IUPAC FAIRSpec-ready aggregations: Recommendations for researchers, authors, and publishers

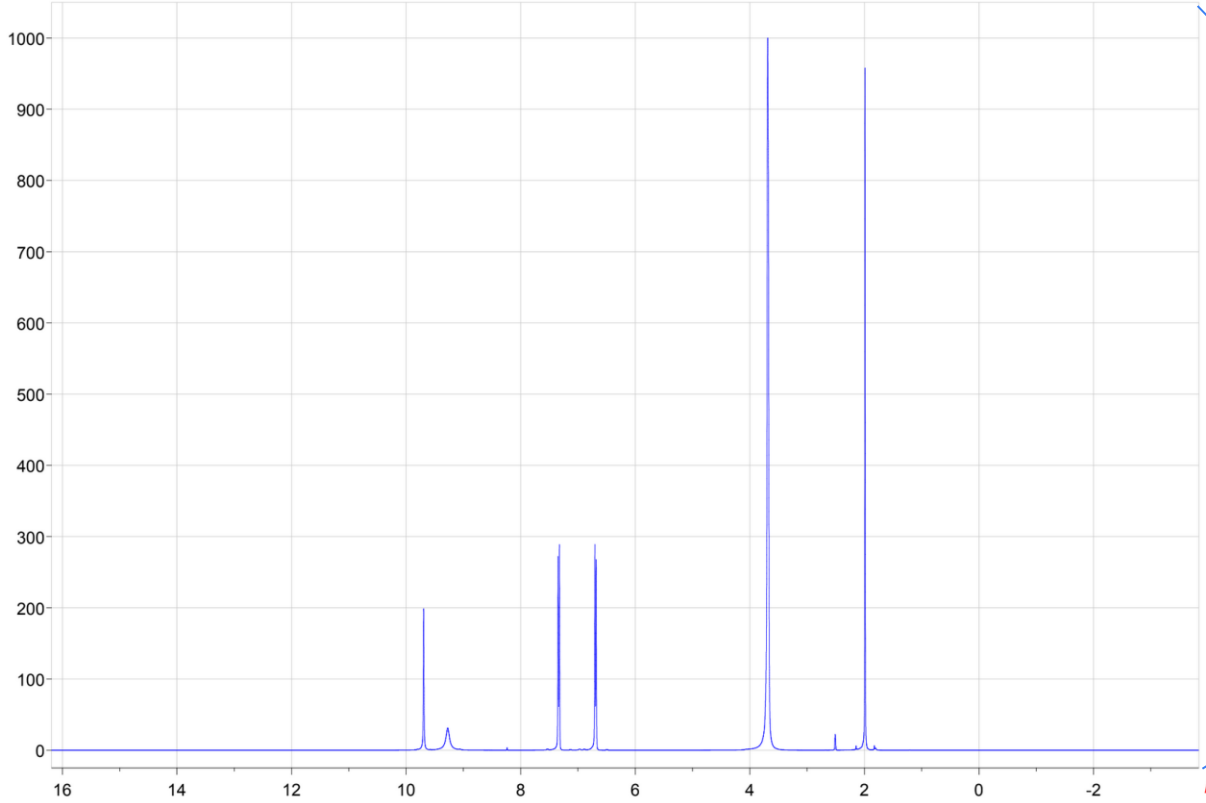
ACS National Meeting, Mar. 27, 2023

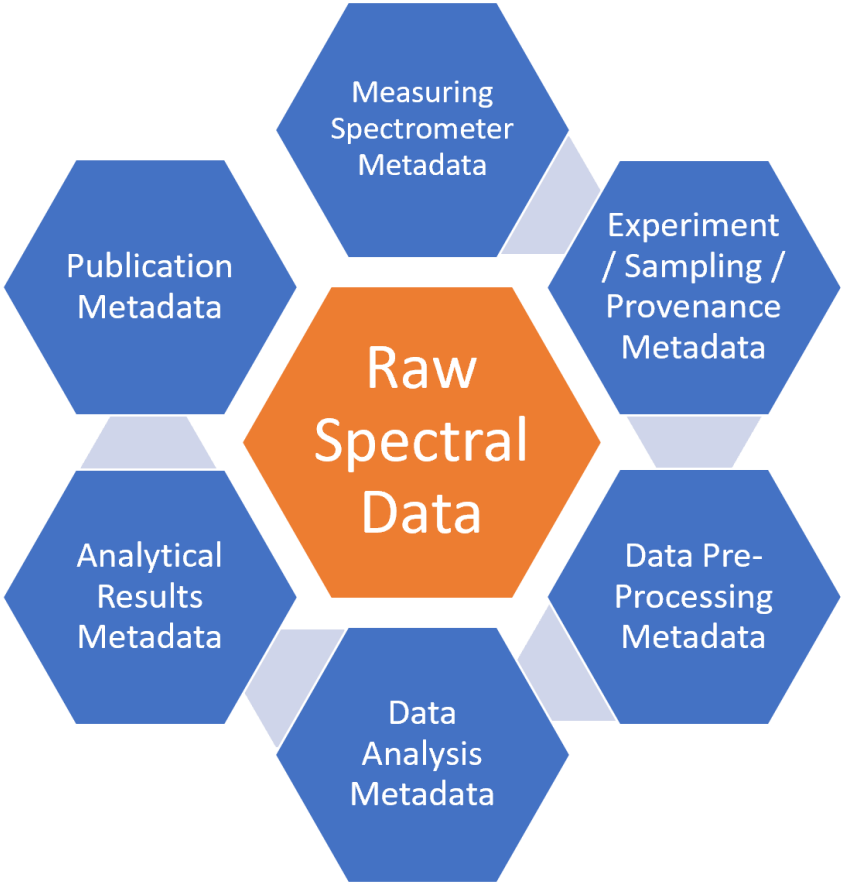
Robert M. Hanson, Mark Archibald, Ian Bruno, Stuart J. Chalk, Anthony N. Davies, Damien Jeannerat, Robert J. Lancashire, Jeff Lang, Henry S. Rzepa

IUPAC Project 2019-031-1-024

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

What’s wrong with this picture?

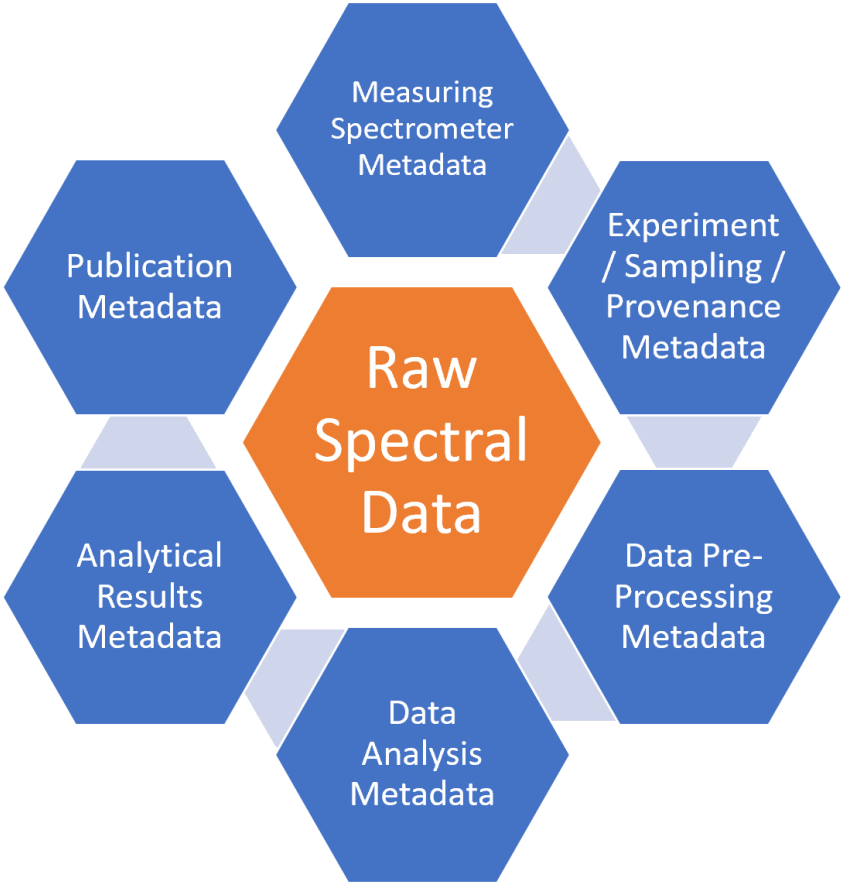




In an *Ideal FAIRSpec World*

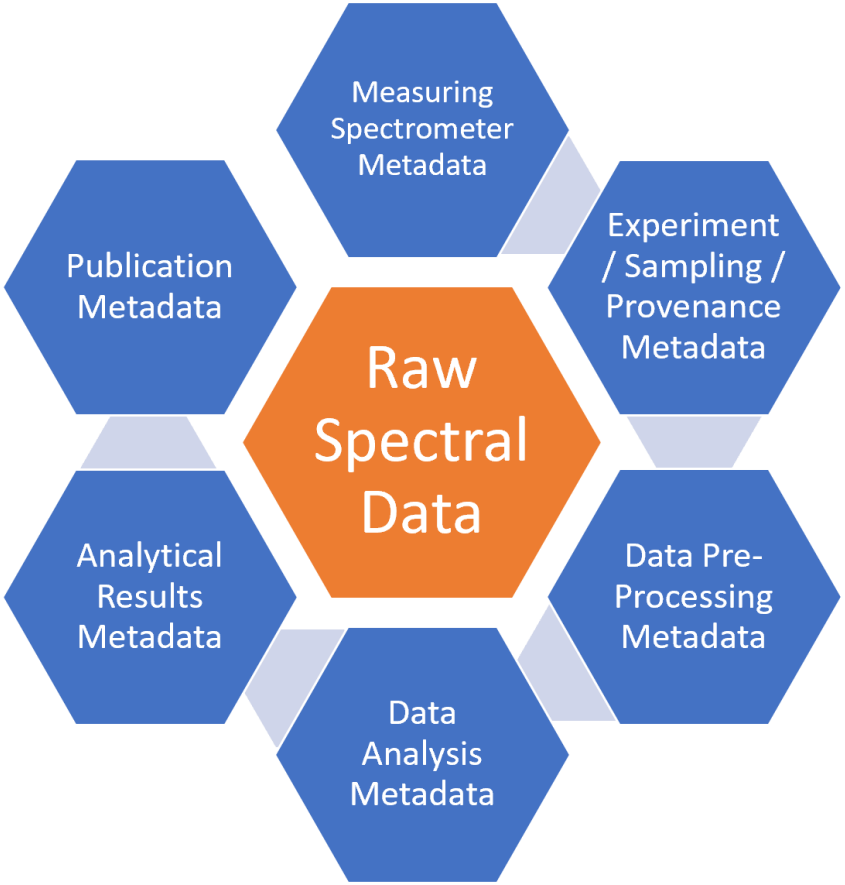
All metadata would be retained –

- Instrument metadata
- Dataset provenance
- Data preprocessing
- Post-acquisition processing
- Analysis
- Publication information
- More?



In an *Ideal FAIRSpec World*

And all this (and the data set itself!) would be *findable.*

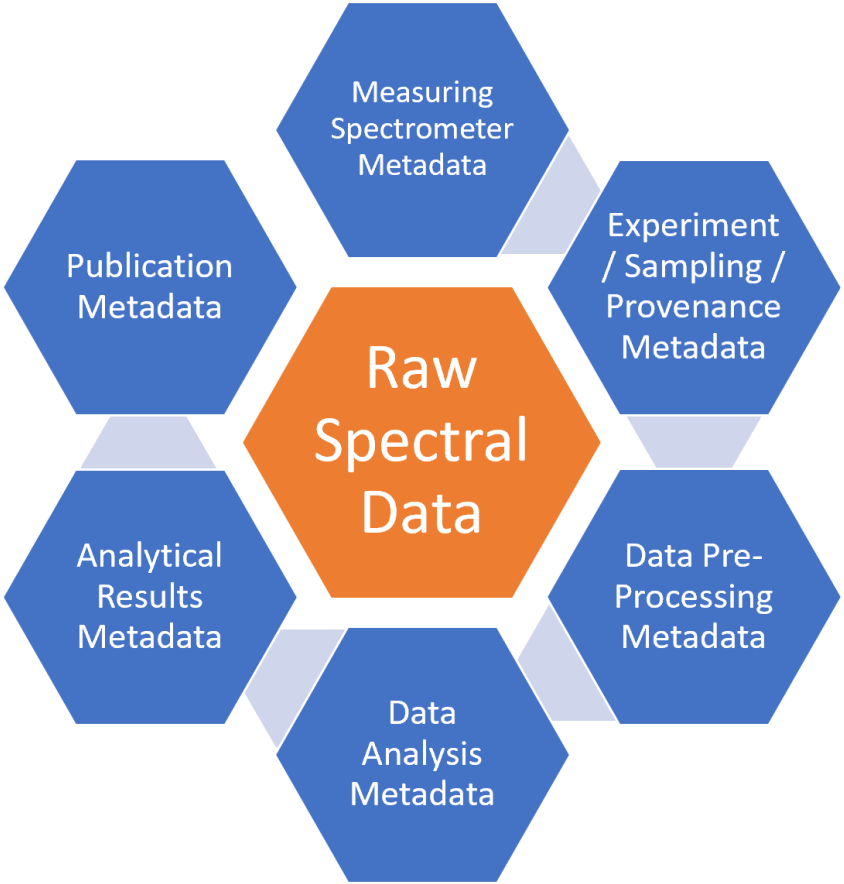


In an *Ideal FAIRSpec World*

And all this (and the data set itself!) would be *findable.*

By humans and machines.

FAIR == "Fully Artificial Intelligence Ready"

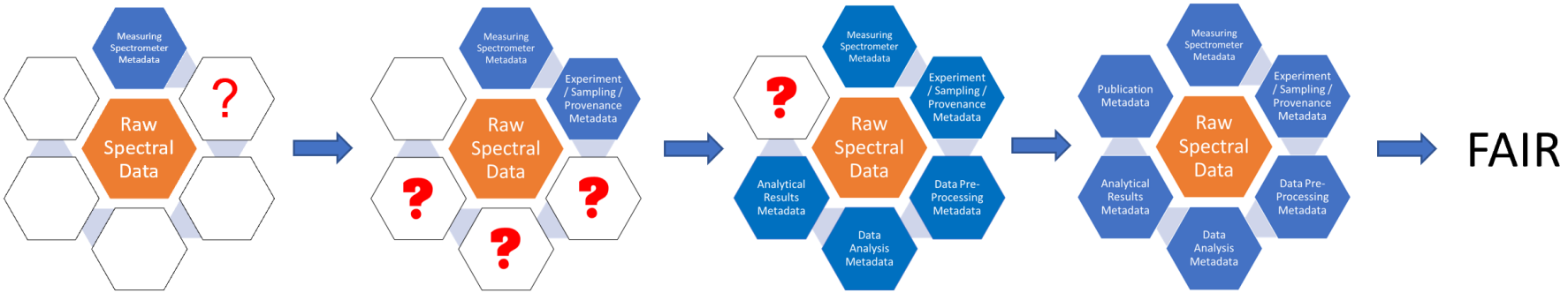


Just to be clear…

We are not here to talk about "FAIR data".

We are here to talk about the **FAIR *management* of data**.

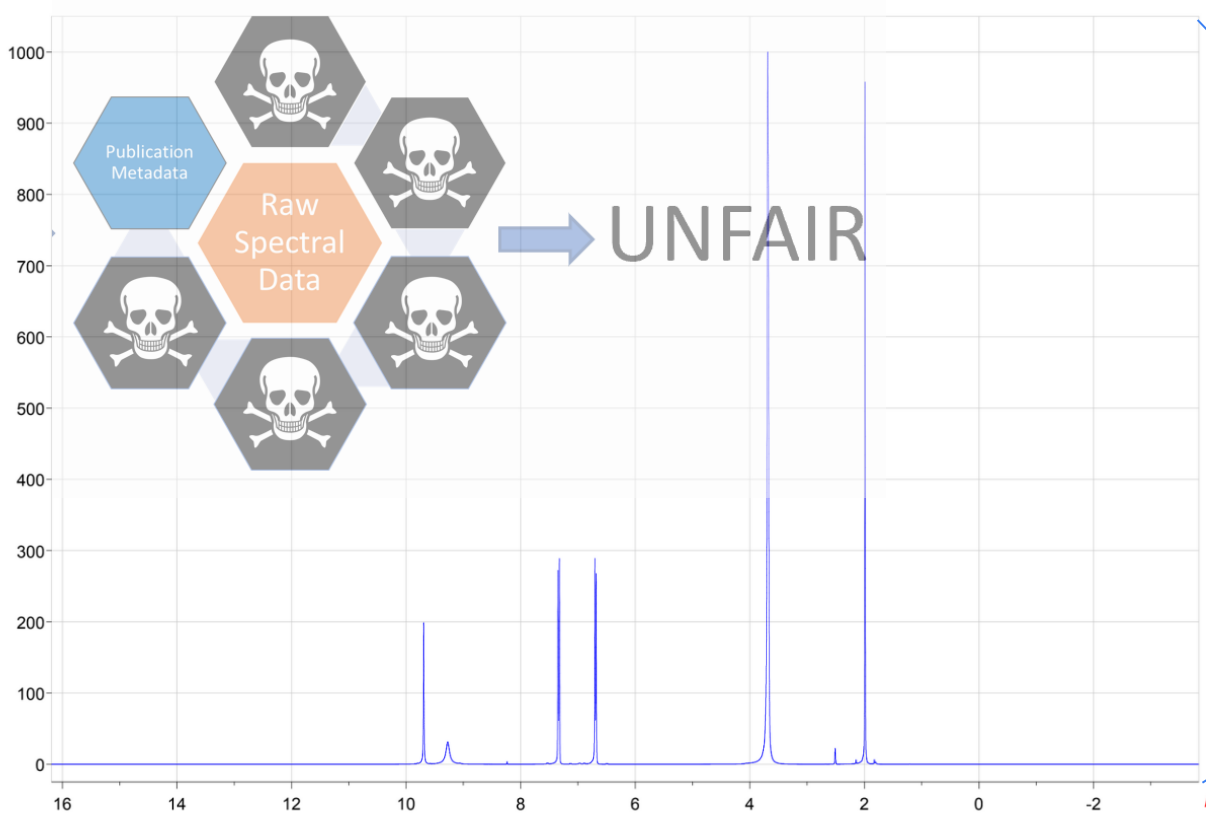
FAIR data management is a *continuous process*



Measure → Analyze → Report → Publish → Cite

Local → **Global**

What we have *here* is a recipe for data and metadata ***death!***



INTERNATIONAL UNION OF
PURE AND APPLIED CHEMISTRY

Bob Hanson

Damien Jeannerat

FAIRSpec PROJECT TEAM

IUPAC Project: 2019-031-1-024

Development of a Standard for FAIR Data Management of Spectroscopic Data

Mark Archibald

Ian Bruno

Stuart Chalk

Tony Davies

Robert Lancashire

Jeff Lang

Henry Rzepa

IUPAC INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

PROJECT DETAILS

DEVELOPMENT OF A STANDARD FOR FAIR DATA MANAGEMENT OF SPECTROSCOPIC DATA

Project No.:	2019-031-1-024
Start Date:	18 March 2020
End Date:	
Cite:	https://iupac.org/project/2019-031-1-024
Division Name:	Committee on Publications and Cheminformatics Data Standards

Objective

The objective of this project 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 standards 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.

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.

IUPAC FAIRSpec Principles

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.

What it means to be *FAIRSpec Ready:*

- Don’t wait until publication time to organize your data!
- Recognize the ongoing value of well-organized data.
- Find (or create!) the right tools for the job.
- Allow for corrections and addition of new information.

IUPAC FAIRSpec Principles

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.

What it means to be *FAIRSpec Ready:*

- Recognize context – a day's work, a project, a team effort.
- Associate spectra with chemical structure, if you can.
- Allow for ambiguity and reconsideration of these associations.
- Find ways to validate your structural and spectral analysis.

IUPAC FAIRSpec Principles

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.

What it means to be *FAIRSpec Ready:*

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

IUPAC FAIRSpec Principles

4. Metadata must be registered and standardized.

A. Register key metadata.
B. Assign a variety of persistent identifiers.
C. Enable metadata crosswalks.
D. Allow for value-added benefits.

What it means to be *FAIRSpec Ready:*

- Findability relies upon proper registration.
- This is not necessarily something you have to do yourself.
- Professionals in your organization will be involved.
- Your publisher will be involved.

IUPAC FAIRSpec Principles

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.

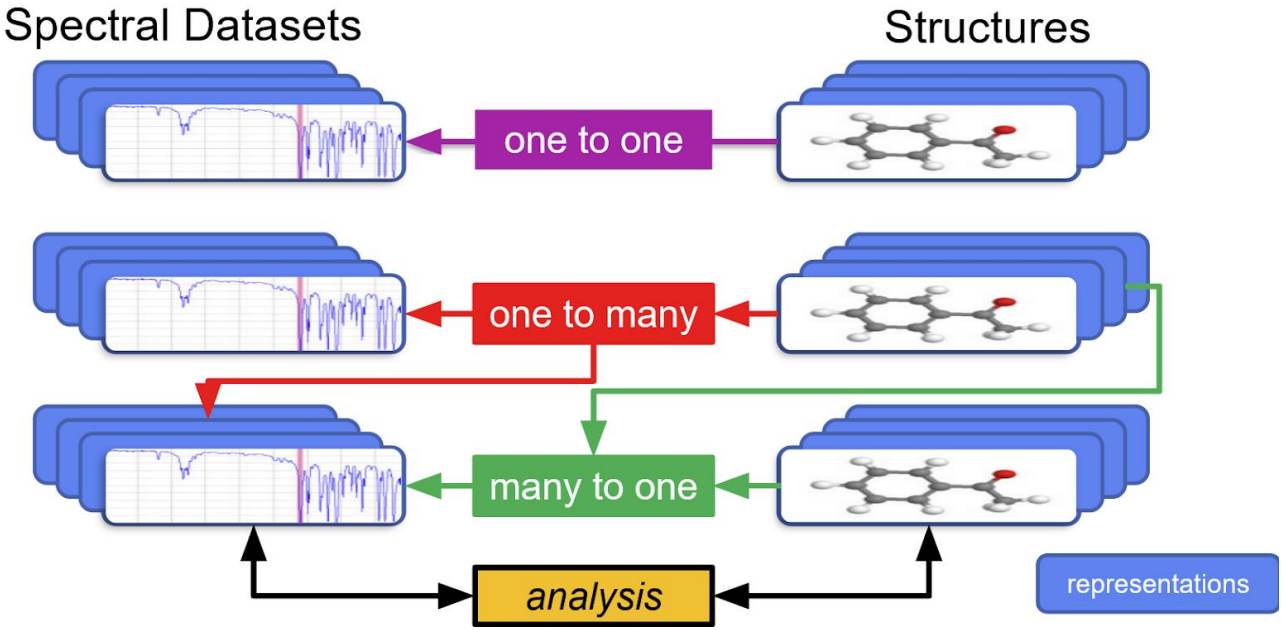
D. All data formats should be valued.

What it means to be *FAIRSpec Ready:*

- How can we make this as simple for you as possible?
- How can we make this useful to you *now?*
- We need your input!

Key Concept: Associations – Relational Metadata

One to One and One to Many FAIR Relationships



Two kinds of metadata

Descriptive metadata

- The solvent used
- The instrument manufacturer
- The type of the representation (raw data set, image, peak listing; MOL file, SMILES, InChI)

Relational metadata

- The associations among sample, structure, spectrum, and analysis within a collection
- The relation of this collection to other data collections
- The relation of this collection to other works (lab notebook, group project, publication)

Two kinds of metadata

Descriptive metadata

- The solvent used
- The instrument manufacturer
- The type of the representation (raw data set, image, peak listing; MOL file, SMILES, InChI)

Relational metadata

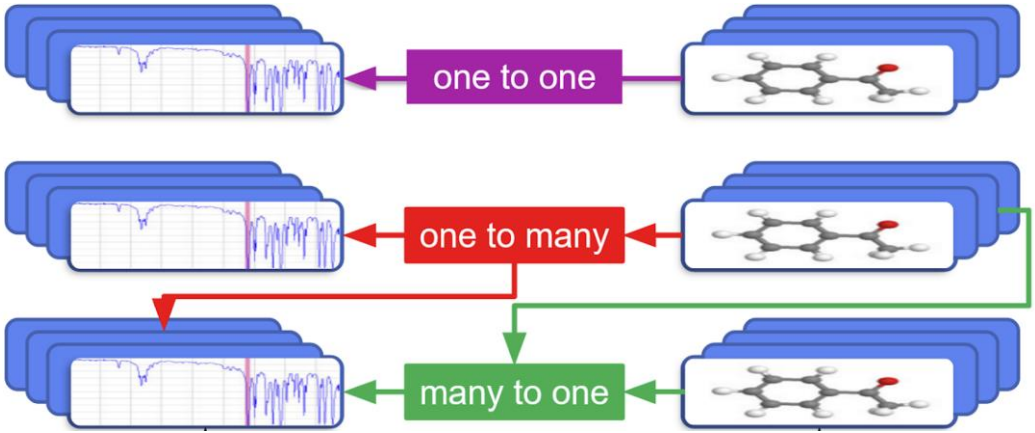
- The associations among sample, structure, spectrum, and analysis within a collection
- The relation of this collection to other data collections
- The relation of this collection to other works (lab notebook, group project, publication)

Descriptive metadata embedded in datasets and structure file formats is a starting point and can be easily “extracted” using automation.

Relational metadata is the real challenge.

We need to know...

- What compound (or at least sample) is associated with this spectrum?
- What (do you think) is the structure of that compound?

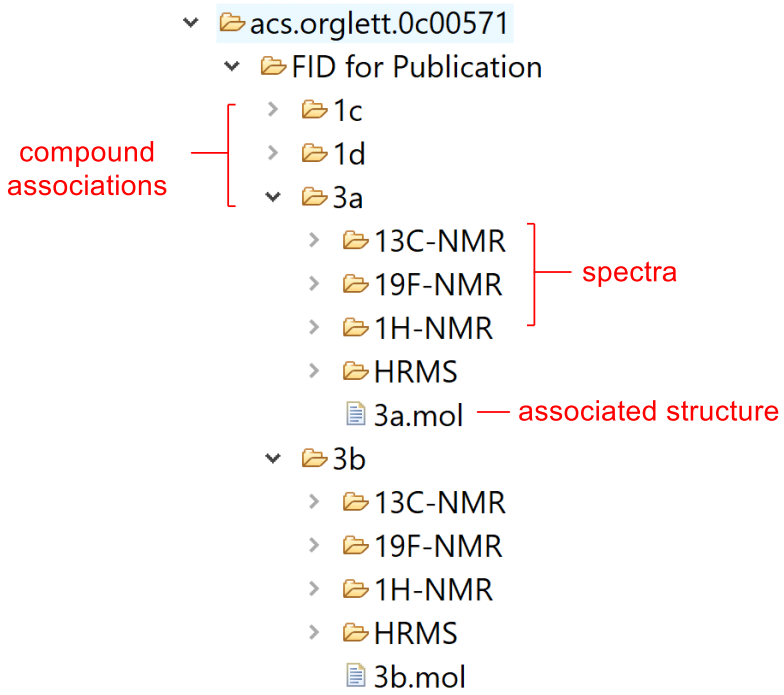


Curation workflow

The starting point is a well-organized ***FAIRSpec-Ready data aggregation***.

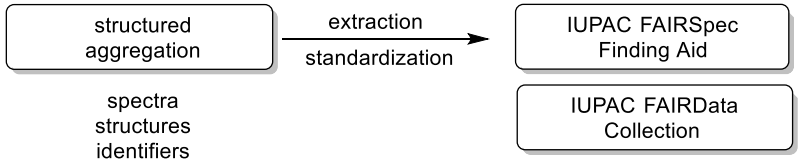
structured aggregation

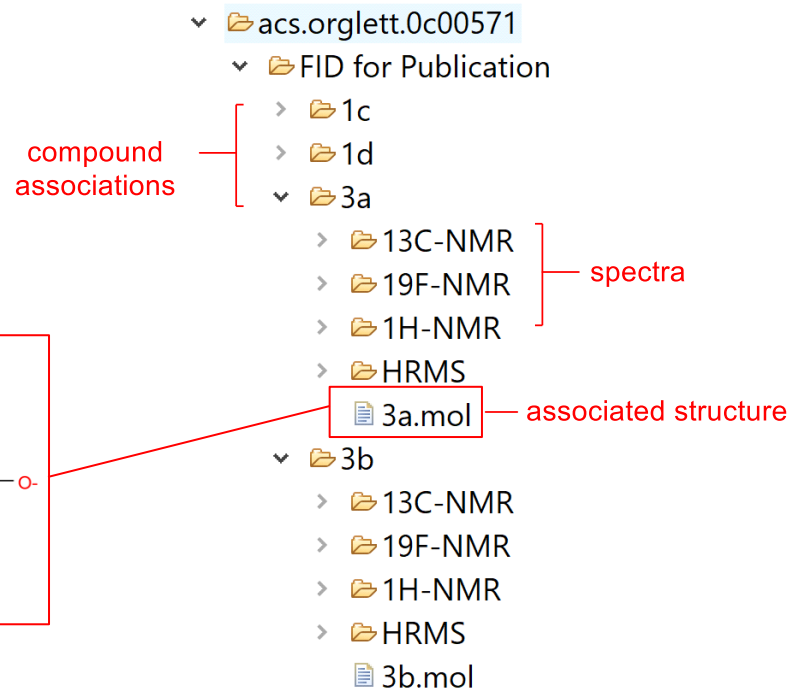
spectra
structures
identifiers



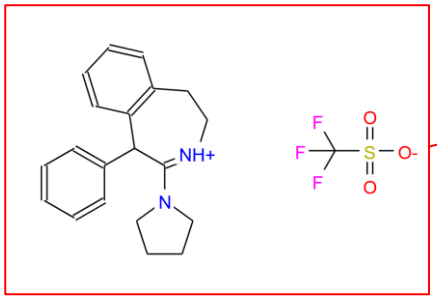
Curation workflow

Automated extraction gets us more structure representations…



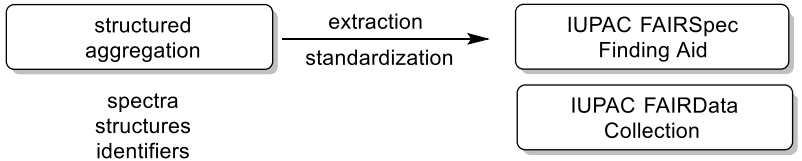


for example,
SMILES, **InChI**, and
PNG image
representations.



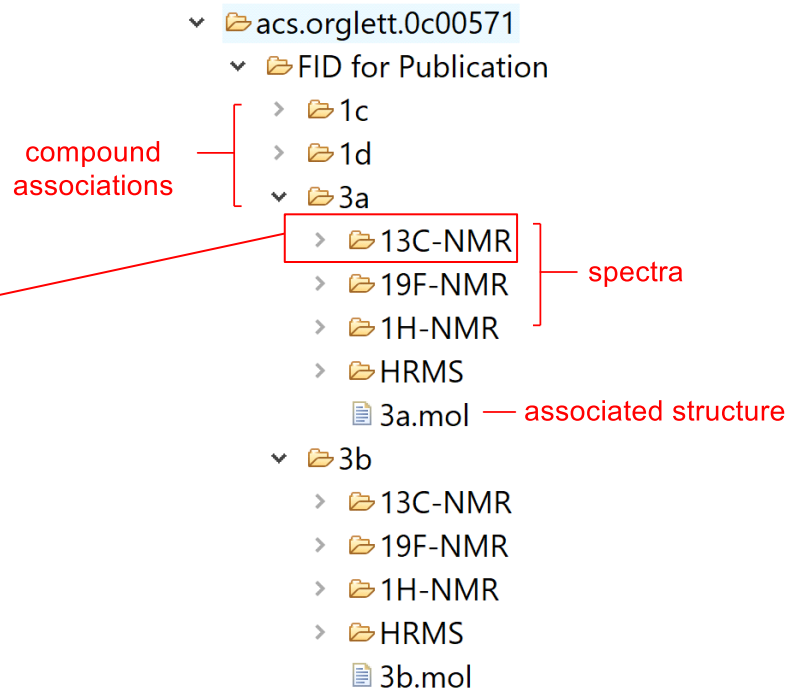
Curation workflow

…and pulls out metadata relating to the experiments…



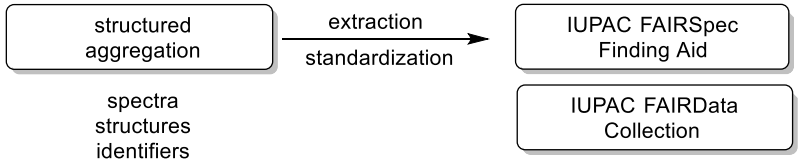
IFD Properties

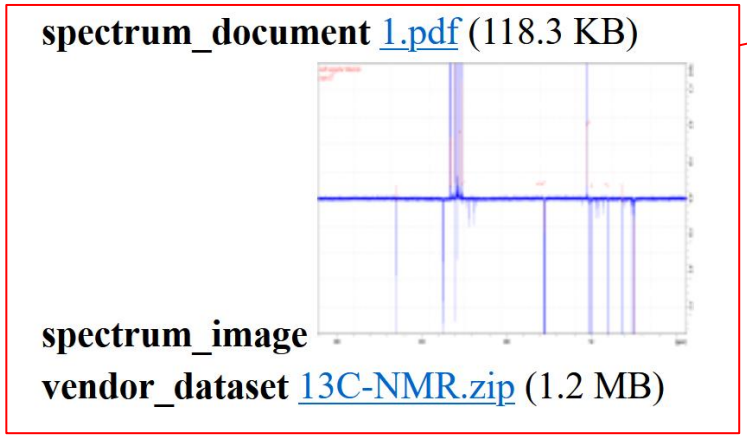
expt_absolute_temperature	298.1494
expt_dimension	1D
expt_nucl1	13C
expt_nucl2	1H
expt_pulse_prog	deptqgpsp
expt_solvent	CDCl3
expt_title	Auftraggeber Maulide hazh37
instr_manufacturer_name	Bruker
instr_nominal_freq	600
instr_probe_type	Z126545_0016 (CPP BBO 600S3 BB-H&F-D-05 Z)
proc_timestamp	2019-07-24T11:12:05+02:00

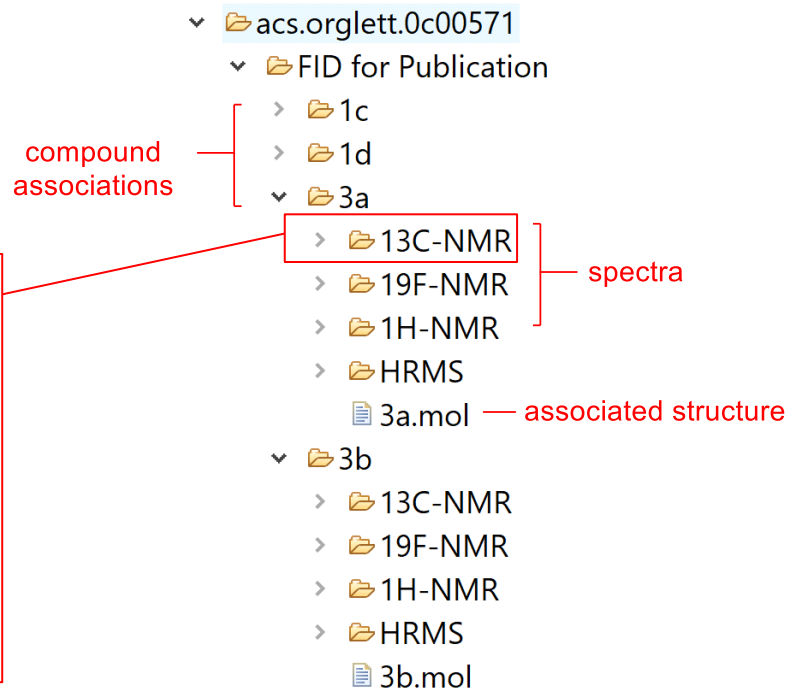


Curation workflow

…as well as multiple spectral data representations

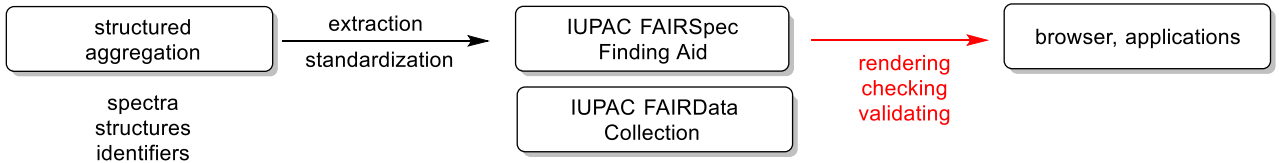






Curation workflow

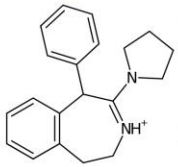
The Finding Aid can be read by stand-alone or browser-based applications



allowing for early-stage error correction as well as **local** structure, substructure, and spectroscopic metadata browsing.

Compound 3a

from SMILES:

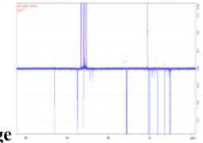


inchikey KZHKHOYBVCYSSO-UHFFFAOYSA-N
molecular_formula H 23 C 20 N 2

mol_2d 3a.mol (3 KB)
smiles c1cccc2c1.C21C2=[NH+1]CCc3c1cccc3.N24CCCC4
standard_inchi InChI=1S/C20H22N2.CHF3O3S
/c1-2-9-17(10-3-1)19-18-11-5-4-8-16(18)12-13-21-20(19)22-14-6-7-15-22;2-1(3,4)8(5,6)7
/h1-5,8-11,19H,6-7,12-15H2;(H,5,6,7)
fixedh_inchi InChI=1/C20H22N2.CHF3O3S
/c1-2-9-17(10-3-1)19-18-11-5-4-8-16(18)12-13-21-20(19)22-14-6-7-15-22;2-1(3,4)8(5,6)7
/h1-5,8-11,19H,6-7,12-15H2;(H,5,6,7)/fC20H23N2.CF3O3S/h21H;/q+1;-1

Spectra 3a/13C-NMR

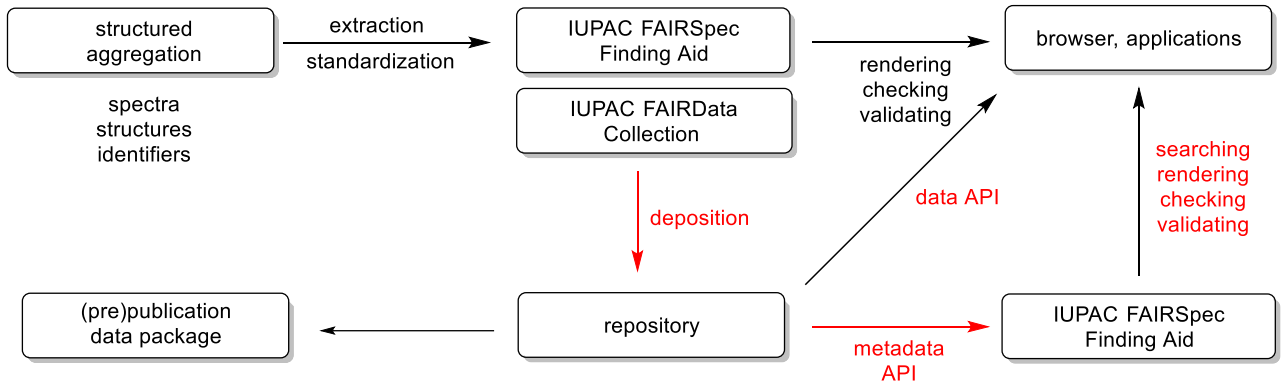
spectrum_document 1.pdf (118.3 KB)



spectrum_image
vendor_dataset 13C-NMR.zip (1.2 MB)

Curation workflow

A public version of the collection and its finding aid can be placed in a repository, given a persistent identifier, connected to a publication, and cited.



In Summary

IUPAC FAIRSpec standardization

- simplifies project management
- provides better continuity of group knowledge
- enables local or remote real-time structure-data validation
- allows distributed access to distributed data
- allows for private as for public collections
- enables standardization of communication between systems (ELN/LMS/repository deposition and query/author-publisher)
- provides more citation pointers to publications and other citable objects
- allows for "above the value line" innovation

Break-Out Room Discussion

- Focus on “FAIRSpec-Ready”
- Discuss these ideas in an informal format, with reference to the slides
- Interactively explore finding aids from the ACS pilot study
- Discuss problematic issues, concerns, complications
- Brainstorm on possible funding possibilities for future implementation

We’re here to listen!!

Thank you!