ACS OMEGA

Article

Cite This: ACS Omega 2019, 4, 3280−3286

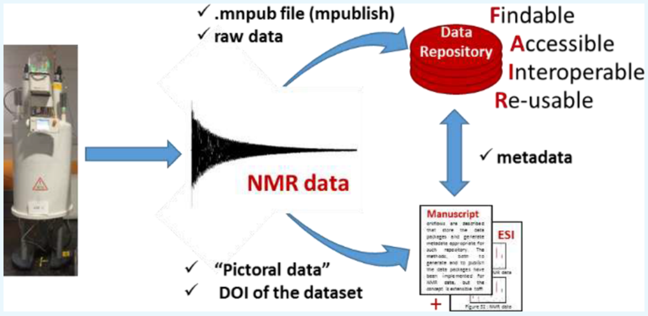
http://pubs.acs.org/journal/acsodf

Workflows Allowing Creation of Journal Article Supporting Information and Findable, Accessible, Interoperable, and Reusable (FAIR)-Enabled Publication of Spectroscopic Data

Agustin Barba,[†] Santiago Dominguez,[†] Carlos Cobas,*,[†] David P. Martinsen,[‡] Charles Romain,*,[§] Henry S. Rzepa,*,[§] and Felipe Seoane[†]

[†] Mestrelab Research, S.L., Feliciano Barrera 9B − Bajo, 15706 Santiago de Compostela, Spain
[‡] David Martinsen Consulting, Rockville, Maryland 20850, United States
[§] Department of Chemistry, MSRH, Imperial College London, 80 Wood Lane, London W12 0BZ, U.K.

Supporting Information



ABSTRACT: There is an increasing focus on the part of academic institutions, funding agencies, and publishers, if not researchers themselves, on preservation and sharing of research data. Motivations for sharing include research integrity, replicability, and reuse. One of the barriers to publishing data is the extra work involved in preparing data for publication once a journal article and its supporting information have been completed. In this work, a method is described to generate both human and machine-readable supporting information directly from the primary instrumental data files and to generate the metadata to ensure it is published in accordance with findable, accessible, interoperable, and reusable (FAIR) guidelines. Using this approach, both the human readable supporting information and the primary (raw) data can be submitted simultaneously with little extra effort. Although traditionally the data package would be sent to a journal publisher for publication alongside the article, the data package could also be published independently in an institutional FAIR data repository. Workflows are described that store the data packages and generate metadata appropriate for such a repository. The methods both to generate and to publish the data packages have been implemented for NMR data, but the concept is extensible to other types of spectroscopic data as well.

1. INTRODUCTION

Recent studies have examined the rationale for greater sharing of primary research data among scientists, including in chemistry.[1] A number of benefits have been suggested including research integrity and replicability and reuse by peers. Some prominent cases of data manipulation and fraudulent data have been uncovered in recent years,[2,3] the expectation is that upon publishing the primary research data associated with an experiment, it becomes more difficult to manipulate data in a way that cannot be detected. The ease with which data can be manipulated or manufactured varies by technique, and although some attempts can be detected by human inspection, others would undoubtedly escape even simple algorithmic techniques. Upon the discovery of manipulated data, the crystallography community recently recommended that CIF data can be accompanied by structure factors in an attempt to reduce fraud.[4] Structure factors had been considered of minimal importance once the CIF format for crystal structures became standard for publishing and sharing crystal data, but it is now realized that these parameters are less susceptible to manipulation than the processed data recorded in the CIF files. Because of the cultural norms in the crystallography community, the expected deposit of CIF files has resulted in relatively straightforward methods to store both CIF and structure factors, to validate the formats and the data using tools such as checkCIF, and to display the data in both publications and interactively via online publication.

Another proposed benefit of making primary data available is the utilization of artificial intelligence/machine learning (AI/ML) approaches to process data, to enhance existing methods for the prediction of properties of chemical structures, to develop new materials, to discover new drugs, and so on.[5] Ultimately, the goal of these AI/ML techniques would be to understand the science well enough to propose new scientific theories. Whether the latter is feasible, the use of AI/ML techniques requires a large quantity of data in an understandable and reusable format.

Archiving and sharing of analytical raw data are also pivotal for the development of new signal processing algorithms aimed at extracting information that has been elusive to existing

Received: October 30, 2018
Accepted: January 17, 2019
Published: February 14, 2019

methods. For instance, although the Fourier transform (FT) method has proved to be the gold standard for the processing of NMR data because of its computational performance and robustness, it also has well-known limitations such as poor resolving power, leakage artifacts (*i.e.* Gibbs oscillations), and phase and baseline distortions. There are presently many other alternatives to the FT method that can overcome these deficiencies or new ones that might be developed in the future. Those new methods could potentially uncover new information (i.e., new signals invisible to old methods) that can be used to, for example, prove or discard a chemical structure as inferred with different methods. Obviously, this can be only done if the primary acquired data is preserved. Furthermore, even raw data is generally stored with some loss of information. For example, NMR data is acquired by accumulating many scans to improve the signal-to-noise ratio. However, only the final averaged FID is kept, whereas all the individual transients are thrown away. Although this is usually not a concern, there are some advanced processing techniques that could be used to take advantage of the information contained within the individual scans.[6]

For experimental data other than that for crystal structures, instruments have the ability to export in standard formats. For spectroscopic data, JCAMP-DX is a veritable common export format for data sharing, although it has yet to achieve widespread adoption.[7] Although the format has some disadvantages in the nonstandard ways in which instrument manufacturers have added custom parameters, it is still useful for understanding the basic data. However, there are two disadvantages with using export formats for spectroscopic or instrumental data. First, the exported data may be processed data and hence subject to manipulation. Second, the steps to generate the data objects for publications either as selected lists of peaks for the experimental section in the body of the article or as figures in the body of the article or in the supporting information are labor intensive and result in some loss of information. Once those traditional publishable objects have been prepared, it would be an additional human effort to generate and organize the machine-readable data files for submission as supporting information. Although many publications allow submission of supporting information data files, most publications do not require the raw data. There have been few other incentives for researchers to do the extra work to prepare the data files for publication, but most do not. This situation is changing. For instance, a new NMR file format has been recently proposed, NMReData,[8] although its scope is limited to structure characterization of small molecules (i.e., NMR assignments). This format is essentially an extension of the existing structure data format (SDF) and includes the so-called NMR record, which is a compressed folder that, in addition to the NMReDATA file, contains all related 1D and 2D spectra including the raw data such as the FIDs and all the acquisition and processing parameters.

Funding agencies are starting to issue policies that require researchers to preserve and share the research data collected during the course of a research grant. Government funding agencies, for example in the United States, U.K., EU, and Australia, as well as private foundations, such as the Howard Hughes Medical Institute, Wellcome Trust, and Gates Foundation, are mandating that research data in support of journal articles be published in a reusable form. In addition, while data repositories have been common in discipline specific areas for many years and in fact have appeared and disappeared, more generic data repositories such as Figshare,[9] Data Dryad,[10] and Dataverse[11] are now available for researchers to store their experimental details. Data journals are also being established to allow researchers to publish and get credit for data collection. These include, for example, Data Science,[12] Nature Scientific Data,[13] MethodsX,[14] and SoftwareX.[15] When publishing in these data journals, researchers can get credit in the form of a citation for data and/or software publication and at the same time also get credit for publishing a traditional article describing the results and conclusions of the research. Community norms are still emerging as to how these publications will be treated in terms of scientific recognition.

The rigor with which researchers actually comply with funding agency mandates is now a subject of much debate in the research data community. Neylon has proposed that the focus shifts from development of data sharing policies to culture change.[16] His argument is that although policies are important, they focus on changing individual behavior. Given the barriers to data sharing, such as lack of understanding of how to go about sharing data and the need to commit extra human and funding resources to sharing data, a change in culture to promote data sharing would help to motivate researchers to change their individual behavior.

Even though one might agree that an ideal state would be for all primary research data in support of scientific publication to be made publicly available, there are still constraints even in a digital environment with vast amounts of inexpensive storage available. A typical organic synthesis article might include three or four NMR experiments for each of 30−40 substances. Generating the supporting information file consisting of graphical representations of the NMR data with headings, structure diagrams, and sampling parameters is particularly time-consuming. Furthermore, such documents can grow to be very large, there are examples now of supporting information approaching or even exceeding 1000 pages.[17] An alternative approach would facilitate ways in which the researcher can organize and store the FID files as a package for submission as a dataset or a fileset. Because the FID files are the basis for both the data package and the graphical SI data, a workflow that allowed for the generation of both forms of SI at the same time would remove one of the barriers to submission of full/raw data. Although this approach is consistent with the general concept that content should be stored in canonical form and rendered as appropriate at the time of delivery to the user,[18] it also achieves one of the hallmarks of publication—that the reader has the option if they wish for a static representation of what the author intended, while at the same time, the interested reader, whether human or machine, has access to both a more interactive and/or a lossless version of the same data.

2. RESULTS AND DISCUSSION

The Mpublish project aims to address the twin challenges of increasing the findability, accessibility, interoperability, and reusability (FAIR) of spectroscopic research data[19] and facilitating the preparation of such data for publication using automated workflows. The first two sections will describe two key features of the Mpublish project, that is,

- an automated workflow for the author to prepare the data for publication and

DOI: 10.1021/acsomega.8b03005

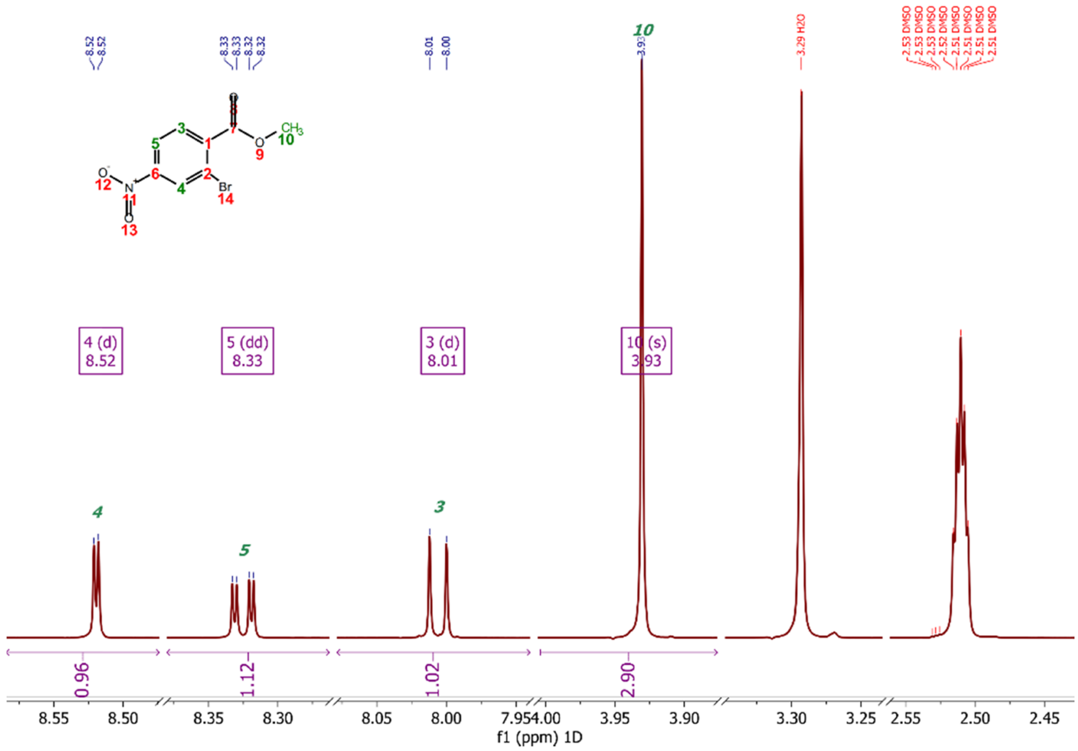


Figure 1. ^{1}H NMR spectrum ready for publication.

- a signing system for the publisher, which allows the reader to review the data while addressing reuse and interoperability of the data.

The last section will describe how deposition in the Imperial College Data Repository ensures data FAIRness.

2.1. Preparation of Primary Research Data for Publication Using an Automated Mpublish Workflow. The Mpublish author workflow has been implemented as a plugin of the Mnova software platform[20] and is now being integrated into the Mgears automation system within Mnova. Although this limits its use to researchers and scientists that have licensed access to the Mnova software for the processing of their NMR, LC/GC/MS, and optical spectroscopic primary data, the Mnova software does have the advantage that it is widely adopted by the chemistry research community.

The Mpublish workflow generates a complete package ready to be submitted to a specific journal from primary datasets (.mnova or raw data files) that has been satisfactorily processed, analyzed, and annotated within the Mnova platform (see further details on the Mpublish author workflow in Supporting Information and Figure S1). Using the Mpublish user interface, the author will be able to

- easily select the data they want to publish, either by following a system of file masks, which match the organization of the data on disk, or by following a system of multiselection or a specific metadata field on the spectral database (Figure S2),
- choose a journal template,
- enter author information (name, organization, etc.).

Subsequent execution of the Mpublish author functionality will run an automated process (implemented as an Mnova script based on ECMAScript)[21] and will generate

- individual spectrum figures with the corresponding file title, acquisition parameters, and peak list (multiplicity, chemical shifts) templated to the selected journal (Figure 1),
- an ODT (open document text) file including all the processed data (figure, peak list, etc.), and
- a folder containing the primary raw data used to generate these processed and analyzed data (subfolder and files renamed and reorganized to identify the datasets and match them with the titles included in the automatically generated aforementioned ODT zip file).

The automatically generated ODT file[22] can be edited and used to prepare the final supporting information (SI) file intended for submission along with the manuscript and along with inclusion of experimental procedures and other appropriate data. The preparation of NMR data for the SI document is thus greatly facilitated by use of the Mpublish automation.

Overall, the advantages of the Mpublish workflow are

- simultaneous generation of a folder including the NMR data (zip file) along with an SI document (ODT file) presenting and describing the data (figure, peak list) in line with a selected journal template,
- easy modification of the SI document (to satisfy referee comments, to change the template) by using the Mpublish user interface, which will allow to select/remove data and modify author information and journal template, and
- generation of primary data available in machine-actionable form, enabling data mining and aggregation.

2.2. Signing System Allowing Reusability and Interoperability of Primary Research Data. A second aspect of the Mpublish project is the ability for the publisher (e.g., the data repository owner) to cryptographically "sign" a dataset, which will be possible for the reader to review using

Mnova software, without the cost associated with the software license. This will allow the reader to analyze the research data more extensively than with a static PDF.

2.2.1. Publisher's Perspective. To enable this feature, the publisher has to provide Mestrelab with a public key, which will be used to sign the data and compose an Mnova publication text file (mnpub). The mnpub file will point to the primary research data (mnova or zip file) and hold both the signature and the public key information. An mnpub file will have to be created for each dataset that the publisher would like to make freely available for review using Mnova software. To this end, the publisher must use an RSA key pair: (i) the public key is sent to MestreLab to allow them to generate a certificate file (.mncrt), which is then made available to the publisher and (ii) the publisher uses their private RSA key to "sign" the dataset, which generates a signature. Overall, the mnpub file will be composed with the signature, the certification file (mncrt) and a URL to the signed dataset (see example of mnpub file).[22] The process is simple and can be executed with very little overhead by the publisher (see Supporting Information for further details). The procedure can be fully automated as a part of the submission process and will not require any additional input from the authors. For example, as described in the last Section 2.3, the procedure can be implemented in a data repository submission web service where the authors only need to upload an mnova file or a zip file, with the mnpub file generated automatically by the web service.

2.2.2. Reader's Perspective. The reader can freely download Mnova software (a purchased license is only needed for general use) and open the .mnpub file. This file will be read by Mnova and verified with the public key information contained in the mnpub file. If the verification is successful, the dataset pointed by the URL field of the mnpub file will be retrieved for use within Mnova with full functionality, without the need for a license. The reader will be able to review, reprocess, and reanalyze the research primary data, benefiting from all the functionality available within the Mnova software package, albeit only for the digitally signed dataset. This effectively makes such signed data open for review and reproduction independently of access to licensed (in this example analytical chemistry) software applications. The impact of such submission of primary research data associated with articles is that it becomes available for detailed review by a wide range of stakeholders such as article reviewers, editors, data curators, publishers, peers, the general public and artificial intelligence, and machine learning systems. Metadata associated with the primary research data as generated by the original instrument and then uploaded to the relevant repository also facilitates subsequent automated or workflow machine processing of the data on as large an aggregated scale as needed.

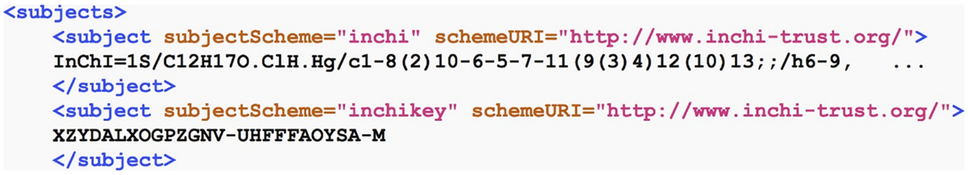
2.3. FAIR-Enabling Mpublish Data[23] by Deposition in a Data Repository. Although the Mpublish procedure provides an excellent mechanism for making complete NMR data readily and appropriately available, the procedure also needs to include mechanisms for ensuring it adheres as completely as possible to the FAIR data principles (findable, accessible, interoperable, and reusable).[11] Indeed, enclosing the data with the rich toolkit provided by Mnova addresses in large measure the I of FAIR. The other attributes can be accomplished by deposition into a data repository, a process that also includes generating metadata and registering it appropriately to ensure standardized publication. Here, we look at the steps taken to achieve this as implemented in the Imperial College Mpublish pilot project.[24,25]

2.3.1. Data Granularity. The first task is to decide upon appropriate granularity for the metadata. Mnova files (generated from Mnova 11.0+ or using the automated Mpublish workflow described above) can be generated for entire collections of spectra in a single file containing many molecules, or they can be restricted to just a single molecule, containing if necessary spectra for different experiments and nuclei. We strongly favor the latter approach, because metadata for single molecules can be generated far more easily and transparently, and we believe also more usefully than for molecule collections. This is largely based on the availability of the InChI molecular identifier[26] and the possibility of generating it for a specific molecule using a simple machine processable workflow procedure. We in fact apply it using OpenBabel[27] in which every file submitted to the data repository is automatically screened by OpenBabel, and if molecular content is identified, then both an InChI string and an InChI key are generated from it.

In our recommendations, every NMR spectrum submitted for deposition is accompanied by a suitable molecular connection table in the form of standard files such as the MDL molfile (.mol) or the ChemDraw .cdx or .cdxml files (Figure S5). This is essential if raw filesets such as folder collections produced directly by instruments are submitted (in the form of a compressed ZIP archive). Connection tables can also be embedded in the Mnova files resulting from the primary processing of such instrumental data, but currently, OpenBabel is not capable of extracting such information from this format and hence in practice, these too require separate inclusion of a molecule description file. We suggest that the best practice is in fact to provide both the raw instrument archive as a .zip archive (for reusers who choose to use different software) together with the Mnova superset, which includes both the FID and all analysis of the far more "user-friendly" frequency domain spectrum (including solvent reference, annotations, assignments, phase and baseline correction, etc.). Optionally, other processed data formats such as the open standard JCAMP-DX or NMReData noted above can also be provided for further interoperability and even a visual format such as PDF to allow readers the option of access to the traditional format currently found in most supporting information files.

2.3.2. Metadata. Further metadata is also generated in a workflow manner (Figure S5), as implemented in the data repository.[25] A minimal set would include the following:

1. The ORCID identifier for the depositor along with any further such identifiers for appropriate collaborators. For the repository,[28] we use ORCID as an intrinsic part of the login process and so its inclusion in the metadata records is automatic, as is the affiliation of the depositor.
2. The deposition date and time are also recorded automatically, together with the identity of the formal publisher of the data, which can be the research institution as in the example here.
3. A title and description are recorded, these being part of the Dublin Core metadata elements.[29] Although these can have any value, we favor including the systematic name of the molecule as part of the title field and a brief description of the experimental procedures used to prepare the compound in the description field.



```
<subjects>
    <subject subjectScheme="inchi" schemeURI="http://www.inchi-trust.org/">
    InChI=1S/C12H17O.ClH.Hg/c1-8(2)10-6-5-7-11(9(3)4)12(10)13;;/h6-9,   ...
    </subject>
    <subject subjectScheme="inchikey" schemeURI="http://www.inchi-trust.org/">
    XZYDALXOGPZGNV-UHFFFAOYSA-M
    </subject>
```

Figure 2. <subject> element deriving from the DataCite V 4.1 schema, populated with InChI metadata.

4. We include a license for reuse and if necessary, reanalysis, in our case the CC0 creative commons license,[30] this directly addressing the R of FAIR.
5. A much less frequently applied but what we consider crucial metadata component is a declaration of a so-called resource map enabling object reuse and exchange (ORE),[31] which addresses the A of FAIR. This in turn would allow any machine-driven automated procedures for retrieving the data files on a scale larger than a human would adopt.
6. A media type is also declared for specific files in any uploaded fileset. For the Mpublish project, these include chemical/x-mnova for the basic Mnova files and chemical/x-mnpub, which is automatically generated using the cryptographic keys, which control the single-use license on which Mpublish is based.

Once assembled, the metadata is then formulated against the DataCite metadata schema (V 4.1).[32] The metadata specific to the molecule is accommodated in the <subjects> element of the schema, an extensible component where subject-specific terms can be appropriately defined. An example is shown in Figure 2 where the subjectScheme attribute is included to avoid clashes with any other domains, which may accidentally use the same vocabulary. In this example, an InChI identifier is included. To assist molecule discovery, we normally include data for a single molecule so that only one InChI identifier need be included and deprecate including data for multiple different molecules in the same dataset.

With this approach, control over the dictionary used in the <subject> element is highly desirable, because uncontrolled and undeclared extensions would inhibit the deployment of discovery tools. To this end, a Data Interest Group in chemistry has been established and regular meetings scheduled to promote communal agreement.[33] The first proposal includes that shown in Figure 2. Other more NMR specific extensions could in future include, for example, a richer set of NMR-related metadata declaration, such the NMR nucleus studied, the accepted name of any pulse sequence used in the experiment, solvent, temperature, and so forth. Workflows specific to extract this information automatically would also need to be developed. The repository used to construct this demonstrator also supports the concept of hierarchical collections, which serve to organize the data associated with a project into subcollections. Subject element metadata for such top-level collections would not normally be included if they contain no explicit molecule-based datasets.

The DataCite registration API is then used to register the metadata with DataCite,[34] a global aggregator of such content. In return, DataCite responds with a persistent identifier for the metadata taking the form of a Digital Object Identifier or DOI. This is deliberately consistent with the DOI schemes used for journal articles, allowing the development of synergies between the two. The metadata itself can be downloaded as an XML file from the landing page for any collection or dataset using a link, which includes the DOI of the data: https://data.datacite.org/application/vnd.datacite.datacite+xml/10.14469/hpc/4751

If a more readily human-readable presentation is desired; then, the option of a style-based transform of this file is available. We also anticipate text-based presentations of the metadata will become available directly from the equivalent DataCite page, as invoked by: https://search.datacite.org/works?query=id:10.14469/hpc/4751

An example of early use of Mpublish deposition in a published article[18] illustrates how the article can formally cite the data (as references 35[35] and 36[18] in this case) and in turn how the article can itself be cited via the metadata describing the data thus establishing bidirectional linking.

2.3.3. Discovery and Findability. The final aspect of FAIR-enabling the Mpublish data addresses F, the findability. The registration of the metadata with DataCite allows the search interface provided there to be used to discover data with the appropriate properties. Machine processable examples are included below to illustrate this.

1. https://search.datacite.org/works?query=media.media_type:chemical/x-mnpub* uses the media type declaration to discover all datasets registered with this type.
2. https://search.datacite.org/works?query=media.media_type:chemical/x-mnpub*+AND+subjects.subjectScheme:inchikey+AND+subjects.subject:XZYDALXOGPZGNV-UHFFFAOYSA-M+AND+media.media_type:chemical/x-gaussian* shows how the media type can be combined with two Boolean AND operations to restrict the search to datasets where a specified InChIKey has also been registered and is also accompanied by a Gaussian computation output. This search only specifies that both media types must be present in the fileset but not necessarily that the two are related.
3. https://search.datacite.org/works?query=contributors.nameIdentifiers.nameIdentifier:*0000-0002-8635-8390+AND+media.media_type:chemical/x-mnpub* illustrates how the search can be combined with a restriction to those Mpublish datasets associated with a specific researcher as defined by their ORCID.
4. We note that the above syntax is not human-friendly and hence the need for development of more accessible forms of the search interface that can take advantage of these rich searches.
5. https://app.dimensions.ai/discover/publication?search_text=10.14469 allows the discovery of datasets assigned the DOI prefix (Imperial College) linked to journal publications registered with the Crossref metadata store, as one example of the synergies noted above.

The inclusion of an ORE resource map in the collected metadata allows the results of a search of the above type to then be scripted to allow direct retrieval of specific files on as large a scale as necessary. It is important to note that knowing just the DOI of any published fileset does not automatically enable this, a DOI normally points to what is called a landing page from which further parochial navigation is required to allow access to individual files is needed. Normally, a human performs this navigation (because it is rarely standard in any predictable sense), but with data, it is essential to have a formal declaration that a machine can traverse automatically. Thus, the A of FAIR ideally relates both to visual access by a human and also to applications such as data mining by machines. One example[18] of the use of an ORE map to automatically retrieve and display data based only on knowledge of the DOI and the desired media type is based on the JSmol molecular visualizer. Such a feature could in principle also be developed for more specific spectroscopic tools such as Mnova.

2.3.4. FAIR Compliance. FAIR compliance can be evaluated using an objective, automated, and community-governed framework.[36] The evaluation[37] for the repository used here[25] is based on the features described above. We recommend that any repository providing access to FAIR-enabled data such as described here should be submitted to such an evaluation.

3. CONCLUSIONS

The Mpublish tools available within the Mnova NMR analysis program allow the workflow generation of publication-ready supporting information (SI) files, which can be submitted to a publisher and/or a data repository. Both the original primary/raw data and the generated SI files can be integrated with cryptographic license keys to allow access to individual datafiles without the requirement of first obtaining full commercial licenses. Further workflows can be used to generate a set of standardized metadata describing the datasets. The registration of such metadata with a global aggregating agency such as DataCite in return for association with a persistent identifier (a DOI) can transform the data into a rich resource where all four attributes of FAIR data are at least partially addressed. These include the provenance of the data and unique descriptors of the molecule associated with the spectra to enhance findability via rich searches of the indexed metadata. A metadata resource map allows rich (machine) access, a toolkit facilitates interoperability, information is included in the metadata about any associated data relating to other aspects such as computational simulations and models and an appropriately declared license facilitates reuse. We suggest this model for spectroscopic publication could serve as a starting point for extension to many other forms of molecular spectroscopy and instrumentally generated primary or raw data.

Although ideally FAIR data might imply software agnostic and software independent tools, realistically the generation of FAIR data must take into account tools available now for researchers to prepare and publish research data. The methods described above do not purport to be a complete solution, they are merely an attempt to suggest some usable tools that can achieve the goals of FAIR data now for a specific type of data and to describe an approach, which could be extended to other types of data by other researchers in related domains.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acsomega.8b03005.

Mpublish workflows, Mpublish specification, and FAIR data collection available free of charge at DOI: 10.14469/hpc/4751 (see ref 22) (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: carlos@mestrelab.com (C.C.).
*E-mail: c.romain@imperial.ac.uk (C.R.).
*E-mail: rzepa@imperial.ac.uk (H.S.R.).

ORCID

Charles Romain: 0000-0002-1851-8612
Henry S. Rzepa: 0000-0002-8635-8390

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

C.R. thanks Imperial College London for his Junior Research Fellowship.

■ REFERENCES

(1) McAlpine, J. B.; Chen, S.-N.; Kutateladze, A.; MacMillan, J. B.; Appendino, G.; Barison, A.; Beniddir, M. A.; Biavatti, M. W.; Bluml, S.; Boufridi, A.; Butler, M. S.; Capon, R. J.; Choi, Y. H.; Coppage, D.; Crews, P.; Crimmins, M. T.; Csete, M.; Dewapriya, P.; Egan, J. M.; Garson, M. J.; Genta-Jouve, G.; Gerwick, W. H.; Gross, H.; Harper, M. K.; Hermanto, P.; Hook, J. M.; Hunter, L.; Jeannerat, D.; Ji, N.-Y.; Johnson, T. A.; Kingston, D. G. I.; Koshino, H.; Lee, H.-W.; Lewin, G.; Li, J.; Linington, R. G.; Liu, M.; McPhail, K. L.; Molinski, T. F.; Moore, B. S.; Nam, J.-W.; Neupane, R. P.; Niemitz, M.; Nuzillard, J.-M.; Oberlies, N. H.; Ocampos, F. M. M.; Pan, G.; Quinn, R. J.; Reddy, D. S.; Renault, J.-H.; Rivera-Chávez, J.; Robien, W.; Saunders, C. M.; Schmidt, T. J.; Seger, C.; Shen, B.; Steinbeck, C.; Stuppner, H.; Sturm, S.; Taglialatela-Scafati, O.; Tantillo, D. J.; Verpoorte, R.; Wang, B.-G.; Williams, C. M.; Williams, P. G.; Wist, J.; Yue, J.-M.; Zhang, C.; Xu, Z.; Simmler, C.; Lankin, D. C.; Bisson, J.; Pauli, G. F. The value of universally available raw NMR data for transparency, reproducibility, and integrity in natural product research. *Nat. Prod. Rep.* **2019**, *36*, 35−107.

(2) Harrison, W. T. A.; Simpson, J.; Weil, M. Editorial. *Acta Crystallogr. E* **2010**, *E66*, e1−e2.

(3) Borrell, B. Fraud rocks protein community. *Nature* **2009**, *462*, 970.

(4) Larsen, S.; Kostorz, G. *Publication standards for crystal structures.* 2011. https://www.iucr.org/home/leading-article/2011/2011-06-02 (accessed Jul 7, 2018).

(5) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547−555.

(6) Taylor, H. S.; Haiges, R.; Kershaw, A. Increasing Sensitivity in Determining Chemical Shifts in One Dimensional Lorentzian NMR Spectra. *J Phys. Chem. A* **2013**, *117*, 3319−3331.

(7) Lampen, P.; Lambert, J.; Lancashire, R. J.; McDonald, R. S.; McIntyre, P. S.; Rutledge, D. N.; Frohlich, T.; Davies, A. N. An Extension to the JCAMP-DX Standard File Format, JCAMP-DX V.5.01. *Pure Appl. Chem.* **1999**, *71*, 1549−1556.

(8) Pupier, M.; Nuzillard, J.-M.; Wist, J.; Schlörer, N. E.; Kuhn, S.; Erdelyi, M.; Steinbeck, C.; Williams, A. J.; Butts, C.; Claridge, T. D. W.; Mikhova, B.; Robien, W.; Dashti, H.; Eghbalnia, H. R.; Farès, C.; Adam, C.; Kessler, P.; Moriaud, F.; Elyashberg, M.; Argyropoulos, D.; Pérez, M.; Giraudeau, P.; Gil, R. R.; Trevorrow, D.; Jeannerat, D. NMReDATA, a standard to report the NMR assignment and parameters of organic compounds. *Magn. Reson. Chem.* **2018**, *56*, 703−715.

(9) *Figshare.* https://figshare.com (accessed Oct 25, 2018).

(10) *DRYAD.* https://datadryad.org (accessed Oct 25, 2018).

(11) *The Dataverse Project*. https://dataverse.org (accessed Oct 25, 2018).

(12) *EPJ Data Science*. https://epjdatascience.springeropen.com (accessed Oct 25, 2018).

(13) *Scientific Data*. https://www.nature.com/sdata/ (accessed Oct 25, 2018).

(14) *MethodsX*. https://www.journals.elsevier.com/methodsx (accessed Oct 25, 2018).

(15) *SoftwareX*. https://www.journals.elsevier.com/softwarex (accessed Oct 25, 2018).

(16) Neylon, C. Building a Culture of Data Sharing: Policy Design and Implementation for Research Data Management in Development Research. *Res. Ideas Outcomes* **2017**, *3*, No. e21773.

(17) Lopchuk, J. M.; Fjelbye, K.; Kawamata, Y.; Malins, L. R.; Pan, C.-M.; Gianatassio, R.; Wang, J.; Prieto, L.; Bradow, J.; Brandt, T. A.; Collins, M. R.; Elleraas, J.; Ewanicki, J.; Farrell, W.; Fadeyi, O. O.; Gallego, G. M.; Mousseau, J. J.; Oliver, R.; Sach, N. W.; Smith, J. K.; Spangler, J. E.; Zhu, H.; Zhu, J.; Baran, P. S. Strain-Release Heteroatom Functionalization: Development, Scope, and Stereospecificity. *J. Am. Chem. Soc.* **2017**, *139*, 3209−3226.

(18) Clarke, J.; Bonney, K. J.; Yaqoob, M.; Solanki, S.; Rzepa, H. S.; White, A. J. P.; Millan, D. S.; Braddock, D. C. Epimeric Face-Selective Oxidations and Diastereodivergent Transannular Oxonium Ion Formation Fragmentations: Computational Modeling and Total Syntheses of 12-Epoxyobtusallene IV, 12-Epoxyobtusallene II, Obtusallene X, Marilzabicycloallene C, and Marilzabicycloallene D. *J. Org. Chem.* 2016, *81*, 9539−9552 and the accompanying interactive FAIR data table at DOI: 10.14469/hpc/1248.

(19) 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.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; Lei, J., van der; Mulligen, E., van; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018.

(20) *Mnova, Version 12.0.3*. Mestrelab Research S.L.: Santiago de Compostela, Spain. www.mestrelab.com/mnova (accessed Oct 2018).

(21) *Ecma International*. https://www.ecma-international.org/ (accessed Oct 25, 2018).

(22) Dominguez, S.; Cobas-Gomez, J. C.; Martinsen, D. P.; Rzepa, H. S.; Romain, C. Workflows allowing creation of journal article Supporting Information and FAIR-enabled publication of Spectroscopic data. *Imperial College Data Repository*. 2018, DOI: 10.14469/hpc/4751.

(23) *FAIRsharing*. https://fairsharing.org (accessed Oct 25, 2018).

(24) McLean, A.; Romain, C.; Bakewell, C.; Harvey, M. J.; Rzepa, H. S. Demonstration of Professional Preview of FAIR (Findable, Accessable, Inter-operable and Re-usable) NMR Data files using Mnova and Mpublish. *Imperial College Data Repository*. **2016**, DOI: 10.14469/hpc/1053 (accessed Oct 25, 2018).

(25) Harvey, M. J.; McLean, A.; Rzepa, H. S. A metadata-driven approach to data repository design. *J. Cheminf.* **2017**, *9*, 4.

(26) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminf.* **2015**, *7*, 23.

(27) Guha, R.; Howard, M. T.; Hutchison, G. R.; Murray-Rust, P.; Rzepa, H. S.; Steinbeck, C.; Wegner, J. K.; Willighagen, E. The Blue Obelisk—Interoperability in Chemical Informatics. *J. Chem. Inf. Model.* **2006**, *46*, 991−998.

(28) *Imperial College Data Repository*. Registry of Research Data Repositories. DOI: 10.17616/R3K64N (accessed Jul 20, 2018).

(29) *Dublin Core Metadata Initiative*. http://dublincore.org/schemas/ (accessed Jul 23, 2018).

(30) *Creative Commons*. https://creativecommons.org/publicdomain/zero/1.0/ (accessed Jul 23, 2018).

(31) *Open Archives Initiative Object Reuse and Exchange*. http://www.openarchives.org/ore/1.0/datamodel (accessed Jul 23, 2018).

(32) *DataCite Metadata Working Group*. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.1. DataCite e.V. 2017, DOI: 10.5438/0014.

(33) *Metadata Recommendations for DataCite Registration*. https://sites.google.com/view/digchem/datacite-recommendations (accessed Jul 23, 2018).

(34) *DataCite*. https://datacite.org (accessed Jul 23, 2018).

(35) Clarke, J.; Bonney, K. J.; Yaqoob, M.; Solanki, S.; Rzepa, H. S.; White, A. J. P.; Millan, D. S.; Braddock, D. C. *Imperial College Data Repository*. 2016, DOI: 10.14469/hpc/1116. The sub-collection that relates specifically to FAIR NMR data in Mpublish form. DOI: 10.14469/hpc/1267.

(36) Wilkinson, M. D.; Dumontier, M.; Sansone, S.-A.; da Silva Santos, L. O. B.; Prieto, M.; Gautier, J.; McQuilton, P.; Murphy, D.; Crosas, M.; Schultes, E. Evaluating FAIR-Compliance through an Objective, Automated, Community-Governed Framework. 2018, *bioRxiv*. 418376, DOI: 10.1101/418376.

(37) *Imperial College Research Data Repository*. DOI: 10.25504/fairsharing.letkjt (accessed Dec 12, 2018).