nmrML: a community supported open data standard for the description, storage, and exchange of NMR data.

Daniel Schober, Daniel Jacob, Michael Wilson, Joseph A Cruz, Ana Marcu,
Jason R Grant, Annick Moing, Catherine Deborde, Luis F de Figueiredo,
Kenneth Haug, et al.

nmrML: A Community Supported Open Data Standard for the Description, Storage, and Exchange of NMR Data

Daniel Schober,*[†] Daniel Jacob,[‡] Michael Wilson,[§] Joseph A. Cruz,[§] Ana Marcu,[§] Jason R. Grant,[§] Annick Moing,[‡] Catherine Deborde,[‡] Luis F. de Figueiredo,[∥] Kenneth Haug,[∥] Philippe Rocca-Serra,[⊥] John Easton,[#] Timothy M. D. Ebbels,[⊗] Jie Hao,[⊗] Christian Ludwig,[$] Ulrich L. Günther,[×] Antonio Rosato,[○] Matthias S. Klein,[¶] Ian A. Lewis,[¶] Claudio Luchinat,[○] Andrew R. Jones,[▽] Arturas Grauslys,[▽] Martin Larralde,[+] Masashi Yokochi,[◆] Naohiro Kobayashi,[◆] Andrea Porzel,[&] Julian L. Griffin,[%] Mark R. Viant,[■] David S. Wishart,[§] Christoph Steinbeck,[∥] Reza M. Salek,*,[∥] and Steffen Neumann[†]

[†]Leibniz Institute of Plant Biochemistry, Department of Stress and Developmental Biology, Weinberg 3, 06120 Halle, Germany

[‡]INRA, Univ. Bordeaux, UMR1332 Fruit Biology and Pathology, Metabolome Facility of Bordeaux Functional Genomics Center, MetaboHUB, IBVM, Centre INRA Bordeaux, 71 av Edouard Bourlaux, F-33140 Villenave d'Ornon, France

[§]Departments of Computing Sciences and Biological Sciences, University of Alberta, Edmonton, Canada T6G 2E8

[∥]European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, U.K.

[⊥]University of Oxford, e-Research Centre1, 7 Keble Road, Oxford OX1 3QG, U.K.

[#]School of Engineering, University of Birmingham, Edgbaston, Birmingham B15 2TT, U.K.

[⊗]Computational and Systems Medicine, Department of Surgery and Cancer, Imperial College London, London, SW7 2AZ, U.K.

[$]Institute of Metabolism and Systems Research, University of Birmingham, Edgbaston, Birmingham B15 2TT, U.K.

[×]Institute of Cancer and Genomic Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, U.K.

[○]Magnetic Resonance Center (CERM) and Department of Chemistry, University of Florence, 50019 Sesto Fiorentino, Florence, Italy

[¶]Department of Biological Sciences, University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada

[▽]Institute of Integrative Biology, University of Liverpool, Bioscience Building, Crown Street, Liverpool L69 7ZB, U.K.

[+]Ecole Normale Supérieure Paris-Saclay, 61 Avenue du Président Wilson, 94230 Cachan, France

[◆]Institute for Protein Research (IPR), Osaka University, 3-2 Yamadaoka, Suita-shi, Osaka, 565-0871, Japan

[&]Department of Bioorganic Chemistry, Leibniz Institute of Plant Biochemistry, 06120 Halle (Saale), Germany

[%]Department of Biochemistry, University of Cambridge, Downing Site, Cambridge CB2 1QW, U.K.

[■]School of Biosciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, U.K.

Ⓢ *Supporting Information*
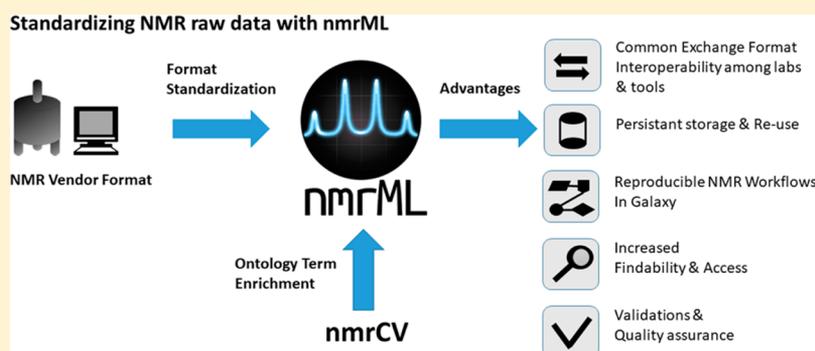
ABSTRACT: NMR is a widely used analytical technique with a growing number of repositories available. As a result, demands for a vendor-agnostic, open data format for long-term archiving of NMR data have emerged with the aim to ease and encourage

continued...

sharing, comparison, and reuse of NMR data. Here we present nmrML, an open XML-based exchange and storage format for NMR spectral data. The nmrML format is intended to be fully compatible with existing NMR data for chemical, biochemical, and metabolomics experiments. nmrML can capture raw NMR data, spectral data acquisition parameters, and where available spectral metadata, such as chemical structures associated with spectral assignments. The nmrML format is compatible with pure-compound NMR data for reference spectral libraries as well as NMR data from complex biomixtures, i.e., metabolomics experiments. To facilitate format conversions, we provide nmrML converters for Bruker, JEOL and Agilent/Varian vendor formats. In addition, easy-to-use Web-based spectral viewing, processing, and spectral assignment tools that read and write nmrML have been developed. Software libraries and Web services for data validation are available for tool developers and end-users. The nmrML format has already been adopted for capturing and disseminating NMR data for small molecules by several open source data processing tools and metabolomics reference spectral libraries, e.g., serving as storage format for the MetaboLights data repository. The nmrML open access data standard has been endorsed by the Metabolomics Standards Initiative (MSI), and we here encourage user participation and feedback to increase usability and make it a successful standard.

Nuclear magnetic resonance (NMR) spectroscopy is an important analytical tool in organic chemistry, biochemistry, natural products research, structural biology, and metabolomics. Recently the need for an open NMR data standard covering the free induction decay (FID) to support data reproducibility has been acknowledged.[1] As instrument vendors typically provide the data processing software and produce evolving data formats together with the instrument hardware, developers of third party NMR analysis software often need to devote considerable effort into reading and writing these vendor-specific formats. This applies both to commercial software and to community developed open-source tools such as the BATMAN R package,[2] Bayesil,[3] NMRProcFlow,[4] rNMR[5] and MetaboLab.[6] With the recent termination of the Agilent/Varian NMR spectrometer range, the question of long-term readability of discontinued vendor formats has become paramount for a growing NMR community. Data in proprietary formats can age quickly, and NMR data stored in such formats can become obsolete, making valuable results inaccessible and irreproducible in the long term. Also, spectra processing and quantification tools would benefit from a standardized storage format for processed NMR data, i.e., serving workflow systems. For NMR data repositories such as MetaboLights,[7] Metabolomics WorkBench,[8] Human Metabolome Database HMDB,[9] and BioMagResBank,[10] key questions regarding long-term data persistence, i.e., on sustainability, usability, and accessibility are arising.

Currently, the most widely used open data exchange format for NMR data is JCAMP-DX version 6.0,[11] but due to the broad scope and complexity of this format, many different vendor-dependent variants exist. Coordinated updating for all variants, in order to reflect the state of the art in NMR methodology, is rarely seen in this 30 year old format. This variability can lead to incompatibilities between different software packages, and as a result no content-based (semantic) validation of JCAMP-DX is available. While JCAMP-DX is likely to remain in use for NMR data capture for many years, it is clear that alternative approaches, such as XML or JavaScript Object Notation (JSON) with peer-maintained ontologies, would be beneficial.

The first efforts toward establishing an XML-based open NMR standard and controlled vocabulary were discussed in 2007 by the ontology working group[12] of the Metabolomics Standards Initiative (MSI)[13] and a consortium of U.K. universities discussing minimal reporting guidelines.[14] In 2011, a series of initiatives by members of the NMR-based metabolomics and biomolecular NMR communities were launched to explore the creation of a new community standard for NMR data exchange and storage. This included meetings attended by NMR stakeholders including metabolomics database representatives and vendors. This initiative and subsequent meetings were then taken over by the COSMOS (COordination of Standards in

MetabOlomicS) EU FP7 consortium,[15] aiming to coordinate the establishment of a persistent NMR data format and open source data analysis tools for the NMR community. The main goals were

(1) *Data sharing in an open vendor-agnostic manner*, so that users, tool developers, and public repositories can import or export data to support integrated (meta-)analysis and secondary data usage.

(2) *Search and retrieval of relevant results*, minimizing alternate ways of encoding the same information, so that data sets with a similar setup can be identified and compared.

(3) *Spreading best practices and evaluation of the results*, whereby the data quality can be assessed in light of intelligibility and completeness along minimum information standards supported by automatic validation aids.

(4) *Improved data persistence and traceability over time*, delivering a self-describing easy-to-use yet robust raw data storage format to support long-term archiving.

From such efforts, it was decided that the new data format would be called nmrML (for NMR Markup Language) and it should

(1) Be compatible with existing vendor formats (Varian/Agilent, Bruker, JEOL) and partially compatible with certain variants of JCAMP-DX.

(2) Be XML-based, so as to be similar to established XML formats by the Proteomics Standard Initiatives (PSI), i.e., mzML for mass spectrometry.[16]

(3) Support the use of controlled vocabularies/ontologies to annotate spectral data and metadata with standardized community descriptors, which can be maintained in a decentralized peer production manner.

(4) Initially focus on the capture of small molecule NMR data with macromolecular NMR data being addressed in succession; but be flexible enough to be expanded in scope.

(5) Be easy to understand and integrate into existing open analysis and processing software.

(6) Contain sufficient spectrometer data, acquisition, and processing metadata to permit the reconstruction of the NMR spectrum and experiment.

(7) Capture coarse-grained spectral assignment data for molecule identification and quantification in chemical mixtures. Capture fine-grained assignment and chemical structure data of pure-compound spectra for use in organic synthesis and natural product studies, medicinal chemistry, and reference NMR spectral libraries.

Under these development constraints, members of the nmrML COSMOS team created the nmrML data standard, the necessary software support, and fostered support from databases to both accept and display nmrML data. Figure 1 summarizes available nmrML compliant tools and functionalities in support
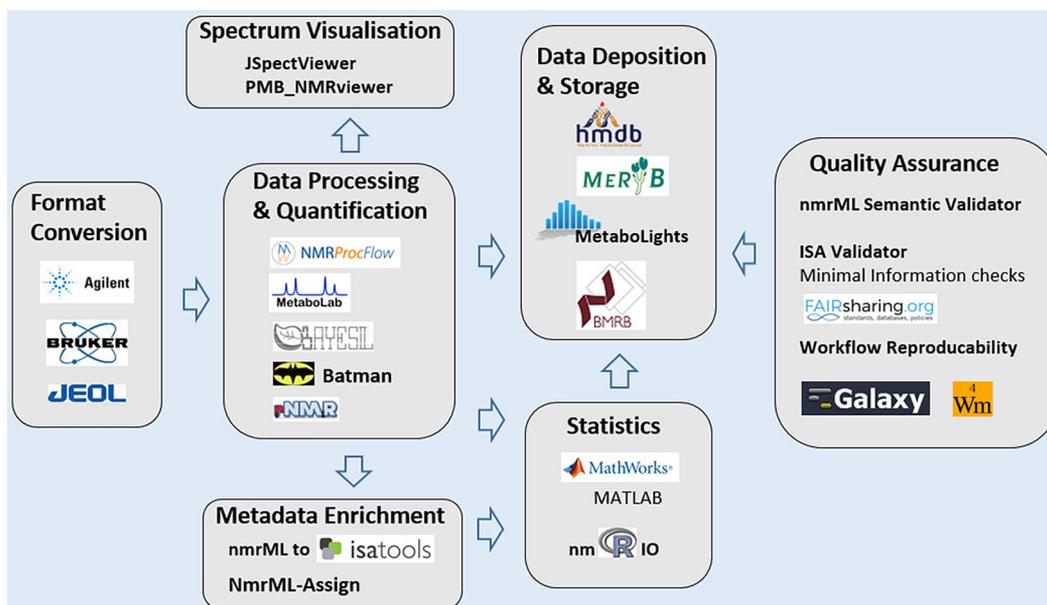
Figure 1. A prototypical metabolomics workflow for NMR data processing and storage is shown and nmrML-aware tools supporting each workflow step are illustrated. Vendor to nmrML converters, NMR data processing, and visualization tools as well as public repositories that accept nmrML as standard data format are highlighted. Parsers for MATLAB and R, which make nmrML data accessible to statistics tools, and content validators that assist in data quality control and workflow reproducibility are shown. Many of our tools already run in Galaxy-based workflow management environments. Bruker logo reprinted with kind permission from Bruker BioSpin Group, Copyright 2017. JEOL logo reprinted with kind permission from JEOL (Germany) GmbH, Copyright 2017. Agilent Technologies Corporate Signature Copyright 1999 Agilent Technologies, Inc. Reproduced with permission, Courtesy of Agilent Technologies, Inc. MATLAB is a registered trademark and reprinted with kind permission from The MathWorks, Inc.

of a typical NMR data handling workflow for a metabolomics or similar experiment.

MATERIALS AND METHODS

The nmrML format specification is composed of an XML Schema Definition (XSD) and an accompanying controlled vocabulary called nmrCV. Leveraging on existing efforts, the nmrML development started by updating a predecessor XSD developed at The Metabolomics Innovation Centre (TMIC, http://www.metabolomicscentre.ca/exchangeformats.htm) in Edmonton, Canada, with additional elements and structures from a BML-NMR XSD developed at the University of Birmingham.[17] Both of these efforts were integrated, expanding the TMIC predecessor, as it was already capturing the basic raw data and had the controlled vocabulary (CV) reference mechanism in place. The nmrML CV referencing mechanism and basic XML architecture was inspired by mass spectrometry markup language (mzML), the PSI standard mass spectrometry data format used in proteomics and metabolomics.[16] The mzML community standard captures raw MS spectral data, instrument parameters, experiment metadata, and peak assignment, as well as compound quantitation data. Given the similarity in data capture, storage, and retrieval between modern MS and NMR experiments, many of the successful features found in mzML were transferred and adapted to nmrML. The NMR.owl CV by the MSI,[12] and a parallel TMIC effort NMR CV, developed to serve the TMIC XSD, were integrated. The merged nmrCV organizes common and essential NMR terms into a simple is-a class hierarchy (taxonomy). The nmrML 1.0.0 format presented here is the outcome of these integration efforts and will serve as the MSI recommended common data standard and terminology for open access NMR data. While the nmrML.xsd mostly covers raw data, it also provides for some NMR data elements computed by open access NMR processing and quantification tools.

Development was coordinated via mailing lists, video conferences, and during multiple workshops and hackathons. The choice of XML was motivated by technical maturity, flexibility and universality of XML in both capturing and presenting scientific data. There is an abundant XML expertise to leverage on, as XML resides at the base of the semantic Web stack. The appearance of all knowledge capture XML elements can be controlled via the XSD (mandatory vs optional) and hence allows for content completeness checks. We implemented converter Web services to generate valid nmrML from vendor raw data files. Links to nmrML compliant databases as well as NMR processing and spectrum visualization software are provided in Table 1. Format parsers, application program interfaces (APIs), and validation Web services have been set up. All code libraries, an issue tracker as well as a file versioning and release policy are available on the developer's GitHub pages at https://github.com/nmrML/nmrML.

RESULTS

The nmrML core specification, including the XSD and nmrCV, can be found at http://nmrml.org. The referenced nmrCV.owl currently contains over 600 terms and is indexed under the National Center for Biomedical Ontology (NCBO) Bioportal ontology library.[18] Our documentation Web site (http://nmrml.org/examples) provides tutorial material and videos, code examples for single compound reference spectra, as well as mixed-compound 1D and 2D NMR spectra.

nmrML Architecture. The nmrML XSD element hierarchy contains multiple sections that organize the information that can appear in an nmrML XML data file in a community-agreed and self-explanatory way. This facilitates understanding of the format by both humans and by data processing software alike. The current top level XSD structure provides high-level base elements for the grouping and capture of NMR data, describing the

651

Table 1. Non-exhaustive list of nmrML compatible open source software, clustered by tool category

tool category	tool name	key functions	URL	developer
format converters	nmrML converter (Java)	converts vendor to nmrML format (*recommended*)	https://github.com/nmrML/nmrML/tree/master/tools/Parser_and_Converters/Java	Institut National de la Recherche Agronomique (INRA), France
	nmrML converter (Python)	converts vendor to nmrML format	https://github.com/nmrML/nmrML/tree/master/tools/Parser_and_Converters/python/pynmrml	The Metabolomics Innovation Center (TMIC), Canada
	nmrML to ISA converter	generates prepopulated ISA files from nmrML files	https://github.com/nmrML/ISA-tools/nmrml2isa	EMBL-EBI, United Kingdom
	BMSxNmrML	converts BMRB metabolomics entries to nmrML format	http://bmrbdep.pdbj.org/en/bmsxnmrml.html	Institute for Protein Research (IPR), Japan
parsers	MATLAB parser	MATLAB functions parsing and decoding nmrML files, and also writing MATLAB data into nmrML format.	https://github.com/nmrML/nmrML/tree/master/tools/Parser_and_Converters/Matlab	Imperial College London (ICL), United Kingdom
	nmRIO	R package for parsing and decoding nmrML files	https://github.com/nmrML/nmrML/tree/master/tools/Parser_and_Converters/R/nmrIO	Leibniz Institute of Plant Biochemistry (IPB), Germany
	nmrGlue	modules for working with NMR data in Python; for processing, analyzing, and inspecting NMR data	https://github.com/jjhelmus/nmrglue	Argonne National Laboratory (ANL), IL, USA
data validators	nmrML semantic validator	XML Schema compliance and rule-based validation of CV term usage	http://nmrml.org/validator	Leibniz Institute of Plant Biochemistry (IPB), Germany
spectrum viewers	JSpectraViewer (JSV)	interactive 1D NMR Spectral viewer used in tools such as Bayesil and nmrML-Assign	http://nmrml.bayesil.ca	The Metabolomics Innovation Center (TMIC), Canada
NMR processing, and identification/quantification tools	NMRProcFlow	interactive 1D NMR spectral viewer, spectral processing and quantification tool dedicated to metabolomics	http://nmrprocflow.org	Institut National de la Recherche Agronomique (INRA), France
	Bayesil	automated compound identification, quantification and annotation from 1D NMR spectra	http://bayesil.ca, http://tmic.bayesil.ca	The Metabolomics Innovation Center (TMIC), Canada
	nmrML-Assign	nmrML conversion, annotation and peak assignment to compounds for reference 1D NMR spectra	http://nmrml.bayesil.ca	The Metabolomics Innovation Center (TMIC), Canada
	Batman	Bayesian deconvolution and automated quantification of metabolites from 1D NMR spectra	http://batman.r-forge.r-project.org	Imperial College London (ICL), United Kingdom
	rNMR	region-of-interest based NMR spectra quantification from 1D and 2D NMR spectra	http://rnmr.nmrfamde.wisc.edu	University of Calgary (U of C), Canada
workflow tools	PhenoMeNal app library	lists containerized nmrML aware tools to build Galaxy NMR workflows	http://portal.phenomenal-h2020.eu/app-library	EMBL-EBI, United Kingdom
	SOMA:tameNMR	NMR data processing and analysis via Galaxy Workflows	https://github.com/pgb-liv/tameNMR	University of Liverpool (UoL), United Kingdom
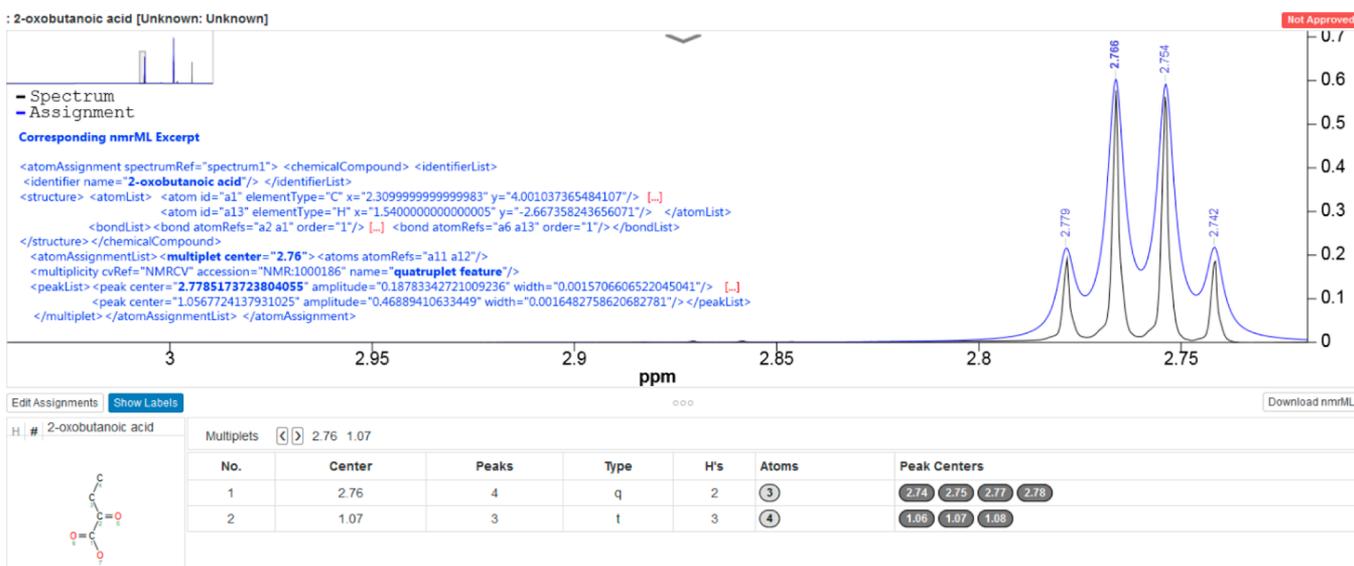
Figure 2. Assignment of an identified molecule in a single compound spectrum, generated in nmrML-Assign and displayed using the JSpectraViewer (JSV). An uploaded raw FID for the 2-oxobutanoic acid reference compound was automatically processed with Bayesil. The resulting interactive JSV spectrum then allows the assignment of peaks to specific atoms, using the nmrML-Assign tool. The assignment metadata is then saved in the nmrML format (see https://github.com/nmrML/nmrML/tree/master/examples/reference_spectra_examples/hmdb). An excerpt view of the corresponding nmrML code (blue code inset) is shown for the quadruplet assignment (Multiplet no. 1) of the second peak (bold code). The corresponding HMDB entry is available from http://www.hmdb.ca/metabolites/HMDB00005, with the ^{1}H spectrum found at http://www.hmdb.ca/spectra/nmr_one_d/1024.

nmrML version, the sources of the controlled vocabularies or ontologies used for metadata annotation, the data depositor contact, source files/formats, software lists, the instrument configuration, sample information (e.g., solvent and reference standards), acquisition settings, and data processing information. This is followed by the spectral FID raw data, as a base64-encoded binary. In addition to such a "minimal" nmrML data file, additional information such as molecule identification/spectral assignment metadata and quantification data can also be included. For example, if the NMR data is for a pure reference compound or a newly isolated/synthesized single chemical, the nmrML file can include data on the chemical structure and corresponding atom-specific peak feature assignments (see example generated by nmrML-Assign in Figure 2 or http://nmrml.org/examples/3). If the NMR data is for a complex mixture, consisting of many different compounds from an analytical setting, the nmrML file can include data on peak positions, integrated peak areas, and putative peak assignments, together with relative or absolute concentrations of some or all of the compounds but no annotation of individual peak features to atom environments (see http://nmrml.org/examples/4). Code examples of a minimal nmrML data XML file as well as for the expanded metadata case are provided on the examples page (http://nmrml.org/examples).

The nmrML structure consists of an XSD that allows it to reference a dedicated NMR controlled vocabulary (nmrCV). The XSD defines the allowed XML structure, whereas the controlled vocabulary provides the terminology to describe the NMR data in detail using standardized textual values for XML-defined tags. In areas where the terminology is likely to change faster than the nmrML XSD can be updated, the representation is branched out from XSD to CV-usage. This approach can accommodate rapid technology/terminology changes in a flexible way, as the CV can be maintained externally by a larger NMR user peer group: for example, terms for new NMR probes

can be represented in a nmrML file by requesting the addition of corresponding new CV terms in the nmrCV, without the need for a full XSD and any subsequent software revisions. The combined usage of XML and a separate CV also allows multiple validation levels to be established (see below). The CV referencing mechanism is explained in detail on the documentation pages.

Tools Compatible with nmrML. We have created Web-based easy-to-use tools to make nmrML more accessible to the broader organic chemistry and metabolomics communities. To ensure that nmrML will be broadly adopted by life sciences and chemical researchers, these tools cover a large fraction of a typical NMR data acquisition, processing and storage workflow to generate, convert, process, validate, and publish nmrML files (Figure 1). Additionally, we have worked closely with open source and commercial tool developers to encourage nmrML format support and adoption. We have summarized efforts already leveraging on the nmrML format in Table 1.

nmrML Converters, Parsers, and Validators. Format converters translate the exchange syntax from vendor raw data formats into XSD-compliant nmrML by means of mappings from Bruker "acqus" or Varian "procpar" raw files to nmrML elements and CV terms. An extensive parameter mapping table is available in the documentation pages. A comprehensive JAVA-based converter automatically generates valid nmrML files from Bruker, Agilent/Varian, and JEOL raw files. It is also available as a Web service (http://nmrml.org/converter) and Docker container. It can be run in batch mode for high-throughput batch conversion of multiple zipped raw data. A Python-based converter that uses the nmrGlue API[19] to access the vendor parameters is also available. Also an nmrML2ISA parser,[20] written in Python, has the ability to read experimental NMR data and metadata from nmrML data files and passing it over to an autogenerated ISA-Tab[21] assay file, i.e., defining a basic metadata backbone ISA-Tab format, i.e., for submission to the MetaboLights repository.[7] In addition, nmrML bindings for multiple programming languages

(Java, Python, Ruby) as well as for widespread data analysis tools like the R statistics package, MATLAB, and open source NMR tools exist. A parser called nmRIO makes nmrML content available to R-based tools such as Batman and rNMR for higher level analysis. A MATLAB parser renders nmrML data available to the MATLAB tool for further statistical processing. An nmrML semantic validator allows the revisal of the correct implementation of manually populated or enriched nmrML files, with regard to XML schema compliance, CV term usage, and allowed term cardinalities. At the core, the XML syntax and structural validity of nmrML XML instances, such as XML element and attribute position, order, and cardinality, can be checked by any validating XML parser against the nmrML.xsd, which defines these allowed elements and their expected characteristics. On the next higher layer, so-called mapping rules can enforce semantic validity[22] of the ontological descriptions used, by testing which CV terms are allowed in which elements. The elements with their allowed CV descriptor hierarchies are outlined in a mapping rule file. The OpenMS/Topp-based[23] nmrML validator (http://nmrml.org/validator) checks that these higher level semantic criteria are being met in a given XML instance. For example, a validation rule file can enforce minimal reporting guidelines such as the MSI-sanctioned Core Information for Metabolomics Reporting (CIMR; http://mibbi.sourceforge.net/projects/CIMR.shtml, accessed November 27, 2017). These validation scenarios make nmrML more easily accessible to quality assurance than JCAMP-DX or other more verbose and equivocal formats that do not rely on controlled vocabularies.

nmrML Data Processors and Viewers. The following tools facilitate NMR data processing and compound identification. nmrML-Assign (http://nmrml.bayesil.ca) is a JavaScript Web application based on Bayesil that allows users to upload vendor formatted 1D NMR raw data or nmrML and to then interactively add compound identification metadata (see Figure 2, Example 3). The Bayesil-generated interactive spectrum allows assigning peaks to specific atoms in a proposed molecule after the Bayesil Web service[3] was used to upload a chemical structure and perform a spectral prediction to help with the assignment process. The assigned atoms are displayed on both the spectrum and the molecule image. Once the assignment process is complete, the annotated file can be saved as enriched nmrML file, which can then be reloaded and interactively viewed and edited or submitted to HMDB. nmrML-Assign works both with ^{1}H and ^{13}C NMR spectra in Bruker or Agilent/Varian format. Bayesil also allows users to upload 1D spectra of biological mixtures (e.g., serum, plasma, cerebrospinal fluid) as shown in Example 4 on our Web site and to perform an automated assignment and quantification of all visible peaks.

The Batman R package estimates metabolite relative concentrations from spectral data and automatically assigns them to metabolites from a target list. Batman can access nmrML data and is using the nmRIO parser. rNMR[5] is a region-of-interest rather than peak-list-based software for visualizing, assigning, and quantifying metabolites in complex 1D and 2D NMR data. The upcoming version of rNMR will read nmrML files directly and can convert them into its native data format. NMRProcFlow is a pipeline tool for the reproducible processing and visualization of 1D NMR data in metabolomics. It allows to pipe processed NMR data as tabular data matrix to statistics workflow tools like bio-statflow.org. It relies on the NMR spectra viewer (https://github.com/nmrML/nmrML/tree/master/tools/Visualizers/PMB_NMRviewer), as its design acknowledges iterative parameter adjustments by means of repeated visual inspection by the user.

nmrML Compatible Databases. A principal objective behind the establishment of nmrML is to ensure data continuity and persistence in NMR repositories and reference libraries. Several key NMR experiment and reference databases now support the upload, storage, display, and download of nmrML data. HMDB, with more than 1500 1D ^{1}H and ^{13}C NMR spectra collected at 500 and 600 MHz ("Human Metabolome Database: Database Statistics", http://www.hmdb.ca/statistics, accessed May 15, 2017), describes more than 1000 reference spectra for pure compounds in the Human Metabolome Library (HML, http://www.hmdb.ca/hml). More than 600 metabolites in HMDB now include NMR reference spectra with complete spectral assignments. These metabolites have 1D NMR annotated spectra available and are downloadable in the nmrML format. Other databases such as DrugBank,[24] YMDB[25] and ECMDB[26] plan to support nmrML compatible reference spectra in the future. BMRB entries are available in XML and RDF, as common open representations of NMR-STAR data format.[27] BMRB has archives of time-domain data and fully assigned nmrML files are accessible, which were generated from BMRB/XML files via the BMSxNmrML converter (see Table 1). In addition to the growing collection of reference spectral libraries, the open access NMR data repository MetaboLights[7] has experimental NMR data archival, which now accepts nmrML data from depositors and allows one to extract basic ISA-Tab metadata from it (see above). It now handles nmrML data from biological mixtures as well as from pure reference compounds. The MeRy-B[28] plant metabolomics NMR knowledge base accepts both JCAMP-DX and nmrML format with the plan to fully adopt nmrML in order to leverage ontological spectra preprocessing terms embedded within nmrML. Work is underway to have the Metabolomics WorkBench[8] accept nmrML data as part of the international MetabolomeXchange initiative (metabolomexchange.org/).

Pipelines and Workflow Support. With the recent push to standardize and facilitate the access to data processing workflows,[29] devoted workflow environments such as Galaxy[30] have gained more weight, the intent here being transparency, traceability, and reproducibility of pipeline-generated data and audit. Galaxy-based metabolomics analysis pipelines are emerging[31] and some are in development for NMR data, such as W4M-NMR[31] (http://workflow4metabolomics.org/the-nmr-workflow) and SOMA:tameNMR (https://github.com/pgb-liv/tameNMR). The NMR processing tool NMRProcFlow[4] uses nmrML as its native spectral data format and containerization of modules for workflow integration is progressing. To foster nmrML as input format for Galaxy workflow pipelines, the PhenoMeNal projects App library portal (http://portal.phenomenal-h2020.eu/app-library already provides nmrML-aware tools (like the nmrML converter) as containers for NMR workflow integration.

■ DISCUSSION

This Perspective describes the first iteration of nmrML (version 1.0.0). We have designed and developed a flexible, open standard data format called the NMR Markup Language (nmrML) for capturing and disseminating NMR data for small molecules. This represents a community-driven effort that involved extensive consultations and many metabolomics, NMR spectroscopy, chemoinformatics, and computing science laboratories from across Europe and North America. Further enhancements are planned for nmrML, and these will include extensions to nD NMR data and the inclusion of macromolecular data in the XML and additional terms in nmrCV. Currently, only basic processed data is captured, e.g., for molecule identification and

quantification, and the latter is equivalent to what mzTab stores for MS data and what is captured in mzIdentML[32] and mzQuantML.[33] The introduction of nmrML hence brings NMR spectroscopy in alignment with existing data standardization efforts in metabolomics, such as mzML for mass spectrometry and will ultimately contribute to cross-technology and multiple omics data comparison. We hope further tools like XEASY[34] for macromolecular NMR analysis and NMRPipe[35] for *n*D NMR will leverage on nmrML in the future. MetaboLab[6] provides high-throughput preprocessing for MATLAB driven NMR statistics and is currently implementing an nmrML parser for standardized data import. In addition, further metadata will be added to nmrML, i.e., as required to store *n*D spectra. In addition to the persistent data storage/exchange standard and CV, we have also described and developed database support and software tools that make use of nmrML. These tools include nmrML viewers, nmrML data converters, processors, and annotators, and these will facilitate the widespread adoption of nmrML and permit the facile generation of nmrML data from proprietary vendor formats. Bruker Corp. indicated willingness to incorporate the nmrML converter into their TopSpin software as nmrML file format export option. Although the benefits to individual users will become more evident as more software supports this open standard, users can already store and archive their NMR data in a persistent format, which stays readable in the long term. Users can extract NMR metadata into the ISA Tab format,[20] e.g., easing submissions to public databases such as MetaboLights. Data in local institutional repositories will gain value through eased reanalysis of old data with future state-of-the-art tools. Furthermore, users can integrate their data into workflow management systems, which eases repetitive processing tasks. Reproducibility and trustability of data is further increased by community data validation, e.g., in terms of minimal information coverage, and will result in increased data quality. The use of nmrML validators will allow users to check nmrML files with regard to consistency and content completeness. Together with ISA-Tab metadata validation, this will greatly contribute to overall quality assurance and traceability of NMR data. The nmrML standard also enables easier multicenter collaborations, e.g., allowing for an interoperable data exchange format when communicating with a regional NMR metabolomics center. It also eases comparison of results among different laboratories, e.g., for the purpose of standardization or SOP development. On the tool developers side, nmrML can save programmers' time and effort to write multiple parsers for all vendor formats. Given cross-communication between the MSI, PSI, and other standardization governance bodies, harmonized data standards will ease community integration, i.e., bridging over different technologies, e.g., by allowing MS and NMR data comparisons or even multiple omics investigations. This would pave the way for more integrative systems biology approaches.

Overall the nmrML specification and the expandable nmrCV will allow for a detailed standardized description of NMR workflow functionalities. The use of nmrML in workflow tools like tameNMR and the reuse of containerized workflow components in recombinable app libraries will allow NMR data processing to be more traceable and rerunnable in different (local or cloud) environments. The capture of selected basic metadata within the same nmrML file as the data eases pipeline development as it reduces the complexity of file tracking in Galaxy, as data moves between modules.

A recent survey (http://phenomenal-h2020.eu/home/wp-content/uploads/2016/09/Deliverable8.1.pdf) on data standards usage among the metabolomics community indicated that 13.5% of NMR practitioners are already using nmrML, about the same number of people indicating that they use JCAMP.

Further testing of the current XSD with diverse experimental configurations is required to increase coverage, fitness of purpose, and future flexibility. We hence welcome any community feedback and engagement via our email list at https://groups.google.com/forum/?hl=en#!forum/nmrml/join to improve and evaluate this first nmrML release. Remarks, suggested changes, and extension requests should be sent to info@nmrml.org or via our Git issue tracker. By standardizing data descriptions, nmrML and its accompanying nmrCV will help make NMR data *Findable, Accessible, Interoperable, and Reusable*, FAIR.[36] This is particularly relevant in light of the recent push by funding bodies to have scientists conduct and publish more reproducible research.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.7b02795.

> Serialization of the nmrML.xsd as Schema description (PDF)
>
> Documentary material with FAQ and nmrML tutorial (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: dschober@ipb-halle.de. Phone: +49 (0)345 5582 1476.
*E-mail: reza.salek@ebi.ac.uk. Phone: +44 (0)1223 48 4502.

ORCID ⬡

Daniel Schober: 0000-0001-8014-6648
Antonio Rosato: 0000-0001-6172-0368
Julian L. Griffin: 0000-0003-1336-7744

Author Contributions

D.S. drafted the manuscript, coordinated the nmrML updates, contacted NMR data repository developers and created tutorial material. D.S., D.J., M.W., and P.R.-S. implemented the nmrCV. D.S. and S.N. set up the semantic validator and created the mapping files. M.W. created and updated the nmrML.xsd, set up the Git and nmrML.org home pages, and wrote the Python parser. D.J. updated the nmrML.xsd and created and maintains the JAVA converter. D.J. created example nmrML files and coordinated NMRProcFlow interactions. J.A.C. created the predecessor nmrML XSD and an nmrCV predecessor. A. Marcu and J.R.G. deployed the nmrML Assign Web server, added assigned spectra for HMDB compounds, and coordinated Bayesil interactions. A. Moing and C.D. supervised the Java converter development, MeryB example generation, and provided feedback as wet-lab NMR and metabolomics database users. L.F.d.F., K.H., and R.M.S. helped integrating nmrML format with the MetaboLights repository. L.F.d.F. contributed to the initial version of the JAVA converter and created example nmrML files. P.R.-S. worked on and advised on CV and ontology reuse and coordinated the nmrML to ISA converter development with M.L. and R.M.S. T.M.D.E. and J.H. contributed the MATLAB parser and BATMAN advice. C. Ludwig, J.E., and U.L.G. were instrumental in aligning the initial XSD to BML-NMR repository needs. C. Luchinat and A.R. were initial driving forces overseeing the overall NMR data standards coordination. M.S.K. and I.A.L. worked toward rNMR integration. A.R.J. and A.G. were contributing to the NMR workflow tool tameNMR.

REFERENCES

(1) Bisson, J.; Simmler, C.; Chen, S.-N.; Friesen, J. B.; Lankin, D. C.; McAlpine, J. B.; Pauli, G. F. *Nat. Prod. Rep.* **2016**, *33* (9), 1028−1033.
(2) Hao, J.; Astle, W.; De Iorio, M.; Ebbels, T. M. D. *Bioinformatics* **2012**, *28* (15), 2088−2090.
(3) Ravanbakhsh, S.; Liu, P.; Bjorndahl, T. C.; Bjordahl, T. C.; Mandal, R.; Grant, J. R.; Wilson, M.; Eisner, R.; Sinelnikov, I; Hu, X.; Luchinat, C.; Greiner, R.; Wishart, D. S. *PLoS One* **2015**, *10* (5), e0124219.
(4) Jacob, D.; Deborde, C.; Lefebvre, M.; Maucourt, M.; Moing, A. *Metabolomics* **2017**, *13* (4), 36.
(5) Lewis, I. A.; Schommer, S. C.; Markley, J. L. *Magn. Reson. Chem.* **2009**, *47* (S1), S123−S126.
(6) Ludwig, C.; Günther, U. L. *BMC Bioinf.* **2011**, *12*, 366.
(7) Haug, K.; Salek, R. M.; Conesa, P.; Hastings, J.; de Matos, P.; Rijnbeek, M.; Mahendraker, T.; Williams, M.; Neumann, S.; Rocca-Serra, P.; et al. *Nucleic Acids Res.* **2013**, *41*, D781−D786.
(8) Sud, M.; Fahy, E.; Cotter, D.; Azam, K.; Vadivelu, I.; Burant, C.; Edison, A.; Fiehn, O.; Higashi, R.; Nair, K. S.; Sumner, S.; Subramaniam, S. *Nucleic Acids Res.* **2016**, *44* (D1), D463−D470.
(9) Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; et al. *Nucleic Acids Res.* **2013**, *41*, D801−D807.
(10) Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioannidis, Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; et al. *Nucleic Acids Res.* **2008**, *36*, D402−D408.
(11) Davies, A. N.; Lampen, P. *Appl. Spectrosc.* **1993**, *47* (8), 1093−1099.
(12) Sansone, S.-A.; Schober, D.; Atherton, H. J.; Fiehn, O.; Jenkins, H.; Rocca-Serra, P.; Rubtsov, D. V.; Spasic, I.; Soldatova, L.; Taylor, C.; Tseng, A.; Viant, M. R. *Metabolomics* **2007**, *3* (3), 249−256.
(13) Fiehn, O.; Robertson, D.; Griffin, J.; van der Werf, M.; Nikolau, B.; Morrison, N.; Sumner, L. W.; Goodacre, R.; Hardy, N. W.; Taylor, C.; et al. *Metabolomics* **2007**, *3* (3), 175−178.
(14) Rubtsov, D. V.; Jenkins, H.; Ludwig, C.; Easton, J.; Viant, M. R.; Günther, U.; Griffin, J. L.; Hardy, N. *Metabolomics* **2007**, *3* (3), 223−229.
(15) Salek, R. M.; Neumann, S.; Schober, D.; Hummel, J.; Billiau, K.; Kopka, J.; Correa, E.; Reijmers, T.; Rosato, A.; Tenori, L.; et al. *Metabolomics* **2015**, *11* (6), 1587−1597.
(16) Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Römpp, A.; Neumann, S.; Pizarro, A. D.; et al. *Mol. Cell. Proteomics* **2011**, *10* (1), R110.000133.
(17) Ludwig, C.; Easton, J. M.; Lodi, A.; Tiziani, S.; Manzoor, S. E.; Southam, A. D.; Byrne, J. J.; Bishop, L. M.; He, S.; Arvanitis, T. N.; Günther, U. L.; Viant, M. R. *Metabolomics* **2012**, *8* (1), 8−18.
(18) Whetzel, P. L.; Noy, N. F.; Shah, N. H.; Alexander, P. R.; Nyulas, C.; Tudorache, T.; Musen, M. A. *Nucleic Acids Res.* **2011**, *39* (Suppl), W541−W545.
(19) Helmus, J. J.; Jaroniec, C. P. *J. Biomol. NMR* **2013**, *55* (4), 355−367.
(20) Larralde, M.; Lawson, T. N.; Weber, R. J. M.; Moreno, P.; Haug, K.; Rocca-Serra, P.; Viant, M. R.; Steinbeck, C.; Salek, R. M. *Bioinformatics* **2017**, *33*, 2598−2600.
(21) Rocca-Serra, P.; Brandizi, M.; Maguire, E.; Sklyar, N.; Taylor, C.; Begley, K.; Field, D.; Harris, S.; Hide, W.; Hofmann, O.; et al. *Bioinformatics* **2010**, *26* (18), 2354−2356.
(22) Montecchi-Palazzi, L.; Kerrien, S.; Reisinger, F.; Aranda, B.; Jones, A. R.; Martens, L.; Hermjakob, H. *Proteomics* **2009**, *9* (22), 5112−5119.
(23) Bertsch, A.; Gröpl, C.; Reinert, K.; Kohlbacher, O. *Methods Mol. Biol.* **2011**, *696*, 353−367.
(24) Wishart, D. S. *Pharmacogenomics* **2008**, *9* (8), 1155−1162.
(25) Jewison, T.; Knox, C.; Neveu, V.; Djoumbou, Y.; Guo, A. C.; Lee, J.; Liu, P.; Mandal, R.; Krishnamurthy, R.; Sinelnikov, I.; Wilson, M.; Wishart, D. S. *Nucleic Acids Res.* **2012**, *40*, D815−D820.
(26) Guo, A. C.; Jewison, T.; Wilson, M.; Liu, Y.; Knox, C.; Djoumbou, Y.; Lo, P.; Mandal, R.; Krishnamurthy, R.; Wishart, D. S. *Nucleic Acids Res.* **2013**, *41*, D625−D630.
(27) Yokochi, M.; Kobayashi, N.; Ulrich, E. L.; Kinjo, A. R.; Iwata, T.; Ioannidis, Y. E.; Livny, M.; Markley, J. L.; Nakamura, H.; Kojima, C.; Fujiwara, T. *J. Biomed. Semantics* **2016**, *7* (1), 16.
(28) Ferry-Dumazet, H.; Gil, L.; Deborde, C.; Moing, A.; Bernillon, S.; Rolin, D.; Nikolski, M.; de Daruvar, A.; Jacob, D. *BMC Plant Biol.* **2011**, *11*, 104.
(29) Weber, R. J. M.; Lawson, T. N.; Salek, R. M.; Ebbels, T. M. D.; Glen, R. C.; Goodacre, R.; Griffin, J. L.; Haug, K.; Koulman, A.; Moreno, P.; et al. *Metabolomics* **2017**, *13* (2), 12.
(30) Goecks, J.; Nekrutenko, A.; Taylor, J.; The Galaxy Team. *Genome Biol.* **2010**, *11* (8), R86.
(31) Giacomoni, F.; Le Corguillé, G.; Monsoor, M.; Landi, M.; Pericard, P.; Pétéra, M.; Duperier, C.; Tremblay-Franco, M.; Martin, J.-F.; Jacob, D.; Goulitquer, S.; Thévenot, E. A.; Caron, C. *Bioinformatics* **2015**, *31* (9), 1493−1495.
(32) Jones, A. R.; Eisenacher, M.; Mayer, G.; Kohlbacher, O.; Siepen, J.; Hubbard, S. J.; Selley, J. N.; Searle, B. C.; Shofstahl, J.; Seymour, S. L.; et al. *Mol. Cell. Proteomics* **2012**, *11* (7), M111.014381.
(33) Walzer, M.; Qi, D.; Mayer, G.; Uszkoreit, J.; Eisenacher, M.; Sachsenberg, T.; Gonzalez-Galarza, F. F.; Fan, J.; Bessant, C.; Deutsch, E. W.; Reisinger, F.; Vizcaino, J. A.; Medina-Aunon, J. A.; Albar, J. P.; Kohlbacher, O.; Jones, A. R. *Mol. Cell. Proteomics* **2013**, *12* (8), 2332−2340.
(34) Bartels, C.; Xia, T. H.; Billeter, M.; Güntert, P.; Wüthrich, K. *J. Biomol. NMR* **1995**, *6* (1), 1−10.
(35) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. *J. Biomol. NMR* **1995**, *6* (3), 277−293.
(36) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; et al. *Sci. Data* **2016**, *3*, 160018.