Progress report of IUPAC Project 2019-031-1-024
Development of a Standard for FAIR Data Management of Spectroscopic Data

Submitted by Bob Hanson (Task Group Chair), Dec. 16, 2021

Project Task Group Members

Mark Archibald, Royal Society of Chemistry
Ian Bruno, Cambridge Crystallographic Data Centre
Stuart J. Chalk, University of North Florida
Antony Davies, University of South Wales
Robert M. Hanson, St. Olaf College (co-chair 5/20 - 10/21; chair 11/21 - present)
Damien Jeannerat, NMRprocess.ch (co-chair 5/20 - 10/21; member 11/21 - present)
Robert J. Lancashire, The University of the West Indies
Jeffrey Lang, American Chemical Society Publications Division
Henry S. Rzepa, Imperial College London

Executive Summary

The project is continuing, and we are requesting a two-year extension, to DEC-31-2023. This is the first major progress report of Project 2019-031-1-024, Development of a Standard for FAIR Data Management of Spectroscopic Data, summarizing our findings from May 2020 through November 2021. We report the (submitted for) publication of a list of guiding principles that form the basis of our work. Along with those principles, we have created a set of working definitions of terms relevant to the project. (Links in blue are bookmarks into this document.) We have defined the scope of the project and have worked out a data/metadata model for what we are calling an "IUPAC FAIRSpec Finding Aid" that will be associated with an "IUPAC FAIRSpec Data Collection."

A valuable exercise we carried out during the summer of 2021 was the analysis of 13 datasets provided by the ACS FAIR Data pilot involving authors of articles in J. Org. Chem. and Organic Letters. This work was instrumental in our development of our models and understanding of the task overall. In connection with that analysis, we created a working "data and metadata extraction" utility that successfully extracted multiple representations of spectra and structures from these datasets, creating a prototypical IUPAC FAIRSpec Data Collection with associated Finding Aid.

We are currently working on fleshing out the details of the data and metadata models, which will be presented at the spring 2022 ACS National Meeting.

In addition, we have identified a number of stakeholders and have been contacting them to introduce our ideas and (hopefully) be able to align our recommendations with their emerging practices.

Origins

IUPAC project 2019-031-1-024 was initiated March 2020; announcement published in *Chem Int* July 2020, p. 28. The project is an outgrowth of a series of IUPAC- and NSF-sponsored discussions, including the IUPAC/CODATA Joint Workshop "Supporting FAIR Exchange of Chemical Data through

Standards Development" in Amsterdam in July of 2018, and the NSF-sponsored workshop “FAIR Publishing Guidelines for Spectral Data and Chemical Structures" in Orlando, Florida, in March of 2019 (See https://osf.io/psq7k/).

Activities

Meetings

The task group met biweekly by Zoom May 2020 - July 2021, then weekly starting Sep 2021.

Online Presence

IUPAC project page: https://iupac.org/projects/project-details/?project_nr=2019-031-1-024

GitHub project: https://github.com/IUPAC/IUPAC-FAIRSpec

Reference implementation/working demonstration: https://chemapps.stolaf.edu/iupac/demo/demo.htm

We also maintain a private GoogleDoc space for task group members only.

Publications

1. Robert M. Hanson and Jeffrey Lang, *FAIRSpec, Finding Aids for Primary Research Data* Charleston Hub/Against the Grain, Apr 16, 2021 This blog-like post is a summary of our initial finding aid idea and some of the basic ideas of FAIR management of data in interview format (Jeff Lang interviewing Bob Hanson). https://www.charleston-hub.com/2021/04/fairspec-finding-aids-for-primary-research-data

2. Robert M. Hanson, Damien Jeannerat, Mark Archibald, Ian Bruno, Stuart Chalk, Antony N. Davies, Jeffrey Lang, Robert J. Lancashire, and Henry S. Rzepa, *FAIR enough?* Spectroscopy Europe, 2021, vol. 33 no. 1 https://doi.org/10.1255/sew.2021.a9

3. Robert M. Hanson, Damien Jeannerat, Mark Archibald, Ian Bruno, Stuart Chalk, Antony N. Davies, Jeffrey Lang, Robert J. Lancashire, and Henry S. Rzepa, *IUPAC Specification for the FAIR Management of Spectroscopic Data in Chemistry (IUPAC FAIRSpec) - Guiding Principles,* Pure and Applied Chemistry (submitted Oct. 22, 2021)

Presentations (presenter in bold)

1. **Robert M. Hanson**, *FAIR Scientific Data Management: Connection to Digital Archiving of Historical Records*, Minnesota Historical Society Digital Archival Group Meeting, Dec. 12, 2019.

2. **Robert M. Hanson**, Damien Jeannerat, Mark Archibald, Ian Bruno, Stuart J. Chalk, Antony N. Davies, Robert J. Lancashire, Jeff Lang, Henry S. Rzepa, *Progress toward developing an IUPAC FAIR*

standard for spectroscopic data description & management, American Chemical Society National Meeting, April 14, 2021

3. Robert M. Hanson, Damien Jeannerat, Mark Archibald, Ian Bruno, Stuart J. Chalk, **Antony N. Davies**, Robert J. Lancashire, Jeff Lang, Henry S. Rzepa, *Progress toward developing an IUPAC FAIR standard for spectroscopic data description & management,* IUPAC General Assembly, August 9, 2021 Committee on Publications and Cheminformatics Data Standards

Progress to Date

Progress to date is divided here into six areas, including:

1. Development of a set of guiding principles
2. Agreeing upon a set of working definitions for terminology we use in our discussions
3. Design of a preliminary data model
4. Design of an associated preliminary metadata model
5. Work with outside organizations to ensure metadata will be findable
6. Development of a prototype reference implementation for working purposes only

Guiding Principles

The FAIRSpec Guiding Principles cover quite a bit of territory and each have nuances that we think make them unique extensions of the now famous FAIR Principles. Specifically:

- An emphasis on "cradle to immortality" management, where we go from the first day that data are collected to beyond the (sometimes thought of) graveyard of publication to the (potentially immortal) realm of reuse.
- A focus on context, recognizing the importance of the concept of a *collection* and desiring to keep relationships between spectra and their related compounds together, not "selling the car for parts."
- The recognition that metadata management requires curation, preferably automated, but, from the very start, requiring a certain amount of human attention.
- Making sure metadata is standardized within and across disciplines as much as possible.
- Designing a system with parts that can be developed in parallel and adapt to future needs.

The five principles and their twenty corollaries 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. Design to adapt to future needs.
C. Respect digital diversity.
D. All data formats should be valued.

Working Definitions

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][32, 33].

chemical structure identifier A meaningful alphanumeric text string that can uniquely identify a chemical compound and facilitate its handling in computer databases string of characters that characterizes a structure. 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 FAIRSpec Data 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 FAIRSpec Data 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.

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.

FAIR Data Management Data management based on the FAIR (Findable, Accessible, Interoperable, and Reusable) Guiding Principles, 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 FAIRSpec Data Collection A curated spectroscopic data collection organized using the principles described in this article and the (developing) IUPAC FAIRSpec Specification[31].

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 FAIRSpec Data Collection[31].

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.

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.

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 graph of persistent identifiers themselves as the basic entities that are linked together; whatever they refer to is left implicit.

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. 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.

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 FAIRSpec 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.

Data Model

The Recommendations will focus on *heterogeneous spectroscopic data collections.* That is, data collections that contain multiple representations, including spectroscopic data, structural models and/or sample descriptions, and post-acquisition structure/spectral analysis. They may be as simple as a single NMR spectrum and its associated structural model, or as complex as a collection covering the collected output of a Ph.D. student over the course of their studies. Central to this model is the *IUPAC FAIRSpec Finding Aid*, which describes the characteristics of the collection in a way that is standardized, predictable, understandable to humans, analyzable by machines, and optimized for reuse. The metadata model follows the structure of the data model, specifying characteristics of representations and their most salient properties.

Note that by "data" we mean data in a broad sense, including not only instrument-derived spectroscopic data, but also sample and structure representations, structure-spectra analyses, and possibly sampling data. Thus, we expect an IUPAC FAIRSpec Collection to contain one or more multiple spectroscopic representations -- "raw" FID files in vendor formats, more standardized data

formats, such as JCAMP-DX and NMR-STAR files, real and imaginary processed data, spectral images, and peak listings, for example.

The term "data model" here refers to an overall description of the sorts of digital objects that we expect to encounter in an IUPAC FAIRSpec Collection and how different data representations are related.

An overview of our basic data model is illustrated in Figure 1.

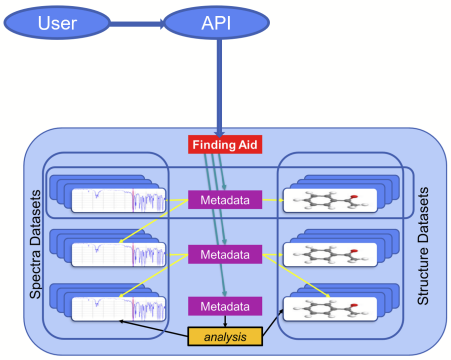

Figure 1. An overview of the data model for an IUPAC FAIRSpec Collection. The collection is heterogeneous, containing several different types of representations (digital objects) grouped into various collection types and associated by metadata. The key data element is the IUPAC FairSpec Finding Aid, which provides access to all subcollections and properties of the representations.

The starting point for our model is the concept of a *collection* of digital objects. These objects take the form of various ***data representations***. Collections can themselves be composed of collections and (being a virtual model) these subcollections can overlap. For instance, the collection of spectra and the collection of structures overlap with a structure-spectra collection (shown here as the top blue outline around a set of spectroscopic datasets and a set of structure representations).

In chemistry, the connection between structure and spectrum is important. (Exceptions may include instances where the structure is not yet known, or when unique molecular structures may not be meaningful -- such as for alloys, polymers, and composites -- in which case we allow for *sample*-based identification.) The relation of structure to spectra is "one-to-one" (a single spectrum for a specific compound), "many-to-one" (mixtures), "one-to-many" (multiple spectra for a single compound), and "many-to-many" (multiple spectra for mixtures of compounds).

One of the roles of the Task Group recommendations will be to specify "best practices" for the sort of structure-related files that accompany spectral data (mol, sdf, cif), ensuring that standardized chemical identifiers (InChI, SMILES, IUPAC preferred names) can be associated with spectra automatically, without human involvement, if necessary.

The four types of objects (representations) we recognize in a collection include:

- spectroscopic representations
    - datasets, as produced by spectrometers and other instruments
    - images of spectra
    - peak listing
    - single-block JCAMP-DX files
    - multi-block JCAMP-DX files*

    - aggregated analyses, such as MestreNova files**
- structure representations
    - mol, sdf, cif data
    - molecular drawing files (such as from ChemDraw)
    - images of structure drawings
- sample representations
    - reviewing relevant standards, including ongoing IUPAC projects
    - (to-date unspecified) representations of physical samples
- analysis representations
    - (to-date unspecified) representations of post-acquisition structure/spectrum analyses

*One of the things we have struggled with is the place for aggregated data+analysis objects such as MestreNova files and multi-block JCAMP-DX files. These can be quite complex.

**It is not clear how one would properly extract meaningful data and metadata from MestreNova files. This issue arose from the ACS pilot, where several authors submitted PDF-like (page formatted) MestreNova files instead of instrument datasets. One of the ACS pilot submissions was a single 155-MB MestreNova file containing 45 spectra for 18 structures. We were able to extract thumbnail images and structures to create more usable representations, but we consider this a preliminary practice only, not best practice. We will be looking into command-line scripting options in relation to MestreNova files.

Metadata Model

By "metadata model" we mean both a standard metadata schema, involving structured key/value pairs (*metadata elements*), and the underlying abstract model (*metadata classes*) that can be "serialized" to produce a digital object representing a collection and containing those key/value pairs.

The underlying abstract metadata model is divided into two basic *metadata classes: IFSObjects* and *IFSRepresentations* (Figure 2). *IFSObjects* are abstract collections (lists), whereas I*FSRepresentations* are pointers to "actual" digital objects. Both *IFSObjects* and *IFSRepresentations* have associated properties.

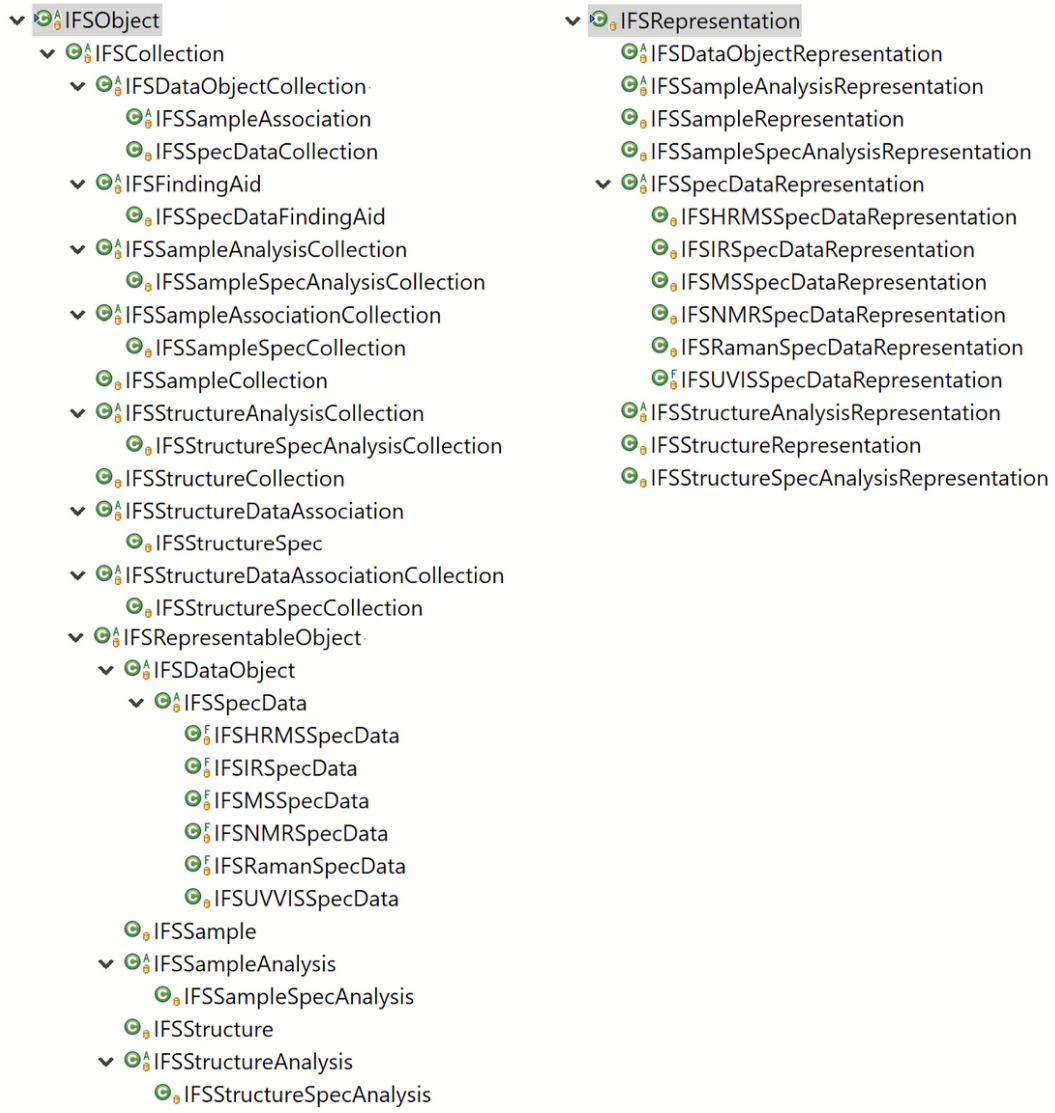

Figure 2. The hierarchical relationships of *IFSObject* and *IFSRepresentation* metadata objects. Note that only *IFSRepresentableObjects have associated actual digital representations. IFSCollection* is a completely abstract metadata-only concept.

Note that the *IFSFindingAId* metadata class is the key *IFSObject* for an IUPAC FAIRSpec Collection and is itself a type of *iFSCollection.* Although it does not technically have an IFSRepresentation, it can be *serialized* to produce a digital object (for example, a JSON or XML document).

The basic structure of a proposed IUPAC FAIRSpec metadata element key is a prefixed set of identifiers that match our metadata model, somewhat along the lines of CIF syntax. Capitalization is proposed to be important, and all IFS metadata keys start with "IFS.representation." or "IFS.property." and then proceed with one or more lowercase class names separated by ".". For example:

IFS.representation.spec.nmr.vendor.dataset
IFS.representation.spec.ir.spectrum.image
IFS.property.collection.source.data.uri

The model allows for uppercase unit suffixes. Currently there is only one such property:

IFS.property.spec.nmr.expt.temperature.K

We are still debating the use of units in property names. The reason for this one is that we want to make it very clear *to a human* that temperature in NMR is to be expressed in Kelvin. But this is still a very preliminary assignment, and we recognize that there are difficulties inherent to adding units to metadata keys.

The IFS.representation.* metadata classes currently include:

.spec.ir
.spec.hrms
.spec.ms
.spec.nmr
.spec.raman
.spec.uvvis

.struc.cdx
.struc.cdxml
.struc.mol
.struc.png
.struc.sdf
.struc.unknown

and are expected to expand as necessary to include a wide variety of data sources and structure representation types. Representation metadata has not been developed yet for samples or analysis.

Property metadata classes are similarly constructed:

IFS.property.collection.*
IFS.property.spec.*
IFS.property.struc.*
IFS.property.sample.*
IFS.property.analysis.*

The full set of NMR properties, for example, currently include:

IFS.property.spec.nmr.expt.dim
IFS.property.spec.nmr.expt.freq.1
IFS.property.spec.nmr.expt.freq.2

IFS.property.spec.nmr.expt.freq.3
IFS.property.spec.nmr.expt.label
IFS.property.spec.nmr.expt.nucl.1
IFS.property.spec.nmr.expt.nucl.2
IFS.property.spec.nmr.expt.nucl.3
IFS.property.spec.nmr.expt.pulse.prog
IFS.property.spec.nmr.expt.solvent
IFS.property.spec.nmr.expt.temperature.K
IFS.property.spec.nmr.instr.freq.nominal
IFS.property.spec.nmr.instr.manufacturer.name
IFS.property.spec.nmr.instr.probe.type

where bold indicates required metadata for any NMR spectrum. (Only for 2D or 3D spectra the corresponding frequency and nucleus metadata in italics would be required.) To the extent that the properties are available, metadata relating to pulse program, temperature, and probe type would be desirable.

We have not yet discussed the exact specifications for the values of any of these metadata keys.

Metadata Referencing

We are working with metadata registration agencies DataCite, CrossRef, and Re3Data, encouraging them to allow the inclusion into their schemas references of discipline-specific metadata schemas such as those we are proposing. It is becoming clear that other disciplines are starting to think along the same lines, and progress is being made on this front.

Metadata referencing can originate in any descriptive article that references the data on which the narrative of the article is based. It can also be referred to within the metadata registered for datasets, in the form of associated or related identifiers to other datasets. To complete the cycle, dataset metadata can itself reference articles or other instances in which the dataset is quoted, presented or discussed. Metadata referencing should ideally be bidirectional, in which say an article and a related published dataset, or two datasets in different locations, each reference the other with additional context added where possible, such as the relationship between the two objects.

The referencing should conform to the schema being used to structure the metadata. We are encouraged that the referencing of data by extending the CrossRef article publishing schema is currently a topic of active development.

Prototype FAIRification Workflow Implementation

We have come to recognize the distinction between a ***digital aggregation*** (a set of ***digital entities*** – think "files") and a ***digital collection*** (a set of ***digital objects*** that are connected by metadata). Thus, we are not suggesting that data sets necessarily be in the form of IUPAC FAIRSpec Data Collections. Rather, we imagine a workflow that starts with a reasonably well organized digital aggregation, and then, through a "FAIRification workflow," that aggregation is turned into an IUPAC FAIRSpec Data Collection. So, for example, given a digital supporting information aggregation supplied by an author, while it would be great if it were already an IUPAC FAIRSpec Data Collection, we imagine (and have

experimented with) a process that involves extraction, standardization, serialization, and packaging (Figure 3).

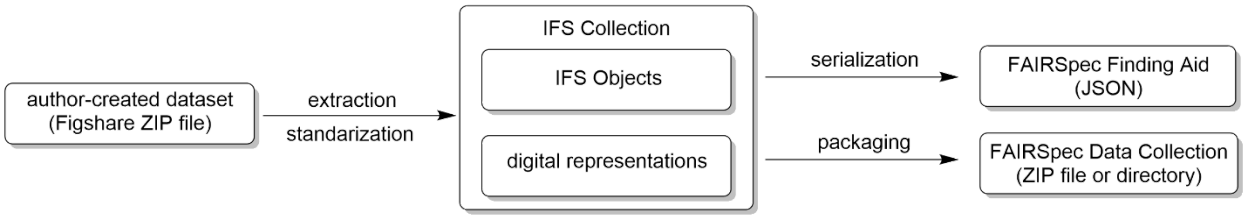

Figure 3. Example FAIRification workflow. A digital aggregation in the form of a ZIP file is standardized and extracted to produce an IFS Collection model with associated representations (digital object). Serialization of the IFS Collection model produces the IUPAC FAIRSpec Finding Aid in JSON format. Packaging of the digital objects produces the associated IUPAC FAIRSpec Data Collection.

Examples of the sort of extraction and standardization that was required for the ACS pilot aggregations include:

- Most of the Bruker datasets that were submitted by authors are not readable in Bruker's own TopSpin system, simply because they do not reside in a directory with a name that is a simple integer. So to correct for this, we added a numeric directory.
- Bruker datasets sometimes contain thumbnail images of spectra, which could be extracted and used as separate representations.
- Several authors submitted MestreNova files rather than instrument datasets. This proved quite problematic, as a MestreNova file is a proprietary publicly undocumented format that can be considered more like an "interactive PDF," with pages and page formatting, rather than a structured dataset. We have not experimented with scripted access to these data via MestreNova scripting. Instead, we created an extraction utility that can scan a binary MestreNova file for pertinent data and metadata. The result was interesting but not fully satisfying, and certainly not to be considered best practice.

Future Plans

Model Development

Both the data model and metadata model are undergoing continued rapid development. There are several missing pieces that still need work, not the least of which is the choice of a system for the cataloging and distribution of the metadata schema. We plan to be active on this front throughout the next year.

Implementation Testing

We expect to continue implementation testing in order to discover hidden issues with the data and metadata models both at St. Olaf College (where we have a fully automated NMR facility and a steady flow of approximately 120 student users) and Imperial College London (where there is an institutional repository, and work is being done to develop workable FAIR data management procedures).

Presentations and Publications

We have submitted an abstract for the spring 2022 ACS National Meeting, where we plan to discuss our proposed data and metadata models. it reads:

> *Much of the science reported in the field of chemistry is based on, or backed by, spectroscopic data and analysis. Synthetic organic chemists, natural product chemists, and many others publish their work along with accompanying details relating to experimental procedures using a proof of structure in the form of NMR, IR, and MS spectra and analysis. Graduate students similarly present their theses with extensive supporting information. Although a widely adopted IUPAC standard for storage and retrieval of spectroscopic data (JCAMP-DX) has been available for many years, to date there has been no standard way to find that data or to describe a correlation of that data with chemical structure. The objective of the IUPAC Project Development of a Standard for FAIR Data Management of Spectroscopic Data is to develop standards for the production and dissemination of digital data objects that contain enough spectral data and metadata that they can be (a) findable through either human or machine-automated searches on the web, (b) available through standard interfaces, (c) interoperable and transferable among systems, and (d) readable and reusable over time, for both humans and machines.*
>
> *In this presentation, we will present progress toward these goals, including proposals for standards that allow for extraction of key metadata from spectroscopic data, metadata for spectral analysis -- relating chemical structure and its corresponding spectral signals, and, especially, metadata for collections of related data (for example, all the spectroscopic data for a thesis or publication).*
>
> *The presentation will draw the connection between the field of digital archival science and the task at hand, emphasizing the importance of accepting data in a wide range of formats, quality, and completeness. We will discuss the critical components required for FAIR management of spectroscopy collections, together with an illustration of how searches of standardized spectral analysis metadata, which has been globally registered, aggregated and indexed, might act as a finding and accessing tool.*
>
> *Finally, we will present our thoughts relative to how a new culture of spectroscopic data sharing can be achieved based on the proposed IUPAC standards.*

We are looking into other relevant meetings where we could publicise our efforts.

Timeline

We are requesting a two-year extension, to DEC-31-2023. Our current timeline is as follows:

- early 2022: Working with the NFDI4Chem project and others, including metadata registration agencies, to see how we can work together or in parallel to make sure our metadata and data models are compatible
- summer 2022: Finalizing the draft models and recommendations
- fall 2022: Seeking wide feedback on the recommendations, adjusting as necessary

- early 2023: PAC submission, or if necessary, second round of adjustments
- summer 2023 (at the latest): Finalization and publication of recommendations