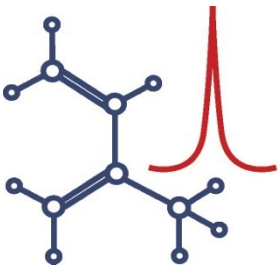


FAIR Chemical Data Publishing Guidelines Workshop Chemical Structures and Spectra

NSF FAIR Chemical Data Publishing Guidelines Workshop on Chemical Structures and Spectra: Major Outcomes and Outlooks for the Chemistry Community

Vincent F. Scalfani[†]* **and Leah R. McEwen**[†][†]*

[†]The University of Alabama, Tuscaloosa, AL, USA

[†][†]Cornell University, Ithaca, NY, USA

*Correspondence to: vfscalfani@ua.edu and lrm1@cornell.edu

March 24, 2020

Abstract

The National Science Foundation Office of Advanced Cyberinfrastructure (NSF-OAC) funded a workshop in March 2019 focused on advancing the sharing of machine-readable chemical structures and spectra. Around 40 stakeholders from the chemistry, chemical information, and software communities took part in the two-day workshop entitled “FAIR Chemical Data Publishing Guidelines for Chemical Structures and Spectra.” Major topics discussed included publishing data workflows and guidelines, FAIR criteria/metadata profiles, value propositions, a publisher implementation pilot, and community support and engagement. This report summarizes the workshop conversations, major outcomes, and target areas for further activities. Primary outcomes from the workshop include identification of key metadata elements for sharing machine-readable structures and spectra, a sample of concise author guidelines, and a publisher proposal to accept enhanced supporting information files including these data types and associated metadata alongside articles. Selected target areas for further activities include the creation of author file and metadata packaging tools to facilitate easy compilation of data, and increased training for stakeholders specifically in the generation and handling of machine-readable file formats. We conclude this report with our outlooks and highlight several related community efforts initiated after the workshop.

Introduction

Application of FAIR to compilation of chemical data

Chemical data is regularly reported to support the characterization of compounds, molecular forms and transformations. Despite thorough indexing of the literature, it remains difficult to find and directly access many types of chemical data. Most chemical data enter the published space in analog form (i.e., human-readable) and require manual curation to integrate into existing tools. Extracting research data from published materials by either manual or automated approaches can be resource intensive, error prone, and costly [1-2]. In addition, most databases operate in siloed environments, which limits discovery and interoperability between databases. As the volume of chemical information continues to expand, unless we can scale and enrich our indexing processes, discovery and utility will diminish. Improvements in standards and protocols for processing data, and automation of ingest and integration routines will facilitate scalability of searching and accessing data across collections and tools. However, these activities depend on the availability of machine-readable data at the original source.

The sharing of research data in machine-readable form is an increasing expectation among funders of scholarly and scientific research [3]. Much has been commented on the cultural challenges that sharing data as first class output presents to current practices in dissemination of research [4]. As service professionals who work with researchers in a variety of capacities in support of the research life cycle, we wanted to directly engage the chemistry community in considering this question. The FAIR Data Principles provide a good starting point for understanding what is required on a technical level to enable data to be effectively shared and (re)-used in the digital environment [5]. Discerning what existing tooling and approaches may be available to support the needs of both upstream data generators as well as downstream data users was a major goal of the NSF FAIR Chemical Data Publishing Guidelines Workshop on Chemical Structures and Spectra.

Where to start: where are the data, where is the intellectual value?

The synthesis of chemical compounds and their discovery in natural sources is a backbone of chemistry research, with over 160 million unique compounds reported in the CAS REGISTRY and 100 million in PubChem at the time of this writing [6-8]. This contribution to scientific knowledge involves the collection and communication of various types of measurement and characterization data to support the claims made in manuscripts. Data classes might include spectra (e.g., NMR, HRMS, IR, XRD), elemental analysis, melting point, and others as specified for various journals. The current practice in most cases is to assemble these data as Supporting Information (SI) along with experimental details [9]. SI files provide logically organized profiles of information readily understood and openly accessible by knowledgeable human readers. However, existing chemistry databases are not integrated into publishing workflows and spectral related data remain largely embedded in PDF documents, inaccessible to machine readers and often not of sufficient quality for human interpretation.

Characterization data are tied to the compounds that they represent, whether discrete small molecules, extended solids, or other substance materials. From these data is derived information about the chemical structure, configuration and other properties. The bulk of published material in chemistry is organized by chemical structure as a logical point of discovery and reuse of literature and data. Including structures in machine-readable forms along with articles and datasets would facilitate automated chemical indexing, automated validation and reuse of chemical data. However, similar to spectral data, chemical structures are commonly conveyed as static figures in the literature [8].

This represents the current “state of the art” in sharing chemical data in organic chemistry research. Given both the prevalence of spectroscopic data in chemistry and the existence of a regular practice for sharing, albeit analog, this scenario was selected as a tractable use case for considering the application of FAIR to chemistry data. Spectroscopic data and chemical structures are also widely generated and used in other subdisciplines and sectors of chemistry and other domains who analyze chemical samples. Further analysis would be facilitated with access to raw data and files at higher resolution than diagrams currently available in SI files. Understanding and articulating the needs, opportunities, challenges and roles of stakeholders to develop more FAIR approaches to these common data types will facilitate dissemination of a critical mass of chemical data and broader engagement with the community to share data more generally.

Where to start: what technical motifs exist, what tooling can be repurposed?

Establishing guidelines to support FAIR chemical data sharing will necessarily be an iterative process and should build on previous experience and best practices emerging in other fields as much as is plausible. Advances in instrumentation, data analysis and downstream applications in cheminformatics provide a number of machine-enabled motifs for both chemical structures and spectral data that could potentially be repurposed for supporting sharing of machine-readable research data up front. Standard criteria for data exchange have been established and adopted in closely aligned fields such as crystallography that chemists and publishers are familiar with using [10]. Workflows for depositing data and metadata have been well mapped for a number of repositories that can serve as informative models [e.g., 7, 11-12].

Balancing the benefits of streamlined end-to-end workflows such as those developed in more focused use cases around a single repository or institution, with the pragmatic need to engage more scientists and stakeholders by lowering barriers, the workshop discussion topics were designed with two general strategies in mind – to consider what can be done now building on current infrastructure to start the transition of practices from handling primarily analog outputs towards managing digital outputs; and to surface areas where further work needs to be done collectively across the community to support FAIR workflows, such as enhancements to technical standards or training in working with machine readable files.

This workshop report summarizes the discussions, major outcomes, and outlooks based on materials, notes and comments from workshop participants as well as cited literature. This

synthesis represents the perspectives of the authors and does not necessarily reflect the specific views of individual participants or affiliated organizations. The interested reader is encouraged to view the workshop project page on the Open Science Framework (OSF), which contains the full workshop notes, presenter slides, and associated information (https://osf.io/psq7k/) [13].

Workshop Agenda and Participants

The NSF FAIR Chemical Data Publishing Guidelines Workshop on Chemical Structures and Spectra was a two day workshop held March 29 – 30, 2019 that brought together around 40 participants in an effort to advance the sharing of machine-readable chemical structures and spectra alongside publications. Participants included chemical researchers from universities and national laboratories, librarians, chemical data repository managers, cheminformatics software developers, scientific societies, and publishers (Appendix A: Workshop Participants).

Day one of the workshop began with presentations from researchers discussing case studies for structure and spectral data reuse along with their pain points and strategies for sharing machine-readable chemical data. The chemical data reuse presentations were followed by a series of presentations that reviewed current chemical data infrastructure for managing and sharing data. The remainder of day one and the majority of day two at the workshop were devoted to strategic breakout group sessions tasked with discussing and evaluating a particular issue around sharing machine-readable chemical structures and spectral data.

The breakout discussions were arranged around three primary themes: workflows involved in publishing data, content and description of data, and stakeholder interests. These three sessions ran in parallel in two sequential tracks to facilitate discussion from broad considerations to practical suggestions. The six breakout groups were led by selected participant facilitators and included: (1) Publishing Workflow and Data Deposit Hack; (2) FAIR Criteria/Metadata Profiles for Spectra and Chemical Structures; (3) Value Propositions for Stakeholders; (4) Planning Workflow Implementation Pilots; (5) Drafting Harmonized Guidelines for Publishing Machine-readable Data; and (6) Community Support and Engagement.

Track 1	Day 1 – Friday	Track 2	Day 2 – Saturday
A1	Publishing data workflow	**A2**	Plan implementation pilot
B1	FAIR criteria	**B2**	Draft harmonized guidelines
C1	Value propositions	**C2**	Community engagement

Summary of Presentations on Reuse and Existing Infrastructure

The reuse case study presentations included a review of the importance and utility of available machine-readable chemical data. For example, Bisson presented on the Raw NMR Data Initiative in Natural Products Research [16], involving over 70 researchers of the field and promoting the value and importance of access to raw NMR spectral data. Access to the original raw machine-readable spectra allows researchers to, for example, detect incorrect structures, promote research integrity, increase reproducibility, and enhance peer-review [16,17]. Related to detecting errors and enhancing peer-review, Hunter outlined current data analysis strategies employed by the journal *Organic Letters* in an effort to improve spectral data quality in journal article SI [18].

The desire for access to raw machine-readable data was bolstered by chemistry researchers discussing their personal strategies for sharing machine-readable data, along with their pain points of reviewing the literature and engaging in research without access to the original data [19].

Appropriate file packaging with associated metadata is key to successful data sharing of machine-readable spectra and structures. As such, several presentations reviewed current workflow initiatives that package raw NMR spectral data with appropriate metadata, including the Mpublish workflow [20-21] and the NMReDATA initiative [22-23]. In addition, progress toward developing a platform independent metadata model for spectral data was discussed by Martinsen, while Hicks emphasized the importance of connecting appropriate metadata such as chemical structure identifiers to journal articles [24].

Finally, current available infrastructures for managing and sharing chemical data were highlighted, including the Chemotion Electronic Laboratory Notebook & Repository [25-26], the MassBank of North America Database (MoNA) [27-28], the Cambridge Structural Database [10,29], Supramolecular.org [30-31], PubChem [7, 32], and ChemSpider [12, 33]. While this was certainly not an exhaustive review of suitable chemical data repositories, it did provide enough of a sampling of the types of infrastructure already available for sharing machine-readable chemical data and provided an ideal segue into the breakout group discussions.

Summary of Breakout Groups

Day one of the workshop consisted of three concurrent breakout groups tasked to review publishing and data workflows (A1), FAIR/metadata criteria for spectral data and structures (B1), and the value propositions for stakeholders (C1).

Participants in **Breakout Group A1** discussed potential publishing workflows for sharing machine-readable spectra and structures alongside primary articles. Generally, it was felt that the ideal situation would be for associated data to be deposited in an appropriate domain repository and then cross-referenced (i.e., permalink) with the primary article. Domain repositories offer many advantages for data sharing such as the ability to standardize/validate data and specialized user search query capabilities for the data (e.g., chemical structure

search), which is generally not available on publisher platforms. However, several participants were also concerned that at this time the barrier is too high for authors and publishers to broadly integrate domain repositories without more clearly established online workflows. Furthermore, few repositories are currently prepared to support ingestion and curation of spectroscopic data.

Further discussion focused on enhancing the current workflow of packaging data within publisher supporting information (SI). As current publisher SI is well integrated into standard author and publisher workflows, the suggested hypothesis was that the first step for the chemistry community could be to incentivize a systematic enhancement of publisher SI files. This enhanced SI file would contain a structured package of data and metadata. By enhancing the SI, we limit workflow disruptions, while simultaneously advancing data sharing. An enhanced SI file with machine-readable data could also be ingested into data repositories in the future. Participants then considered what contents and file types the enhanced SI package file should contain. Discussions here were preliminary, but some themes began to emerge, such as including compound identifiers (i.e., InChI [34]), NMR FID data, and author information.

Breakout Group B1 discussed key metadata elements needed for describing machine-readable spectra and structure data. Three main categories were identified including bibliographic metadata, chemical structure information, and spectral metadata. Bibliographic metadata included, for example, a list of contributors and ORCIDs, title, publication reference, funding information, and data types. Including InChI identifiers was the core criteria for referencing structure information in metadata as the InChI facilitates cross database indexing.

For spectral files, participants considered the types of metadata fields auto-generated with NMR software such as file name, solvent, temperature, pulse sequence and number of scans [20]. Several metadata fields were discussed as a priority, including the nucleus type, experiment (1D, 2D, etc.), frequency, solvent, temperature, and pulse-sequence.

Finally, the importance of registering the metadata for discovery was discussed, and one potential solution that emerged was to deposit the metadata with DataCite [35] as it is available now and can exchange with Crossref for appropriate linking to publications. What to include in the registered metadata vs. what further information can be detailed in the data package was also a topic of discussion.

Breakout Group C1 was tasked with enumerating stakeholder value propositions for sharing machine-readable chemical structures and spectra. The discussions brainstormed ideas on post it notes related to a Value Proposition Canvas [36] for each stakeholder. Participants reviewed the jobs, pains, and gains associated with sharing machine-readable chemical data for several stakeholder roles (Table 1).

Table 1. Selected stakeholder jobs, pains, and gains discussed in Breakout Group C1 for sharing machine-readable chemical data (wording edited for clarity).

Stakeholder	Jobs	Gains	Pains
Funders	● Require research standards, quality, and accountability ● Fund reusable research ● Determine appropriateness of proposed research ● Ensure research data available is FAIR	● Advance science and society ● Serve the public ● Increased value and prestige from citation and data reuse. ● Discover new knowledge more efficiently	● Counseling and supporting researchers ● Disciplinary data expectations and FAIR parameters differ across disciplines ● Enforcement challenges
Researchers	● Maintain accurate lab notebook ● Establish reproducible methods ● Organize and manage data ● Secure funding	● Earn reputation for data integrity and quality of reported research ● Ability to discover and access more data ● Receive credit faster for shared research	● Retrieving past research data. ● Time and needed training involved with data sharing ● Data management
Authors	● Communicate data ● Support hypotheses and claims ● Meet grant requirements	● Establish priority ● Extend usefulness of reported work ● Ability to reuse published data	● Time needed ● Managing data across projects/collaborators ● Lack of clear credit for data publishing
Publishers	● Disseminate information ● Facilitate peer-review ● Provide reliable access to information and data	● Increase reputation ● Increase readership ● Attract authors and readers	● Time with handling and processing data ● Author submitted metadata is challenging
Editors and Reviewers	● Ensure integrity of results ● Providing feedback ● Advance journal goals ● Stewards of scientific results	● Reproducible science ● Ensure quality/reputation of authors and journal ● More citations to journal ● Learn about the latest research first	● Additional time required to review data ● More knowledge required to review data ● Getting recognition for additional work
Repositories	● Provide access to data ● Discovery tools for finding data ● collect/organize metadata ● Preserve data	● Enable and contribute to discovery ● Buy-in from community ● Maximize integration of information	● User training ● Curation ● Understanding priority data classes

Such enumeration of customer profiles can serve as a starting point for developing data sharing support services, documentation, training, and products that support needed work and realize gains for stakeholders while reducing challenges. As a result, early discussions began regarding potential needed products and documentation to help facilitate machine-readable data sharing such as identifier and data validation tools, reputation/recognition services, and metadata standards. Moreover, data management training for stakeholders emerged as a clear trend.

The discussions from the first round of Breakout Groups were more fully elaborated in subsequent discussions (*vide infra*) the following day. After a review of the progress made on day one with Breakout Groups A1, B1, and C1, discussions continued on day two with planning a workflow implementation pilot (A2), drafting guidelines for publishing machine-readable structures and spectra (B2), and determining needed community support and engagement (C2).

Breakout Group A2 participants were asked to sketch out an implementation pilot for sharing machine-readable chemical structures and spectra alongside publications that could be implemented in the near term. The use-case of organic synthesis research was considered as chemical structures in this domain can generally be well-defined by current machine-readable linear notations (e.g, InChI [34] and SMILES [37]). The key idea that emerged from this breakout group was a proposal for a Publisher Pilot, where publishers would accept an enhanced supporting information file package (e.g., zip, BagIt [38]) containing organized machine-readable spectra, structure identifiers, and associated metadata for all compounds synthesized in the article. In return, the publisher would provide incentives with a recognition system for authors that participate in the Pilot. For example, publishers would help promote the author's work with custom FAIR data badging and additional promotion of the article.

A variety of assessment measurements for a Pilot were discussed including metrics to determine "FAIRness" of data, tracking article/data views, tracking citations, and surveying authors and peer-reviewers participating in the Pilot effort. Limitations with packaging the data were identified including the current lack of support tools for automated assembly of the package, data validation, increased time requirements for authors and reviewers, and the lack of discoverability of the dataset if the metadata of the package file is not registered with a DOI service. Despite these limitations, there was strong support for the enhanced supplementary package file from participants as it was seen as a simple next step for what the community can accomplish today.

More specific information about the contents of enhanced supplementary data packages were discussed by **Breakout Group B2**, which sought to draft harmonized guidelines for publishing machine-readable chemical spectra and structures. A data submission checklist list was developed that outlined the data and metadata components to include in an enhanced supplementary information package based on discussion in previous groups:

1. Bibliographic Information (via README file):
    a. Title
    b. Author, Contributors (e.g., names, ORCIDs, affiliations)
    c. Associated Publication
    d. Funding Information (if it exists)
    e. Date of creation
    f. Version Number
    g. Data Types
    h. Formats
    i. Software
    j. License (may depend on repository)

2. Spectra
    a. Raw NMR data (e.g., FID files)

    b. Processed data in standard text-based format (e.g., JCAMP-DX, [39-40]), including the available metadata fields
        i. Instrument model
        ii. Frequency
        iii. Nucleus (for each dimension)
        iv. Experiment (1D, 2D)
        v. Solvent
        vi. Temperature

3. Chemical structures
    a. File names of spectra linked to structures
    b. InChI identifier (if available)
    c. Structure representation (e.g., molfile [41,42] and/or SMILES)

Data and metadata files would then be arranged in a logical directory order (e.g., one compound and associated spectra and metadata per folder), archived together in one main folder with the bibliographic metadata, and then sent to the publisher alongside the article.

To help assemble the package file, participants felt that a web-based “Toolkit Wizard” should be developed for authors that facilitates the organization of the files, captures metadata in a standardized format, and validates the data. Validation could be implemented in stages, for example low level validation can check file types, while a more advanced validation could include a structure checker. Conversations also surfaced regarding the need for author training tutorials on chemical file format generation and spectra data export. These conversations were also elaborated in Breakout Group C2, which led discussions around needed community support and engagement.

The first main topic discussed within **Breakout Group C2** was a brainstorm of needed technical infrastructure tools and desired features. For example there was a desire for a common publisher agnostic web service tool where authors can submit machine-readable data and the tool would assist with packaging the data with associated metadata for publication in supporting information. This then led to discussions on how to integrate packaged machine-readable data into repositories and publisher workflows. Other topics discussed included a desire for more institutional support, training, and personnel to help researchers with data sharing and machine-readable file format generation (e.g., InChI, SMILES, JCAMP-DX). Attendees recommended additional researcher training on data standards, workflows, and local leaders to help facilitate data sharing. In addition, it was recognized that funders and publishers can help with basic level data literacy and training.

Key Themes

Several common themes emerged from the discussions that were further discussed by the full group.

Stakeholder interests:
Many different stakeholders are involved in the publication and dissemination of chemistry research output and while there are not many venues for cross-perspective discussion of these workflows, all of these parties play critical roles in the current landscape and are juggling many competing priorities and challenges. Common priorities for stakeholders include timely dissemination of quality work that is replicable/reproducible, advances science and is rewarded through reputation, recognition and funding opportunities for further research. Common themes that challenge stakeholders include overall pressure on time and infrastructure, as well as lack of familiarity with concepts of machine-readability and FAIR management of data and metadata.

Sharing data:
There are advantages that can be realized by aligning data sharing practices with the more common and familiar practice of publication of articles where appropriate, including maximizing current workflows, infrastructure and resources. Improvements can be made in current practices for developing supplemental material to articles that can facilitate sharing of data files in the near team and enhance their future utility as online data handling workflows continue to improve. Depositing data into domain repositories where available can provide needed expert scientific and technical curation, such as validation checks, and further opportunities for advanced searching and analysis so valued in chemistry research.

Metadata description:
To facilitate preparation, publishing and further reuse, data and metadata components should be packaged as a dataset. Considering how to describe these data types through machine-readable metadata is as critical both scientifically and technically for reusing the data as the files themselves. For spectroscopic data, this involves information about both the spectroscopic measurement and the chemical species being studied that needs to be consistently structured in machine-readable form. Current services for registering high level bibliographic metadata about articles through the DOI mechanism that facilitate citation and discovery are well adopted in the publishing community and now being adapted for datasets. This is an opportunity to consider what domain specific information may also be useful to include to facilitate similar cross-linking between datasets, articles and other chemical information resources.

Working Towards FAIR

In the closing session, the full group identified several target areas for further activities supporting transition towards FAIR workflows:

- Development of drop-in file packaging services, with inclusion of standard file formats, metadata generation and validation, as these are further refined in the community.
- Training for stakeholders, particularly researchers in generating and working with machine readable data and overall data management.

- Refinement of existing and emerging standard formats for chemical structures and spectroscopic data.
- Articulation of critical provenance and scientific metadata for discovery and analysis of chemical structures and spectroscopic data.
- Identification / development of authoritative domain repositories and minimum required validation and curation for chemical structures and spectroscopic data.
- Spread the word about FAIR: what the current state is in chemistry, what use cases have been described so far, what data are available, how can local institutions support researchers, etc.
- Organize further workshops across stakeholders to address outstanding challenges.

The International Union of Pure and Applied Chemistry (IUPAC) and the GO FAIR Chemistry Implementation Network (ChIN) concluded the workshop with their perspectives on supporting ongoing efforts as two international organizations active in chemical data and FAIR. The IUPAC Committee on Publications and Cheminformatics Data Standards [14] is charged to "promote interoperable and consistent transmission, storage, and management of digital [chemical information] content through the development of standards." The committee is launching a number of projects to formulate machine-processable technical descriptions building on the authoritative scientific definitions of IUPAC to enable the application to global problems in digital science. ChIN [15] is developing a series of personas and use cases involved in sharing FAIR data to help assess the current landscape of resources and tools and highlight gap areas for further development.

Post Workshop Outcomes and Continuing Initiatives

Conversations and initiatives related to FAIR chemical structures and spectra have continued over the past year, following the workshop. It is clear that "FAIR" resonates with stakeholders, and more support and activity are needed across the board.

Publisher Pilot

One of the major outcomes of the workshop was the proposed Publisher Pilot [43], a program seeking publisher participation to encourage author submission of machine-readable chemical structures and spectra as a package file alongside their article submission. We developed some basic author instructions for compiling this enhanced supporting information package file [44] and the Royal Society of Chemistry posted a short "how-to" generate machine-readable structure data blog post [45].

Conversations about implementing the Pilot are ongoing with Publishers and have been productive thus far. For example, the American Chemical Society (ACS) Publications Division announced an initiative in February 2020 to encourage submission of machine-readable spectra and structures as a package file along with article submissions in the *Journal of Organic Chemistry* and *Organic Letters* [46], building off of the outcomes of the workshop. We plan to

continue the Pilot conversation with publishers and related stakeholders at future chemistry meetings and events.

Packaging Datasets

We note that the file package proposed at the workshop by attendees is somewhat of a simplified version of the NMReDATA initiative [23]. Shortly after the Orlando workshop, in September 2019, several spectroscopy software providers presented implementations of the NMReDATA format at the 1st NMReDATA symposium in Porto, Portugal [47,48]. Open-source code and tools are also available that support structure elucidation and 3D visualization based on the NMReData format [49-51]. As the NMReDATA initiative progresses and extends to other types of chemistry data [52], this presents one option for authors to formulate machine-readable package files containing chemical structure and spectral data.

The ACS Research Data Center has also created an openly accessible packager tool to assist authors with compiling the machine-readable data package, which was noted as a critical needed tool at the workshop. The tool is open and anyone can use it to package data; results do not need to be submitted to ACS journals [53].

Metadata Guidelines

The need for an easy to use file format for capturing spectroscopic data with robust scientific metadata emerged as a key theme from the workshop discussion. IUPAC has provided the JCAMP-DX standard “Data Exchange” format for spectroscopic data for many years [39,40], but it is not aligned with modern Internet protocols or the FAIR Data Principles. Energized by the workshop and previous review sessions, IUPAC is launching a new project to focus specifically on metadata guidelines that will facilitate better processing of raw and derived spectroscopic datasets [54]. These guidelines would be applicable to any files or datasets that incorporate metadata related to spectroscopic data. The project will also provide validation criteria to enable systems to check files for machine readable and interoperable representation based on the new standard.

Data Sharing Guide

Workshop discussion directly informed the outline for a new chapter on sharing chemical data in the recently revised *ACS Guide to Scholarly Communication* (f.k.a the *ACS Style Guide*), released online in January 2020 [55]. The goal is to help authors get a head start on improving data management for sharing in alignment with research funder requirements and expectations. While it is apparent that there are very few regular and consistent practices across the field, the chapter details several of the suggestions from the workshop that can be followed in the more immediate term to ensure data outputs are more readily discoverable, usable and attributable. The chapter provides further background on some of the exemplar workflows described in the workshop and includes current availability of relevant scientific terminologies, file formats, and repositories for preparing and sharing machine-readable data.

Other Initiatives

Several additional activities related to supporting the publication of sharing chemical are in the pipeline, as reported by workshop participants:

- Chemistry librarians at Association of Research Library member institutions are brainstorming approaches to researcher education and training in creating machine-readable chemical structure and spectra files and “train the trainer” opportunities.
- A new project to develop a primer for curating chemical data associated with publications has been approved by the Data Curation Network [56].
- The e-research group at the University of Geneva is using chemistry as a test-case for the development of a model academic repository (called yatera), including domain-specific input forms, validation procedures, harvesting tools and visualization features. Underlying parameters and formats can inform further development of standards in the field of chemistry.
- The German National Research Data Infrastructure is launching a Chemistry Consortium (NFDI4Chem) to support development of open and FAIR infrastructure for research data management in chemistry [57].
- ChIN is developing a white paper on the current state of FAIR Data in Chemistry and participating in GO FAIR activities to facilitate combining of data across domains [58].
- Preliminary planning is in the works in IUPAC to convene a workshop of repositories that handle chemistry data to articulate barriers and review potential approaches for increasing submission of chemical research data types.

Outlooks and Conclusions

A primary motivation for the workshop was to engage key stakeholders to help reach a broader “data frontier” in chemistry and realize a critical mass of shared data in the domain. We are endeavoring to accomplish this by addressing areas where workflows can be enhanced towards supporting the FAIR Data Principles. Sharing package files containing machine-readable structures and spectra alongside chemistry articles as an outcome of the workshop is a tractable first step toward shifting the chemistry community from almost exclusive human-only readable supporting information files to reusable machine-readable supporting information files. Building on a familiar process to collate data as supporting information minimizes the barrier for both authors and publishers to manage the change. The more expediently we focus on standardizing the guidance for machine-readable supporting information accepted across publishers, the easier it will be for researchers to generate and reuse these materials.

As researchers become more comfortable with sharing machine-readable data alongside publications as regular practice, it is important that the workflow continues to shift towards deposition of data with domain repositories. Data deposited within repositories offers several advantages over a package file including additional level of findability (e.g., chemical structure search), additional metadata and data linking, validation checks, ongoing maintenance/curation,

and application of advanced analysis techniques. Many of the types of concerns raised in the workshop discussions around lack of familiarity with technical issues are handled already by repositories as part of their stewardship and curation role. Connecting into domain repositories and aligning needs for supporting spectroscopic and structure data and metadata will ultimately lower barriers and greatly facilitate access and utility of these data for the broader scientific community.

As the sharing of machine-readable data increases in the chemical community, it is critical for all stakeholders to be mindful of interoperability. Publishing well-organized, FAIR data will continue to be characterized by a number of different paths and workflows. Procedures, tools, and outputs may vary depending on the types of chemistry, data classes, reporting conventions in a given field, research culture of the organization and lab, and individual preferences. Clear communication for humans and machines needs to accommodate diverse and creative application and interpretation of scientific study anchored in common agreed principles of practice. Documentation in the form of structured metadata is important for data to be interpreted in their scientific meaning and to avoid introduction of artifacts from different conventions or software variations. In addition to providing information about experiments necessary to interpret data, it is also important to document subsequent activities involving data, including description of software packages (e.g., names, sources and versions) used in analysis and visualization.

The FAIR Data Principles emphasize the ability for machine processes as well as human users to discover and access data. Most approaches to discovery and access are still very much biased towards human navigation and further effort will be necessary to support automated access. While specific approaches to discovery may vary between systems or local needs, establishing consistent metadata that includes standard identifiers and leverages common protocols will be critical to broadly enable chemical data to be FAIR for machines across these needs. This should involve at minimum, registration of datasets with DOI metadata services such as DataCite and persistent linking to associated published articles. It will facilitate greater computational activity to maximize the utility of the object level metadata available through the DOI mechanism in balance with richer scientific description provided along with the data files. Inclusion of InChIs as a standard compound identifier for discrete molecules will facilitate indexing of data sets and cross-linking with other resources among the vast corpus of chemical information [59,60].

As a community of practice, it will be critical within the chemistry domain to collectively understand the gaps towards reaching FAIR, particularly issues of discovery, archiving and other curation essential for establishing the criteria for quality on which the discipline has come to depend. This high level of rigor in correspondence with potential for advanced analyses and computational application underlies the field of chemical information. It is the collective responsibility of the practitioners in this field to continue to address the various “meta” issues around handling FAIR data on behalf of the research community and the scientific knowledgebase. The FAIR Data Principles can provide a general direction, but harmonization of

guidelines that are appropriate for chemistry and realize the scientific goals of the discipline will depend on clarity of stakeholder roles to work in tandem to support this process.

Acknowledgements

We are grateful to the NSF OAC for funding the workshop: Award No. 1838958 and 1838960. Nominal sponsorship included the ACS Division of Chemical Information, the IUPAC Committee on Cheminformatics and Data Standards, and the GO FAIR Chemistry Implementation Network. Our workshop advisors, Angie Hunter (ACS Publications), Ian Bruno (CCDC), Guy Jones (Royal Society of Chemistry), and Dave Martinsen (Martinsen Consulting) all provided a tremendous amount of their time and expertise in helping to plan and execute a successful workshop. In addition, we thank the breakout group facilitators and all the attendees for their enthusiasm and participation in the workshop.

References

[1] Clark, A. M.; Williams, A. J.; Ekins, S. Machines First, Humans Second: On the Importance of Algorithmic Interpretation of Open Chemistry Data. *J. Cheminf.* **2015**, *7*, 1–20. https://doi.org/10.1186/s13321-015-0057-7.

[2] Krallinger, M.; Rabal, O.; Lourenço, A.; Oyarzabal, J.; Valencia, A. Information Retrieval and Text Mining Technologies for Chemistry. *Chemical Reviews* **2017**, 117,7673–7761. https://doi.org/10.1021/acs.chemrev.6b00851.

[3] Open Data at NSF. https://www.nsf.gov/data/

[4] Callaghan, S.; Donegan, S.; Pepler, S.; Thorley, M.; Cunningham, N.; Kirsch, P.; Ault, L.; Bell, P.; Bowie, R.; Leadbetter, A. Making Data a First Class Scientific Output: Data Citation and Publication by NERC's Environmental Data Centres. *International Journal of Digital Curation*, **2012**, *7*, 107-113.doi.org/10.2218/ijdc.v7i1.218

[5] Wilkinson, M. D.; Dumontier, M.; Aalbersberg, Ij. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* **2016**, *3*, 160018. https://doi.org/10.1038/sdata.2016.18.

[6] CAS REGISTRY - The gold standard for chemical substance information. https://www.cas.org/support/documentation/chemical-substances

[7] PubChem. https://pubchem.ncbi.nlm.nih.gov/

[8] Southan, C. Opening up Connectivity Between Documents, Structures and Bioactivity. *ChemRxiv. Preprint.* **2019** https://doi.org/10.26434/chemrxiv.10295546.v1

[9] Scalfani, V.F. (2019): RDA Publisher Forum - Chemistry Journal Data Submission and Sharing Policies Checklist 2017. figshare. Dataset. https://doi.org/10.6084/m9.figshare.8870144.v1

[10] Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallographica Section B* **2016**, *72* (2), 171–179. https://doi.org/doi:10.1107/S2052520616003954.

[11] Steinbeck,C.; Krause,S.; Kuhn,S. NMRShiftDB - Constructing a Free Chemical Information System with Open-Source Components. *Journal of Chemical Information and Computer Sciences* **2003,** *43*, 1733-1739.

[12] ChemSpider. http://www.chemspider.com/

[13] Scalfani, V.F; McEwen, L.R. NSF OAC 2019 Workshop: FAIR Publishing Guidelines for Spectral Data and Chemical Structures OSF, **2019**, https://osf.io/psq7k/

[14] IUPAC Committee on Publications and Cheminformatics Data Standards. https://iupac.org/who-we-are/committees/committee-details/?body_code=024

[15] GO FAIR Chemistry. https://www.go-fair.org/implementation-networks/overview/chemistryin/

[16] Bisson, J. The Raw NMR Data Initiative in Natural Products Research Presentation. https://osf.io/8xstq/

[17] 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.; et al. The Value of Universally Available Raw NMR Data for Transparency, Reproducibility, and Integrity in Natural Product Research. *Natural Product Reports* **2018**. https://doi.org/10.1039/C7NP00064B.

[18] Hunter, A. Supporting Information Review at Organic Letters Presentation. https://osf.io/ctnq4/

[19] Rupar, P. Sharing Chemical Data at The University of Alabama Presentation. https://osf.io/3fwba/

[20] Rzepa, H. FAIR Criteria/metadata profile for spectral data and chemical structures Presentation. https://doi.org/c3k6

[21] Barba, A.; Dominguez, S.; Cobas, C.; Martinsen, D. P.; Romain, C.; Rzepa, H. S.; Seoane, F. Workflows Allowing Creation of Journal Article Supporting Information and Findable, Accessible, Interoperable, and Reusable (FAIR)-Enabled Publication of Spectroscopic Data. *ACS Omega* **2019**, *4* (2), 3280–3286. https://doi.org/10.1021/acsomega.8b03005.

[22] Jeannerat, D. NMReData and reuse of Spectral Data Presentation. https://osf.io/w9sxz/

[23] Pupier Marion; Nuzillard Jean-Marc; Wist Julien; Schlörer Nils E; Kuhn Stefan; Erdelyi Mate; Steinbeck Christoph; Williams Antony J; Butts Craig; Claridge Tim D. W; et al. NMReDATA, a Standard to Report the NMR Assignment and Parameters of Organic Compounds. *Magnetic Resonance in Chemistry*. https://doi.org/doi:10.1002/mrc.4737.

[24] Martinsen, D. Progress on Data Model and Metadata for Spectral Data Files Presentation. https://osf.io/3qpfk/

[25] Tremouilhac, P. Chemotion Electronic Laboratory Notebook & Repository for Research Data Presentation. https://osf.io/fa2b3/

[26] Tremouilhac, P.; Nguyen, A.; Huang, Y.-C.; Kotov, S.; Lütjohann, D. S.; Hübsch, F.; Jung, N.; Bräse, S. Chemotion ELN: An Open Source Electronic Lab Notebook for Chemists in Academia. *Journal of Cheminformatics* **2017**, *9* (1), 54. https://doi.org/10.1186/s13321-017-0240-0.

[27] Mehta, S.S. MassBank of North America (MoNA): An open-access, auto-curating mass spectral database for compound identification in metabolomics presentation. https://osf.io/sc6zb/

[28] Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; et al. MassBank: A Public Repository for Sharing Mass Spectral Data for Life Sciences. *J. Mass Spectrom.* **2010**, *45* (7), 703–714. https://doi.org/10.1002/jms.1777.

[29] Bruno, I. CCDC Workflows Presentation. https://osf.io/9rztq/

[30] Thordarson, P. Opendatafit project & Supramolecular.org Presentation. https://osf.io/ab8jg/

[31] Thordarson, P. Determining Association Constants from Titration Experiments in Supramolecular Chemistry. Chem. Soc. Rev. 2011, 40 (3), 1305–1323. https://doi.org/10.1039/C0CS00062K.

[32] Bolton, E. PubChem Data Repository Capabilities Presentation. https://osf.io/xj8y2/

[33] Jones, G. ChemSpider Data Repository Capabilities Presentation. https://osf.io/k7dwr/

[34] Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J Cheminform* **2015**, *7* (1), 23. https://doi.org/10.1186/s13321-015-0068-4.

[35] DataCite. https://datacite.org/

[36] Strategyzer: Value Proposition Canvas: A Tool to Understand What Customers Really Want. https://www.strategyzer.com/blog/value-proposition-canvas-a-tool-to-understand-what-customers-really-want

[37] Daylight Theory Manual v4.9. https://www.daylight.com/dayhtml/doc/theory/
[38] The BagIt File Packaging Format (V1.0). https://tools.ietf.org/html/rfc8493

[39] Davies, A. N.; Lampen, P. JCAMP-DX FOR NMR. *Applied Spectroscopy* **1993**, *47* (8), 1093–1099. https://doi.org/10.1366/0003702934067874.

[40] Grasselli, J. G. JCAMP-DX, A STANDARD FORMAT FOR EXCHANGE OF INFRARED-SPECTRA IN COMPUTER READABLE FORM. *Pure and Applied Chemistry* **1991**, *63* (12), 1781–1792. https://doi.org/10.1351/pac199163121781.

[41] Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *Journal of Chemical Information and Modeling* **1992**, *32* (3), 244. https://10.1021/ci00007a012.

[42] Biovia CTFile format definitions available on request (registration required). https://www.3dsbiovia.com/products/collaborative-science/biovia-draw/ctfile-no-fee.html

[43] Enhanced Supporting Information Publisher Pilot - NSF FAIR Chemical Data Publishing Guidelines Workshop: Chemical Structures and Spectra. https://osf.io/fqd82/wiki/home/

[44] Author Guidelines for Sharing Enhanced Supporting Information (Machine-readable Spectra and Chemical Structures). https://osf.io/fqd82/wiki/Author%20Instructions/

[45] Tips and tricks: generating machine-readable structural data from a ChemDraw structure. https://blogs.rsc.org/chemspider/2019/11/08/tips-and-tricks-generating-machine-readable-structural-data-from-a-chemdraw-structure/

[46] Hunter, A. M.; Carreira, E. M.; Miller, S. J. Encouraging Submission of FAIR Data at *The Journal of Organic Chemistry* and *Organic Letters*. *J. Org. Chem.* **2020**, *85* (4), 1773–1774. https://doi.org/10.1021/acs.joc.0c00248.

[47] NMReDATA Compatible Software. https://nmredata.org/wiki/Compatible_software

[48] NMReDATA Symposium 2019. https://nmredata.org/wiki/Symposium2019

[49] Java programs for the NMReDATA format. https://github.com/NMReDATAInitiative/javatools

[50] NMReDATA Javascript Parser. https://github.com/cheminfo/nmredata

[51] NMReDATA J_reader (Angel Herráes). http://www3.uah.es/nmr_e_data/reader/reader.htm

[52] CHEMeDATA Initiative. https://chemedata.github.io/

[53] ACS Research Data Center Data Packaging Tool. https://researchdata.acs.org/pkgtool/#/home

[54] IUPAC Project No.: 2019-031-1-024. Development of a Standard for FAIR Data Management of Spectroscopic Data. https://iupac.org/projects/project-details/?project_nr=2019-031-1-024

[55] McEwen, L. ACS Guide to Scholarly Communication. Chapter 3.1 Data Sharing. https://pubs.acs.org/doi/full/10.1021/acsguide.30101

[56] Data Curation Network. https://datacurationnetwork.org/

[57] NFDI4Chem – Chemistry Consortium in the NFDI. https://www.nfdi4chem.de/

[58] Save the Date: International FAIR Convergence Symposium & CODATA General Assembly in Paris on 22-24 October 2020. http://codata.org/blog/

[59] Chambers, J.; Davies, M.; Gaulton, A.; Hersey, A.; Velankar, S.; Petryszak, R.; Hastings, J.; Bellis, L.; McGlinchey, S.; Overington, J. P. UniChem: A Unified Chemical Structure Cross-

Referencing and Identifier Tracking System. *Journal of Cheminformatics* **2013**, *5* (1), 3. https://doi.org/10.1186/1758-2946-5-3.

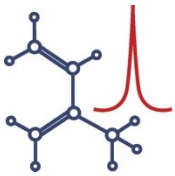
[60] InChI FAQ: What is the purpose of InChI? https://www.inchi-trust.org/technical-faq-2/#2.3

Appendix A: Workshop Participants *(partial, with permission)*

Name	Organization affiliation
Angie Hunter	American Chemical Society
Damien Jeannerat	University of Geneva / NMReDATA Initiative
David Van Craen	Indiana University Bloomington
Donna Wrublewski	Caltech
Evan Bolton	NCBI / NLM / NIH / DHHS
Fiona Shortt de Hernandez	Thieme Publishers
Gregory M. Banik	Bio-Rad Laboratories
Guy Jones	Royal Society of Chemistry
Henry Rzepa	Imperial College London
Huan Chen	National High Magnetic Field Laboratory
Ian Bruno	CCDC
Jacob Boes	Stanford University
Jake Yeston	Science/AAAS
Jeff Lang	American Chemical Society
Jessica Freeze	Yale University
Johannes Hachmann	University at Buffalo - SUNY
Jonathan Bisson	University of Illinois at Chicago
Josef Eiblmaier	InfoChem GmbH
Kent Griffith	Northwestern University
Leah McEwen	Cornell University

Name	Organization affiliation
Martha L. Chacón-Patiño	National High Magnetic Field Laboratory
Martin G. Hicks	Beilstein-Institut
Matthew G. Donahue	University of Southern Mississippi
Michael Qiu	ACS Publications
Pall Thordarson	UNSW Sydney Australia
Paul A Rupar	University of Alabama
Pierre Tremouilhac	Karlsruhe Institute of Technology
Richard Hartshorn	IUPAC/University of Canterbury
Sajjan Singh Mehta	University of California, Davis
Samantha MacMillan	Cornell University
Simon Coles	University of Southampton
Steffen Pauly	Springer Nature
Steve Heller	NIST
Stuart Chalk	University of North Florida
Ty Abshear	Bio-Rad Laboratories
Valery Tkachenko	SCIENCE DATA SOFTWARE, LLC
Vincent Scalfani	University of Alabama
Ye Li	MIT

Appendix B: Workshop Agenda



FAIR Chemical Data
Publishing Guidelines Workshop
Chemical Structures and Spectra

Agenda

March 29–30, Orlando, FL
Orange County Convention Center
Rooms W312B and W312C
Limited Attendance Workshop

#FAIRchemDATA
Workshop Chairs: Leah McEwen and Vincent Scalfani
Advisors: Angie Hunter, Ian Bruno, Guy Jones, and Dave Martinsen
NSF OAC – Award No. 1838958 and 1838960
Updates and Supporting Material – osf.io/psq7k

Day 1: Friday, March 29, 2019

Welcome & Introductions

7:30 AM – 8:00 AM	Coffee and Continental Breakfast
8:15 AM – 8:30 AM	**(If available) Beth Plale, NSF OAC, Science Advisor for Public Access** Welcome
8:30 AM – 9:00 AM	**Leah McEwen and Vincent Scalfani** Premise for workshop and review of goals and outcomes

Chemical Data Reuse Case Studies

9:00 AM – 9:15 AM	**Damien Jeannerat** NMReData and reuse of spectral data
9:15 AM – 9:30 AM	**Jonathan Bisson** The NMR raw data initiative
9:30 AM – 9:45 AM	**Pall Thordarson** Opendatafit project and supramolecular.org
9:45 AM – 10:00 AM	**Henry Rzepa** Mpublish and persistent identifiers
10:00 AM – 10:15 AM	**Matthew Donahue** Researcher story: Data reuse challenges
10:15 AM – 10:30 AM	**Paul Rupar** Researcher story: Open Science Framework
10:30 AM – 11:00 AM	Coffee Break

Day 1: Friday, March 29, 2019

Background Presentations

11:00 AM – 11:10 AM	**Angie Hunter** Current journal spectral and chemical structure review practices
11:10 AM – 11:20 AM	**Martin Hicks** Addressing the F in FAIR: Freeing up meta-data in journal publications
11:20 AM – 11:30 AM	**Dave Martinsen** Progress on data model and metadata for spectral data files
11:30 AM – 11:40 AM	**Pierre Tremouilhac** Chemotion IR
11:40 AM – 11:50 AM	**Sajjan Singh Mehta** MoNA database
12:00 PM – 12:10 PM	**Ian Bruno** CCDC workflow
12:10 PM – 12:20 PM	**Evan Bolton** PubChem data repository capabilities
12:20 PM – 12:30 PM	**Guy Jones** ChemSpider data repository capabilities
12:30 PM – 1:30 PM	Lunch

Work-group Breakouts

1:30 PM – 4:30 PM	**Pierre Tremouilhac and Ian Bruno, Facilitators** Track A1 – Publishing Work-flow and Data Deposit Hack **Henry Rzepa and Dave Martinsen, Facilitators** Track B1 – FAIR Criteria/Metadata Profiles for Spectra and Chemical Structures **Jeff Lang, Stuart Chalk, and Elsa Alvaro, Facilitators** Track C1 – Value Propositions for Stakeholders
3:00 PM	Coffee Available
4:30 PM – 5:30 PM	Work-group Breakout Report and Discussion

Day 2: Saturday, March 30, 2019

Welcome and Debrief

7:30 AM – 8:30 AM	Coffee and Continental Breakfast
8:30 AM – 9:00 AM	**Leah McEwen and Vincent Scalfani** Day 1 debrief and review goals

Work-group Breakouts

9:00 AM – 12:00 PM	**Martin Hicks and Angie Hunter, Facilitators** Track A2 – Plan Work-flow Implementation Pilots **Guy Jones and Ye Li, Facilitators** Track B2 – Draft Harmonized Guidelines for Publishing Machine-readable Data **Donna Wrublewski and Richard Hartshorn, Facilitators** Track C2 – Community Support and Engagement
10:00 AM	Coffee Available
12:00 PM – 1:00 PM	Lunch
1:00 PM – 2:00 PM	Work-group Breakout Report and Discussion

Next Steps: Community Structures for Moving Forward

2:00 PM – 2:15 PM	**Richard Hartshorn** IUPAC community
2:15 PM – 2:30 PM	**Simon Coles** GO FAIR Chemistry Implementation Network (ChIN)
2:30 PM – 3:00 PM	Final Discussion
3:00 PM – 3:30 PM	Goodbyes and Coffee
3:30 PM	Workshop Adjourns
3:30 PM – 5:00 PM	Organizing Team Debrief and Planning

Workshop Nominal Cosponsors

ACS Division of Chemical Information
IUPAC Committee on Publications and Cheminformatics Data Standards
GO FAIR Chemistry Implementation Network