Draft Technical Report: FAIRSpec-Ready Spectroscopic Data Collections – Preliminary Advice for Researchers and Authors. Part 1: Guidelines

Mark Archibald,[a] Ian Bruno,[b] Stuart Chalk,[c] Antony N. Davies,[d] Robert M. Hanson,[e] Stefan Kuhn,[f] Robert J. Lancashire,[g] and Henry S. Rzepa.[h]

[a]Royal Society of Chemistry, Thomas Graham House, Science Park, Milton Road, Cambridge, CB4 0WF, UK, [b]Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK, [c]Department of Chemistry and Biochemistry, University of North Florida, Jacksonville, FL, USA, [d]SERC, Sustainable Environment Research Centre, Faculty of Computing, Engineering and Science, University of South Wales, UK, [e]Department of Chemistry, St Olaf College, Northfield, Minnesota, USA, [f]University of Tartu, Institute of Computer Science, Narva mnt 18, 51009 Tartu city, Tartumaa EST, [g]Department of Chemistry, The University of the West Indies, Kingston 7, Mona Campus, Jamaica, [h]Department of Chemistry, Imperial College, Molecular Sciences Research Hub, White City Campus, Wood Lane, London W12 0BZ, England.

Abstract

In this first of a two-part series, we introduce the concept of a FAIRSpec-ready spectroscopic data collection – that is, a collection of instrument data, chemical structure representations, and related digital items that is ready to be automatically or semi-automatically extracted for metadata that will allow the production of an IUPAC FAIRSpec Finding Aid. Associating this finding aid with the collection produces an IUPAC FAIRSpec Data Collection. The challenge we set for researchers is relatively simple: to maintain their data in a form that allows critical metadata to be extracted in a discipline-specific way, increasing the probability that the data will be findable and reusable both during the research process and after publication. We focus on a few specific suggestions that researchers can use to maximize the “fairness” of their spectroscopic data collection. Most importantly, following these guidelines ensures that instrument datasets are unambiguously associated with chemical structure. The guidelines promote the inclusion of the instrument dataset itself in the collection and describe ways of organizing the collection such that automated metadata creation is possible. In these guidelines we emphasize the importance of systematically organizing data throughout the entire research process, not just at the time of publication.

1. Introduction

1.1 The IUPAC "FAIRSpec" Project

To promote the adoption of the FAIR principles in chemical sciences throughout the data production, publication, and post-publication process, the IUPAC Project *Development of a Standard for FAIR Data Management of Spectroscopic Data*[1] was organized in 2019. The overall goals of this ongoing project include:

- the development of a clear set of FAIR Data Management Principles specific to chemistry-related spectroscopic data;
- the design of a means of describing a spectroscopic data collection that distills critical metadata associated with the digital items in the collection, thus providing a standardized, accessible method of exploring the contents of a data collection without the need to individually access items in the collection (the *IUPAC FAIRSpec Finding Aid* [2,3]);
- providing researchers and authors with guidelines for the organization of spectra and associated chemical structure information that allows machine-assisted curation of the data, creating the necessary link between chemical structure and spectroscopic data that is often key to its analysis and discussion; and
- the specification of a standardized set of metadata keys and values that will allow a broad range of services that can efficiently search for and retrieve spectroscopic datasets of interest.

Principles for the FAIR management of spectroscopic data in chemistry have been published[4] and provide a foundation for the work described in this Technical Report. The five main principles and their associated corollaries are shown in Fig. 1.

1. **FAIR Management of data should be an ongoing concern.**
    a. FAIR management of data must be an explicit part of research culture.
    b. FAIR management of data should be of intrinsic value.
    c. Good data management requires distributed curation.
    d. Experimental work is by nature iterative.
2. **Context is important.**
    a. Digital objects are generally part of a collection.
    b. Chemical properties are related to chemical structure.
    c. Data relationships are diverse and develop over time.
    d. FAIR management of data should allow for validation.
3. **FAIR management of data requires curation.**
    a. Data reuse relies upon practical findability.
    b. Data has to be organized to be accessible.
    c. Data interoperability requires well-designed metadata.
    d. Value is in the eye of the reuser.
4. **Metadata must be standardized and registered.**
    a. Register key metadata.
    b. Assign a variety of persistent identifiers.
    c. Enable metadata crosswalks.
    d. Allow for value-added benefits.
5. **FAIR data management standards should be modular, extensible, and flexible.**
    a. Modularity allows specialization.
    b. Allow for future needs.
    c. Respect format and implementation diversity.
    d. All data formats should be valued.

Figure 1. The five FAIRSpec principles and their corollaries.

In this Technical Report, we focus on the penultimate goal of our project – providing guidelines for the creation of what we are calling *FAIRSpec-ready* data collections. As summarized in Fig. 2, we consider the five FAIRSpec principles as applied to the general practice of developing and managing spectroscopic data collections that optimizes their ability to be machine- as well as human-readable.

1. FAIR Management of data should be an ongoing concern.

- Don’t wait until publication time to organize your data.
- Recognize the ongoing value of well-organized data.
- Allow for corrections and addition of new information.

2. Context is important.

- Associate spectra with chemical structure as much as possible.
- Allow for ambiguity and the reconsideration of these associations.
- Find ways to validate your structural and spectral analysis.

3. FAIR management of data requires curation.

- Accept that you are going to have to do part of the work.
- Optimize opportunities for data citation.
- Do not presume to know how people will utilize your data.

4. Metadata must be registered and standardized.

- Findability relies upon proper registration.
- Work with data management professionals in your organization.
- Include discipline-specific metadata.

5. FAIR data management standards should be modular, extensible, and flexible.

- FAIR data management should be as simple as possible.
- Find (or create!) the right tools for the job.
- Find ways to make data management useful to you and your project *now.*

Figure 2. Best practices in spectroscopic data management based on the five FAIRSpec principles.

Of primary relevance to this report is the recognition that ***context is important***. Spectroscopy data objects are typically part of a collection that relate to one or more chemical structures through diverse relationships that may develop over time. Further, data must be organized to enable accessibility, and associated metadata must be well-designed to facilitate interoperability. It is important to value all data formats, including commonly used proprietary formats as well as recognized standards published by national or international standardization bodies such as IUPAC.

We have chosen to discuss these guidelines prior to describing specifications for the details of our proposed *IUPAC FAIRSpec Finding Aid* because we feel that one of most important components of the curation necessary to produce an *IUPAC FAIRSpec Data Collection* is a critical minimal level of curation that can be applied to essentially any working collection of spectroscopic data, whether or not an IUPAC FAIRSpec Finding Aid is ever generated. Effective data management starts immediately after spectroscopic data are generated within a laboratory. If this minimal curation is done well, over time, many of the goals of FAIR data management can then be accomplished efficiently or even automatically in later stages of the process, however that workflow might ultimately be defined. If done poorly, the effort of data preservation may be painfully time consuming, and meaningful extraction of metadata may even become impossible.

While these guidelines may seem extensive upon first sight, we wish to emphasize that the creation of a FAIRSpec-ready collection can be simplicity itself. Successful efforts can be as involved as implementing a fully "data-aware" laboratory management system or as simple as just maintaining a set of file directories on an instrument, *provided chemical structure representations are added appropriately.*

1.2 Organization of This Report

The guidelines presented here cover four specific areas:

- guidance for the generation of structural representations that accompany those datasets,
- guidance for best practices in the preparation of spectroscopic datasets, particularly in regard to their associated descriptive and relational metadata,
- guidelines for the organization of digital items in the collections containing spectroscopic data and their associated structural representations, and
- guidance for maximizing the potential for registering metadata with recognized metadata management agencies.

These guidelines are not intended to cover every possible sort of data or structure representation. In several cases involving structures, such as organometallic compounds and polymers, we recognize that there is no perfect solution. As such, these guidelines should be taken as starting points for development of further guidance.

Some of the terminology used here is defined more fully in the appendix to this Technical Report.

1.3 Intended Audience

The primary intended audience of this Technical Report includes:

- practicing researchers who create and work with experimental spectroscopic datasets and are interested in best practices relating to the management of their growing data collections
- principal investigators with an interest in maximizing their data's potential to be found, used, and referenced both internally within their institution and by members of their community
- institutional staff responsible for working with researchers to utilize instruments that collect spectroscopic data
- librarians working with researchers to develop best practices in relation to data management within their institution
- institutional repository managers with responsibilities that include ensuring the highest level of FAIR data management within their institutions
- journal editors and publishing house staff tasked with developing guidance for authors and reviewers in the creation and review of electronic supporting information datasets associated with scientific publications involving spectroscopic data
- developers of electronic laboratory notebooks (ELNs) and other services associated with the scientific enterprise such as integrated laboratory management platforms and data validation services
- funding agencies interested in providing guidance to their grantees in developing practical wide-reaching FAIR data management plans, particularly in the field of chemistry
- anyone working outside the context of chemistry-related spectroscopy interested in developing similar guidelines for FAIR data management within their own discipline

1.4 Management of Spectroscopic Data

Spectroscopic data are key components of many chemistry endeavors both in academia and industry. Spectroscopic analysis forms a principal function in the "proof of structure" in most areas of experimental chemistry, answering questions such as: "What have I made?", "How pure is it?", and "Why didn't this reaction work the way I expected it to?" Spectroscopic data provides evidence required for publication of scientific results by journal publishers and is the basis for subsequent experimental replication or extrapolation of results.

Unfortunately, up until very recently, it has been the norm that such data are provided only in packaged document form, vendor-proprietary formats, or in reduced or processed form as a “spectrum”, perhaps even just an image of a spectrum, rather than as the primary or raw instrumental output. The result is that published data are often significantly less useful than they could potentially be.

As primary products of funded research, plans for the maintenance and sharing of spectroscopic data are increasingly becoming a requirement of funding agencies, although guidelines as to how this should be achieved tend to be minimal, with language such as

> *The Data Management Plan…may include…the standards to be used for data and metadata format and content.* [5]

and

> *Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections, and other supporting materials created or gathered in the course of work under NSF awards. Recipients are expected to encourage and facilitate such sharing.*[6]

In addition, in strongly regulated industries such as pharmaceutical research and manufacturing there are long-standing legal requirements to conserve and be able to produce on demand original analytical data such as spectra for auditors.[7][8] These data may well be required long after the measuring instruments themselves have been retired.

In both contexts – academic and industrial – this issue is compounded by the international chemistry community’s lack of an agreed-upon, standardized mechanism for organizing and sharing *collections* of spectroscopic data. Instead, every research group is left to their own devices to save and share data in whatever organizational scheme they choose to implement, often as a single monolithic document supplementing a publication, the ESI. In this form, the data cannot be said to be easily discoverable or findable other than by direct association via the landing page of the journal article reporting the results. The "data" are likely accessible to humans only via copy/paste operations, if these are even permitted by the format. When raw or minimally processed data are provided, they tend to be in the form of an idiosyncratically organized archive file. Such "collections" are not sufficient for optimal broad-context findability or ease of re-use.

To a practicing chemist, it should be obvious that digital entities coming from a laboratory instrument constitute "primary data" (referred to herein as "datasets"). However, the word *data* as used here is a broader term. For our purposes, data includes processed data such as spectra, and derived data and analyses. These might include peak lists or chemical shift splitting, and integration descriptions in NMR spectroscopy, as well as 1D and 2D spectral assignments in relation to molecular structure. Chemical structure representations (MOL and SDF, for example) also fall in this category.

1.5 FAIR Management of Spectroscopic Data

Relatively recent discussions of data management have emphasized what are referred to as FAIR (findable, accessible, interoperable, and reusable) data management guidelines.[9] It is important to emphasize that what we are referring to as "FAIR" here is not so much the data themselves as it is the *management* of that data, and in particular, the production and organization of the *metadata* associated with experimental data. *Metadata* are digital items that play the role of documentation for data, resource discovery, and contextualization. Metadata accompanies electronic data in describing the data, making internal relationships among related digital items in a collection, and relating that collection to other work.

Thus, we prefer to refer to "FAIR data management" rather than "FAIR data" itself as a way of emphasizing that we are not just referring to specific file formats (experimental "datasets") when using these terms, but rather to how those datasets are described and organized. We focus here on the development of a rich metadata context associated with experimental datasets.

It is important to note at the outset that these widely discussed FAIR guidelines were designed not only to facilitate access to such data by humans, but also to allow autonomous discovery and re-use by machines in contexts such as artificial intelligence and machine learning. ("FAIR" has been said to also refer to *Fully Artificial-Intelligence Ready*.) Unfortunately, in the chemical spectroscopies, the current forms of ESI as representations of data are rarely if at all conformant with FAIR guidelines or optimal for automated processing.

An important aspect of FAIR data management relates to how a metadata record is stored and shared. There are two main models for the use of metadata. The one most associated with *findable, accessible, interoperable, and reusable* focuses on the public distribution of data and its associated metadata. The general FAIR guidelines focus on a largely discipline-blind model where a digital object is registered with an appropriate registration agency in exchange for a persistent identifier (PID, normally in the form of a digital object identifier, or *DOI*, from the DOI Foundation [https://www.doi.org/]), and that record is then aggregated into a metadata store, where it can also be indexed and searched. The metadata records include full file paths (URLs) or database references to the components or collections as held in an appropriate repository. This first model is primarily focused on the post-publication finding and relating of data and their collections on the web.

1.6 Not Just for Publication: The FAIR Data Workflow

A more nuanced model for FAIR data management is a private, locally relevant model discussed in detail below. This second discipline-specific model, which we introduce in these guidelines, is one where the metadata are stored on a file system where they can be summarized in a local *IUPAC FAIRSpec Finding Aid* (discussed more specifically in Section 4). This highly structured document describes the accumulating data collection in a machine-readable format. The IUPAC FAIRSpec Finding Aid can then serve as a starting point for both the mostly discipline-blind PID-based model as well as a much more fine-grained private or public discipline-specific purposes.

In the everyday activities of a research laboratory, it is common to acquire numerous spectroscopic data taken by multiple researchers, over multiple timeframes, of samples containing compounds having known or unknown structure. Many researchers already spend significant time curating their data collections, whether that be as complex as a dedicated institutional system or as simple as a hand-written laboratory notebook or spreadsheet. ELNs already serve an important role in the organization of the metadata associated with spectroscopic data. However, with few exceptions[10] such tools (and the metadata they collect) are largely unstandardized, and their adoption is variable and can be very minimal.

Thus, well before any publication is in sight, there is a need for more standardized workflows for the organization of spectroscopic data collections, particularly in relation to their associated samples and the putative structures of their associated compounds. Every experimental chemist understands the importance of both organizing and communicating such information. The guidelines we present here suggest simple ways that researchers can better organize their data and associated metadata to provide added value throughout the research project lifetime. Indeed, even well past the point of project completion and publication, when digital data collections are stored in and shared by repositories, there is a great need for standardization of the metadata associated with spectroscopic data. These guidelines provide the basis for a workflow to produce data collections that are ready to be processed – and, indeed, might be regularly processed privately and locally – to make FAIRSpec-ready collections that are useful throughout the research process.

1.7 Descriptive and Relational Metadata

Metadata is central to this discussion. We distinguish two types of metadata – *descriptive* and *relational.*

Descriptive metadata in the context of spectroscopy constitute information such as the type of instrument used, the temperature, the solvent, and other details of the analytical methods involved in acquiring and processing the actual "experimental data." The description also includes declarations of the media types that the data are held in, which itself will help identify whether it is primary or raw (lossless) data directly captured from an instrument, or whether it has already been processed (with possibly some loss of data) in some form, as for example in a conversion of a time domain or induction form into a frequency domain or spectral form.

Relational metadata include information that ties or associates experimental spectroscopic data to their relevant context or provenance – notebook page and sample references, compound numbers in a publication, researcher and organizational identifiers, proposed chemical structure details and references to both related spectra and other analytical techniques. Relational metadata can also be at the collection level, indicating relationships to other collections such as other datasets or to journal publications. Relational metadata can also provide licensing information together with broader information about sponsoring and research organizations and help locate related metadata records that are registered with agencies such as DataCite[11] specifically for the purposes of improving wide-ranging findability and accessibility. Relational metadata items at the collection level are also often associated with individual PIDs.

Relational metadata are dynamic and are likely to increase with time as the project evolves, with the addition of new and related spectra, association with other forms of data such as computational models, and as connections to chemical structure are made or revised. Even after publication the relationships amongst the items in a collection are likely to change as related publications appear, and additional metadata may be needed to refer to later publications or datasets as well. Descriptive metadata, in contrast, are likely to be more static (and, in many cases, unchangeable for ethical reasons), being produced once and then not changed again.

2. Guidance for Digital Chemical Structure Representations in a FAIRSpec-Ready Data Collection

2.1 Digital Chemical Structure Representations

We start with guidelines for creating chemical structure representation specifically in relation to spectroscopic data within a FAIRSpec-ready collection.

There are many ways in which a structure can be represented, ranging from an image or diagram, through electronic formats that capture detailed atomic coordinates and connectivities, to linear representations, chemical names, and other standard identifiers. In practice, such representations perform many functions. Authors use images, for example, to convey aspects of chemical structure and bonding that relate to their finding. Cheminformaticians use digital representations to reference chemical properties. An important purpose of a digital structure representation in the context of spectroscopic data collections is to provide the critical digital metadata that allow finding and discussing the relationship between structure and spectroscopy. In cases where detailed post-acquisition spectral analysis has been carried out, specific structure representations are necessary to correlate specific atoms and/or functional groups in a compound's structure with specific signals in the spectroscopic data.

Digital structure representations conveying the chemical structures of compounds associated with spectroscopic data might include one or more of the file types given in Table 1, where pros and cons of each are mentioned. The most useful structure representations in our context describe the molecule in terms of atoms and bonds with 2D or 3D coordinates

(e.g. MDL-MOL[12,13] or CDXML[14] allowing automated generation of representations that can be included in metadata such as SMILES[15,16,17] and InChI[18] if appropriate. Although simple images are potentially valuable representations in a variety of contexts, they must not be the sole representation for a compound's structure, as they cannot generally be reliably converted to any of the other formats.

The key point here is not that one representation is inherently better than another, rather that we value the presence of *multiple* representations of a structure within a collection. We encourage researchers to provide as many structure representations as they are reasonably able to – 2D drawing file and an image, or a 2D structure as well as a 3D structure. In terms of a collection being FAIRSpec-*ready*, we do not need all the possible representations, only enough to generate additional ones through automated workflows – for example, just a simple SMILES or CDXML representation. We do not (and in many cases could not) seek to dictate a single correct method of structural representation; rather we have tried to present here methods most likely to preserve chemical information after machine processing and to highlight common pitfalls where information would be scrambled or lost.

Table 1: Common digital chemical structure representations — pros and cons

Representation type	**Considerations**
MDL-MOL Version 2000	Benefits: high interoperability, can contain 2D or 3D coordinates Limitations: limited ability to convey bonding
MDL-MOL Version 3000	Benefits: all the features of Version 2000, with expanded capabilities to describe special bonding types and multi-atom connection bonding (as in ferrocene) Limitations: less widely implemented to date in toolkits and informatics platforms.
CDXML	Benefits: well-specified structural format generated and read by popular drawing programs, allows for the expression of "nicknames" such as Ph and TBS, highly versatile in expressing nuances of bonding, may provide warnings of inappropriate bonding or charge states Limitations: generally, less interoperable than MOL (at least at the time of this writing); ambiguity can arise from using bonding and atom elements for nonmolecular depictions, such as titles, labels, and drawn lines; specification is no longer formally maintained, but last published working specification version is available (ref. 12)
CDX	Benefits: well-specified binary equivalent of CDXML, easily converted to CDXML with open-source tools Limitations: binary format lacks the human readability aspects of MOL and CDXML; see note for CDXML in relation to specification
SMILES	Benefits: 1D (character string) compact format, generally well-

	specified standard, allows for stereochemical ambiguity, allows explicit double bond or aromatic bond descriptions, easily interconvertible with 2D- or 3D-formats within the range of its applicability Limitations: there is no, universally accepted, "canonical" form; different toolkits and toolkit versions differ in interpretations, particularly of what "aromatic" means and where it is applicable. An IUPAC project to formalize guidelines for reliable use of SMILES is currently in progress.[19]
InChI	Benefits: 1D (character string) compact format, canonical, options can include or not include hydrogen and stereochemical "layers", easily derivable from 2D- or 3D-formats within the range of its applicability Limitations: depending upon the presence of layers, cannot always be converted unambiguously to 2D or 3D structure representations or to handle tautomeric isomers; does not currently encode bonding details for metal-containing compounds or some advanced stereochemistry features. Projects to address priority limitations are currently in progress.[20,21]
Image	Benefits: provides a readily interpretable and displayable option for finding aid viewers Limitations: does not generally allow for reliable error-free conversion to any of the other representations; **generally, not acceptable as the sole structure representation in a collection**
Chemical Name	Benefits: particularly the IUPAC preferred name for a compound, when appropriate, is the gold standard for unambiguous description of a chemical entity. Limitations: prone to errors that are not easily discoverable; requires complex processes for converting to other forms of chemical identifiers.

Thus, all chemical structure representations involve priorities and trade-offs. When a chemist draws a chemical structure, the purpose is generally to allow unambiguous communication with other chemists. Additionally, though, the structures we draw could be used also to communicate unambiguously with machines. However, often the qualities of a drawn structure that make it unambiguous to a human may introduce ambiguity when interpreted by a machine. The trade-offs and considerations described below are often necessary because the ecosystem of tools to draw, interchange, and process chemical structures electronically lacks the functionality (or broadly agreed upon methods) to handle some of the more nuanced aspects of structural representation. Ultimately, in the context of FAIR data management, the goal is to use a chemical structure to generate the metadata that enables both humans *and* machines to communicate efficiently and unambiguously with each other. Thus, the focus on unambiguous machine interpretation may lead to slightly different priorities for drawing a structure than chemists may be used to.

2.2 Generating Digital Chemical Structure Representations

What follows is a set of suggestions for best practices in producing the chemical structure representations associated with spectroscopic datasets.

1. **Provide multiple representations of the same structure when feasible.** The IUPAC FAIRSpec Principles embrace a variety and multitude of structure representations. Many of these structure representations are interconvertible. Thus, if a FAIRSpec-Ready data collection includes only a CDXML or MOL representation, that can generally be sufficient to generate all the other representations for inclusion in an IUPAC FAIRSpec Data Collection, such as one or more forms of SMILES or InChI. Nonetheless, the presence of multiple representations for a given structure allows for increased interoperability and improved opportunities for data and metadata validation.

2. **Include only one structure per file.** Drawing files containing multiple structures (such as reaction schemes or tables or figures for publication) generally cannot be processed by automated methods. Particularly in relation to spectroscopic collections, correlating specific structures with specific spectra requires that individual structures have their own digital representations (i.e. files). In the same vein, generic labels such as "R" or "X" ("Markush" drawings) should not be used to represent multiple compounds that differ only at specific locations, because doing so does not allow the sort of direct association with spectroscopic data we need in this context.

3. **Produce structure representations free of annotations.** In preparation for extraction of metadata from a structure, do not annotate structures with labels or text boxes. Numbers providing compound numbers or adjacent to atoms to identify specific atoms are generally not interpretable by software. Hydrogen bonds and other “weak” bonds, though perhaps important to a discussion, are not advisable for these structure-data associations. Notations such as *(R), (S), racemic, scalemic, 93% ee, 3:1 dr,* etc. should not be part of the structure representation itself. If such annotations are important to the discussion, provide a separate representation that includes them. Chemical names should not be included with structures, as this can sometimes result in errors in processing, and can lead to unmanageable image widths. In general, chemical names are not necessary in FAIRSpec-ready collections, as they can be generated from well-made structure or drawing files.

   To the extent that it is useful within the given context, we suggest that *the place for structure-specific annotation is in metadata, not within the structural representation itself.* As such, we describe in Section 5.1 several simple ways of adding annotations in a way that allows them to be associated with structural representations without being hidden within them.

   Most importantly, it is advisable to adhere to IUPAC-recommended structure depictions as much as possible. IUPAC recommendations are available for graphical representation of chemical structures[22] and the depiction of stereochemistry.[23] Ongoing IUPAC efforts are expected to expand upon these guidelines specifically in relation to machine-readable formats[24] .

4. **Take care when using abbreviations in a structure.** Abbreviated atom labels (e.g. ‘OTHP’) may cause problems when converting from a drawing program’s native format to other structure representations, such as MOL, SMILES, and InChI. It is advisable to look for errors in the drawing program - an abbreviated group flagged as a possible error by the program typically means that the drawing program cannot interpret the chemical meaning of the abbreviation. If in doubt, expand all abbreviations, at least temporarily, just to check. Or, if the program allows, check that the calculated molecular formula matches the intended structure. (Is “PMB” *phosphorus-metal-boron* or *para-methoxybenzyl*?)

5. **For mixtures of compounds, consider the spectroscopic context.** Here we wish to distinguish between *structure representation*, *compound,* and *sample.* For our purposes here, a structure *representation* is a digital item, a series of bytes, whether that be a SMILES in the form of a string of characters or a MOL file or an image. We contrast that to a *compound*, which may or may not have an associated chemical structure or spectrum. Thus, we speak of "the structure of a compound", or "this compound's spectrum." Essentially, the term *compound* has an associative nature. It is the connecting link between a spectrum and its associated structure. Chemical *samples*, on the other hand – the actual starting point for experimental spectroscopy – are generally mixtures of compounds, and the "compounds" themselves may even be mixtures. It is not uncommon, for example, to see in publications the phrase *Compound 3c was a 10:1 mixture of diastereomers.* Whether or not this is correct usage of the term "compound" is not for us to say. After much discussion within our project group, we have come to the following context-based consensus:

    a. For a single compound for which the NMR spectrum shows (in the author’s opinion) minor impurities or residual solvent, only include the structure of the principal component. In the case of a compound reported with “high enantiomeric excess”, include only the structure of the major enantiomer if, in the given context, the minor enantiomer is nothing more than an undesired impurity.

    b. In contrast, if the context emphasizes the stereochemical nature of the mixture (for example, the analysis is from chromatography that separates and identifies enantiomers, or the NMR spectrum clearly indicates signals from two diastereomers and is used to determine their ratio), multiple structures should be included in individual files. Additional compound association-level descriptive metadata could provide information regarding the nature of the mixture.

    c. For racemates, include only the structure of one of the enantiomers unless the enantiomers are distinguishable by the associated spectroscopic data. Representation of racemates is a special case that is trivially depictable for human consumption but much more difficult to achieve for machine-readability. Current machine-readable formats lack a reliable, consistent way to store the information that the structure is racemic. We note that InChI has a flag for racemates, but it is not part of standard InChI. MOL files may use the "chiral flag" set to zero to indicate a racemic mixture, but this feature has not been implemented reliably in the past.[25]

Differentiation between use cases a,b, and c, may well be driven by the environment in which the work was carried out. For example, in pharmaceutical research and manufacturing it may well be a regulatory requirement that the spectroscopic data is measured specifically to identify and quantify extremely low-concentration compounds within a mixture such as when reporting toxic metabolite levels in a product. Here not only must the compound identification be carried through to the final documentation, it is essential that chiral information, where appropriate, is correctly reported as this can have a direct bearing on the toxicity of compound identified.

The underlying principle here is that data and metadata should not be mixed in the same representation unless the representations themselves demand it as part of its standard. Providing metadata separate from data makes the metadata itself significantly more findable. The FAIRSpec Finding Aid allows for any amount of additional annotation to be associated with structures as described in Section 5.1.

6. **Do the best you can with compounds that present unsolved challenges for structure description.** Few structure representations properly represent coordination and dative bonds or allow for multicenter attachments (as for many inorganic and organometallic compounds). Even when they do, it is not always obvious how to generate the proper representation for machine readability. At the very least, these structure representations should be accompanied by an image representation, as it is quite likely that neither SMILES nor InChI can properly describe them. Other cases exist (such as atropisomerism) where generating suitable machine-readable representations remains an unsolved challenge. Again, the guiding principle here is that the best representation is one that conveys the intention of the creator, whatever that representation might be. To the extent that this can be more than an image, all the better. For example, if a drawing program is used to produce the image, provide both the drawing program's native representation and the image.

7. **Use standard descriptions for macromolecules, supplemented with images.** Macromolecules such as proteins are always best represented by established formats such as PDB[26], mmCIF[27], or BinaryCIF[28], (the latter two being more extensible and more actively maintained). Additionally, complex biomolecules may be represented in a SMILES-like linear string using the Hierarchical Editing Language for Macromolecules (HELM),[29] a machine-readable linear notation supported by IUPAC and the Pistoia Alliance.[30] Nonetheless, images are welcome additional representations that can be used to supplement and provide meaningful additional interpretation and annotation of these formats. As for organic and inorganic polymers, and network solids, we give no specific guidance other than to provide "appropriate and meaningful" digital representations, whatever that might mean in the context of the collection, even if that is only an image. Additional metadata can be used to be more descriptive.

3. Guidance for Instrument Dataset Representations in a FAIRSpec-Ready Data Collection

In this section we outline ways in which the instrument dataset itself can be optimized for incorporation into an IUPAC FAIRSpec Data Collection. It is not for these guidelines to elevate one digital representation over another at the instrument level. On this point, we refer to the original recommendations for data representations as given in the *FAIR Guiding Principles for scientific data management and stewardship* as elaborated by GO FAIR[31]. Specifically:

F2. (Findability) Data are described with rich metadata

R1.3. (Reusability) Data should meet domain-relevant community standards

"Data" in the GO FAIR context means much more than just instrument "datasets". It includes chemical sample, structure, and analysis representations, as well as all the metadata associated with the collection. Nonetheless, with these goals in mind, specifically in relation to instrument datasets, we suggest:

1. **The original instrument dataset is an important representation.** There is no question that the potentially most important digital representation of an instrument dataset is the one that came from the instrument itself. If it is practical to provide this primary data, it should be provided. For NMR this implies that the FID is made available.
2. **Multiple representations are valued.** If we consider the likely prospects for data reuse, no single data representation will always be the most valuable in all contexts (Table 2).
3. **Generally, do not combine a molecular structure with its associated spectrum within a single representation**. While this might seem to be the obvious thing to do, and it can be handled in certain cases during metadata extraction, we suggest that it is bad practice to create "hard-coded" relationships between structure and instrument datasets that may or may not be the final story. The IUPAC FAIRSpec Data Collection separates structures from spectra to associate them in more flexible ways, for example when the need is for the structure only. Keeping these aspects separate digitally allows for easier later-stage reinterpretation of the data.
4. **Package only one dataset per file.** While some digital formats allow combining multiple instrument datasets for the important work of comparative analysis, to be FAIRSpec-Ready, just as for structures, there should only be one item per representation. For example, it is not sufficient to have a single file that contains all the spectra in a collection, as each collection may have its own specific associations (all the spectra for *this* compound, all the spectra of *this* type, etc.), and the ability to repackage data into (initially unknowable) different collections is an important aspect of the IUPAC FAIRSpec Data Collection. In the case where one file contains both a spectrum and structure representations, it is particularly critical that there be only one spectrum with only one structure (and that both are extractable independently).

Table 2: Digital dataset representation examples

Representation type	**Considerations**
Original instrument data (for example, the NMR FID and associated parameter files)	Benefits: highest integrity with no information loss; high potential for reuse; allows for alternative and/or automated (re)analysis; potential for generation of all other representations; allows possibility of fraud detection; allows the most reliable automated metadata extraction. Limitations: vendor-specific format may require licensing access to a vendor-specific reader; may not express important aspects of the data processing used in the analysis; must contain enough of the key parameters for further processing; may no longer be documented; shortest expected lifetime.
Original data exported to an alternative standardized format, such as JCAMP-DX(FID)[32] NMR-Star[33] or nmrML[34]	Benefits: highest possibility of reuse and interoperability; vendor-agnostic open format; allows for automated production of additional representations; recognized by regulators as potentially the longest expected lifetime. Limitations: Depending on implementations potential for loss of data resolution or precision; potential for loss of some metadata fields.
Instrument-processed data such as transformed NMR data in the form of a spectrum or spectra	Benefits: most generally and immediately informative to the practicing scientist or educator; can be examined in detail and repurposed. Limitations: does not allow for early-processing adjustments, such as in NMR spectroscopy phasing, line broadening or use of non Fourier-transform methods; may require proprietary software to read; information loss compared to original data; less scope for fraud detection.
Third party-processed data	Benefits: concise; convenient if this is the standard process in each laboratory; may include meaningful annotation added by the originator such as integration or peak identification"; May be the only method of getting FAIR metadata to be associated with the data. Limitations: format may require proprietary software to read; possibly limited ability to extract key metadata from proprietary formats; may disallow alternative analysis; may be more subject to fraud or other sorts of spectral editing (removal of solvent peaks, for example)
Inline string description (for example, in NMR spectroscopy) a string describing field strength, solvent, chemical shifts, coupling constants, and integration	Benefits: concise; a common requirement for publication as part of the experimental details; distills the essential features of the spectrum. Limitations: minimally informative.
peak listing or peak table	Benefits: easily machine-readable; possibly all that is needed for some forms of reuse.

	Limitations: minimal semantic information; requires additional context and metadata for interpretation.
image	Benefits: most immediately identifiable and informative to working chemists, educators, and students; excellent for accompanying more robust representations. Limitations: almost certainly reduced resolution; no additional processing possible; susceptible to crude data editing; does not generally allow for conversion to any of the other representations; not acceptable as the sole structure representation in a collection; often cannot represent the details used by automated processing software.

4. Guidance for the Organization of Digital Items in a FAIRSpec-Ready Data Collection

The essential aspect of a FAIRSpec-ready data collection is that it is organized in a systematic manner that allows (1) machine-based extraction of key metadata and (2) unambiguous association of specific spectra with specific chemical structures (to the extent that such an association is possible). More specifically, the data and associated metadata should, with perhaps a small amount of additional curation, allow for the creation of an IUPAC FAIRSpec Finding Aid. We briefly introduce this metadata document first, then discuss the key elements of a FAIRSpec-ready collection.

4.1 The IUPAC FAIRSpec Finding Aid

An IUPAC FAIRSpec Finding Aid is a document that describes in detail the contents of a spectroscopic collection. The structure of the document is based on the *IUPAC FAIRSpec Metadata Object Model*[35], which describes how abstract objects (for example, “structures”, “spectra”, “compounds”, and “analyses”) and their various types of digital representations (CDXML and MOL files, instrumental data sets, PDF reports, etc.) are related and how they are to be described by structured metadata.

A key feature of the IUPAC FAIRSpec Finding Aid is that it is *extensible.* While it is expected to contain certain metadata in an IUPAC-specified format, the FAIRSpec Metadata Object Model allows researchers to add whatever additional metadata they wish to add, depending upon their needs. Thus, a key feature of a FAIRSpec-Ready collection is that it provides any additional ("non-extractable") metadata in a standardized format. We will see how this additional metadata can be represented in Section 5.

A *FAIRSpec-ready* data collection provides a means to create such a finding aid and its associated collection and landing page *via automation.* We want to be able to pass the FAIRSpec-ready data collection to a software tool (perhaps a public or private web site or a local software application), that can read the collection's digital items, extract the key descriptive and relational metadata, and create what we are calling an *IUPAC FAIRSpec Finding Aid*. In the process, the extractor may generate additional representations, such as

images of structures, predicted spectra, or peak listings. The extractor would only be limited by its sophistication, *but it could only work if the FAIRSpec-ready collection is properly organized*. This section describes how this organization might be achieved.

Without going into extensive detail here, the essence of an IUPAC FAIRSpec Finding Aid is that it consists of the following parts.

- A **mandatory header** section that
    - identifies itself as an IUPAC FAIRSpec Finding Aid
    - identifies the target repositories and pointers to local or remote data items
    - identifies related work such as publications or ELNs
    - specifies licensing and other access-related aspects of the collection
- An optional section listing **individual samples**, each with its own set of representations, key metadata, and identifier
- An optional section listing **individual structures**, each with its own set of representations, key metadata, and identifier
- An optional section listing **individual experimentally or computationally derived datasets**, each with its own set of representations, key metadata, and identifier (the specification allows for predicted, simulated, and experimental spectra to be represented, as long as they are identified as such in their associated metadata.)
- An optional **compounds** section making one-to-one, one-to-many, or many-to-one associations among items on two or more of the above lists. For example, a collection of sample-structure, sample-spectra or structure-spectrum relationships.
- An optional section listing individual **structure-spectral analyses**, each with its own set of representations and relationships to specific spectra and their related structures
- Additional custom sections as needed.

In principle, except for the header, any combination of the additional sections is possible. Thus, an IUPAC FAIRSpec Data Collection could be as simple as a single structure representation and a single spectrum. In that case, its IUPAC FAIRSpec Finding Aid might look like this (figuratively):

> Header: IUPAC FAIRSpec Finding Aid
> Structures: [#1 (points to a CDXML file)]
> Spectra: [#1 (points to a spectrum)]
> Compounds: [#1, ["Structure#1", "Spectrum#1"]]

Importantly, the collection should not consist simply of a set of spectra with no structures or other meaningful associations.

4.2 The FAIRSpec-ready collection

The IUPAC FAIRSpec Finding Aid describes an IUPAC FAIRSpec Data Collection. The guidelines presented here for ongoing research data collections or collections submitted to agencies as bona fide FAIRSpec-ready collections have the following features:

1. **The most important characteristic of a FAIRSpec-ready data collection is that it is organized systematically and consistently.** In principle, any systematic organization that conveys appropriate relationships can be processed to become an IUPAC FAIRSpec Data Collection with an associated IUPAC FAIRSpec Finding Aid. Nonetheless, the organizational principles described herein specifically illustrate a small number of suitable FAIRSpec-ready data collection organizations.

2. **Organization will depend upon context.** The FAIRSpec-ready Data Collection created on the day a dataset is generated will likely look quite different from one that ultimately is used in creating the IUPAC FAIRSpec Data Collection associated with a publication. At the beginning of an endeavor, for example, there is typically just a physical sample. A spectrum is taken. There may be the *expectation* of a certain structure, but perhaps not. If nothing else, it is hoped that the spectrum will at least support a hypothesis relating to chemical structure. It may be the case that the structure is truly unknown, and it is only after the spectrum is analyzed that it is "known". (More precisely, only with the help of spectroscopy can the structure be *hypothesized*.) Thus, an initial collection may be just one or more spectra and their associated sample identifier. One or more (plausible) structures are added a week later. More spectra are taken. More samples are created and analyzed. Everything is sample-based at first, but then, later, the key organizing principle shifts to chemical structure – chemical "compounds". Ultimately, upon publication, the key organizing principle will be "compound number in the article," and only selected spectra will be included. Importantly, *the underlying data have not changed. (We hope!) Only the organization has changed.* And, importantly, the FAIRSpec-ready collection has probably not changed at all, except for some key metadata.

3. **Extractor utilities may impose their own conventions.** Individual tools developed to extract metadata from a data collection to create an IUPAC FAIRSpec Data Collection and its associated IUPAC FAIRSpec Finding Aid may develop their own specific requirements that go beyond what is suggested here. However, an FAIRSpec-ready data collection must follow the conventions described here. Metadata tools that refer to themselves as FAIRSpec-ready extractors may enforce occurrences of "should" in what follows as "must" but may not impose additional restrictions that contradicts these conventions. For example, an extractor might allow Unicode characters outside the specified range for identifiers, but if it does, it must change those characters to the allowed set for identifiers described for IUPAC FAIRSpec Data Collections.

4. **Unique identifiers should be used for samples, structures, and datasets.** One characteristic of the IUPAC FAIRSpec Finding Aid is that it involves "pointing" to digital representations. This is analogous to file names in a file system – and, for that matter, may be exactly that. For example, a compound might have a name "C3" as its identifier; an NMR dataset might be referred to as C3_H1_NMR. It is not important that these identifiers be semantic – that is, that they convey meaning, such as this example does – and, in fact, it is often desirable that they *not* be descriptive. They simply need to be *unique*. The following guidelines for identifiers should be followed:

a. Characters used throughout the metadata should be represented in the UTF-8 character set. This requirement allows for the full range of international language characters and is consistent with current web standards.

b. Characters specifically in relation to *identifiers* should be limited to 7-bit ASCII characters. IUPAC FAIRSpec Finding Aids utilize string-based identifiers for cross-referencing digital objects. For flexibility of processing, identifiers should only utilize simple alphanumeric [A-Za-z0-9] ASCII characters along with a limited set of punctuation, namely "+-'.,_()[]". Single spaces are allowed only if not leading or trailing; multiple sequential spaces characters, tab, and new-line characters are not allowed within identifiers.

5. **Data should be organized by sample or compound identifier.** IUPAC FAIRSpec Finding Aids primarily are intended to be used in the context of chemical compounds with defined chemical structure. Even when they are not, the associations referred to as "Compound Associations" are the preferred organizing principle. This allows for the natural relationship of "structure *of a compound*", and "spectrum *of a compound",* with a direct relationship to typical publications in the field. For example:

```
1a/
        NMR/
                1H-NMR/
                13C-NMR/
2a/
        NMR/
                1H-NMR/
                13C-NMR/
```

Here we have unique identifiers of the form 2a/NMR/1H-NMR/xxxx. The file system-like hierarchy provides uniqueness. Note that these "file paths" are themselves semantic. The file path itself is conveying the relationship between structure and spectrum that will become codified in the IUPAC FAIRSpec Finding Aid.

6. **Identifiers need not be related to compounds.** We use the term "compound association" to include associations of any number of spectra with any number of structures (including zero in each case). When a collection is being created at a stage in the research endeavor when the compound has not been identified, any meaningful identifier (meaningful to the researcher, that is) can be used as a proxy. For example, "RMH-III-23.rf-0.65" might be sufficient immediately after chromatographic isolation, with no structure representations provided. At a later stage in the research, when the identity of the isolated compound is better known, that might be changed to "RMH00123" and associated using that ID with additional spectra and a specific CDXML or MOL representation. At the point of publication, this might be changed to "3b". If names are being assigned by multiple researchers, a metadata document describing these associations (typically a spreadsheet or a line-based comma- or tab-separated values text file) is appropriate. For instance, we might have data organized by instrument and compound:

```
NMR/
```

RMH-IV-13b/
RMH-IV-13c/
RMH-IV-13d/
IR/
HSR-II-112/
HSR-II-113/
HSR-II-114/

or the other way around:

RMH-IV-13b/
NMR/
IR/
RMH-IV-13c/
NMR/
IR/
RMH-IV-13d/
NMR/
IR/
HSR-II-112/
NMR/
IR/
HSR-II-113/
NMR/
IR/
HSR-II-114/
NMR/
IR/

etc., where unique laboratory sample identifiers are used instead of compound identifiers, because at the stage in the process, that is all that is available.

Whatever system is used, it should be systematic, with unique identifiers (think "file paths" for each item in the collection. An example of such an organized collection is the one associated with our development work for a publication in inorganic chemistry[36] which is one of the implementations highlighted in Part 2 of this series. In that case, the top level comprises a list of the instrumental or computational techniques applied, one of them being NMR spectroscopy itself, rather than being based on a top-level list of compound identifiers. This alternative mode arose in this case because the sources of the data were from specific ELNs used to create automated collections of datasets associated with the specific technique supported by the ELN.

Organizing by compound or sample identifier could be facilitated by ELNs that allow this mode. We feel that the benefits of doing so, including the ELN-based creation of the IUPAC FAIRSpec Finding Aid based on chemical compounds, would be considerable. This principle could also be included in the specifications for future designs of "IUPAC FAIRSpec-Compliant ELNs.

7. **The organization of files may reflect the types of spectroscopy used.** Subdirectories "NMR", "IR", "UVVIS", and "HRMS" (high-resolution mass spectrometry) can assist in both human and machine readability.

8. **The organization of files should minimize the duplication of information.** Collections could be used in multiple contexts. Ongoing research collections are generally created prior to any knowledge of final publication compound numbers. Data published for one manuscript, with a specific set of compound numbers might also be referred to in another manuscript, with other compound numbers. If maximum flexibility is desired, paths such as "3a/3a-NMR/3a-1HNMR" should be avoided, using simply "3a/NMR/1HNMR" instead. In this way, if "3a" is not the final compound number in the publication, or the collection is used in multiple contexts, there is only one directory to rename.

9. **Structure representations can be added along the path to a dataset.**

   One of the simplest ways to associate structures with spectra is to place a structure representation within the directory path leading to the spectral data. Thus, using the example given above, we might have:

```
RMH-IV-13b/
        structure.cdxml
        NMR/
        IR/
RMH-IV-13c/
        structure.cdxml
        NMR/
        IR/
RMH-IV-13d/
        structure.cdxml
… etc.
```

   This is enough to make the key association between structure and spectrum in simple cases.

10. **Structure representations can be integrated into datasets if desired.**

    Some vendors allow one or more structure files to be added to an instrument dataset. For example:

```
RMH-IV-13b/
        NMR/
                NMR-2024.04.15a/
                        structure.mol
                NMR-2024.04.15b/
                        structure.mol
        IR/
RMH-IV-13c/
```

NMR/
structure.mol
IR/
…etc.

This can work, but it suffers from the issue that there is considerable duplication. Should the structural hypothesis be modified, then we have the error-prone challenge of modifying all of the structural representations within this directory path as well.

11. **Structure representations can be placed in a parallel set of directories.** Perhaps the cleanest way to associate structures with spectra is to keep them separate, but to provide associated identifiers. This could be sample based, as in:

data/
RMH-IV-13b/
NMR/
IR/
RMH-IV-13c/
NMR/
IR/
… …
structures/
RMH-IV-13b/
structure.cdxml
RMH-IV-13c/
structure.cdxml
…

This is, in fact, exactly the way the IUPAC FAIRSpec Finding Aid itself is laid out. It has the advantage that there is no duplication. Unique identifiers (e.g., RM-IV-13b) make the metadata connection between structures and their associated spectra. A single modification of a structure will be registered for all its associated spectra automatically.

12. **Compound associations that are mixtures preferably should identify as a mixture using a "+" sign and include separately identified structure representations, one for each component of the mixture, within a "structures" subdirectory.** Thus:

3c+3c'/
structures/
3c.cdxml
3c'.cdxml
NMR/
…

This is enough to make it clear that the multiple CDXML files are of different structures, as opposed to simply alternative representations of the same compound.

13. **Alternatively, and less preferred, if the context demands, isomeric mixtures may be described by ambiguous structure representations.** In special cases, there may be little or no practical relevance of the exact stereochemistry of a structure. In that case, no special effort should be expended to detail it. For example, if the manuscript refers to "Compound 3c" as having a THP group (which has a stereocenter) and no analysis was carried out that determined the stereochemistry of this group, a single CDXML file with no stereochemistry indicated (just the nickname "THP") can be provided. However, if the manuscript refers to "Compound 3c" as a "3:1 mixture of E/Z isomers", then two separate CDXML files, one Z and one E should be provided. It is not necessary to describe the extent of the mixture within the structural context. Metadata can be added later to the compound association itself annotating the specifics of the mixture. Future guidance may allow for a more systematic way of describing mixtures, for example, with MinChI[37,38] with a method designed specifically for IUPAC FAIRSpec Finding Aids.

5. Addition of Curated Metadata to a FAIRSpec-Ready Collection

5.1 Internal metadata records

While structure and instrument data representations in a spectroscopic data collection can be significant sources for the automated extraction of *descriptive* metadata, not all descriptive metadata can be gathered this way. In addition, *relational* metadata records, such as related sample identifiers, are less easily produced or extracted. In addition, by its very nature, a relational metadata record is not associated with one specific digital item. Thus, its proper place is not "within" any of the digital items it relates to. For example, we have seen how a file structure itself can be the basis of relational metadata. ("This structure is associated with this spectrum because they are in the same directory.") But, in general, addition of relational metadata will require *curation.*

Ultimately, key/value metadata associated with an IUPAC FAIRSpec Data Collection (not just a FAIRSpec-*Ready* one) will be expressed primarily in terms of the IUPAC FAIRData Model-specified standard. Thus, for example, descriptive metadata associated with an NMR spectrum in an IUPAC FAIRSpec Finding Aid might appear as shown in Fig. 3, where the full key name for nmr.expt_solvent, for example, is

IFD.property.dataobject.fairspec.nmr.expt_solvent

```
propertyPrefix:                  "IFD.property.dataobject.fairspec"
▾ properties:
    nmr.expt_absolute_temperature:   296.0438
    nmr.expt_dimension:              "1D"
    nmr.expt_freq1:                  100.63037
    nmr.expt_freq2:                  400.2
    nmr.expt_id:                     "c13"
    nmr.expt_nucl1:                  "13C"
    nmr.expt_nucl2:                  "1H"
    nmr.expt_pulse_program:          "zgpg30"
    nmr.expt_solvent:                "MeOD"
    nmr.expt_title:                  "TM-VI-251, Na+PMB-pyr 13C"
    nmr.instr_manufacturer_name:     "Bruker"
    nmr.instr_nominal_freq:          400
    nmr.instr_probe_type:            "5 mm PABBO BB/19F-1H/D Z-GRD Z108618/0621"
    nmr.proc_timestamp:              "2022-01-24T23:05:38Z"
```


Figure 3. Descriptive metadata associated with a Bruker NMR instrument dataset as found in JSON format in an IUPAC FAIRSpec Finding Aid.

It is not expected that a FAIRSpec-ready collection utilizes such standardized keys. And, in fact, it is preferred that metadata such as these, that can be extracted automatically from an instrumental dataset via automation, not be provided explicitly in a FAIRSpec-ready collection at all.

It is important to understand that the IUPAC FAIRSpec Finding Aid allows for both standardized and “ad hoc” properties. Thus, additional metadata that does not fit easily into the standard can always be added. (Such metadata simply would not have any *IFD.property* prefix.) To the extent that *ad hoc* metadata key/value pairs can be mapped to current or future IUPAC FAIRData Standard pairs, that mapping would be done later, during the automated or semi-automated curation process, when metadata extraction is carried out.

We provide here suggestions for adding additional *ad hoc* metadata based on our implementation tests. These points will be elaborated more extensively in Part 2 of this series.

1. **Consistency is important.** In all cases, identifiers and keys should be unique and consistently expressed (including spelling and capitalization). Values should use a consistent, if not standardized, vocabulary.

2. **Point-specific metadata files can provide additional descriptive or relational metadata not easily otherwise conveyed.** Key:value metadata pairs can be listed in a simple text file accompanying structures or data within a collection. For instrument datasets, this is already the practice of some spectrometer vendors, who have adopted a format resembling the IUPAC JCAMP-DX format, storing metadata

specific to that dataset in the form of "private" ##$KEY=VALUE pairs. A generalized format could just be a collection of single lines of the form *key=value*, (popularized in Java properties files) where the keys should be systematically defined and consistent throughout the collection. For example:

If the file *metadata.properties* (in the form of a Java properties file) were added to an instrument dataset containing only

sample_id=RMH-IV-23c

then later automated curation could convert that to an IUPAC FAIRData Sample object with IFD.property.sample.id “RMH-IV-23c”, and associate the specified sample with all spectra contained in this dataset.

Such a file (perhaps created automatically by an ELN) for a specific structure representation might contain metadata such as:

smiles=CC1=CCC(CC1)C(=C)C
inchi=InChI=1S/C10H16/c1-8(2)10-6-4-9(3)5-7-10/h4,10H,1,5-7H2,2-3H3
chebi_id=CHEBI:15384
mf=C10H16

that could be used to complement automated extraction. In addition, such explicitly added metadata can be used to validate accompanying structure representations such as MOL or CDXML files, confirming that the machine reading of a structure has been successful.

Such a file containing

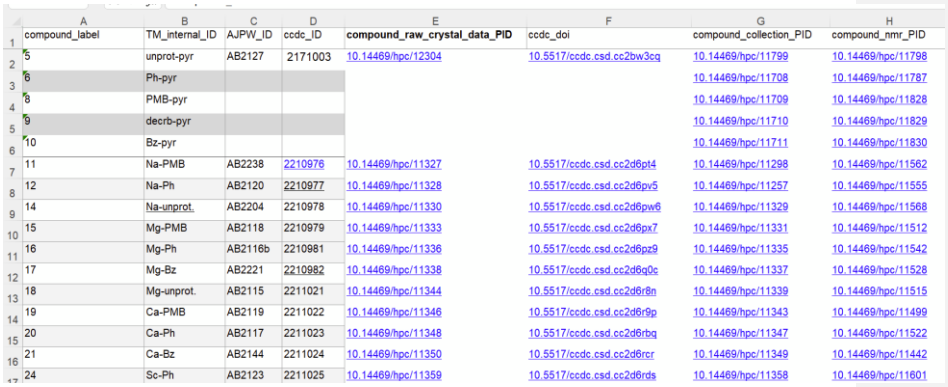
structure_stereochemistry=relative

contained in a structure directory or even in one of the top levels of the collection would be enough to convey the understanding that all chiral structures unless otherwise indicated are to be considered racemic.

3. **A well-organized "primary" spreadsheet can efficiently convey metadata relationships.** Metadata items are essentially key:value pairs. A convenient way to represent such pairs is the standard spreadsheet practice where column headers are keys, and each row contains the set of values associated with a given item (generally identified in the first column). Particularly for a whole collection, a primary internal metadata record in the form of a spreadsheet can be efficient. Standard open file formats such as CSV (comma-separated values)[39], TSV (tab-separated values)[40], ODS (OpenDocument spreadsheet)[41], or XLSX (Office Open XML SpreadsheetML file format)[42], are easily extractable via automation for the metadata they contain.

Thus, we might have something like what is shown in Fig. 4, associating publication compound identifiers with lab-local identifiers, database identifiers, and repository persistent identifiers. When (or if) incorporated into an IUPAC FAIRSpec Finding Aid, these properties

would be either mapped to standardized IUPAC FAIRSpec metadata keys or included as additional properties.



	A	B	C	D	E	F	G	H
1	compound_label	TM_internal_ID	AJPW_ID	ccdc_ID	compound_raw_crystal_data_PID	ccdc_doi	compound_collection_PID	compound_nmr_PID
2	5	unprot-pyr	AB2127	2171003	10.14469/hpc/12304	10.5517/ccdc.csd.cc2bw3cq	10.14469/hpc/11799	10.14469/hpc/11798
3	6	Ph-pyr					10.14469/hpc/11708	10.14469/hpc/11787
4	8	PMB-pyr					10.14469/hpc/11709	10.14469/hpc/11828
5	9	decrb-pyr					10.14469/hpc/11710	10.14469/hpc/11829
6	10	Bz-pyr					10.14469/hpc/11711	10.14469/hpc/11830
7	11	Na-PMB	AB2238	2210976	10.14469/hpc/11327	10.5517/ccdc.csd.cc2d6pt4	10.14469/hpc/11298	10.14469/hpc/11562
8	12	Na-Ph	AB2120	2210977	10.14469/hpc/11328	10.5517/ccdc.csd.cc2d6pv5	10.14469/hpc/11257	10.14469/hpc/11555
9	14	Na-unprot.	AB2204	2210978	10.14469/hpc/11330	10.5517/ccdc.csd.cc2d6pw6	10.14469/hpc/11329	10.14469/hpc/11568
10	15	Mg-PMB	AB2118	2210979	10.14469/hpc/11333	10.5517/ccdc.csd.cc2d6px7	10.14469/hpc/11331	10.14469/hpc/11512
11	16	Mg-Ph	AB2116b	2210981	10.14469/hpc/11336	10.5517/ccdc.csd.cc2d6pz9	10.14469/hpc/11335	10.14469/hpc/11542
12	17	Mg-Bz	AB2221	2210982	10.14469/hpc/11338	10.5517/ccdc.csd.cc2d6q0c	10.14469/hpc/11337	10.14469/hpc/11528
13	18	Mg-unprot.	AB2115	2211021	10.14469/hpc/11344	10.5517/ccdc.csd.cc2d6r8n	10.14469/hpc/11339	10.14469/hpc/11515
14	19	Ca-PMB	AB2119	2211022	10.14469/hpc/11346	10.5517/ccdc.csd.cc2d6r9p	10.14469/hpc/11343	10.14469/hpc/11499
15	20	Ca-Ph	AB2117	2211023	10.14469/hpc/11348	10.5517/ccdc.csd.cc2d6rbq	10.14469/hpc/11347	10.14469/hpc/11522
16	21	Ca-Bz	AB2144	2211024	10.14469/hpc/11350	10.5517/ccdc.csd.cc2d6rcr	10.14469/hpc/11349	10.14469/hpc/11442
17	24	Sc-Ph	AB2123	2211025	10.14469/hpc/11359	10.5517/ccdc.csd.cc2d6rds	10.14469/hpc/11358	10.14469/hpc/11601

Figure 4. A page in a spreadsheet with metadata that can be extracted using automation for additional metadata attributes associated with compounds in a published IUPAC FAIRSpec Data Collection. The spreadsheet provides both human- and machine-readable content.

5.2 Registered metadata records

Quite possibly, each collection ultimately might be associated with a primary registered metadata record conforming to a declared schema such as the DataCite Schema[43]. An example of a DataCite metadata record can be found at https://data.datacite.org/application/vnd.datacite.datacite+xml/10.14469/hpc/10703. This metadata record then allows the connection to be made to additional metadata records as appropriate both within the collection and to associated works. These metadata records should be as chemically rich as possible. An example of how a *single* DOI reference can be use to generate a full IUPAC FAIRSpec Finding Aid for a collection is given on the FAIRSpec GitHub pages[44].

Registering a metadata record for individual instrumental datasets allows for the formation of a unique persistent identifier (PID) to be associated with individual parts of the dataset, enabling a higher probability of findability via automated search engine “bots”. Such registered records should include the unique path or the database reference to the exact location of the dataset itself - whether located on an institutional, specialist, or generalist repository - thus enabling potentially widespread accessibility to the dataset itself. This also allows for distributed data storage, and it also can include selective privacy settings for both pre- and post-publication.

The criteria for selecting an appropriate repository for registering the dataset metadata record should include some consideration of whether the entry of rich chemical metadata is adequately supported either by the repository human user interface or by a repository application programming interface (API). APIs (in particular) are essential for use by automated systems such as an ELN. The repository should include processes for automatic

generation and then registration of the primary metadata record with an appropriate authority, where the master copy will be kept and indexed to facilitate findability. Any local repository copy of the master metadata record should automatically be kept synchronized with the registered version.

Ideally, the repository would produce IUPAC FAIRSpec Finding Aids for its various collections, and these would also be registered with an agency as part of the overall collection. It could also produce IUPAC FAIRSpec Finding Aids in response to search queries as a way of standardizing API calls among various repositories, building them only as needed in response to queries.

A more in-depth discussion of metadata registration optimized for spectroscopy can be found in Part 2 of this series.

6. Summary

We have provided a set of guidelines for the FAIR management of spectroscopic data collections specific to the domain of chemistry. These guidelines cover the generation of machine-readable structure and dataset representations, the organization of what we refer to as the *FAIRSpec-ready collection*, and suggestions for ways to incorporate additional metadata into the collection. This collection optimally would be able to be curated automatically by software to create an *IUPAC FAIRSpec Finding Aid*.

We have emphasized that private FAIRSpec-ready collections can be developed automatically or semi-automatically concurrently with ongoing laboratory research, not just at the time of publication. The potential benefits of this “best practice” include the ability to carry out *on an ongoing basis,* with or without use of an ELN, structure or substructure searches for related spectroscopic data, filtering a collection for spectra of a certain sort or with a given set of property values. In fact, the local benefit of even the most minimal curation (just associating instrument datasets and analysis with specific samples or chemical structures) can be significant.

Researchers do not have to wait until the overall IUPAC FAIRSpec Metadata Model is formalized, nor do they ever have to become experts in its implementation to benefit from these simple guidelines. The key premise here is that a small amount of forward-thinking organization and consistency can go a long way to enabling the findability, interoperability, accessibility, and reusability of spectroscopic data collections, as well as the individual datasets contained within them. As a bonus, publishing of IUPAC FAIRSpec Data Collections allows for longer-term and more diverse exposure for both researchers and their publications.

Appendix: Terminology

In this appendix, we refer to several terms that have multiple meanings in common usage but specific meanings within this context:

compound association
Definitions of "compound" abound. The National Cancer Institute defines a compound as “a substance made from two or more different elements that have been chemically joined”.[45] PubChem describes a "Compound record" as pointing to “at least one Substance record”.[46] The IUPAC Gold Book defines a "racemic compound" as a “crystalline racemate in which the two enantiomers [chirally related "molecular entities"] are present in equal amounts in a well-defined arrangement within the lattice of a homogeneous crystalline addition compound”.[47] Common parlance in published works in the area of organic and inorganic chemistry refer to "Compound XXX, which was a mixture of diastereomers" and "pure compounds" (implying “a compound” can be “impure”). A search of the international patent site WIPO IP Portal for "polymer compound" within the “front page” field returns over 1300 results[48]. Notice that none of these definitions define specifically what makes one compound different from another. None answer some of the most basic questions: Can a compound be a mixture of compounds? Is a mixture of polymers a compound? We opt in these guidelines for a practical, inclusive definition of compound. In this context, a compound association is a metadata object with a unique identifier associating one or more spectra with one or more structures. No more, no less. We leave it to the originator to provide a meaningful identifier within whatever context the collection is found, and the components of the association to speak for themselves. For example, an ID of "2a+2a"' suggests that two different compounds are associated with one or more spectra. Within that association, the structure representations will provide the details of that relationship – whether these are diastereomers, enantiomers, or completely constitutionally different compounds.

dataset
A *digital item* or a collection of digital items that is derived from laboratory analysis (see *Instrumental Dataset*, below) or from computation. The two general forms of computed datasets are spectroscopic predictions based on one or more proposed chemical structures (such as can be generated at the nmrdb.org website[49]) and simulations based on experimental or predicted parameters, as NMR frequencies, chemical shifts, and coupling constants.

digital item
A collection of bytes that may be local to a digital collection or may be part of a remotely accessed collection, pointed to by a URL. Commonly implemented as a "file" on a filesystem or a named item within a compressed archive (for example, an archive with ZIP, TAR, or TGZ format).

digital object
A *digital item* that has associated metadata. In the current context, we make the distinction between the digital items in FAIRSpec-ready data collections, which do not have associated

IUPAC FAIRSpec metadata, and digital objects in IUPAC FAIRSpec Data Collections, which do.

finding aid
In general, a finding aid is a document that assists users in locating items in an archive. A widely used standard (EAD-3) developed by the US Library of Congress exists for archive-related digital finding aids and is used extensively throughout the archival community.[50] We extend that concept to locating *digital items* in a data collection and refer to our specification as the IUPAC FAIRSpec Finding Aid.

instrument dataset
A *dataset* that comprises the raw or minimally transformed data arising from laboratory analysis. Examples would be a file directory or ZIP archive containing NMR free-induction decay data or real- and imaginary-valued transformed data, parameter files, and accompanying metadata.

persistent identifier
A publicly registered character string, such as *10.1515/pac-2021-2009*, providing a long-lasting reference to a *digital object* and its associated metadata.

representation
A *digital object* present within a collection and identifiable using an identifier that is unique within an appropriate context. Representations may be individual *digital items* within a collection referenced by the IUPAC FAIRSpec Finding Aid. Alternatively (or additionally), if they are relatively short, they can be character strings (including Base64-encoded[51] byte arrays) that are contained within a metadata document. For example, a structure file may be present within the collection and have the unique file name "3aa/structures/3aa.mol"; a SMILES string representation need not be in a file of its own; it can be the provided in the finding aid itself as the data value of an IFD.representation.structure.SMILES representation. Similarly, a PNG image may end up represented within an IUPAC FAIRSpec Finding Aid as a Base64-encoded string so that is more easily accessible or because it was originally within a more complex *dataset* file.

List of Abbreviations

API
application programming interface

ASCII
American standard code for information interchange

CDX
ChemDraw exchange format

CDXML
ChemDraw XML format

CSV
comma-separated values

DOI
digital object identifier

EAD
encoded archival description

ELN
electronic laboratory notebook

ESI
electronic supplementary information

expt
experiment

FAIR
findable, accessible, interoperable and reusable

FID
free induction decay

HELM
hierarchical editing language for macromolecules

HRMS
high-resolution mass spectrometry

IFD
IUPAC FAIRData

IR
infrared

JCAMP-DX
Joint Committee on Atomic and Molecular Physical Data

JSON
JavaScript object notation

MDL
Molecular Design Limited

mmCIF
macromolecular crystallographic information file

MOL

Commented [MA1]: What does the 'DX' stand for? Data exchange?

molfile

NMR
nuclear magnetic resonance

NSF
National Science Foundation

ODS
OpenDocument spreadsheet

OTHP
tetrahydropyranyl ether

PDB
Protein Data Bank

PDF
portable document format

Ph
phenyl

PID
persistent identifier

PNG
portable network graphics

SDF
structure-data file

SMILES
simplified molecular input line entry system

TBS
tert-butyldimethylsilyl

TSV
tab-separated values

URL
uniform resource locator

UTF-8
Unicode transformation format – 8-bit

UVVIS
ultraviolet–visible spectroscopy

WIPO
World Intellectual Property Organization

XLSX
Office Open XML SpreadsheetML

XML
extensible markup language

References

[1] *Development of a Standard for FAIR Data Management of Spectroscopic Data*. Project Details. https://iupac.org/project/2019-031-1-024 (accessed 2024-03-26).

[2] *IUPAC/IUPAC-FAIRSpec GitHub Project Site*. https://iupac.github.io/IUPAC-FAIRSpec (accessed 2024-12-03).

[3] Several examples of IUPAC FAIRSpec Finding Aids and their associated landing pages can be found at *IUPAC/IUPAC-FAIRSpec GitHub Project Site Web Pages*. https://github.io/IUPAC/IUPAC-FAIRSpec (accessed 2024-12-03).

[4] Hanson, R. M.; Jeannerat, D.; Archibald, M.; Bruno, I. J.; Chalk, S. J.; Davies, A. N.; Lancashire, R. J.; Lang, J.; Rzepa, H. S. IUPAC specification for the FAIR management of spectroscopic data in chemistry (IUPAC FAIRSpec) – guiding principles. *Pure Appl. Chem.* **2022**, *94*(6), 623–636. https://doi.org/10.1515/pac-2021-2009.

[5] *Chapter II: Proposal Preparation Instructions - Proposal & Award Policies & Procedures Guide (PAPPG) (NSF 23-1) | NSF - National Science Foundation*. 2023. https://new.nsf.gov/policies/pappg/23-1/ch-2-proposal-preparation (accessed 2024-05-29).

[6] *Dissemination and Sharing of Research Results | NSF - National Science Foundation*. https://www.nsf.gov/bfa/dias/policy/dmp.jsp http://web.archive.org/web/*/https://www.nsf.gov/bfa/dias/policy/dmp.jsp (accessed 2020-10-27).

[7] *US Government Code of Federal Regulations 21 CFR 11.10 Electronic Records – Controls for Closed Systems* https://www.ecfr.gov/current/title-21/chapter-I/subchapter-A/part-11/subpart-B/section-11.10 (accessed 2024-12-03)

[8] Patel, K. T.; Chotai, N. P. Documentation and Records: Harmonized GMP Requirements. *J Young Pharm* **2011,** 3(2), 138-150. Available as a National Library of Medicine online article as https://pmc.ncbi.nlm.nih.gov/articles/PMC3122044 (accessed 2024-12-03).

[9] 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.; 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.; van der Lei, J.; van Mulligen, E.; 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*(1), 160018. https://doi.org/10.1038/sdata.2016.18.

[10] 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. *J. Cheminformatics* **2017**, *9*(1), 54. https://doi.org/10.1186/s13321-017-0240-0.

[11] *DataCite: Helping you to find access and reuse research data*. http://www.datacite.org (accessed 2024-12-03).

[12] 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. *J. Chem. Inf. Comput. Sci.* **1992**, *32*(3), 244–255. https://doi.org/10.1021/ci00007a012.

[13] Dassault Systèmes. *CTfile Formats*. Dassault Systèmes. 2020. https://discover.3ds.com/ctfile-documentation-request-form (accessed 2022-01-12).

[14] *CDX Format Specification*. https://iupac.github.io/IUPAC-FAIRSpec/cdx_sdk, (accessed 2024-10-23), which was salvaged from https://web.archive.org/web/20190910135652/https://www.cambridgesoft.com/services/documentation/sdk/chemdraw/cdx/General.htm and related documents (accessed 2024-05-29).

[15] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28*(1), 31–36. https://doi.org/10.1021/ci00057a005.

[16] *Daylight Theory: SMILES*. https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html (accessed 2022-01-12).

[17] *OpenSMILES specification*. http://opensmiles.org/opensmiles.html (accessed 2022-04-14).

[18] Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI - the worldwide chemical structure identifier standard. *J. Cheminformatics* **2013**, *5*(1), 7. https://doi.org/10.1186/1758-2946-5-7.

[19] *IUPAC SMILES+ Specification*. IUPAC | International Union of Pure and Applied Chemistry. https://iupac.org/project/2019-002-2-024 (accessed 2022-04-14).

[20] *InChI Requirements for Representation of Organometallic and Coordination Compound Structures*. IUPAC | International Union of Pure and Applied Chemistry. https://iupac.org/project/2009-040-2-800 (accessed 2024-05-30).

[21] *Enhanced recognition and encoding of stereoconfiguration by InChI tools*. IUPAC | International Union of Pure and Applied Chemistry. https://iupac.org/project/2019-017-2-800 (accessed 2024-05-30).

[22] Brecher, J. Graphical representation standards for chemical structure diagrams (IUPAC Recommendations 2008). *Pure Appl. Chem.* **2008**, *80*(2), 277–410. https://doi.org/10.1351/pac200880020277.

[23] Brecher, J. Graphical representation of stereochemical configuration (IUPAC Recommendations 2006). *Pure Appl. Chem.* **2006**, *78*(10), 1897–1970. https://doi.org/10.1351/pac200678101897.

[24] *Committee on Publications and Cheminformatics Data Standards* https://iupac.org/body/024/ (accessed 2024-12-03).

[25] Apodaca, R. L. *Stereochemistry and the V2000 Molfile Format*. 2021. http://depth-first.com/articles/2021/12/29/stereochemistry-and-the-v2000-molfile-format (accessed 2024-06-12).

[26] *Atomic Coordinate Entry Format Version 3.3*. https://www.wwpdb.org/documentation/file-format-content/format33/v3.3.html (accessed 2022-04-19).

[27] Bourne, P. E.; Berman, H. M.; McMahon, B.; Watenpaugh, K. D.; Westbrook, J. D.; Fitzgerald, P. M. D. Macromolecular crystallographic information file. In *Methods in Enzymology*; Macromolecular Crystallography Part B; Academic Press, 1997; Vol. *277*, pp 571–590. https://doi.org/10.1016/S0076-6879(97)77032-0.

[28] *molstar/BinaryCIF*, 2024. https://github.com/molstar/BinaryCIF (accessed 2024-06-12).

[29] Tianhong Zhang, Hongli Li, Hualin Xi, Robert V. Stanton, and Sergio H. Rotstein. HELM: A Hierarchical Notation Language for Complex Biomolecule Structure Representation. *Journal of Chemical Information and Modeling* **2012,** *52*(10), 2796–2806. https://doi.org/10.1021/ci3001925.

[30] IUPAC Subcommittee on HELM. https://iupac.org/body/803 (accessed 2024-12-03).

[31] *FAIR Principles*. GO FAIR. https://www.go-fair.org/fair-principles (accessed 2024-06-12).

[32] Grasselli, J. G. JCAMP-DX, a Standard Format for Exchange of Infrared Spectra in Computer Readable Form (Recommendations 1991). *Pure and Applied Chemistry 1991, 63*(12), 1781–92. https://doi.org/10.1351/pac199163121781.

[33] Ulrich, Eldon L., Kumaran Baskaran, Hesam Dashti, Yannis E. Ioannidis, Miron Livny, Pedro R. Romero, Dimitri Maziuk, et al. NMR-STAR: Comprehensive Ontology for Representing, Archiving and Exchanging Data from Nuclear Magnetic Resonance Spectroscopic Experiments. *Journal of Biomolecular NMR* **2019**, *73*(1), 5-9. https://doi.org/10.1007/s10858-018-0220-3.

[34] *nmrML - Home* https://nmrml.org/ (accessed 2024-06-12).

[35] *IUPAC_FAIRSpec_Specification_draft.pdf* in https://github.com/IUPAC/IUPAC-FAIRSpec/tree/main/documents/specifications (accessed 2024-12-03).

[36] Mies, T, White, A. J. P., Rzepa, H. S., Barluzzi, L., Layfield, R. A., Barrett, A. G.M., Syntheses and Characterization of Diverse Main Group, Transition, Lanthanide and Actinide Metal Complexes of Ethyl-3-Oxo-2,3-dihydro-1H-pyrazole-4-carboxylate and Related Bidentate Ligands, Inorg. Chem., **2023**, *62*, 13253-76. https://doi.org/10.1021/acs.inorgchem.3c01506.

[37] *InChI extension for mixture composition*. IUPAC | International Union of Pure and Applied Chemistry. https://iupac.org/project/2015-025-4-800 (accessed 2024-06-12).

[38] Clark, A. M.; McEwen, L. R.; Gedeck, P.; Bunin, B. A. Capturing mixture composition: an open machine-readable format for representing mixed substances. *J. Cheminformatics* **2019**, *11*(1), 33. https://doi.org/10.1186/s13321-019-0357-4.

[39] *Common Format and MIME Type for Comma-Separated Values (CSV) Files*. https://www.ietf.org/rfc/rfc4180.txt (accessed 2024-06-12).

[40] *Definition of tab-separated-values (tsv)*. https://www.iana.org/assignments/media-types/text/tab-separated-values (accessed 2024-06-12).

[41] *OpenDocument Spreadsheet (ODS)*. https://esco.ec.europa.eu/en/about-esco/escopedia/escopedia/opendocument-spreadsheet-ods (accessed 2024-06-12).

[42] *Structure of a SpreadsheetML document*. 2023. https://learn.microsoft.com/en-us/office/open-xml/spreadsheet/structure-of-a-spreadsheetml-document (accessed 2024-06-12).

[43] DataCite Metadata Working Group. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.5. **2024**. https://doi.org/10.14454/G8E5-6293.

[44] *examples/v5-icl-repository-DOI-crawl.* https://iupac.github.io/IUPAC-FAIRSpec/#v5 (accessed 2024-12-03)

[45] *compound.* https://www.cancer.gov/publications/dictionaries/cancer-terms/def/compound (accessed 2024-12-03)

[46] *PubChem Compounds* https://pubchem.ncbi.nlm.nih.gov/docs/compounds (accessed 2024-12-03)

[47] *IUPAC Gold Book racemic compound.* https://goldbook.iupac.org/terms/view/R05027 (accessed 2024-12-03)

[48] WIPO Patentscope https://patentscope.wipo.int/search/en/search.jsf (accessed 2024-12-03)

[49] *nmrdb.org Tools for NMR Spectroscopists https://www.nmrdb.org* (accessed 2024-12-03)

[50] *EAD: Encoded Archival Description* https://www.loc.gov/ead (accessed 2024-12-03)

[51] *Base64 https://developer.mozilla.org/en-US/docs/Glossary/Base64* (accessed 2024-12-03)