Invited paper

Robert M. Hanson*, Damien Jeannerat, Mark Archibald, Ian J. Bruno, Stuart J. Chalk, Antony N. Davies, Robert J. Lancashire, Jeffrey Lang and Henry S. Rzepa

IUPAC specification for the FAIR management of spectroscopic data in chemistry (IUPAC FAIRSpec) – guiding principles

https://doi.org/10.1515/pac-2021-2009

Abstract: A set of guiding principles for the development of a standard for FAIR management of spectroscopic data are outlined and discussed. The principles form the basis for future recommendations of IUPAC Project 2019-031-1-024 specifying a detailed data model and metadata schema for describing the contents of an "IUPAC FAIRData Collection" and the organization of digital objects within that collection. Foremost among the recommendations will be a specification for an "IUPAC FAIRData Finding Aid" that describes the collection in such a way as to optimize the *findability*, *accessibility*, *interoperability*, and *reusability* of its contents. Results of an analysis of data provided by an American Chemical Society Publications pilot study are discussed in relation to potential workflows that might be used in implementing the "IUPAC FAIRSpec" standard based on these principles.

Keywords: Cheminformatics; data management; FAIR data; FAIR data management; FAIR principles; spectroscopic data; spectroscopy; standards.

Introduction

Reader note: In this article we use several terms that have multiple meanings. Terms first introduced ***in bold italics*** are terms defined in the glossary, later in this article.

Article note: A collection of invited papers on Cheminformatics: Data and Standards.

***Corresponding author: Robert M. Hanson**, Department of Chemistry, St Olaf College, Northfield, MN, USA, e-mail: hansonr@stolaf.edu
Damien Jeannerat, NMRprocess.ch, Geneva, Switzerland. https://orcid.org/0000-0001-7018-4288
Mark Archibald, Royal Society of Chemistry, Cambridge, UK. https://orcid.org/0000-0001-8687-7134
Ian J. Bruno, Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK. https://orcid.org/0000-0003-4901-9936
Stuart J. Chalk, Department of Chemistry, University of North Florida, Jacksonville, FL, USA. https://orcid.org/0000-0002-0703-7776
Antony N. Davies, SERC, Sustainable Environment Research Centre, Faculty of Computing, Engineering and Science, University of South Wales, Newport, UK. https://orcid.org/0000-0002-3119-4202
Robert J. Lancashire, Department of Chemistry, The University of the West Indies, Mona Campus, Kingston 7, Jamaica. https://orcid.org/0000-0002-6780-3903
Jeffrey Lang, American Chemical Society Publications Division, Washington, DC, USA. https://orcid.org/0000-0003-4895-0278
Henry S. Rzepa, Department of Chemistry, Molecular Sciences Research Hub, Imperial College London, White City Campus, Wood Lane, London W12 0BZ, UK. https://orcid.org/0000-0002-8635-8390

In this article, we report on the preliminary work of the IUPAC Project 2019-031-1-024, Development of a Standard for FAIR Data Management of Spectroscopic Data [1]. We outline and discuss here a set of guiding principles for the development of a standard for FAIR management of spectroscopic data. The principles form the basis for future recommendations of IUPAC Project 2019-031-1-024, which will specify a detailed data model and metadata schema for describing the contents of an "IUPAC FAIRData Collection" and the organization of digital objects within that collection. Foremost among the recommendations will be a specification for an "IUPAC FAIRData Finding Aid" that describes the collection in such a way as to optimize the *findability*, *accessibility*, *interoperability*, and *reusability* of its contents. We offer these principles to help institutional data managers and repository developers design their data collections and (ultimately) to assist authors in optimizing the findability and reusability of their spectroscopic data.

Initiated in March of 2020, the project's stated objective is:

> *… to apply FAIR data principles to spectroscopic* ***data*** *in the field of chemistry building on IUPAC's extensive expertise in this area. The project will develop a standard for the production and dissemination of digital data objects that contain enough spectral data and* ***metadata*** *that they can be (a) findable through semantic searches on the web, (b) available through standard interfaces, (c) interoperable and transferable between systems, and (d) readable and reusable over time, for both humans and machines.*

The problem being addressed, particularly in relation to publication of results in the form of supporting information, was succinctly summarized by David Martinsen in 2017 [2] where he states:

> *Even though the transmission of digital data can be accomplished more easily, many researchers continue to provide their research data as word processing documents or PDF files. While these are adequate for human consumption, they are not necessarily useful for importing data into a software package for visualization or for reanalysis.*

In fact, "not necessarily useful" is putting it diplomatically. For all practical purposes, *essentially useless* might be a better characterization of typical supplementary data's utility for automated, machine-based reuse.

The IUPAC FAIRSpec project is an outgrowth of a series of IUPAC- and NSF-sponsored discussions that followed that publication, including the IUPAC/CODATA Joint Workshop "Supporting FAIR Exchange of Chemical Data through Standards Development" in Amsterdam in July of 2018 [3], and the NSF-sponsored workshop "FAIR Publishing Guidelines for Spectral Data and Chemical Structures" in Orlando, Florida, in March of 2019 [4]. Many of the ***FAIR Data Management*** principles [5] presented during these meetings and in the discussion below were elaborated by Leah McEwen in 2020, again primarily in the context of publication, in the ACS guide to scholarly communication Section 3.1, Data Sharing [6].

The project directly addresses the lack of standardization across chemical data formats. It is not an attempt to develop yet another standard data format in spectroscopy. The project will produce recommendations of what information and metadata should be available in all formats used by software vendors irrespective of how they wish to develop their own data structures to meet their own individual requirements. The recommendations made will be a living specification that will dramatically increase interoperability across systems and across time.

The issue at hand concerns more than just the practice surrounding the publication of scientific articles. It also concerns overall ***data management*** in the overall research context. It is becoming more and more common for funding agencies to require that data involved in a study be managed using a ***data management plan*** that ultimately allows the data acquired within the timeframe of the project to be made available to the public [7–11]. To date, though, these requirements have been fulfilled with only minimal consideration for FAIR principles, hindering the ability of the data to be found and used by the broader science practitioner and science education community. For the most part, it has been considered enough to produce what is effectively a "data dump," sometimes included as supporting information with a publication; sometimes simply residing somewhere on a local institutional or generalist ***data repository***, available for the asking – provided one knows whom to ask and what to ask for. These ***digital aggregations*** are generally provided as monolithic ZIP files containing hundreds or thousands of files (even worse, combined into just a single file) that may be useful to the reader of the article, but are not easily reusable in other contexts.

For example, in a recent pilot study, ACS Publications encouraged authors of the two journals *Journal of Organic Chemistry* and *Organic Letters* to submit digital data along with the standard PDF supporting information for their manuscript [12]. Authors were not directed in any specific way to organize their data. Over the course of 20 months, authors published more than 350 manuscripts in these journals supported with primary research data files. Our Task Group analyzed thirteen of those publicly available submissions [13]. In all, 443 unique chemical compounds were referenced in these collections. Out of 16 200 files found, we were able to extract and catalog 1022 NMR ***datasets***, 30 high resolution MS reports, 85 ChemDraw structure files, 30 MDL Molfile structures, and 495 PNG images. It was also possible to extract 18 additional Molfile structures from the 141 MNova datasets in the collection. Extraction of metadata produced 8016 ***metadata elements***. Notably, all of the NMR data was in proprietary formats.

Particularly problematic was the very low occurrence of ***chemical structure identifiers*** associating the spectroscopic data with the structures of the compounds analyzed. Importantly, only 44 InChIs could be associated with these spectroscopy datasets (by calculation from Molfile data), leaving 399 structures only identified by their associated compound number in the manuscript. Thus, although all of these data are, in principle, *findable* (but only in the sense that one first has to find the article they are associated with), only a few of them are generally *reusable*, in the sense that the full spectral dataset could be found and explored based on a chemical structure search.

Thus, despite the stated requirement for data management and the noble goal of data sharing, the current lack of standards and infrastructure for chemical data means that the resulting depositions are only findable by a reader following a link from the corresponding journal article or through private communication with the author. This serves certain (human) use cases – and is certainly better than no data being available – but it does not enable broader findability, linking to chemical structure, indexing, or significant reuse, especially by automated machine-based processes. Standardized metadata registered with a relevant authority would increase the utility and FAIRness of mandated spectral data depositions, and therefore significantly increase the return value of funded research.

IUPAC has a long history in the development and support of scientific data standards. In 1995, IUPAC took over responsibility for the JCAMP-DX range of scientific standards from the Data Exchange Task Force of the Joint Committee on Atomic and Molecular Physical Data (JCAMP). An early objective of the "JCAMP-DX" task force in 1983 was to design a standard file format for exchange of infrared spectra between vendor data systems that used different proprietary file formats [14, 15]. This was subsequently extended to MS, NMR, IMS, EMR, and CD *etc*. The JCAMP-DX standard has been implemented by most of the major instrument vendors and repository managers and has gone a long way to enabling the reuse of spectroscopic data in the field of chemistry. An overview of the JCAMP-DX format development and its extensions and applications is the subject of another article [16] in this special edition.

Importantly, the JCAMP-DX standard is a standard for spectroscopic ***data representation***. That is, for the storage and transmission of mostly individual raw spectroscopic datasets. While the standard does include a modicum of metadata relating to the data – for example, title, author, creation date, temperature, and instrument information – it was developed prior to the existence of the Internet as we know it, before the term "metadata" was widely popularized, and its designers did not envisage the need for FAIR data management. Nonetheless, JCAMP-DX continues to serve an important role in providing an open alternative to proprietary dataset formats in a wide variety of spectroscopies, leading to widespread interoperability. JCAMP-DX also allows for a concise single-file collection of spectra associated with a given compound. These early IUPAC approaches to the standardization of data content between research publication and reference data collections was very much aligned to the concept of minimum information standards such as have been used in molecular biology to a far more advanced level [17].

One of the most pressing needs that we have identified is for the standardization of metadata in relation to *heterogeneous* ***digital collections*** – those containing a variety of data, such as spectroscopic datasets, structural models and descriptions, sample information, and post-acquisition analysis. Specialist repositories such as MassBank [18] (for mass spectrometry) and the CSD [19] (for crystallography) have made important advances in the formalizing of data formats and metadata schemas in their distinct areas. However, they do not address the overarching standardization of metadata associated with heterogeneous digital collections. Generalist repositories, by definition, collect only general metadata (author, institution, publication details, funding agencies, *etc.*), resulting in discipline-specific metadata often being lost or at least not findable. For example, generalist repositories do not have ways of encoding chemical formulas, ***InChI*** [20], or ***SMILES*** [21] – essential information for chemistry-related searches – in anything other than generalized categories, such as "keyword" or "topic". By standardizing discipline-specific metadata associated with heterogeneous digital spectroscopic collections, the IUPAC FAIRSpec standard will allow for a more consistent discipline-specific approach to the meaning of "FAIR data management."

The bigger picture

Data representations, funder requirements, supporting information for publications – these are all pieces of a bigger puzzle: the overall *FAIR management of data*, from start to finish, from first laboratory reaction to publication and reuse. That is, the standard we are proposing interprets the phrase "FAIR Data Management" as "FAIR (Data Management)" rather than "(FAIR Data) Management." It is this overall management perspective our project aims to facilitate. In particular, our focus is on data *collections*. This article sets forth a set of guiding principles that underlie the standard that we are currently developing for the description and cataloging of data (of any kind, but spectroscopic data in particular) and its associated metadata in ways that are practical, relatively simple to implement, modular, intuitive, easily extended, and flexible. The scope of the project is *spectroscopic* data within a *chemical* context, but the approach described could be fully extendable to any sort of data, within any context.

Ultimately, the proposed standard will involve:

- a set of guiding principles underlying what we mean by "FAIR" in relation to spectroscopic data,
- a detailed ***data model*** for describing the contents of an ***IUPAC FAIRData Collection*** in terms of objects and relationships of objects,
- a recommendation for the organization of ***digital objects*** within a collection,
- a specification for describing properties of digital objects within the metadata records of a ***digital finding aid*** describing the collection,
- a specification for the ***serialization*** of that finding aid,
- a proposal for methods of ***data and metadata extraction*** and the generation of IUPAC FAIRData Finding Aids, and
- recommendations for ***metadata registration*** in order to obtain ***persistent identifiers*** for use in data citations and machine search and accession mechanisms.

In this article, we present just the first of these: guiding principles for the FAIR management of spectroscopic data.

Spectroscopic data in the area of chemistry are intimately connected to both samples and chemical structure. Spectroscopic measurements are made generally to discover the identity of pure compounds, to identify and quantitate the relative amounts of chemical components in a mixture, and to determine the structure of new compounds. Thus, the long-term value of spectroscopic data is generally strongly coupled to its association with chemical structure, or more generally, to a sample. We value the ability to find a spectrum, for example, based on a chemical structure which has been assigned to a spectrum using a chemical structure identifier such as a standardized IUPAC name, ***InChI***, or ***SMILES***.

In addition, there are contexts where it is more appropriate to refer to spectroscopic data in relation to ***samples***. We typically know the identity of a sample (where it came from, for instance), but we don't yet know for sure its chemical structure. Or, perhaps we are working with a material that is not characterizable *per se* as a chemical compound. An overall FAIR data management plan should cover the moment a sample is created all the way through the hypothesis stages of analysis to the "proof of structure" required for publication. The proposed standard needs to apply to all such cases.

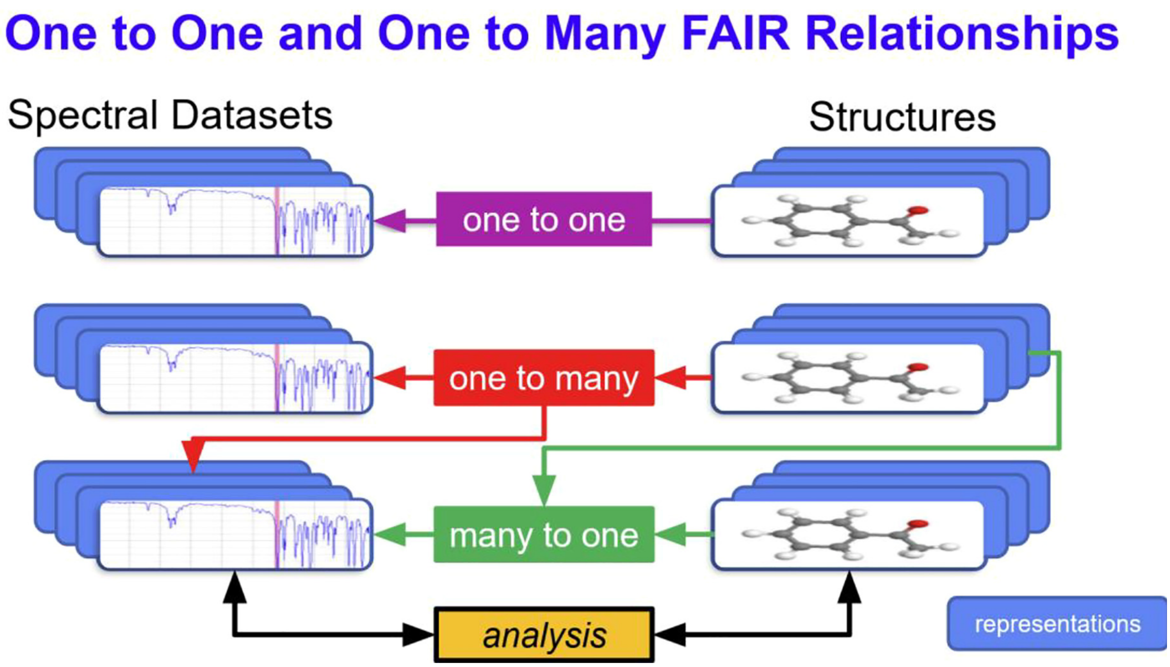

Fig. 1: Key relationships among FAIRSpec digital objects.

A key concept in the ***IUPAC FAIRSpec Data Model*** is the *IUPAC FAIRData Collection*, which uses standardized metadata to make meaningful connections among digital objects and allows for a variety of digital representations of those objects (Fig. 1).

Thus, the IUPAC FAIRSpec Data Model is not about standardizing instrument data file formats (though it benefits from that) or requiring specific representations (though it might make specific recommendations in this regard). Rather, it is about providing a baseline specification for the storage, transmission, and description of the contents of a digital collection derived from spectroscopic measurements or calculations. The ultimate goal is to be able to provide concise descriptions of complex datasets associated with manuscripts, ongoing laboratory work, and teaching, that can be findable, accessible, interoperable, and reusable by machines and humans alike.

Guiding principles for the FAIR management of spectroscopic data

The five principles that underlie development of the IUPAC FAIRSpec standard are given below.

(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.

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

A. *FAIR management of data must be an explicit part of research culture.* FAIR data management should be a part of the design of all ongoing scientific endeavors from the very beginning and throughout its course, not something that is applied just at the time of publication. The key word here is *management*. All too often data are collected and organized in an arbitrary and idiosyncratic way. Management of the data is an afterthought, if at all. The IUPAC FAIRSpec standard is designed to be simple, practical, and, for the most part, automated, in order for it to be adopted and utilized throughout the data workflow, from collection to publication and beyond.

B. *FAIR management of data should be of intrinsic value.* The value of spectroscopic data management has to have real-time value for the originating research group(s), providing value-added benefits within and between those groups that go far beyond the need to satisfy granting agencies.

C. *Good data management requires distributed curation.* By ***curation*** we mean the association of meaning to a digital aggregation, thus turning its ***digital entities*** into *digital objects* and making them findable. Some of this curation can be done automatically; some must be done by hand. Simply designing an electronic filing system for data or associating data with an electronic laboratory notebook (ELN) are examples of curation. While much of what we describe here can be discovered programmatically, ultimately it is the task of researchers themselves to properly prepare and maintain their data and to make the connections between spectroscopic data and chemical structure or sample identity that make for intrinsic value. The FAIRSpec standard is designed to allow for this sort of management from the very beginning of data collection, allowing for data representations that are vendor specific, for example.

D. *Experimental work is by nature iterative.* FAIR management of data associated with a research endeavor must allow for cycles of data *generation*, (re)*processing*, and (re)*analysis*. The IUPAC FAIRSpec standard is designed to allow a progression of associations from sample to structure.

(2) Context is important.

A. *Digital objects are generally part of a collection.* Spectra are taken for a reason, generally associated with a grant or project. They are maintained in relation to a student, researcher, research group, or institution. They are presented as part of a publication or presentation. Taking a cue from the field of digital archiving, FAIR management of data standards should emphasize the value of a *collection*, however that may be defined. A FAIR data management standard should describe the relationship among the different components of a collection, making sure that the finding of one component can lead to the finding of the whole.

B. *Chemical properties are related to chemical structure.* This includes spectroscopic properties as discerned from the collection and analysis of spectroscopic data. This is why interpretation of spectra from structure and *vice versa* is an integral part of undergraduate chemistry courses. As such, spectroscopic data without reference to chemical identifiers of some kind (a compound name, a drawing, an InChI or SMILES), is unlikely to be useful to anyone, including its creator. The IUPAC FAIRSpec standard optimizes the connection between structure and spectra.

C. *Data relationships are diverse and develop over time.* The association of spectroscopic data with chemical structure is not something that happens automatically or immediately. Indeed, initially there is a specific *sample* – generally the result of experimentation. The process involves the collection of spectroscopic data associated with that sample in a "one to many" relationship. That is, for any given sample, there may be several associated spectroscopic data items. FAIR management of data should allow for contexts in which it may not yet be possible (or not *ever* possible) to connect a specific chemical structure to a spectrum and, by association, structure to sample. In addition, a "many to one" relationship must also be possible, either because the sample is a mixture or because the material involved does not lend itself to such a single-structure description in the first place.

D. *FAIR management of data should allow for validation.* When data is made available in machine-readable form, it allows for the possibility of automated processes to be developed for use in validation of the data before, during, and after publication. Simply publishing an image of a spectrum in a PDF document does not allow any such validation. One of the most important potential impacts of the IUPAC FAIRSpec standard is that it will allow not just the routine validation of spectral data in relation to proposed structure, but also a potentially enormous data resource for the development of better validation tools, particularly in the area of NMR spectroscopy.

(3) FAIR management of data requires curation.

A. *Data reuse relies upon practical findability.* A 180 MB zip file containing 1200 files may be "findable," but unless there is a set of key registered metadata describing the data within that zip file, just having the file in hand does not constitute "found". Again, drawing from the area of digital archiving, a FAIR data management standard should describe a digital finding aid that exposes key metadata and allows a ***reuser*** to quickly ascertain whether additional scrutiny of the data collection is warranted, and/or the ability to conduct a rich fielded search of the database store where the metadata describing the collection have been registered.

B. *Data has to be organized to be accessible.* It is probably obvious that data cannot be reused unless it is accessible. This accessibility can be seen at two levels. The IUPAC FAIRSpec standard does not assume that data are ***open***. Owners of data have the right of restricting access to their data under whatever contract it was collected. But even for authorized reusers, data can be effectively inaccessible. For example, the reuser should not have to download and unpack a zip file containing a spectrum that is within a zip file for a compound that itself is within a zip file associated with a publication just to check to see if that spectrum is of immediate value. Thus, the IUPAC FAIRSpec standard allows for the automated or semi-automated repackaging of data and metadata extraction from an original dataset in order to provide a better reuse experience. Accessibility also means having sufficient information in the metadata elements to allow unsupervised machine-based retrieval of the data from a complex collection.

C. *Data interoperability requires well-designed metadata.* The IUPAC FAIRSpec standard must be clearly defined and, as much as possible, mappable onto other metadata standards (*e.g.*, bibliographic standards or instrument identifiers) that are in use or will be in future use in the area of machine-based knowledge discovery. That is, truly "FAIR" as in "Fully Artificial Intelligence Ready" [22].

D. *Value is in the eye of the reuser.* The IUPAC FAIRSpec standard respects the fact that data can have multiple *representations*. One might argue that only the "raw" data from an instrument – a free induction decay in the case of NMR spectroscopy – is the essential representation of the data that would allow for those data to be described as FAIR. However, that is not necessarily the case. The *reuse* of data relies upon data being in a form that is meaningful *for the reuser*. For a practicing spectroscopist, it may be that retrieving the raw spectroscopic dataset is critical, but for others – reviewers, readers, students – the "real" 1D or 2D spectrum may be the only representation they are able to utilize. Sometimes an image is more appropriate, and sometimes just a simple peak list is desired. Key to this principle is that the full nature of reuse cannot be predicted by the data originator.

(4) Metadata must be standardized and registered.

A. Register key metadata. A certain amount of “globally significant” metadata must be registered with a ***metadata registration agency***. While the finding aid for a collection provides metadata that serve the purpose of describing the contents of the collection once the reuser arrives at the site, full findability requires that the most salient aspects of the collection’s metadata be separately registered with a registration agency, giving it an existence that is independent from the finding aid and, in fact, independent of the data itself, and allowing the metadata to be aggregated on a global scale. This registration requires dialog between registration agencies and standards developers to ensure that the sort of metadata being produced under the standards can be integrated into the schemas developed by those agencies.

B. Assign a variety of persistent identifiers. All registered metadata should be assigned a persistent identifier. The persistent identifier serves to allow a persistent link to the ***landing page*** for data at some future point in time, regardless of whether or not the data are in their original location. The landing page is typically an HTML-based webpage designed for human access, whereas a digital finding aid in the form of XML or JSON may be more appropriate for machine-based or artificial intelligence-based applications. The fact that XML-based finding aids (with associated style sheets) can serve for both purposes has made them popular in the digital archival community [23]. Thus, it could be that just the finding aid itself needs to be registered, with multiple entry points into a digital collection for subcollections of data objects. The most common form of a persistent identifier is the DOI, or ***Digital Object Identifier.***

C. Enable metadata crosswalks. Metadata must be standardized and allow for ***metadata* crosswalks** that allow the interoperability and exchange of metadata among different systems. The registration of metadata ensures that it will conform to a predictable structure specified by one or more ***metadata schemas.*** Currently popular schemas for scientific data include those established and curated by the DataCite organization [24], Schema.org [25], and the DDI Alliance [26]. Such schemas may overlap to a considerable extent, but each has nuances that make it especially suited to particular needs or interests. For example, Schema.org was launched by several large commercial search engine providers to help improve the quality of their own data indexing, and the focus of the DDI Alliance is social, behavioral, economic, and health sciences. The process of mapping the metadata from one schema to another, referred to as “crosswalking”, allows metadata standardized for one schema to be also representable by another. The IUPAC FAIRSpec standard must be mappable to one or more of these schemas, at least in part.

D. *Allow for value-added benefits*. Once one or more digital finding aids are registered with persistent identifiers (*e.g.*, DOI’s or handles [27]), they become part of a ***metadata store.*** Registering the finding aid with a DOI will enable it to become part of a queryable system of metadata, which can undergo ***metadata harvesting*** to provide a range of value-added services. For example, a service might index the metadata in ways that make it searchable on a much larger scale. Citation and access statistics can be provided, and ***PID graphs*** can be constructed that provide information about how one digital collection is related to another. A PID graph can be used to discover connections between diverse entities in the research landscape, connecting foundations and instrument manufacturers with information about their market and impact. But for that to be possible at the granularity that we require, harvesting methods must allow for standardized discipline-specific metadata such as we are proposing in this project to be represented in discipline-aware ways.

(5) FAIR data management standards should be *modular*, *extensible*, and *flexible*

A. Modularity allows specialization. By this, we mean that it should allow for core standards to be developed in different subdisciplines in parallel and sequentially, with different emphases and nuances that are unique to those subdisciplines. The goal of the IUPAC FAIRSpec project is to provide the templates for current and future subdisciplines to expand upon in whatever way is appropriate to their area.

B. *Allow for future needs.* FAIR data management standards should be extensible. That is, they should express clear versioning and allow for the inclusion of metadata descriptions that will meet future needs.

C. *Respect format and implementation diversity.* Taking another cue from the digital archivist community, FAIR data management standards should respect variety and should not require (though IUPAC might *recommend*) one data format over another. Our motto is: *The collection is what it is. Work with it.* The IUPAC FAIRSpec standard must allow for workflows that start with data aggregations that contain a wide variety of possible data formats and organizational structure. Thus, while standardized, open, nonproprietary data formats are ideal in many contexts, the IUPAC FAIRSpec standard does not assume knowledge of a reuse context and does not in any way require any specific data format, open or otherwise. What is required is that the format of the data be identified. In addition, the exact format of the rendering of IUPAC FAIRData Finding Aids is not defined, allowing for different modes of expression, such as JSON or XML.

D. *All data formats should be valued.* To whatever extent is possible, preserving the original native instrument format of data is to be valued. This value derives from the need to have a clear pathway of ***data provenance***, minimizing the loss of information and context, maximizing the ability of others to verify the data and analyses, and allowing modified analysis that can be used in contexts not considered by the original data creators. Unlike efforts that have sought to transform raw data into common formats such as JCAMP-DX [16, 28], nmrML [29], NMR-STAR [30], mzML [31] and AnIML [32], the IUPAC FAIRSpec standard accepts data in any such format, including native instrument formats, opting instead for automated metadata and data extraction that allow for multiple data representations for optimum reusability.

Implementation of the principles

We have experimented with a workflow that allows the conversion of an author-supplied supplemental dataset associated with a publication into a (prototype) IUPAC FAIRData Collection with its associated IUPAC FAIRData Finding Aid (Fig. 2).

In each case, the process started with extraction and standardization of one or more ZIP files downloaded from figshare [33], producing a set of standardized digital representations described by an abstract "instance" of the IUPAC FAIRSpec Data Model. Packaging of the representations produced the FAIRData Collection; serialization of the data model instance generated the FAIRData Finding Aid. The process was semi-automatic, requiring the initial manual creation of a small template describing the overall organization of each publication dataset, identifying the structure of the dataset as created by the author. Using this template, an automated process then completed the tasks of extraction, standardization, serialization, and packaging.

While carried out only on a small (thirteen-dataset, 1 GB) sampling of available publication datasets, the experiment provides a proof-of-concept for the development of such a workflow. In addition, it aided us in our refining of the data model associated with the developing standard. Results of the experiment in the form of prototypical landing pages generated on the fly from JSON-serialized finding aids are available on a demo page

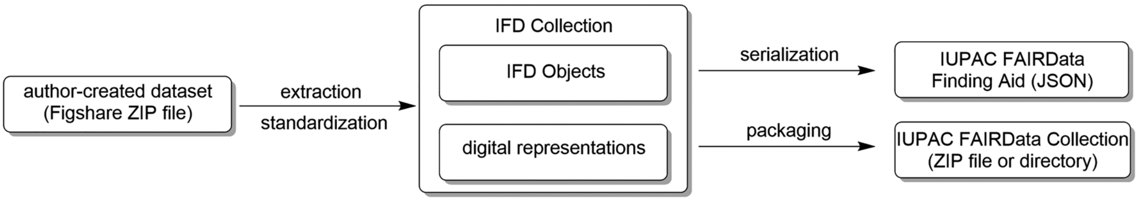

Fig. 2: Experimental workflow for creating IUPAC FAIRData Collections and their associated finding aids.

[34]. Java source code in the form of an Eclipse project can be found on the IUPAC GitHub [35]. The working IUPAC FAIRSpec specification is also available [36].

Looking to the future

We have presented five principles – twenty corollaries – for guiding the development of a FAIR standard for the management of spectroscopic data. The recommendations we expect to make will not be of the nature of yet another new data format. Rather, they will be specifications for how data in any format can be described using metadata that is discipline- and subdiscipline-specific yet at the same time standardized across disciplines in order to create a common framework for improving the findability, accessibility, interoperability, and reusability of spectroscopic data.

Our intention is not to supersede other efforts in this area, but rather to complement and enable them. Thus, for those developing prepublication services, such as institutional repositories and electronic laboratory notebooks, we hope that our recommendations will allow for real-time in-house FAIR management of data and seamless transition to publication. For authors and publishers, we aim to make the process of sharing data a value-added incentive, not a required chore. For data scientists and developers of technology still to be discovered, we hope to enable the distributed development of a corpus of data that can be reused in ways we can not foresee today. By emphasizing context and the association of chemical structure and spectra within collections, we hope to usher in a new era of validated, interconnected data resources that will serve the field of chemistry in ways we cannot yet imagine.

We recognize that data management of any sort, particularly FAIR management, will require effort. Our goal is to enable as much automation as possible in that effort, extracting metadata from data in natural, intuitive ways that make data-intensive research efforts easier, not harder. In order to reach the widest range of reusers (both spatially and temporally), we will be proposing methods for maximizing the visibility of discipline-specific metadata through registration and standardization.

Finally, our recommendation will be modular, extensible, and flexible, allowing for concerted parallel development today and adaptation over time to meet the data needs of the future. We sincerely hope that our recommendations will resonate across not just the discipline of chemistry, but also throughout the broader landscape of scientific research, providing a blueprint for the development of widespread FAIR management of data throughout science well into the future.

Glossary

The glossary below is intended only to clarify what these terms mean in the context of this paper. It is not intended to fully define the terms in all contexts. Definitions from the Research Data Alliance Data Foundation and Terminology (RDA DFT) Work Group are indicated as [RDA] [37, 38].

chemical structure identifier A meaningful alphanumeric text string that can identify a chemical compound. Examples include InChI and SMILES.

curation The process of maintaining, preserving and adding value to data throughout its lifecycle. One aspect of curation is the design and creation of metadata associated with a digital collection. Curation can involve automated machine-based processes as well as manual or semi-automated cataloging of digital objects.

data To a practicing chemist, it should be obvious that digital entities coming from a laboratory instrument constitute "raw data" (referred to herein as "datasets"). However, the word *data* as used here is a broader term. For our purposes, *data* includes the digital entities associated with *spectroscopic data analysis*. 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 graphs (MOL, SDF, for example) also fall in this category.

data and metadata extraction The primarily machine-based act of curation of one or more digital entities associated with a (spectroscopic) dataset carried out in order to generate value-added digital representations of that dataset. For example, the creation of a spectrum in JCAMP-DX format from an instrument-derived "raw" dataset and the creation of a PNG image or peak table from that spectrum, or the extraction of temperature, probe, and pulse sequence information from a dataset.

data management The overall activity of organizing, maintaining, and cataloging data assets. We interpret this to be not just the activity of professional data managers, but also all the curation of data that takes place in the field during data collection and analysis.

data management plan A type of plan usually described in a formal document that outlines how data are to be handled both during a research project and after the project is completed.

data model [RDA] A data model is an abstract model that specifies the structure or schema of a data set. We extend this definition to relate to the full set of digital objects associated with an IUPAC FAIRData Collection.

data provenance [RDA] A type of historical information or metadata about the origin, location or the source of a digital object, or the history of the ownership or location of a digital object.

data repository A service operated by organizations where data assets are stored, managed and made accessible. The repository contains data organized as digital objects and digital entities and is accompanied by descriptive metadata for these items. The three primary types of data repository are *generalist* (not domain-specific), *specialist* (domain-specific), and *institutional* (based at a research institution).

data representation A digital object that may take any one of a number of forms that allow for various levels of data reuse. For example, an IUPAC FAIRData Collection might include data representations in the form of the full raw spectroscopic dataset, a spectrum or free induction decay (FID) stored in JCAMP-DX format, an image, and a text description of the spectrum in a standardized journal-ready format. Each of these data representations has intrinsic value that, for a given re-user, might be the most appropriate or desirable.

dataset (spectroscopic) The "raw" data representation collected by an instrument in whatever native format that instrument creates. This could be a single file or a zip file or folder containing multiple parameter files along with one or more raw or processed data files. In this article, we distinguish between the more general term, *data*, and the more specific terms *spectroscopic dataset*.

digital aggregation [RDA] A bundle of digital entities.

digital collection A digital collection is an aggregation which contains digital objects and digital entities. The collection is described by metadata. A digital collection is an organized, systematic form of purposeful aggregation, grouping or arrangement of elements, that has an identity of its own separate from the identity of the elements. RDA defines a "Data Collection" as "a type of collection formed by some agent-driven aggregation or grouping process whose parts/elements are made of data/datum. A data collection is identified by a PID and described like other types of DOs by metadata" with essentially the same meaning. In addition, we recognize the term *heterogeneous digital collection* to refer to a digital collection that includes a variety of data types (in our context, for example: NMR, IR, MS, X-ray diffraction, polarimetry, cyclic voltammetry, chromatography data) as well as structural or sample properties and representations.

digital entity [RDA] Anything that can be represented by a bitstream (which is a sequence of bits that encodes a specific content, either stored on some media or being transferred under control of protocols).

digital finding aid A digital object that is a description typically consisting of contextual and structural information about an archival resource [39].

digital object [RDA] A digital entity composed of a structured sequence of bits/bytes. As an object it is named. The bit sequence realizing the object can be identified and accessed directly or indirectly via a unique and persistent identifier or by use of referencing attributes describing its properties.

Digital Object Identifier (DOI) A unique character string form of a persistent identifier, such as "10.1021/acsguide" (more precisely referred to as a *DOI Name*) that can be part of a URL such as "https://doi.org/10.1021/acsguide". The distribution and management of DOIs are carried out by a federation of registration agencies under the auspices of the International DOI Foundation [40].

FAIR Data Management Data management based on the FAIR (Findable, Accessible, Interoperable, and Reusable) Guiding Principles [41], recognizing that there are many degrees of "FAIRness", some more aspirational than realized.

InChI or ***International Chemical Identifier*** A textual identifier for chemical substances, designed to provide a standard way to encode molecular information and to facilitate the search for such information in databases and on the web generated using the algorithm as defined by IUPAC.

IUPAC FAIRData Collection A curated spectroscopic data collection organized using the principles described in this article and the (developing) IUPAC FAIRSpec Specification [36].

IUPAC FAIRSpec Data Model (IFS Data Model) The abstract data model currently under development by the Task Group. This model describes the structure and format of data and metadata associated with an IUPAC FAIRData Collection [36].

landing page The endpoint for the resolution of a persistent identifier, typically in HTML or XML format. If the endpoint is changed, this change must ideally be reflected in any registered metadata for that identifier.

metadata [RDA] Data that contains descriptive, contextual and provenance assertions about the properties of a Digital Object. Metadata are data that play the role of documentation for data/resource discovery, description/documentation, contextualization. Metadata can conform to a declared schema that sets out the vocabulary and properties of the metadata. The schema may specify control or constraints on the values of both.

metadata crosswalk A well-defined mapping that translates elements and values from one metadata schema to those of another. Crosswalks facilitate interoperability between different metadata schemas and serve as a base for metadata harvesting and record exchange [42].

metadata element [RDA] An aspect of a digital object generally characterized by a key/value pair. To the extent that the metadata are part of a defined metadata schema, the element will be designated by a unique controlled-vocabulary key, and its value will adhere to the description of that key within the schema.

metadata harvesting The automated collection of metadata records from different sources to create useful aggregations of metadata and the related services that are enabled by this process [43].

metadata registration The process of associating a digital object (quite possibly a collection) with a persistent identifier assigned by a recognized metadata registration agency, allowing URL resolution back to the original digital object. If the location of the digital object is changed, then this change must be recorded in the metadata that has been registered, thus ensuring its persistence. If the data repository where the digital object is stored ceases to operate, the metadata records associated with that repository will continue to be available via the agency where they were registered.

metadata registration agency An organization that provides persistent identifiers for various types of digital objects and/or research outputs in exchange for the registration of a metadata record, allowing these outputs and their associated metadata to become discoverable. DataCite is one example of a metadata registration agency, providing managed curation of an extensive metadata schema. Metadata registration agencies can also provide various services that take advantage of their stored metadata records, including the capability of rich fielded searches and analyses of these records when combined with metadata from authorities specializing in other types of persistent identifiers, such as people (ORCID), research organizations (ROR), data (DataCite), journal articles and funders (CrossRef).

metadata schema [RDA] A type of data schema or structure organized by a logical plan that shows the relationships between metadata elements.

metadata store A queryable database of metadata records.

open data [RDA] Open data are data available/visible to others and that can be freely used, reused, shared, republished and redistributed by anyone, within the parameters defined by license. We note that FAIR management of data does not necessitate open data, and that the act of curation has a cost that might be shared with reusers.

persistent identifier (PID) [RDA] A character string (functioning as a symbol) that identifies a digital object. The identifier can be persistently resolved (digitally actionable) to meaningful metadata state information about the identified digital object.

PID graph A semantic graph in which the nodes are persistent identifiers [44].

property A key:value pair that describes a characteristic of an object.

reuser The person or entity that has accessed a digital object for purposes, quite possibly completely different from any imagined by the originator of the data.

sample A portion of material selected from a larger quantity of material [45]. More specifically, the physical sample that was the source of the spectroscopic dataset in a collection. We note that efforts are underway to uniquely identify and register samples in a persistent manner [46].

serialization (of a finding aid) The generation of a byte sequence in a machine- and potentially human-readable form such as JSON or XML. The IUPAC FAIRSpec standard does not specify a preferred serialization of the IUPAC FAIRData Finding Aid, only that the serialization must preserve the specified structure and vocabulary of the finding aid and its associated collection.

SMILES (Simplified Molecular Input Line Entry System) A linear representation of a molecular graph in character string form, used for searching for, matching, and atom-atom mapping of chemical structures and models.

Acknowledgments: RMH thanks St. Olaf College students Kha Trinh and Lecheng Lyu for their assistance obtaining and unpacking the ACS pilot datasets early on in the development of our workflow prototype. This project is supported by IUPAC, Project 2019-031-1-024.

References

[1] R. M. Hanson, D. Jeannerat, M. Archibald, I. J. Bruno, S. J. Chalk, A. N. Davies, R. J. Lancashire, J. Lang, H. S. Rzepa. Development of a Standard for FAIR Data Management of Spectroscopic Data, https://iupac.org/projects/project-details/?project_nr=2019-031-1-024.
[2] D. Martinsen. *Chem. Int.* **39**, 35 (2017).
[3] A. N. Davies. *Spectrosc. Eur.* **30**, 21 (2018).
[4] V. F. Scalfani, L. McEwen. in *NSF OAC 2019 Workshop*, FAIR Publishing Guidelines for Spectral Data and Chemical Structures, OSF Storage, United States (2019), https://osf.io/psq7k/.
[5] GFISCO FAIR Principles, https://www.go-fair.org/fair-principles/.
[6] L. McEwen. *(Chapter 3.1.4) Res. Data Rep. Chem.* (2020), https://doi.org/10.1021/acsguide.30104.
[7] NIH Final NIH Policy for Data Management and Sharing, https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html.
[8] Q. Schiermeier. *Nature* **591**, 20 (2021).
[9] NSF Division of Chemistry – Advice to Principal Investigators on Data Management Plans, https://www.nsf.gov/bfa/dias/policy/dmpdocs/che.pdf.
[10] UKRI Common principles on data policy – UK Research and Innovation, https://www.ukri.org/funding/information-for-award-holders/data-policy/common-principles-on-data-policy/.
[11] Wellcome Data, software and materials management and sharing policy, https://wellcome.org/grant-funding/guidance/data-software-materials-management-and-sharing-policy.
[12] A. M. Hunter, E. M. Carreira, S. J. Miller. *Org. Lett.* **22**, 1231 (2020).
[13] IUPAC Analysis of thirteen submissions to the ACS Publications digital data pilot, https://github.com/IUPAC/IUPAC-FAIRSpec/tree/main/results.
[14] J. G. Grasselli. *Pure Appl. Chem.* **63**, 1781 (1991).
[15] IUPAC Digital Standards: JCAMP-DX, https://iupac.org/what-we-do/digital-standards/jcamp-dx/.
[16] A. N. Davies, R. M. Hanson, P. Lampen, R. J. Lancashire. *Pure Appl. Chem.* (2022), in press.
[17] FAIRsharing.org MIBBI – Minimum Information for Biological and Biomedical Investigations, https://fairsharing.org/3518.
[18] M. Europe. MassBank: High Quality Mass Spectral Database, https://massbank.eu/MassBank/.
[19] C. R. Groom, I. J. Bruno, M. P. Lightfoot, S. C. Ward. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **72**, 171 (2016).
[20] S. Heller, A. McNaught, S. Stein, D. Tchekhovskoi, I. Pletnev. *J. Cheminf.* **5**, 7 (2013).
[21] Daylight Software Simplified Molecular Input Line Entry System, https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html.
[22] B. Mons. *Nature* **578**, 491 (2020).
[23] LOC Encoded Archival Description, https://www.loc.gov/ead/.
[24] DataCite DataCite: International Data Citation Initiative, https://datacite.org.

[25] W3C Schema.org, https://schema.org.

[26] DDI Data Documentation Initiative Alliance, https://ddialliance.org.

[27] CNRI The Handle System, https://www.handle.net/.

[28] R. S. McDonald, P. A. Wilks. *Appl. Spectrosc.* **42**, 151 (1988).

[29] D. Schober, D. Jacob, M. Wilson, J. A. Cruz, A. Marcu, J. R. Grant, A. Moing, C. Deborde, L. F. de Figueiredo, K. Haug, P. Rocca-Serra, J. Easton, T. M. D. Ebbels, J. Hao, C. Ludwig, U. L. Günther, A. Rosato, M. S. Klein, I. A. Lewis, C. Luchinat, A. R. Jones, A. Grauslys, M. Larralde, M. Yokochi, N. Kobayashi, A. Porzel, J. L. Griffin, M. R. Viant, D. S. Wishart, C. Steinbeck, R. M. Salek, S. Neumann. *Anal. Chem.* **90**, 649 (2017).

[30] E. L. Ulrich, K. Baskaran, H. Dashti, Y. E. Ioannidis, M. Livny, P. R. Romero, D. Maziuk, J. R. Wedell, H. Yao, H. R. Eghbalnia, J. C. Hoch, J. L. Markley. *J. Biomol. NMR* **73**, 5 (2018).

[31] HUPO-PSI, mzML – Reporting Spectra Information in MS-based experiments, https://github.com/HUPO-PSI/mzML.

[32] AnIML the Analytical Information Markup Language, https://www.animl.org/.

[33] Digital Science figshare.com, https://figshare.com.

[34] IUPAC FAIRData Finding Aid, https://chemapps.stolaf.edu/iupac/demo/demo.htm.

[35] IUPAC GitHub Repository for the FAIRSpec Project, https://github.com/IUPAC/IUPAC-FAIRSpec.

[36] IUPAC FAIRSpec Working Draft Specification, https://github.com/IUPAC/IUPAC-FAIRSpec/blob/main/doc/IUPAC_FAIRSpec_Specification_draft.pdf.

[37] G. Berg-Cross, R. Ritz, P. Wittenburg. in *RDA Data Foundation and Terminology DFT: Results RFC*, Research Data Alliance (2015), https://doi.org/10.15497/06825049-8CA4-40BD-BCAF-DE9F0EA2FADF (see file 'DFT Core.pdf').

[38] RDA DFT IG Term Definitions Version 3.0, https://smw-rda.esc.rzg.mpg.de/dft-3.0.html.

[39] UTL Metadata Basics: finding aid, https://dictionary.archivists.org/entry/finding-aid.html.

[40] IDF Digital Object Identifiers, https://www.doi.org/.

[41] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. ’t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons. *Sci. Data* **3**, 160018 (2016).

[42] UTL Metadata Basics: crosswalk, https://guides.lib.utexas.edu/metadata-basics/crosswalk.

[43] UTL Metadata Basics: harvesting, https://guides.lib.utexas.edu/metadata-basics/harvesting.

[44] H. Cousijn, R. Braukmann, M. Fenner, C. Ferguson, R. van Horik, R. Lammey, A. Meadows, S. Lambert. *Patterns* **2**, (2021), https://doi.org/10.1016/j.patter.2020.100180.

[45] IUPAC Gold Book – ‘*sample*, in analytical chemistry’, https://doi.org/10.1351/goldbook.S05451.

[46] IGSN e.V. International Geo Sample Number: IGSN, https://www.igsn.org.