"), freeware (as in "free beer"), and commercially available solutions can handle the above proprietary formats. The Free software, NMRGlue, can convert between different formats. It is available as a Python module and can be readily integrated into an online database. Increasingly user-friendly NMR software also supports peer review by providing options of generating survey and fully processed spectra rapidly.

One viable model to support the construction and regular update of an NMR FID repository entails mandatory deposition prior to the submission of any manuscript. Such a system has already been installed in molecular biology since 1971 with the mandatory deposition of any raw data (e.g., X-ray, NMR beginning in 1989, and electron microscopy) enabling the structural determination of proteins, peptides, nucleic acids, and their complexes (Protein Data Bank, http://www.rcsb.org). Following the submission to PDB, an identification code is assigned per deposited structure and referenced in the manuscript. This allows the reviewers, and any researcher once published, to access the raw data, evaluate the structural correctness, and compare the results for the elucidation of congeneric and/or analogous structures.[8] The sole existence of these repository systems highlights the feasibility and benefit of such a project. This should, therefore, encourage researchers to establish a consortium, independent from the academic publishing outfits, and to work on a unified standardized process for the deposition of raw structural data in an open access database. Expanding existing electronic dissemination mechanisms in journals to the distribution of FIDs as Supporting Information (e.g., as open source ZIP files) could be a productive and rapidly achievable first step that requires only minimal effort. Public access FID archives remain most desirable long-term, and existing sharing models developed and/or mandated by funding agencies might serve as blueprint models.

C. FIDs Have Unique Application Potential

1. Complex NMR Signals, Dereplication, and FID Use

One vital role of describing new structures in the literature is enabling others to facilitate the interpretation of their NMR data and foster the dereplication process (Figure 2). Success rests on the adequate documentation and (Boolean) peer-review of the original spectra.[1, 5] For efficient reproducibility, reported tabulated data must be as complete as possible by showing all relevant parameters of all involved spin systems,

and as precisely as possible to match the resolution and chemical shift accuracy of the underlying spectrum.

As FID sharing can cover the entire gamut of 1D and 2D NMR, its effect extends well beyond dereplication, opening unprecedented opportunities for data re-use and re-purposing. Examples are: (i) comparison of published with newly isolated or synthesized compounds, and recognition of congeners and structural motifs; (ii) reprocessing of spatial (NOESY, ROESY) and long-range (e.g., HMBC) correlation spectra that provide critical information, often more than can be tabulated or depicted; (iii) resolution of ambiguities and clarification of interpretations; (iv) mining for features that were rendered unimportant or suddenly become interpretable due to technological progress; (v) community-based projects that build on digital NMR data.

Even complex HNMR spectra can be interpreted fully by means of QM theory. In fact, signals (too) commonly designated as "multiplets" and resonances with higher order effects are highly diagnostic and can provide valuable or even essential structural information.[5] The complete and accurate description of such fingerprint signals can only be achieved through computer-assisted QM extraction of all δ, J, and $\omega_{1/2}$ values via ^{1}H iterative Full Spin Analysis (HiFSA),[2] which requires the original data. Importantly, this information allows the simulation of the entire complex HNMR spectra, including all characteristic fingerprint signals, at any magnetic field strength. This simplifies end-user comparison of spectra and opens unprecedented applications for low-field NMR instruments that are currently being (re-)introduced into the chemists' toolbox.

Spectral simulation beyond data comparison assessing structural identity also facilitates the structural elucidation of congeners and analogues. Fully interpreted HNMR spectra enhance the accuracy and speed of dereplication. Publication of FIDs provides the best approach for the dissemination of data from which the scientist can decode critical information, e.g., by analyzing complex fingerprint signals. As such, the use of well documented FID repositories represents a very simple, highly specific, and rapid means of disseminating and exchanging structural information. Dereplication can then be performed visually or by software comparison of the spectra, as opposed to difficult-to-read tables or listings which are prone to confusion, technical errors, and incompleteness – and easily overlooked during peer-review. Notably, this unique application potential of FIDs applies not only to purified compounds, but also to mixtures, including complex metabolomic samples. In all these instances, even at high levels of information density (overlap), the FID represents the most genuine information and remains the best source of valid data prior to any interpretation.

2. Purity Determination and Residual Complexity

Original FIDs encode not only the qualitative (structural) and quantitative information of the target compound, but also of any other detected component, i.e., an impurity profile that altogether characterizes the described sample. In fact, the metabolic origin of natural products explains why they are associated inherently with impurity profiles, which interestingly have been understudied. Previously described as residual complexity (RC; for an overview see ref [9] and references therein, as well as go.uic.edu/residual complexity) these profiles reflect the various synthetic routes and/or diverse isolation pathways that lead to the production of the structurally described compounds. Therefore, the analysis and documentation of impurity profiles and RC is another way to disclose the traceability of isolated or synthetic compounds, thereby enhancing the transparency of chemical outcomes and guiding the choice for the optimal methods of production (Figure 2). Subsequently, enabled by FID sharing, the (re-)processed NMR spectra can facilitate the identification of potential impurities and the evaluation of their relative molar concentration via quantitative HNMR (qHNMR)[9] by comparison with the compound of interest.

FID archiving facilitates an accurate and comprehensive documentation of structural identity, while promoting transparency with regard to potential impurities (Figure 2). The public diffusion of FIDs, combined with appropriate NMR processing, can guide efforts towards establishing important biological links via the determined purity of structurally identical, or related compounds, obtained through different processes. This empowers the cross-validation of future reference materials. Collectively, the dissemination of FIDs for structural correctness and purity determination will favor the reproducibility not only of chemical results but also of biological data. "*The relevance of minor constituents cannot be overlooked when assigning pharmacologically active principles,*" as already exemplified in studies describing biological activities wrongly attributed to ursolic acid or epiquinamide.[9, 10] Both structural correctness and description of RC can guide the interpretation of biological results, particularly in exploratory drug discovery.

Conclusions

FIDs are time domain original/raw NMR data, which contain a wealth of information for accurate structural identification and quantitation. FIDs can be easily stored and exchanged, an advantage that can be utilized to promote careful evaluation and validation of new as well as known structures. In addition, access to FIDs facilitates accurate structural comparison and fosters the dereplication process in terms of speed and specificity. The deposition of FIDs for structures that are newly described, or for which no or insufficient FIDs are available, in unified open access digital repositories will allow the unrestricted exchange of critical experimental information as a resource to the scientific community. Better access to data also facilitates a more comprehensive understanding and utilization of structural data, which improves the usefulness of NMR information. In addition, HNMR FIDs also document the purity of the investigated material via qHNMR, which is especially important if the materials are the subject of biological evaluations. Altogether, dissemination of FIDs, in addition to interpreted spectra, can improve the transparency of results in chemistry and natural product research, promoting the reproducibility and subsequent validation of published data.

Acknowledgements

The authors kindly acknowledge support by grant U41 AT008706 from NCCIH and ODS/NIH.

References

1. G. F. Pauli, M. Niemitz, J. Bisson, M. W. Lodewyk, C. Soldi, J. T. Shaw, D. J. Tantillo, J. M. Saya, K. Vos, R. A. Kleinnijenhuis, H. Hiemstra, S.-N. Chen, J. B. McAlpine, D. C. Lankin and J. B. Friesen, *J. Org. Chem.*, 2016, **81**, 878-889.
2. J. G. Napolitano, D. C. Lankin, T. N. Graf, J. B. Friesen, S.-N. Chen, J. B. McAlpine, N. H. Oberlies and G. F. Pauli, *J. Org. Chem.*, 2013, **78**, 2827-2839.
3. W. F. Reynolds and R. G. Enriquez, *J. Nat. Prod.*, 2002, **65**, 221-244.
4. K. C. Nicolaou and S. A. Snyder, *Angew. Chem. Int. Ed.*, 2005, **44**, 1012-1044.
5. G. F. Pauli, S.-N. Chen, D. C. Lankin, J. Bisson, R. Case, L. R. d. Chadwick, T. Gödecke, T. Inui, A. Krunic, B. U. Jaki, J. B. McAlpine, S. Mo, J. G. Napolitano, J. Orjala, J. Lehtivarjo, S.-P. Korhonen and M. Niemitz, *J. Nat. Prod.*, 2014, **77**, 1473-1487.
6. J. M. Saya, K. Vos, R. A. Kleinnijenhuis, J. H. van Maarseveen, S. Ingemann and H. Hiemstra, *Org. Lett.*, 2015, **17**, 3892-3894.
7. S. R. Johnson and B. M. Lange, *Front. Bioeng. Biotechnol.*, 2015, **3**, 1-10.
8. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235-242.
9. G. F. Pauli, S.-N. Chen, C. Simmler, D. C. Lankin, T. Gödecke, B. U. Jaki, J. B. Friesen, J. B. McAlpine and J. G. Napolitano, *J. Med. Chem.*, 2014, **57**, 9220-9231.
10. R. W. Fitch, G. D. Sturgeon, S. R. Patel, T. F. Spande, H. M. Garraffo, J. W. Daly and R. H. Blaauw, *J. Nat. Prod.*, 2009, **72**, 243-247.