Demonstration of the production and use of IUPAC FAIRSpec Data Collections and IUPAC FAIRSpec Finding Aids from "FAIRSpec-ready" data collections

ACS National Meeting, Aug. 20, 2025

Robert M. Hanson, Henry S. Rzepa, Stefan Kuhn, John R. Cort, Swagat Malla

IUPAC Project 2019-031-1-024

Division of Chemical Information:
Framing FAIR: Scientific Research Data Sharing Policies, Frameworks and Principles

Our mandate

From the *NSF Guide to Proposal Preparation*:

Proposals must include a document of no more than two pages uploaded under "Data Management Plan" …. This supplementary document should ***describe how the proposal will conform to NSF policy on the dissemination and sharing of research results*** *… and may include … the standards to be used for data and metadata format and content.*

Investigators are expected to share with other researchers … the ***primary data****, samples, physical collections,* ***and other supporting materials*** *created or gathered in the course of work under NSF awards.*

Our mandate

From the *CHE Data Management and Sharing Plan Guidance*:

Access to data and data sharing practices and policies. …Describe your plans, if any, for providing such general access to data, including **websites maintained by your research group**, and **direct contributions to public databases or software repositories** (e.g., **NMRShiftDB**,…. Consider using **digital object identifiers (DOI)** … for **suitably-archived, publishable data sets**.

Archiving of Data. …Will there be **an easily accessible index that documents where all archived data are stored and how they can be accessed**?

Unaddressed issues: findability and reusability

The key missing term here is **collection**

Unaddressed issues: findability and reusability

How do we organize it? How do we make it available?

Our approach:

Our approach:

Guiding Principles for the FAIR Management of Spectroscopic Data

IUPAC Specification for the FAIR Management of Spectroscopic Data in Chemistry (IUPAC FAIRSpec) - Guiding Principles

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

Pure and Applied Chemistry, 2022

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

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.

just out…

https:///doi.org/10.26434/chemrxiv-2025-vqhnd-v2

A must read!

The key is to provide well-crafted structure representations.

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

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

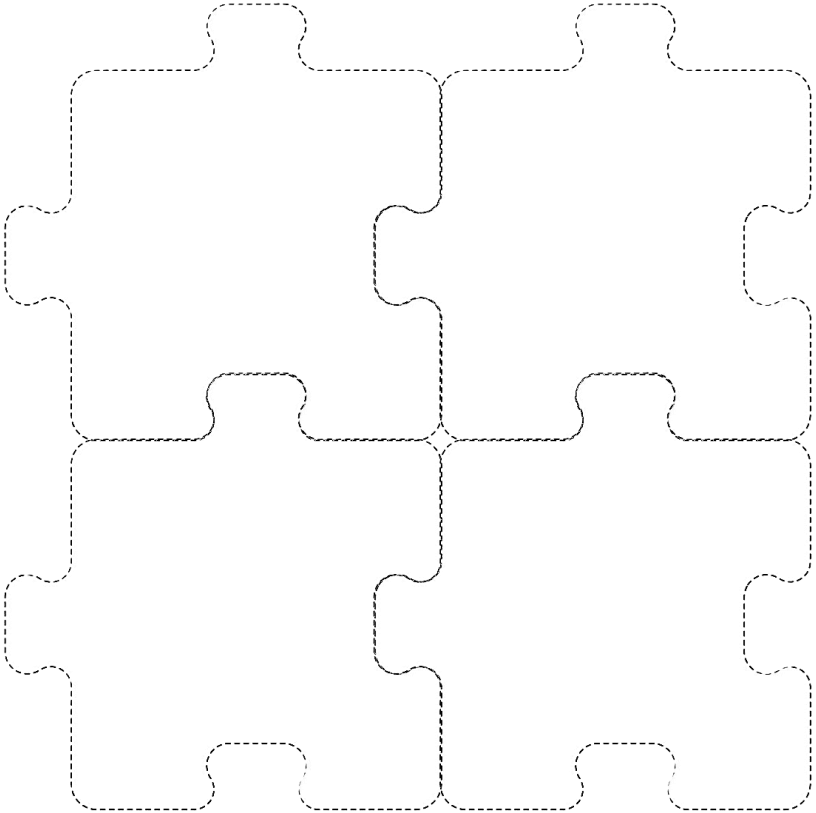
A must read!

…and appropriate dataset representations.

Table 2: Digital dataset representation examples

Representation type	Considerations
Original instrument data (for example, the NMR FID and associated parameter files)	Benefits: highest integrity with no information loss; high potential for reuse; allows for alternative and/or automated (re)analysis; potential for generation of all other representations; allows possibility of fraud detection; allows the most reliable automated metadata extraction. Limitations: vendor-specific format may require licensing access to a vendor-specific reader; may not express important aspects of the data processing used in the analysis; must contain enough of the key parameters for further processing; may no longer be documented; shortest expected lifetime.
Original data exported to an alternative standardized format, such as JCAMP-DX(FID)[38] NMR-Star[39] or nmrML[40]	Benefits: highest possibility of reuse and interoperability; vendor-agnostic open format; allows for automated production of additional representations; recognized by regulators as potentially the longest expected lifetime. Limitations: Depending on implementations potential for loss of data resolution or precision; potential for loss of some metadata fields.
Instrument-processed data such as transformed NMR data in the form of a spectrum or spectra	Benefits: most generally and immediately informative to the practicing scientist or educator; can be examined in detail and repurposed. Limitations: does not allow for early-processing adjustments, such as in NMR spectroscopy phasing, line broadening or use of non Fourier-transform methods; may require proprietary software to read; information loss compared to original data; less scope for fraud detection.
Third party-processed data	Benefits: concise; convenient if this is the standard process in each laboratory; may include meaningful annotation added by the originator such as integration or peak identification"; May be the only method of getting FAIR metadata to be associated with the data. Limitations: format may require proprietary software to read; possibly limited ability to extract key metadata from proprietary formats; may disallow alternative analysis; may be more subject to fraud or other sorts of spectral editing (removal of solvent peaks, for example)

The data management puzzle

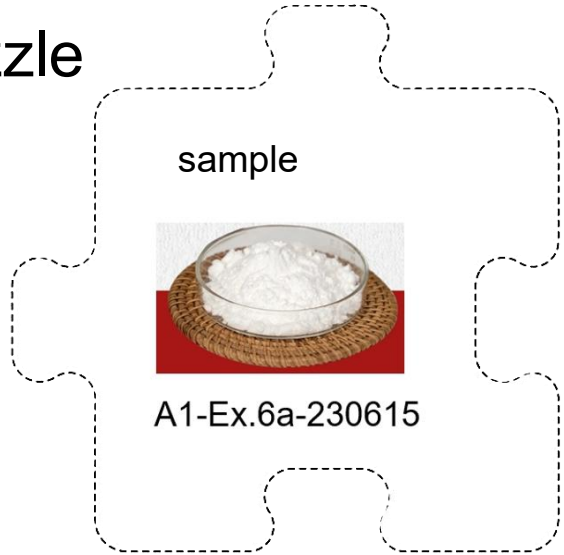


7	1 5 8 9	5 6 8	1 2 6 9	1 3 5 6 9	2 3 5 6 9	4	1 2 5	1 2 3
1 4 5 6 9	2	4 5 6	1 4 6 9	7	3 4 5 6 9	1 5 6	8	1 3
1 4 5 6	1 4 5	3	1 2 4 6	1 5 6	8	1 2 5 6 7	1 2 5 7	9
2 4 8 9	4 7 8 9	2 4 7 8	5	1 6 8 9	4 6 9	3	1 2 4 7	1 2 4 7 8
3 4 5 8	6	4 5 7 8	1 4 8	2	3 4	1 5 7 8	9	1 4 7 8
2 3 4 5 8 9	4 5 8 9	1	4 8 9	3 8 9	7	2 5 8	2 4 5	6
1 2 5 6 8	1 5 7 8	2 5 6 7 8	3	5 6 8	2 5 6	9	1 2 4 7	1 2 4 7 8
1 2 5 8	3	2 5 7 8	2 7 8 9	4	2 5 9	1 2 7 8	6	1 2 7 8
2 4 6 8	4 7 8	9	2 6 7 8	6 8	1	2 7 8	3	5

The data management puzzle

We start with a sample.

It has various characteristics, most importantly an ***identifier***

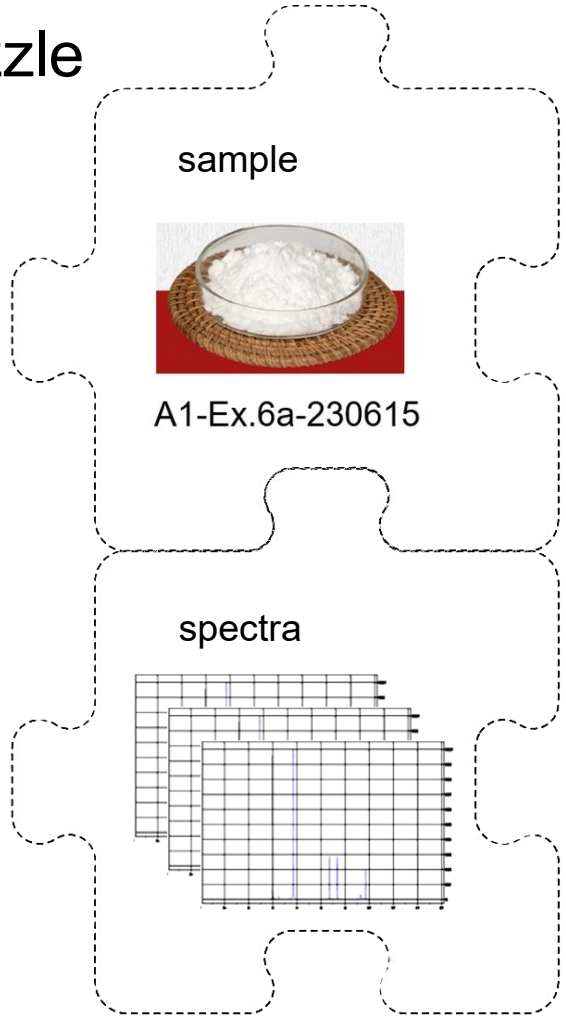


The data management puzzle

We run tests on it, creating instrument datasets and reports.

Now we have a ***data collection***.

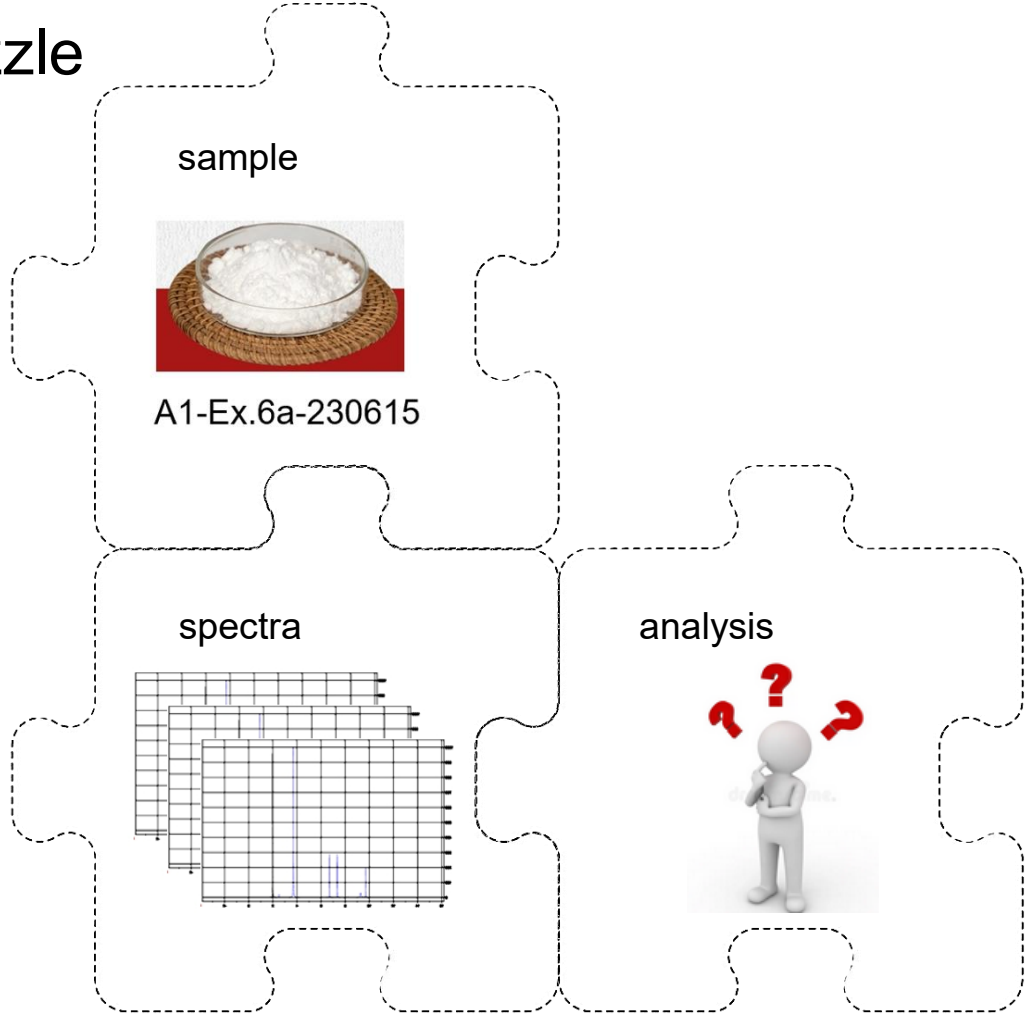
The sample identifier is the *metadata* that establishes the connection between the sample and its spectra.



The data management puzzle

We think on this.

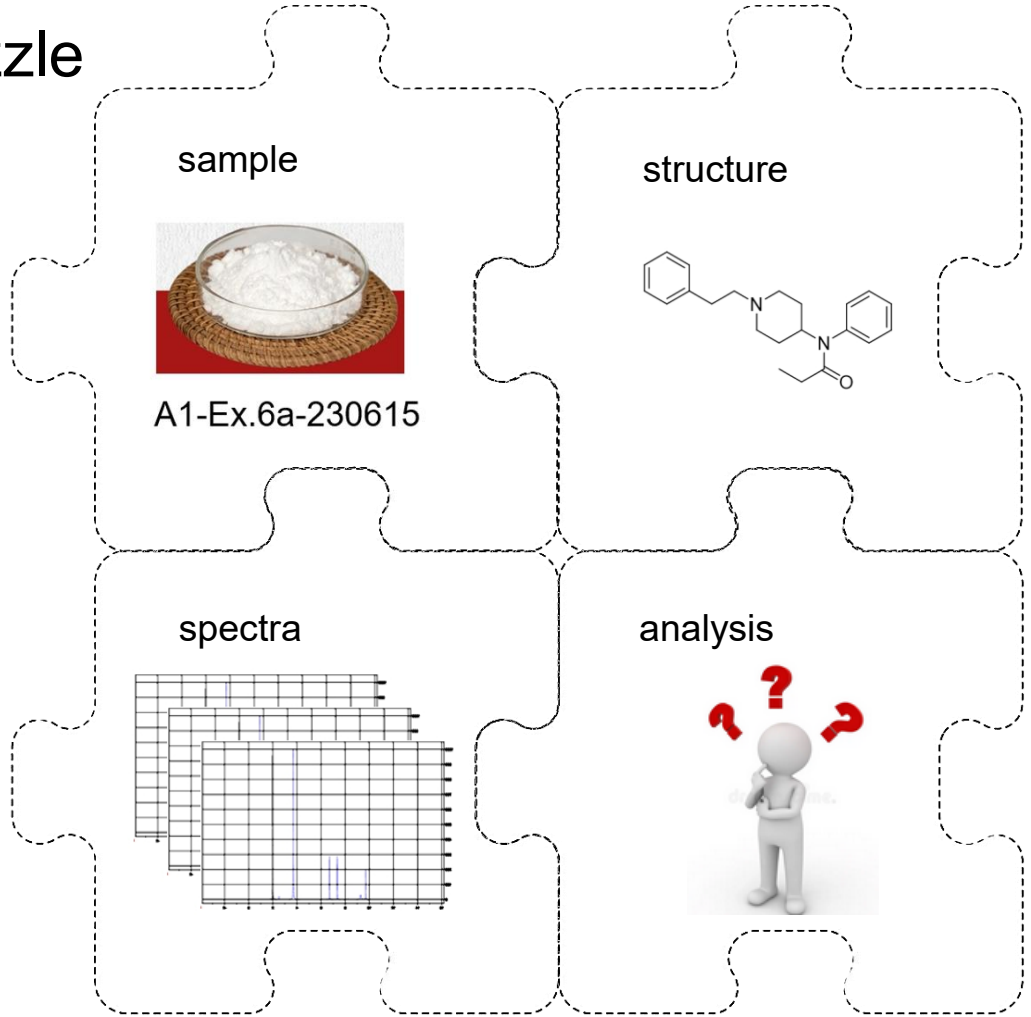
(Or pass it to AI??)



The data management puzzle

Our goal is to make a connection between sample and structure.

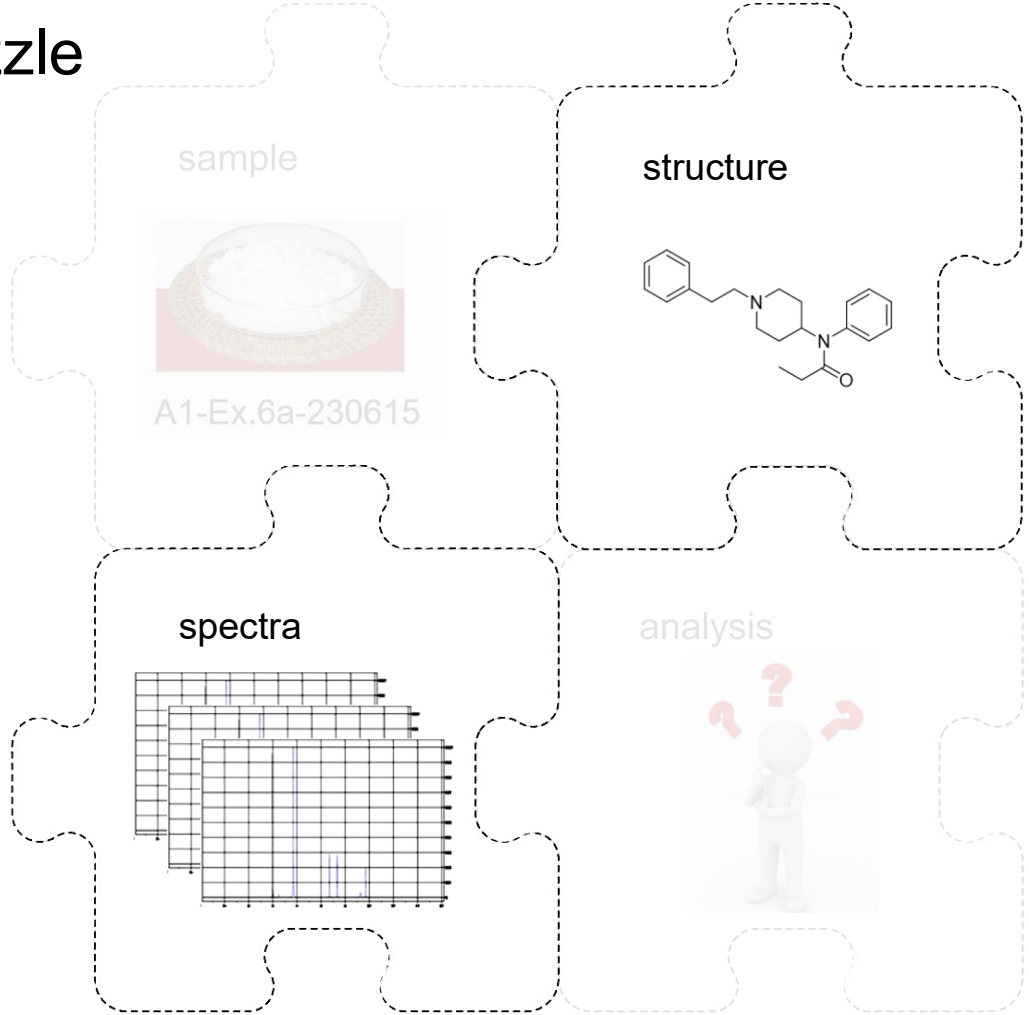
Note how indirect this really is!



The data management puzzle

But, in the end, we often just deliver two of the four pieces – the structure and the spectra.

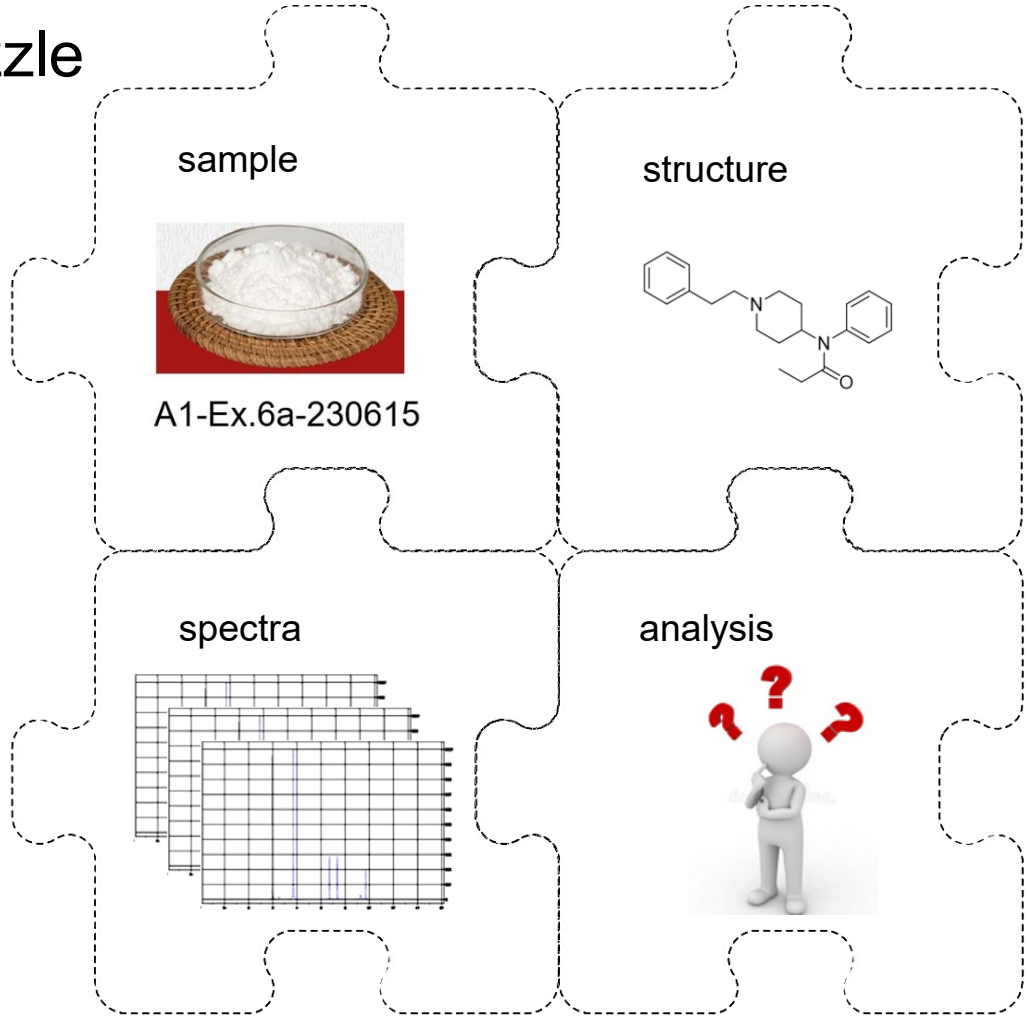
We call this a ***compound association***



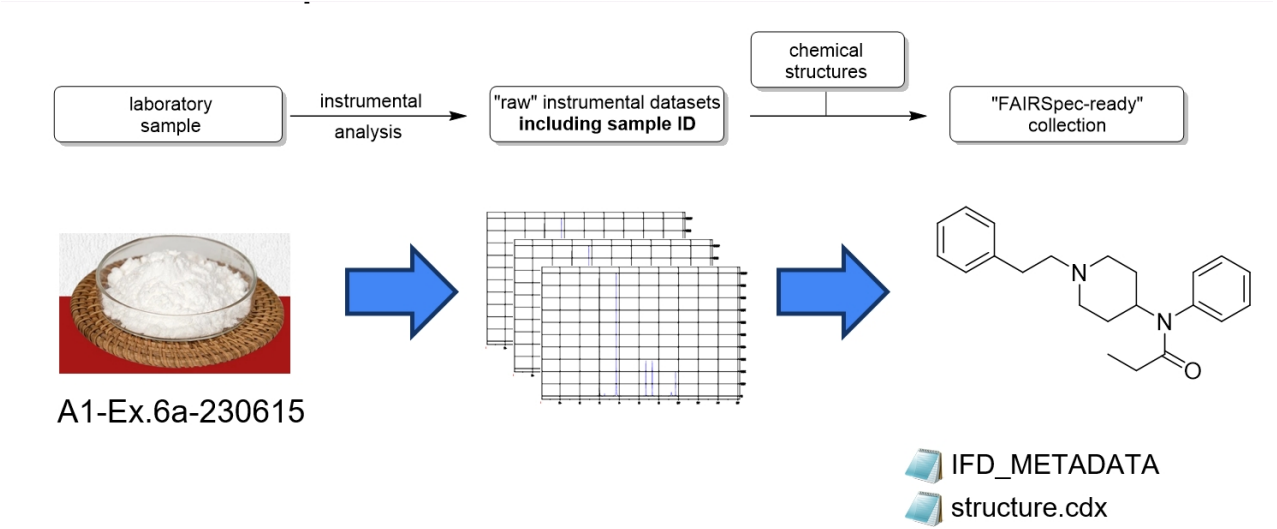
The data management puzzle

The goal of IUPAC Project 2019-031-1-024 is to develop standards for the maintenance and presentation of these relationships.

The overall context is much larger than just “SI for a paper”.



The data management puzzle

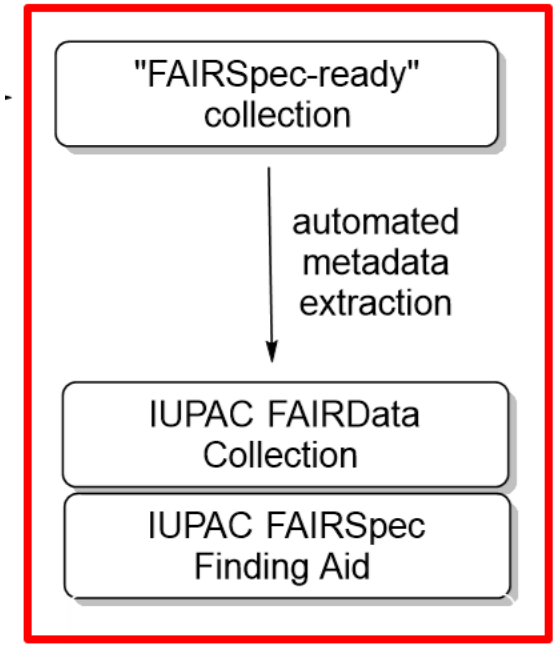


The goal of IUPAC Project 2019-031-1-024 is to develop standards for the maintenance and presentation of these relationships *throughout the entire research workflow*.

The IUPAC FAIRSpec Solution

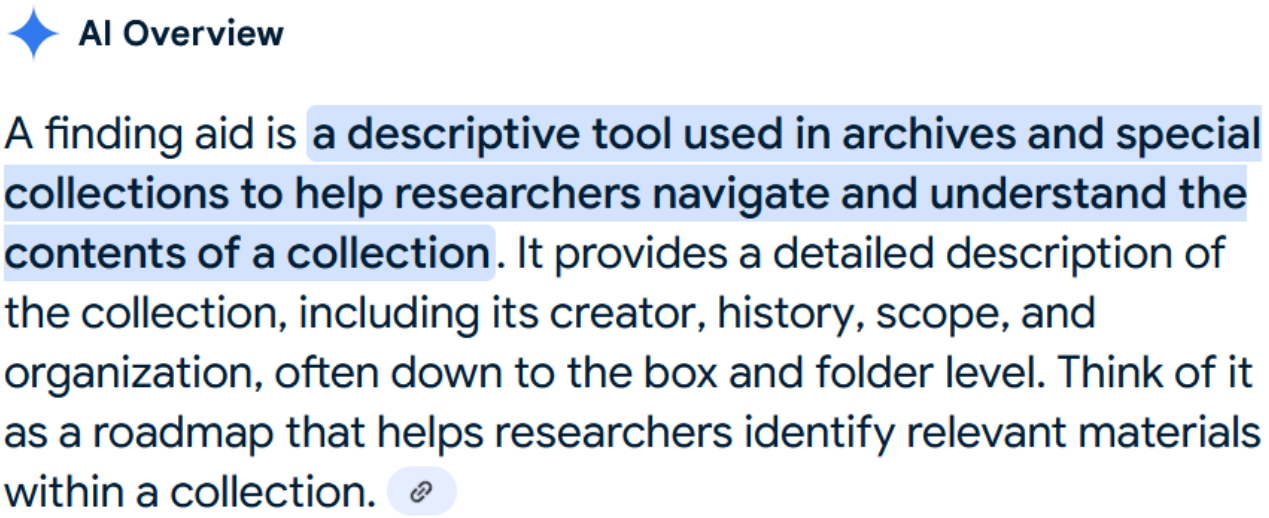
Our solution is to enable authors and data managers to create what we are calling a **“FAIRSpec-ready” collection.** Then let automation take that the rest of the way to an **IUPAC FAIRSpec Collection** with its associated **IUPAC FAIRSpec Finding Aid.**

The good news is that **we have tested this with over a dozen datasets**, *and it works*.



Key concept – *Finding Aid*

a term borrowed from the field of archival science:

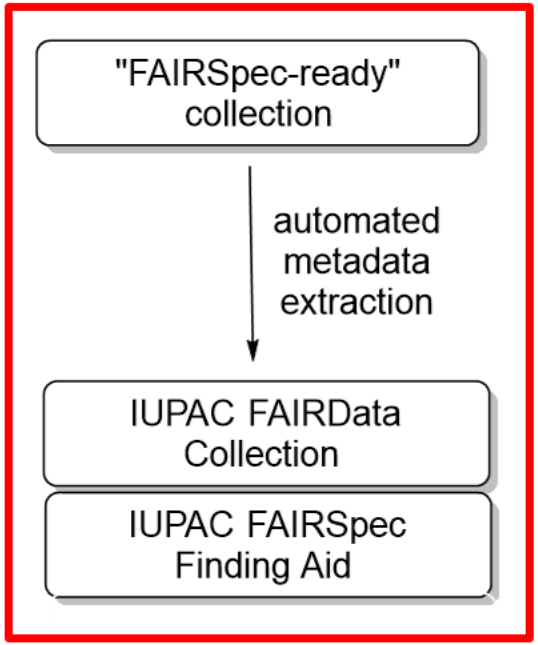


Key concept – *Finding Aid*

An IUPAC FAIRSpec Finding Aid can be represented as a JSON document.

The IUPAC FAIRSpec solution

The focus of this presentation is to present several implementations of this process in different contexts.

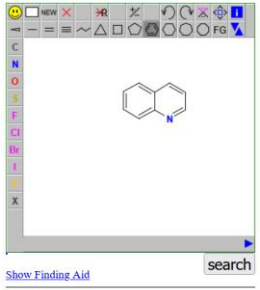


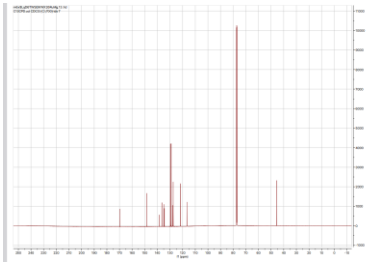
Demonstrations

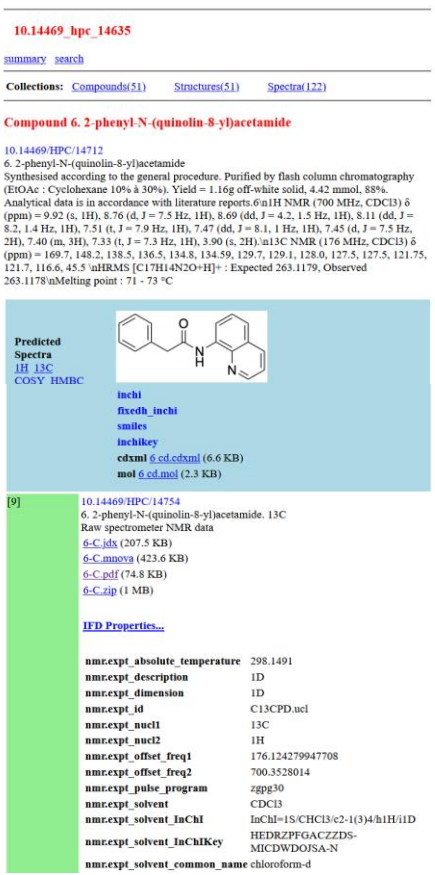
What follows are three demonstrations, in three very different contexts:

1. An undergraduate NMR laboratory using automation.
2. A set of publication spectroscopic supporting information package.
3. A repository holding shared data related to a publication.

All use the same HTML landing page; only the IUPAC FAIRSpec Finding Aids are different.







Demonstrations

Note that everything from here on in this presentation is something you can do yourself by going to our GitHub project:

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

or its associated GitHub pages website

https://iupac.github.io/IUPAC-FAIRSpec

README | MIT license

IUPAC

last updated 2025-06-24 (BH)

IUPAC-FAIRSpec

Welcome to the GitHub project for IUPAC Project 2019-031-1-024, Development of a Standard for FAIR Data Management for Spectroscopic Data.

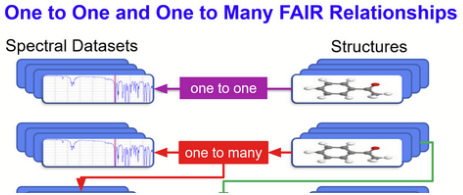
Check out the GitHub web pages site for this project to see what we are talking about.

This is a an active project. Our (rather minimal, admittedly) progress reports can be found in documents/reports

Our current working specification can be found at this site under documentation/specifications.

This GitHub project provides public copies of all presentations and publications of the IUPAC Project, as well as a reference Java implementation of the Standard as a Java library as well as a reference Java implementation of an "IUPAC FAIRSpec data and metadata extractor". This GitHub project is currently under intensely active development. It is reasonable well tested (as of 11/2024) and, though public, is only meant for demonstration purposes. **implement these preliminary standards with this in mind** as they periodically change as new issues emerge.

The principal goal of the project is to define standardized metadata associated with complex collections of spectroscopic data in the area of chemistry -- NMR, IR, Raman, MS, etc. The specification is modular and has been worked out primarily in the area of NMR spectroscopy at this time.



IUPAC FAIRSpec GitHub Project pages

Welcome to the GitHub project web pages for IUPAC Project 2019-031-1-024, Development of a Standard for FAIR Data Management for Spectroscopic Data. At this site we highlight our development of IUPAC FAIRSpec Finding Aids.

See publications for published and submitted publications.

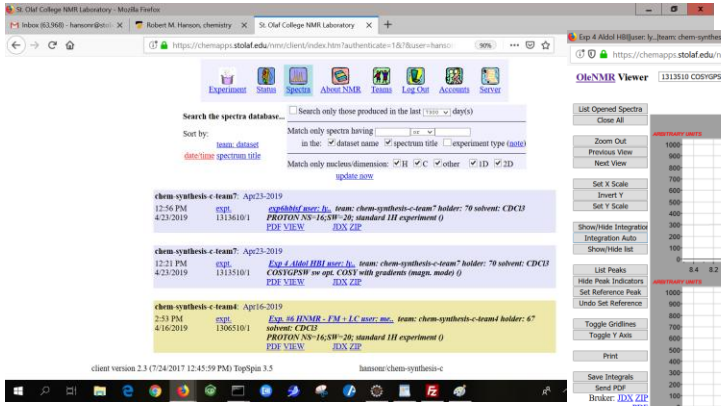
CDX/CDXML specification

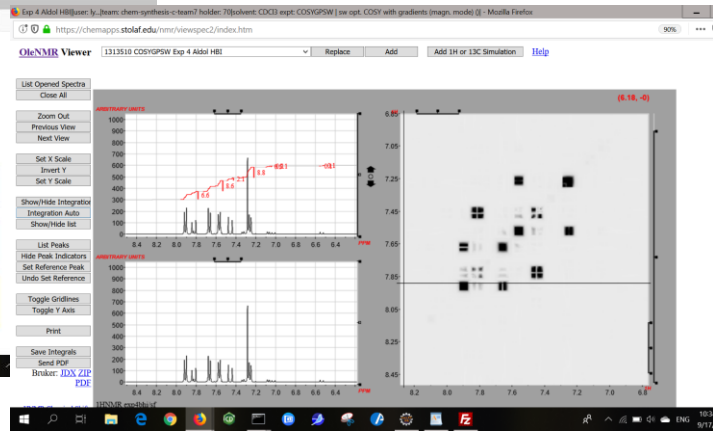
But, first, 352 pages retrieved from multiple points on the wayBack machine preserving the CDX/CDXML specification. (Note that some images were not retrievable.) This specification is important, as it is the basis for one of the most widely used and accessible formats in the chemistry community for communicating structural information. The specification is detailed and extensive, and has been implemented in Jmol, allowing an extractor to use it as a basis for creating value-added structure representations such as molecular formulas, InChI strings, SMILES, and MOL-format descriptions, which can be used both for display and validation. The pages are also available as a ZIP file.

Demonstration 1

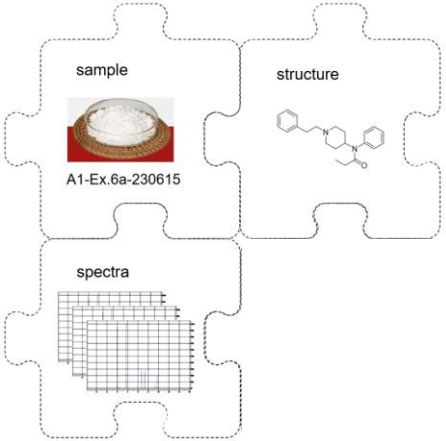
St. Olaf College Organic Chemistry I laboratory course.

Fully automated web-based interface to IconNMR





Demo 2024.1 examples2/v6-stolaf



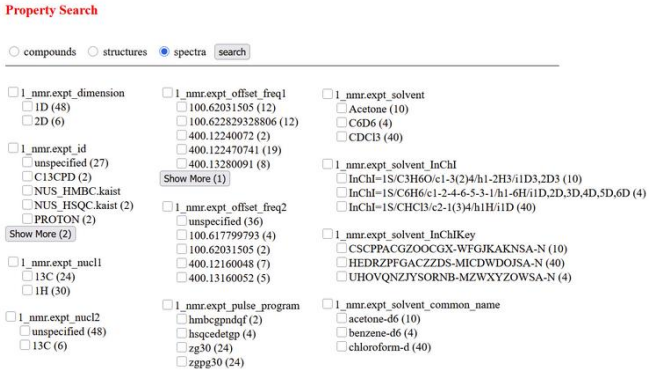
https://iupac.github.io/IUPAC-FAIRSpec

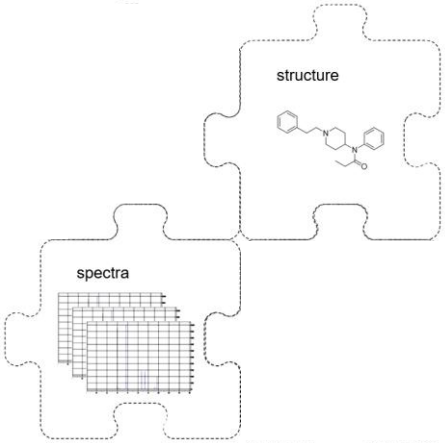
Demonstration 2

Fourteen datasets deposited as supporting information for publications.

Each finding aid is fully searchable for text, properties, and substructures.

Demo 2025.1 [examples2/v6-acs]





This page and its associated IUPAC FAIRSpec Finding Aid were automatically generated by IFDExtractor.java (GitHub site). It is a demonstration page for IUPAC Project 2019-031-1-024, *Development of a Standard for FAIR Data Management of Spectroscopic Data.*

Select a Finding Aid

acs.joc.0c00770 view IFD.findingaid.json IFD.collection.zip (25.2 MB)

Collections:	Compounds(11) Structures(11) Spectra(54)
Title	*Total Synthesis of (+)-Pestalofone A and (+)-Iso-A82775C*
Authors	Geon Kim; Taevan Kim; Sunkyu Han
DOI	https://doi.org/10.1021/acs.joc.0c00770 (metadata)
Data Origin	https://ndownloader.figshare.com/files/22567817 (24.5 MB)
FAIRSpecDataCollection	/acs.joc.0c00770/.IFD.collection.zip (25.2 MB)

acs.orglett.0c00571 view IFD.findingaid.json acs.orglett.0c00571.IFD.collection.zip (199.6 MB)

Collections:	Compounds(30) Structures(30) Spectra(114)
Title	*Synthesis of Novel Heterocycles by Amide Activation and Umpolung Cyclization*
Authors	Haoqi Zhang; Margaux Riomet; Alexander Roller; Nuno Maulide
DOI	https://doi.org/10.1021/acs.orglett.0c00571 (metadata)
Data Origin	https://ndownloader.figshare.com/files/21975525 (189.9 MB)
FAIRSpecDataCollection	https://chemapps.stolaf.edu/iupac/examples2/v6-acs/acs.orglett.0c00571.IFD.collection.zip (199.6 MB)

acs.orglett.0c00624 view IFD.findingaid.json IFD.collection.zip (15.9 MB)

Collections:	Compounds(41) Structures(42) Spectra(80)
Title	*Intermolecular Vicinal Diaminative Assembly of Tetrahydroquinoxalines via Metal-free Oxidative [4 + 2] Cycloaddition Strategy*
Authors	Dangui Wang; Huaibin Yu; Shaohan Sun; Fangrui Zhong
DOI	https://doi.org/10.1021/acs.orglett.0c00624 (metadata)
Data Origin	https://ndownloader.figshare.com/files/21947274 (15.2 MB)
FAIRSpecDataCollection	/acs.orglett.0c00624/.IFD.collection.zip (15.9 MB)

acs.orglett.0c00755 view IFD.findingaid.json IFD.collection.zip (105.7 MB)

Collections:	Compounds(26) Structures(28) Spectra(78)
Title	*Synthesis of a P-Glycoprotein Inhibitor and Its High-Energy (Z)-Isomer by Carbenoid Eliminative Cross-Coupling*
Authors	Subhash D. Tanpure; Mohamed F. El-Mansy; Paul R. Blakemore
DOI	https://doi.org/10.1021/acs.orglett.0c00755 (metadata)
Data Origin	https://ndownloader.figshare.com/files/22150197 (105.6 MB)
FAIRSpecDataCollection	https://chemapps.stolaf.edu/iupac/examples2/v6-acs/acs.orglett.0c00755/IFD.collection.zip (105.7

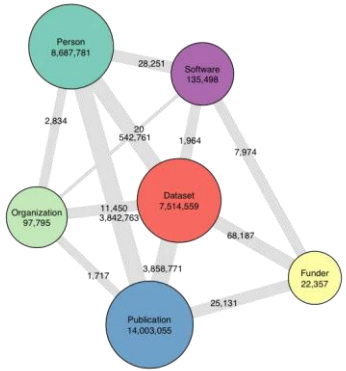
https://iupac.github.io/IUPAC-FAIRSpec

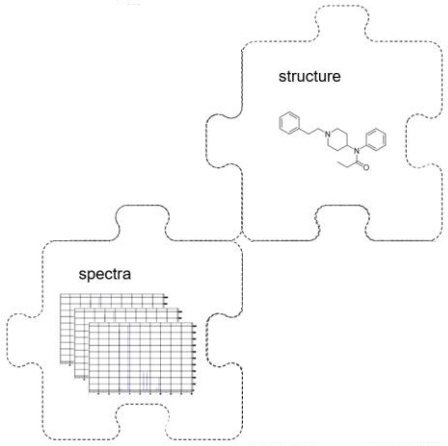
Demonstration 3

A finding aid created by “crawling” a chain of metadata documents at DataCite.

The metadata was richly annotated manually with links to related documents.

Demo 2025.6 [examples2/v6-crawler2]





IMPERIAL

Research Data Repository

Browse | Add Collection | Deposit Data | Help | Search Engine: Google | Search Query:

Syntheses and Characterization of Main Group, Transition Metal, Lanthanide and Actinide Complexes of Bidentate Acylpyrazolone Ligands

DOI: 10.14469/hpc/10386 Metadata

10.14469_hpc_10386

summary search

Collections:	Compounds(57) Structures(57) Spectra(228)
Title	*Syntheses and Characterization of Main Group, Transition Metal, Lanthan… Bidentate Acylpyrazolone Ligands*
Authors	**Thomas Mies, Andrew J. P. White; Henry S. Rzepa, Luciano Barluzzi Layfield, Anthony G. M. Barrett**
DOI	https://doi.org/10.1021/acs.inorgchem.3c01506 (metadata)
Data Title	*Syntheses and Characterization of Main Group, Transition Metal, Lanthan… Bidentate Acylpyrazolone Ligands*
Data DOI	https://doi.org/10.14469/HPC/10386 (metadata)
Description	The synthesis of acylpyrazolone salts and their complexes of main group e… lanthanides, and actinides are described and characterized inter alia by mea… crystallography, NMR, and IR spectroscopies. The complexes consist of tw… ligands bound to the metal atom, resulting in a structurally diverse set of c… (distorted) octahedral, pentagonal-bipyramidal, or antiprismatic arrangeme…

https://iupac.github.io/IUPAC-FAIRSpec

Thank you!

Note that everything from here on in this presentation is something you can do yourself by going to our GitHub project:

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

or its associated GitHub pages website

https://iupac.github.io/IUPAC-FAIRSpec

README MIT license

last updated 2025-06-24 (BH)

IUPAC-FAIRSpec

Welcome to the GitHub project for IUPAC Project 2019-031-1-024, Development of a Standard for FAIR Data Management for Spectroscopic Data.

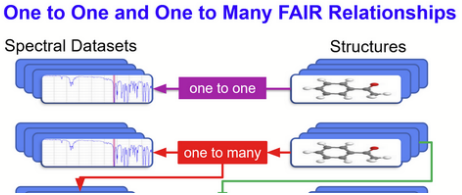
Check out the GitHub web pages site for this project to see what we are talking about.

This is a an active project. Our (rather minimal, admittedly) progress reports can be found in documents/reports

Our current working specification can be found at this site under documentation/specifications.

This GitHub project provides public copies of all presentations and publications of the IUPAC Project, as well as a reference Java implementation of the Standard as a Java library as well as a reference Java implementation of an "IUPAC FAIRSpec data and metadata extractor". This GitHub project is currently under intensely active development. It is reasonable well tested (as of 11/2024) and, though public, is only meant for demonstration purposes. **implement these preliminary standards with this in mind** as they periodically change as new issues emerge.

The principal goal of the project is to define standardized metadata associated with complex collections of spectroscopic data in the area of chemistry -- NMR, IR, Raman, MS, etc. The specification is modular and has been worked out primarily in the area of NMR spectroscopy at this time.



IUPAC FAIRSpec GitHub Project pages

Welcome to the GitHub project web pages for IUPAC Project 2019-031-1-024, Development of a Standard for FAIR Data Management for Spectroscopic Data. At this site we highlight our development of IUPAC FAIRSpec Finding Aids.

See publications for published and submitted publications.

CDX/CDXML specification

But, first, 352 pages retrieved from multiple points on the wayBack machine preserving the CDX/CDXML specification. (Note that some images were not retrievable.) This specification is important, as it is the basis for one of the most widely used and accessible formats in the chemistry community for communicating structural information. The specification is detailed and extensive, and has been implemented in Jmol, allowing an extractor to use it as a basis for creating value-added structure representations such as molecular formulas, InChI strings, SMILES, and MOL-format descriptions, which can be used both for display and validation. The pages are also available as a ZIP file.