Progress toward developing an IUPAC FAIR standard for spectroscopic data description & management

ACS National Meeting, April 14, 2021

Robert M. Hanson, Damien Jeannerat, Mark Archibald,
Ian Bruno, Stuart J. Chalk, Antony N. Davies,
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

Today’s presentation

1. The Vision
2. The Problem
3. The Task
4. The Project
5. Progress to Date
    - Six Key Design Decisions
    - Three Preliminary Experiments
6. Going Forward

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

Today’s presentation

1. The Vision
2. The Problem
3. The Task
4. The Project
5. Progress to Date
    - Six Key Design Decisions
    - Three Preliminary Experiments
6. Going Forward

The FAIRSpec Vision

To enable a world where we can all...

- draw a structure or substructure and find **all published spectra** - of a given type - related to that compound/fragment, filtered by data format; quality; journal; author; date, or other common characteristics
- quickly find linked data **associated with those spectra**

The FAIRSpec Vision

To enable a world where we could...

- validate assignments prior to publication
- submit "raw" (lossless) spectral data with publications, generating various forms of data representation
- implement direct "ELN-to-publish" systems
- automatically add to/harvest spectral data for AI-based global spectroscopic analysis projects

The FAIRSpec Vision

- spectra will be found based on key aspects of the data and methods (e.g NMR frequency & nuclei, IR method)
- spectra will be found by standard compound identifiers (structure, substructure, SMILES, InChI, etc.)
- smart methods of rendering spectral information associated with journal publications will be possible
- new technologies will be built based upon the standards

Today’s presentation

1. The Vision
2. The Problem
3. The Task
4. The Project
5. Progress to Date
    - ○ Six Key Design Decisions
    - ○ Three Preliminary Experiments
6. Going Forward

The Problem

- Too much reliance on published PDF “supplemental information” without concern for interoperability
- No central community-based effort to archive and make available spectroscopic data
- No standards for describing or relating that data to chemical structure

Today’s presentation

1. The Vision
2. The Problem
3. The Task
4. The Project
5. Progress to Date
   - ○ Six Key Design Decisions
   - ○ Three Preliminary Experiments
6. Going Forward

The Task – To do what IUPAC does best

- Develop a standard vocabulary and structure in the area of chemistry
- Enable others to implement area-specific value-added services
- Enable services to work together using a shared set of data descriptors and protocols

Today’s presentation

1. The Vision
2. The Problem
3. The Task
4. The Project
5. Progress to Date
   - ○ Six Key Design Decisions
   - ○ Three Preliminary Experiments
6. Going Forward

The Project

- (2018-19) IUPAC Committee on Publications and Cheminformatics Data Standards (CPCDS) sponsored two FAIR data workshops (Amsterdam and Orlando)
- (March 2020) Initiation of IUPAC Project 2019-031-1-024 two-year time frame
- first year **design**; second year **build**

Today’s presentation

1. The Vision
2. The Problem
3. The Task
4. The Project
5. Progress to Date
    - ○ Six Key Design Decisions
    - ○ Three Preliminary Experiments
6. Going Forward

Design Decision #1

This project is not about the creation of any new file formats.

- not a new NMR-STAR or nmrML
- not an extension to JCAMP-DX
- not about repackaging binary FID data into new a "standard" format

Design Decision #2

We will not limit ourselves to one specific spectroscopic technique.

- start with a focus on NMR because of its significance
- provide a framework for inclusion of other techniques

Design Decision #3

We recognize four key pieces of the puzzle.

- the spectroscopic data itself
- associated chemical identifier/structure-related (meta)data
- associated structure-spectrum analysis (meta)data
- associated general key/value pair metadata
  (authors, associated DOIs, provenance, licenses, etc.)

Design Decision #4

We recognize the importance of multiple representations.

- drawing from successes in earth science and archival science
- varieties of spectroscopic data representations
- key aspects of acceptable chemical identifiers and structure formats

Design Decision #5

*We recognize the importance of a **collection** and its associated **finding aid**.*

- drawing specifically from archival science
- an isolated manufacturer data set has no intrinsic value
- connection to an appropriate chemical identifier is critical
- connection to related spectra and compounds is valuable
- key element is a structured finding aid

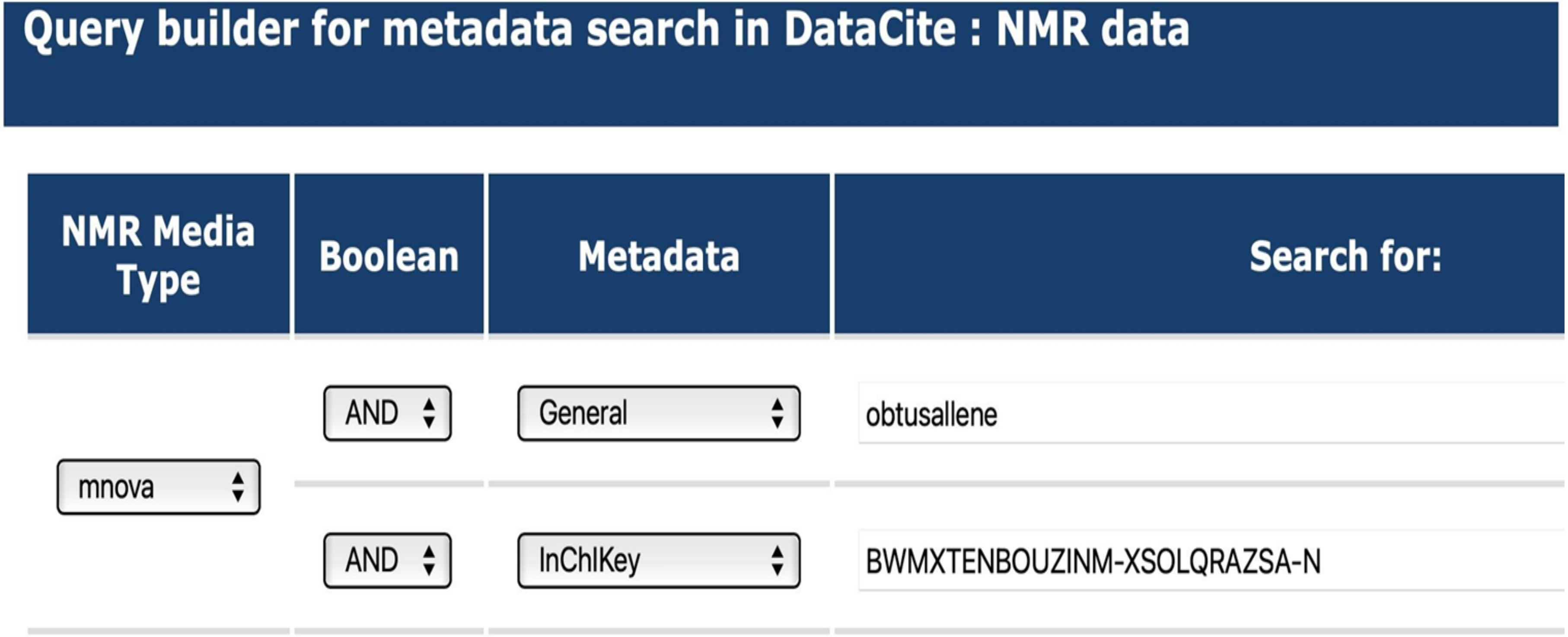
Design Decision #6

We will work closely with known (meta)data managers and other stakeholders, ensuring that whatever we do is mappable to their metadata as much as possible.

- publishers and authors (ACS, RSC)
- repository and database managers (HMDB, BMRB, NP-MRD, NMRShiftDB, nmrdb)
- chemical information services (PubChem)

Preliminary Experiment #1

NMR metadata registered with DataCite



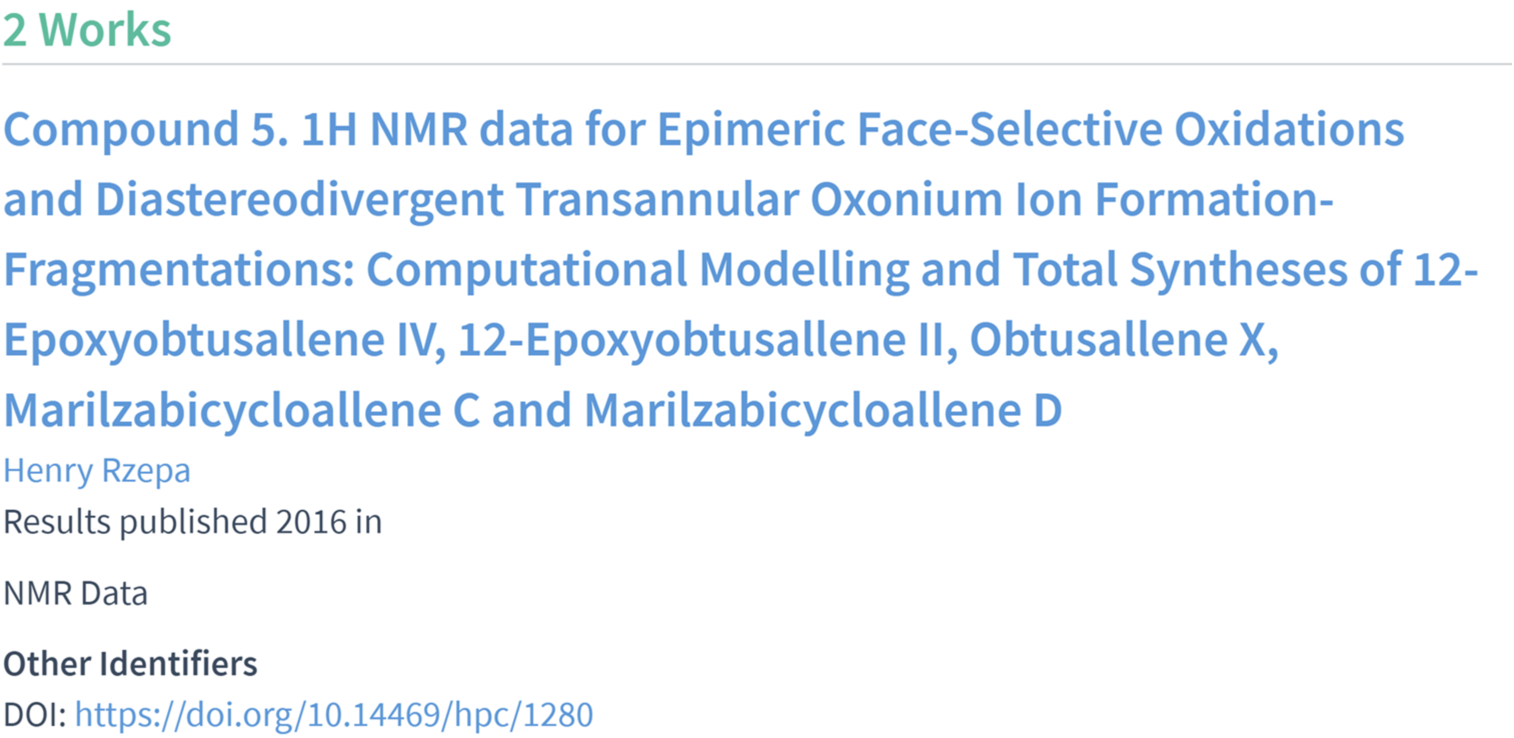
Query builder for metadata search in DataCite : NMR data

NMR Media Type	Boolean	Metadata	Search for:
mnova	AND	General	obtusallene
	AND	InChIKey	BWMXTENBOUZINM-XSOLQRAZSA-N

https://doi.org/f357

Preliminary Experiment #1

NMR metadata registered with DataCite



2 Works

Compound 5. 1H NMR data for Epimeric Face-Selective Oxidations and Diastereodivergent Transannular Oxonium Ion Formation-Fragmentations: Computational Modelling and Total Syntheses of 12-Epoxyobtusallene IV, 12-Epoxyobtusallene II, Obtusallene X, Marilzabicycloallene C and Marilzabicycloallene D

Henry Rzepa

Results published 2016 in

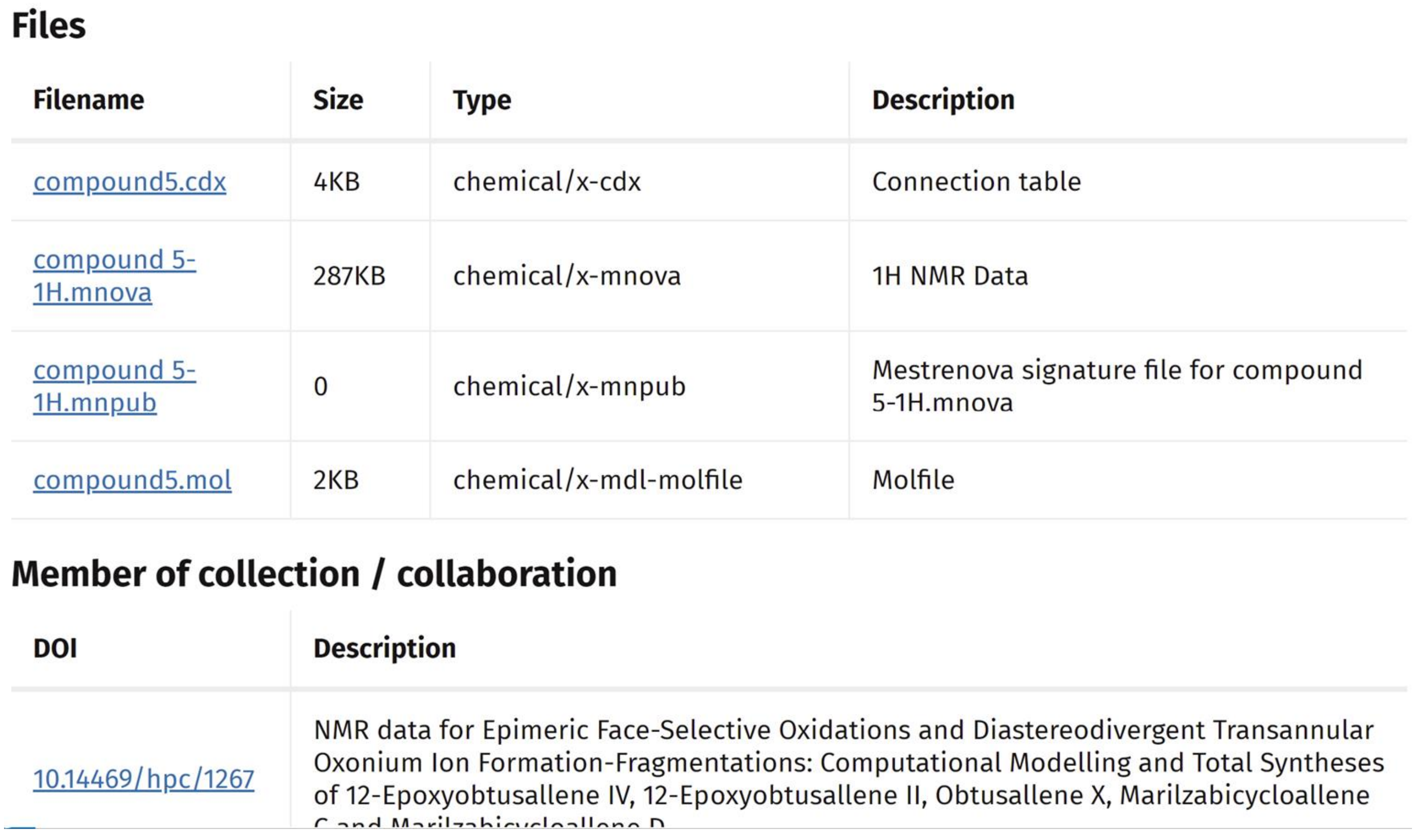
NMR Data

Other Identifiers

DOI: https://doi.org/10.14469/hpc/1280

Preliminary Experiment #1

NMR metadata registered with DataCite



Files

Filename	Size	Type	Description
compound5.cdx	4KB	chemical/x-cdx	Connection table
compound 5-1H.mnova	287KB	chemical/x-mnova	1H NMR Data
compound 5-1H.mnpub	0	chemical/x-mnpub	Mestrenova signature file for compound 5-1H.mnova
compound5.mol	2KB	chemical/x-mdl-molfile	Molfile

Member of collection / collaboration

DOI	Description
10.14469/hpc/1267	NMR data for Epimeric Face-Selective Oxidations and Diastereodivergent Transannular Oxonium Ion Formation-Fragmentations: Computational Modelling and Total Syntheses of 12-Epoxyobtusallene IV, 12-Epoxyobtusallene II, Obtusallene X, Marilzabicycloallene C and Marilzabicycloallene D

Preliminary Experiment #1

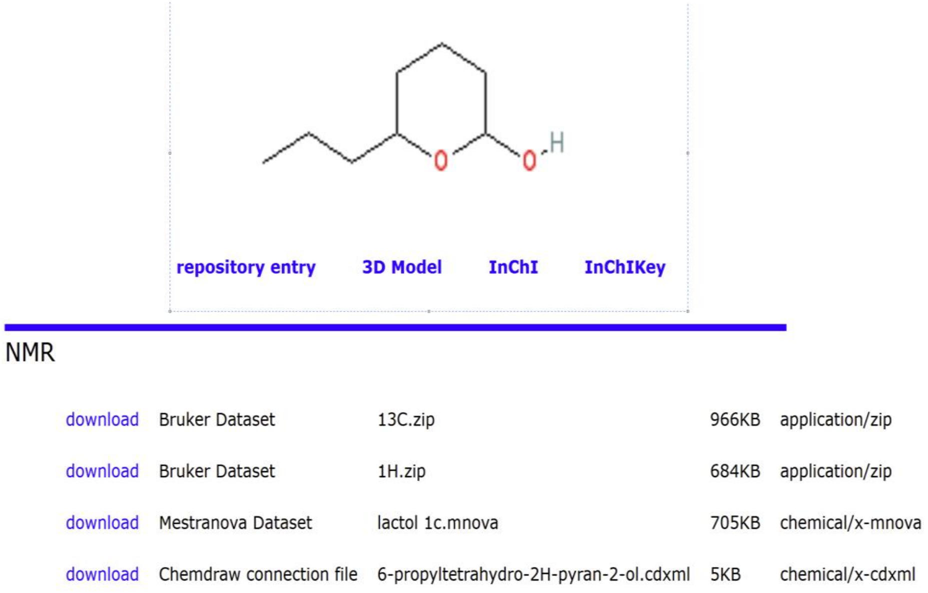
NMR metadata registered with DataCite

Conclusions:

- registration can work within the DataCite system
- institutional repositories need to use PIDs (persistent identifiers) at whatever granularity is desired
- key “deliverables” include landing pages, spectral data packages, and compound information

Preliminary Experiment #2
XML Finding Aids

- Hand-generated a crude finding aid for a paper.
- Used the Library of Congress EAD XML format and simple XML style sheet, creating structure drawings on the fly.



NMR

download	Bruker Dataset	13C.zip	966KB	application/zip
download	Bruker Dataset	1H.zip	684KB	application/zip
download	Mestranova Dataset	lactol 1c.mnova	705KB	chemical/x-mnova
download	Chemdraw connection file	6-propyltetrahydro-2H-pyran-2-ol.cdxml	5KB	chemical/x-cdxml

https://chemapps.stolaf.edu/test/fairspec/sample/example1/findingaid.xml

Preliminary Experiment #2

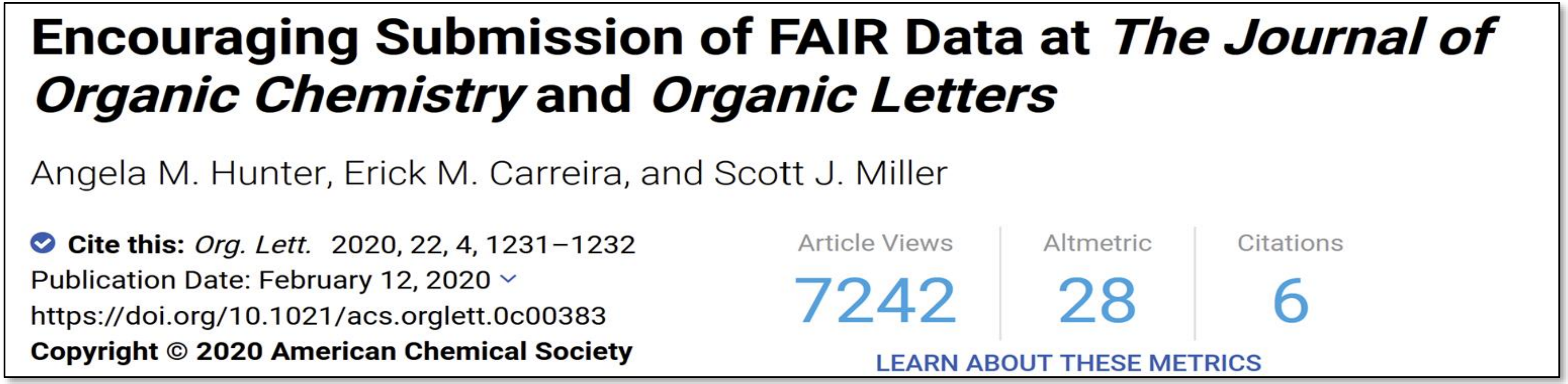
XML Finding Aids

Conclusions:

- XML + structured style sheet could work for implementation
- Was possible to link SMILES to on-demand structure representations created by other services
- EAD is a well thought out archival structure
- Good example of nested collections of related “data”
- Good example of how to map schemas

Preliminary Experiment #3

ACS Publications FAIR Data Pilot



Encouraging Submission of FAIR Data at *The Journal of Organic Chemistry* and *Organic Letters*

Angela M. Hunter, Erick M. Carreira, and Scott J. Miller

Cite this: *Org. Lett.* 2020, 22, 4, 1231−1232
Publication Date: February 12, 2020
https://doi.org/10.1021/acs.orglett.0c00383
Copyright © 2020 American Chemical Society

Article Views	Altmetric	Citations
7242	28	6

LEARN ABOUT THESE METRICS

- Authors submitted data as supporting information
- Over 200 submissions to date
- 13 submissions unpacked at St. Olaf College and analyzed

Preliminary Experiment #3 ACS Pubs FAIR Data

ACS Collection	Size (MB)		digital entities	
	(zip)	(raw)	files	type
joc.0c00770	25	37	720	11 cmpd dirs; 24 Bruker datasets & 12 .mnova files
orglett.0c00874	27	40	1616	36 cmpd dirs; 76 Bruker datasets
orglett.0c00967	29	41	1354	33 cmpd dirs; 62 Bruker datasets
orglett.0c01022	15	52	66	2 dirs; 64 .mnova files
orglett.0c01197	79	101	61	2 dirs; 59 .mnova files
orglett.0c01277	52	74	2463	63 cmpd dirs; 124 Bruker datasets
orglett.0c01297	57	73	1544	29 cmpd dirs; 58 Bruker datasets

Preliminary Experiment #3 ACS Pubs FAIR Data

Observations:

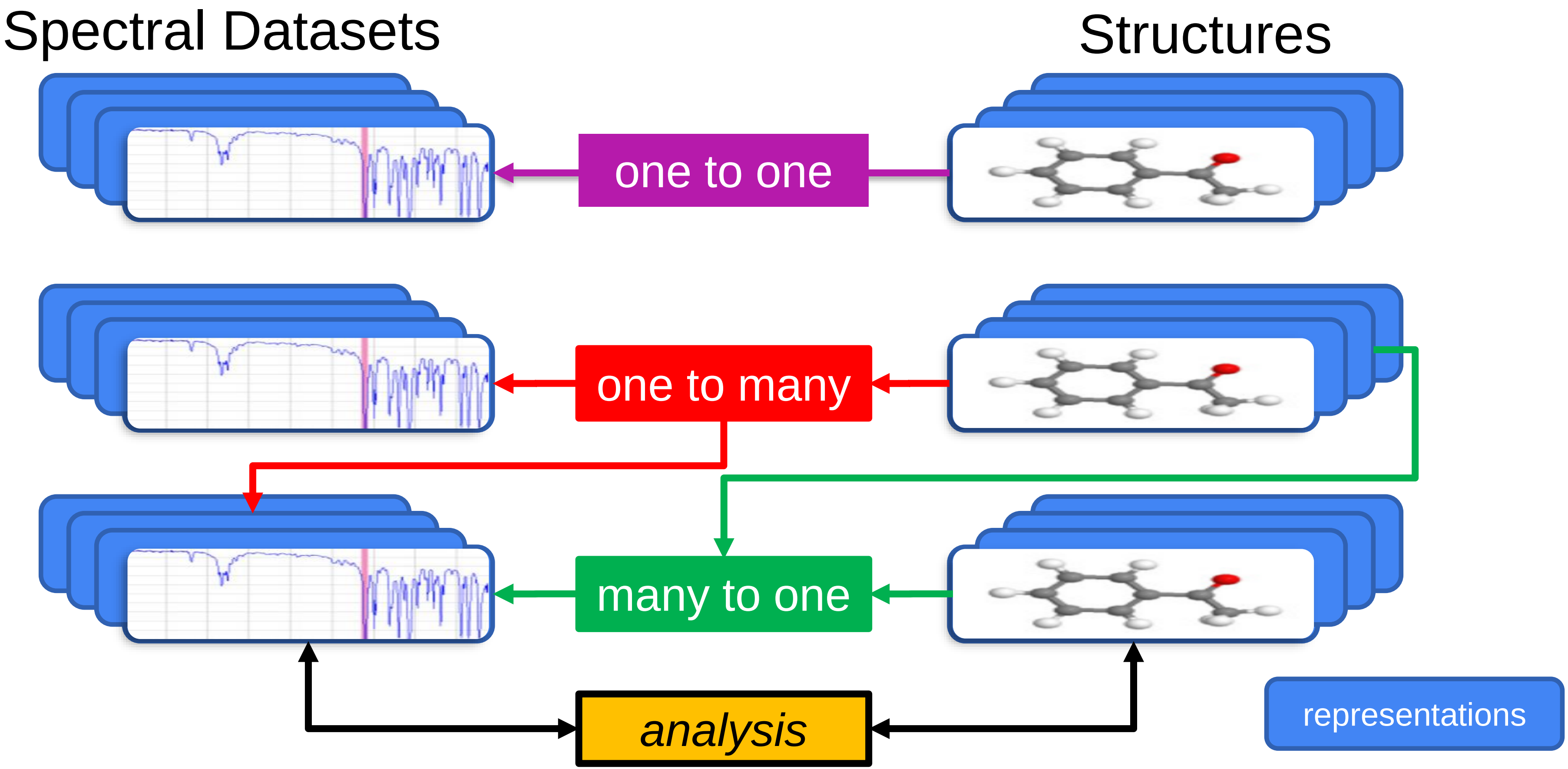
- authors are interested - demand is there
- one-to-one and one-to-many (structure-to-spectrum)
- only one author included structural representations
- proprietary formats only (no long-term stable JCAMP-DX)
- no analyses (mnova?)

Preliminary Experiment #3 ACS Pubs FAIR Data

Conclusions:

- “data representation” and “structure representation”
- implementation workflow will be critical here
- essential to connect one or more chemical identifiers with one or more NMR datasets
- valuable to have at least a minimum analysis (e.g. the "journal description")

One to One and One to Many FAIR Relationships



Levels of NMR data reusability

data representations and reuse level		possible processing		viewing and analysis facilitated (* with additional processing)				
		full	near-full	interactive	enhanced viewing	non-interactive viewing	visual comparison	machine comparison
raw data (FID + parameters)	10	yes	yes	yes*	yes*	yes*	yes*	yes*
minimally processed data, (r+i spectra)	9		yes	yes*	yes*	yes*	yes*	yes*
fully processed data (real spectrum)	8			yes	yes*	yes*	yes*	yes*

Levels of NMR data reusability

data representations and reuse level		possible processing		viewing and analysis facilitated (* with additional processing)				
		full	near-full	interactive	enhanced viewing	non-interactive viewing	visual comparison	machine comparison
peak tables, shifts, integration, and splitting	7			yes*	yes*	yes*	yes	yes
PDF	6				yes	yes	yes	
journal-style description	5				yes*	yes*	yes	yes#
image (PNG)	4					yes	yes	
peak table -- shifts only	3						yes	yes#

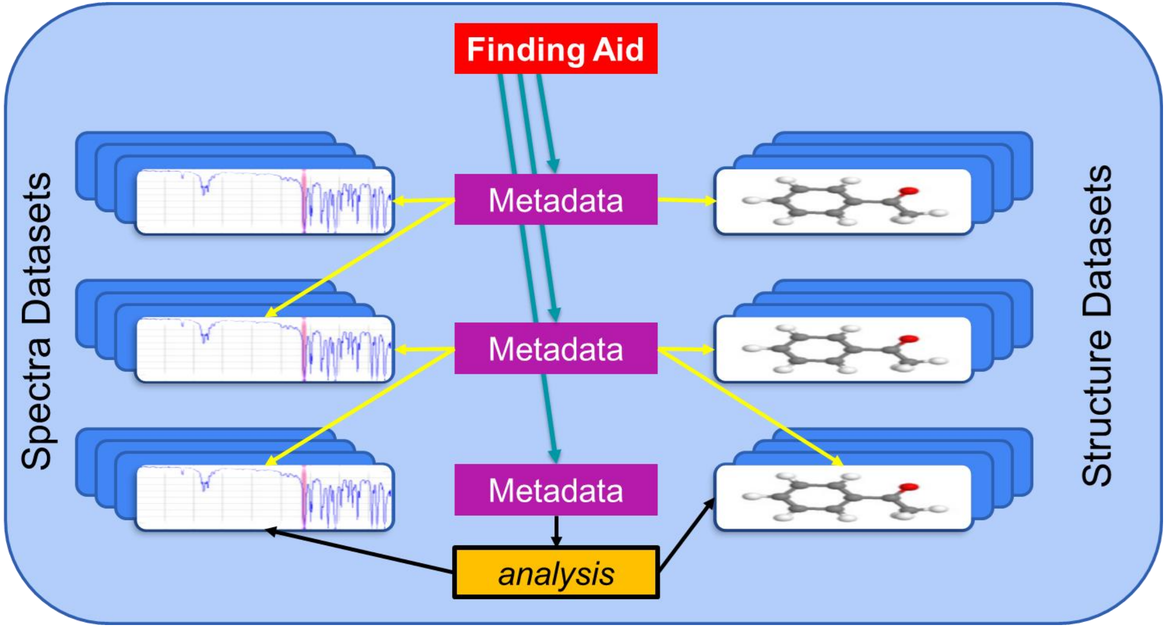
# to some extent (involves lossiness, human error or bias)

Today’s presentation

1. The Vision
2. The Problem
3. The Task
4. The Project
5. Progress to Date
    - ○ Six Key Design Decisions
    - ○ Three Preliminary Experiments
6. Going Forward

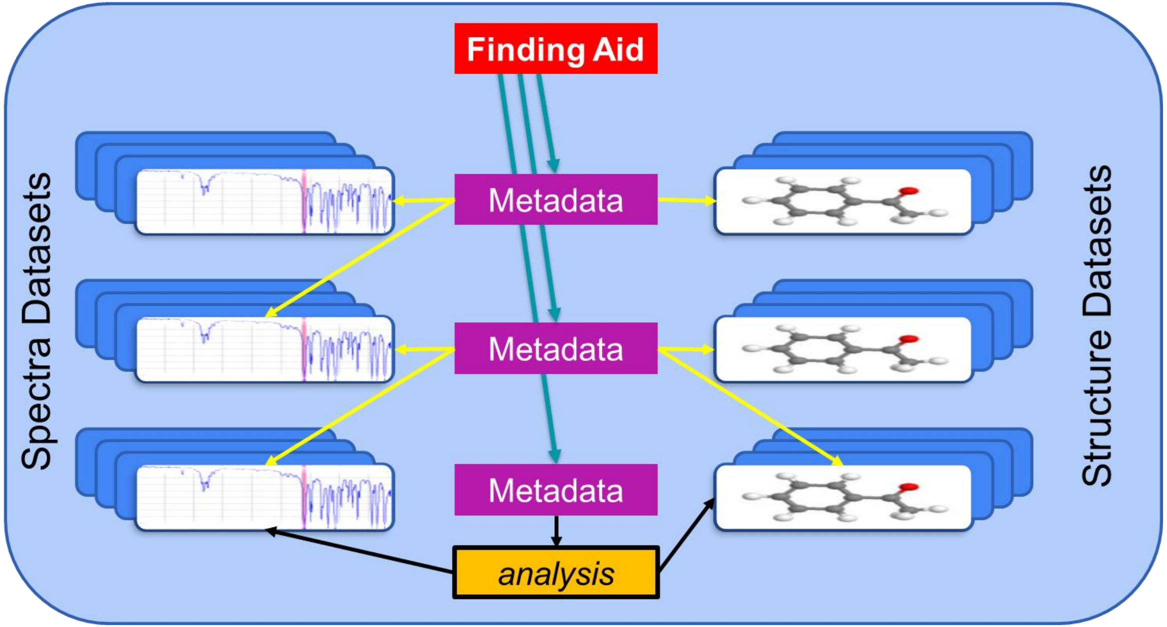
Summary

We believe we have a basic outline of the issues.



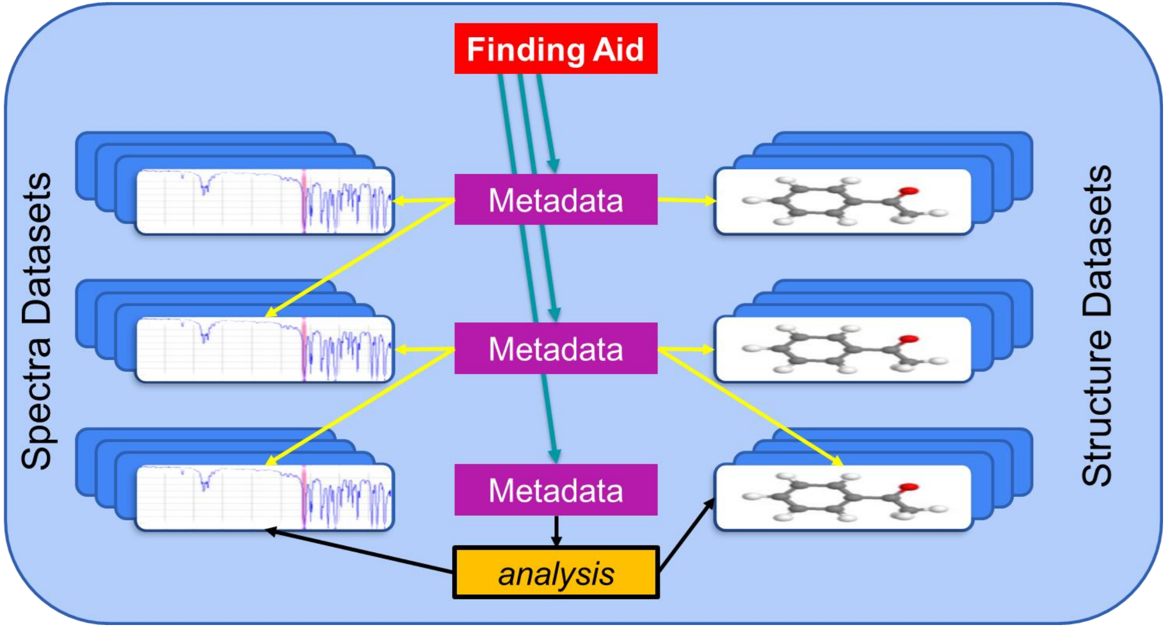
Summary

The task now is to develop realistic metadata standards that can be accepted and widely implemented.



Summary

We have identified stakeholders and are starting to work with them.



Going Forward

Focus on the task at hand

Build a set of metadata specifications that:

1. describes multiple spectroscopic **data representations**,
2. describes **structure and analysis representations** relating to that data, and
3. describes the **contents of a spectroscopic data collection**

Keeping in mind that it must:

1. connect all of this using **standardized mappable metadata**;
2. allow for **selective retrieval** of a variety of spectral data representations, structural models, and analyses; and
3. allow for **metadata to be managed independently** from the data itself

Going Forward

Open/Good Questions

1. How does one distinguish data from *meta*data? (Is that important?)
2. Are there examples of any of this already out there?
3. What about predicted vs. experimental vs. simulated spectral “data”?
4. What about experimental data manipulation? Hacked data? Deep fakes?
5. Data validation? AI ideas?
6. Community efforts? Funding?

Going Forward
How you can help

1. Identify yourself as an interested party – join the discussion
2. Express an interest in collaborating – be and early adopter
3. Work along side us to set up a reference implementation
4. Suggest additional stakeholders
5. Comment and suggest – issues and solutions (please!)

hansonr@stolaf.edu damien.jeannerat@protonmail.com

Appendix

supplemental slides follow

References

1. ACS FAIR Data Pilot https://pubs.acs.org/doi/abs/10.1021/acs.joc.0c00248
2. Biomagnetic Resonance Data Bank (BRMB) https://bmrb.io/
3. “FAIR Enough?” Spectroscopy Europe (16 Mar 2021) https://www.spectroscopyeurope.com/td-column/fair-enough
4. Human Metabolome Database (HMDB) https://hmdb.ca
5. Library of Congress Encoded Archival Description (EAD) https://www.loc.gov/ead
6. NASA Earth Science Data Processing Levels https://earthdata.nasa.gov/collaborate/open-data-services-and-software/data-information-policy/data-levels
7. Natural Products Magnetic Resonance Database (NP-MRD) http://www.npmrd-project.org
8. NMRDB https://www.nmrdb.org/
9. NMReData https://nmredata.org/
10. NMR Markup Language/Controlled Vocabulary (nmrML/nmrCV) http://nmrml.org/
11. NMRShiftDB https://nmrshiftdb.nmr.uni-koeln.de/
12. NMR-STAR https://link.springer.com/article/10.1007/s10858-018-0220-3
13. PubChem https://pubchem.ncbi.nlm.nih.gov
14. Research Data Aliance Data Foundation Terminology (RDA-DFT) https://www.rd-alliance.org/system/files/DFT%20Core%20Terms-and%20model-v1-6.pdf

Some key definitions

“data” vs. “metadata”

“one person’s data is another person’s metadata”

(we are probably not going to agree on this)

Whether a phased NMR spectrum is data – or is just the FID?

Whether a spectral analysis of any sort is data?

…moving on…

Mike Taylor, 2004 http://www.miketaylor.org.uk/tech/metadata.html

Some key definitions

RDA Terminology for Digital “Entities" vs “Objects"

Digital Entity (DE) – Anything that can be represented by a sequences of 0s and 1s.

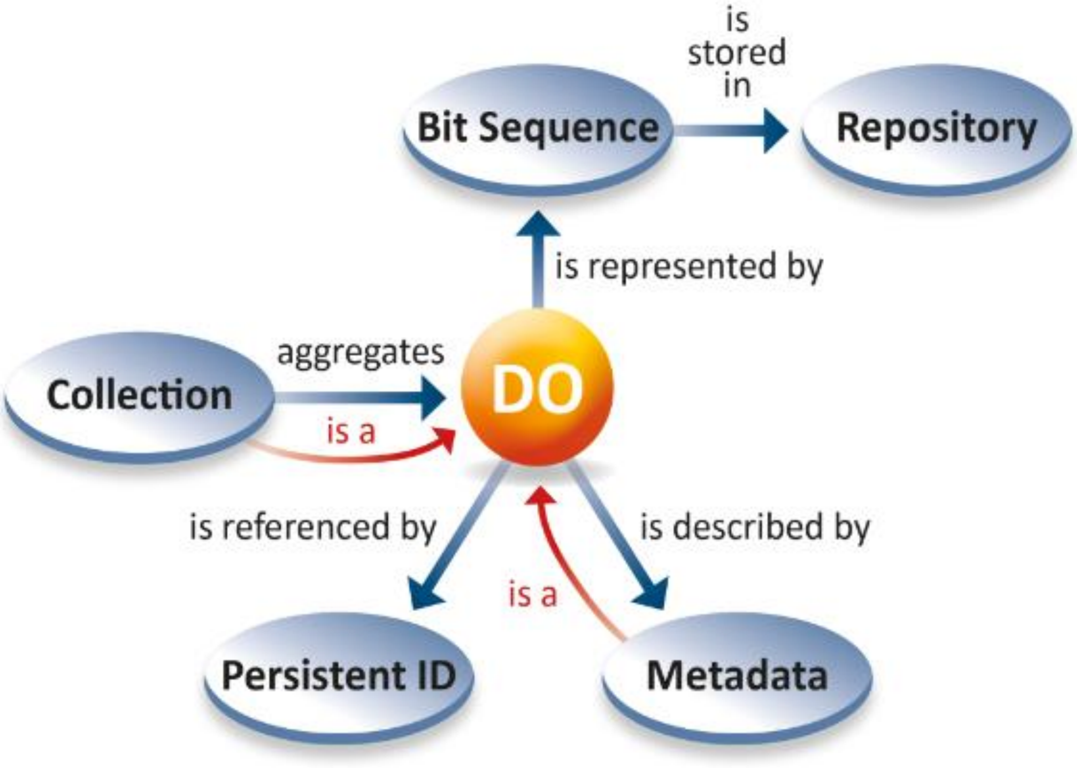
Digital Object (DO) – A structured **DE** that is named with associated attributes that can be used to reference it.

Digital Aggregation – A bundle of **DE**s.

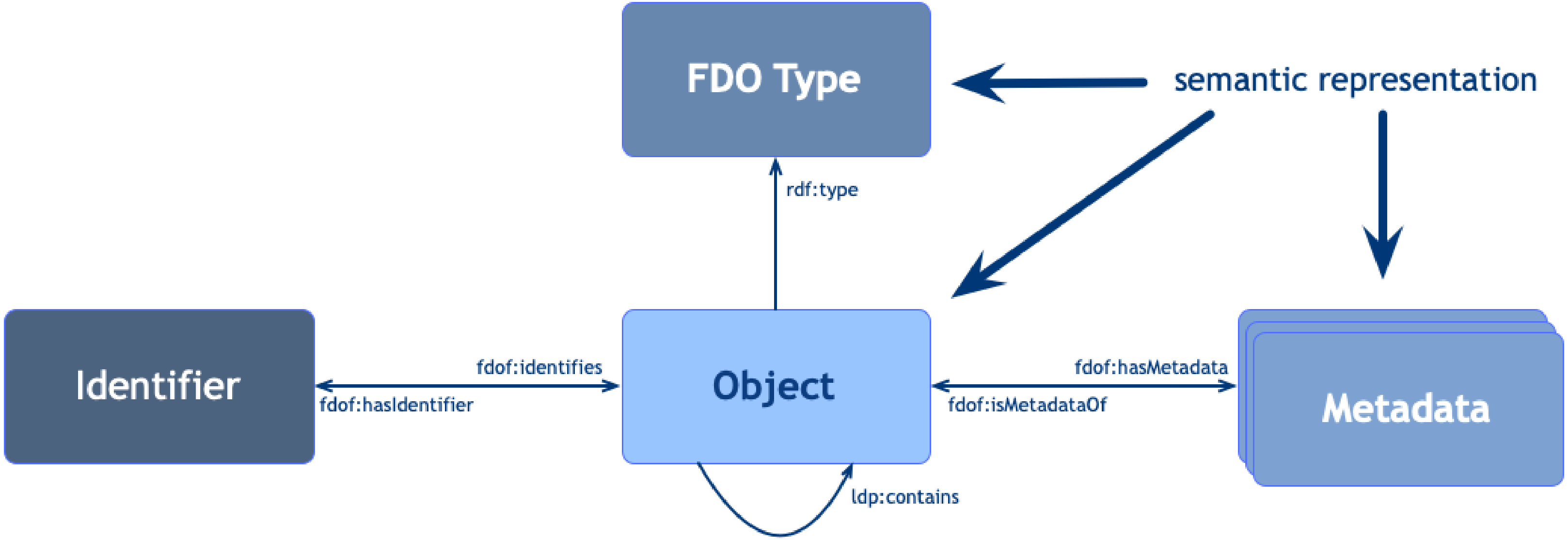
Digital Collection – An aggregation which contains **DO**s and **DE**s identified by a PID and described by metadata.

https://www.rd-alliance.org/system/files/DFT%20Core%20Terms-and%20model-v1-6.pdf

Terminology - Digital "entities" and "objects"

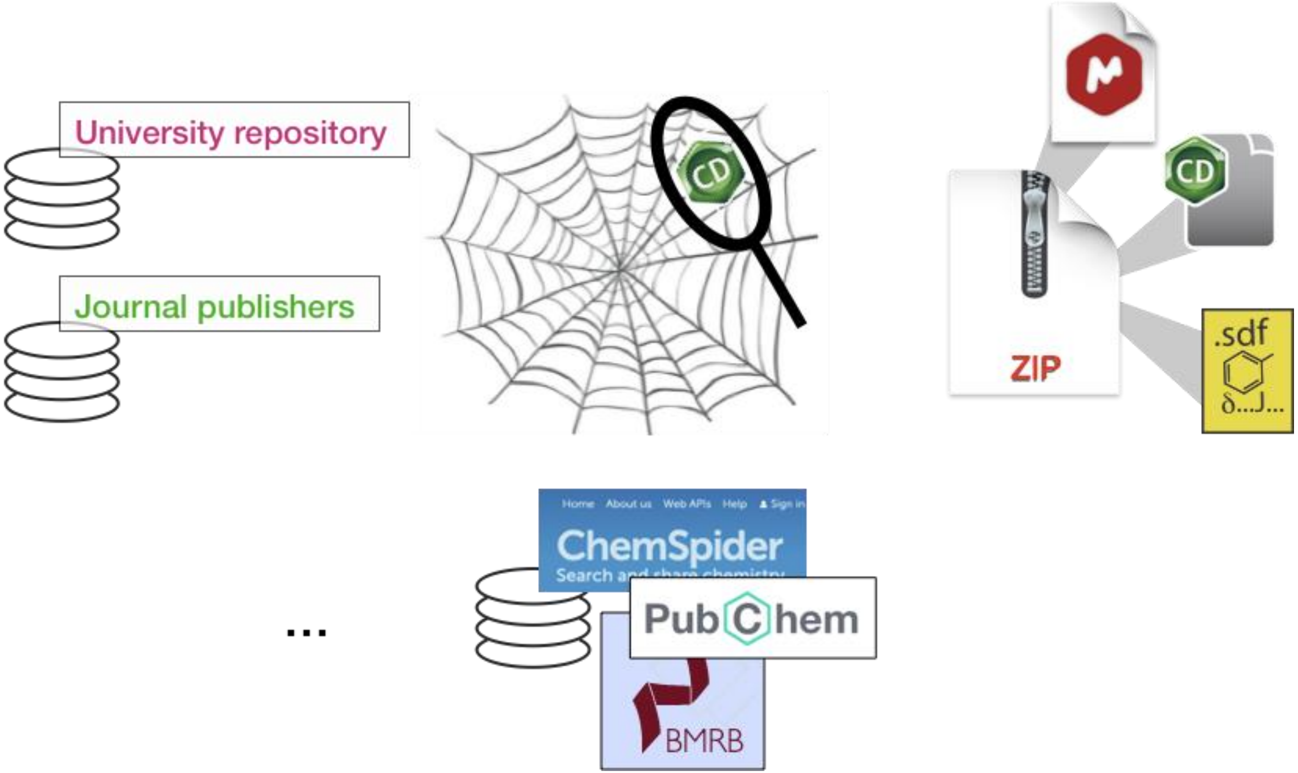


FAIR Digital Object Representation



Metadata is essential for the findability of the data

How can specialized databases find relevant data?



University repository

Journal publishers

...

CD
ZIP
CD
.sdf
δ...J...
Home About us Web APIs Help Sign in
ChemSpider
Search and share chemistry
PubChem
BMRB

More quotes on (meta)data

http://statmath.wu.ac.at/courses/data-analysis/itdtHTML/node54.html

7.3 Metadata

There are many confusing definitions of metadata, not helped by the fact that one man's data is another man's metadata. We will use a very simple definition by exclusion.

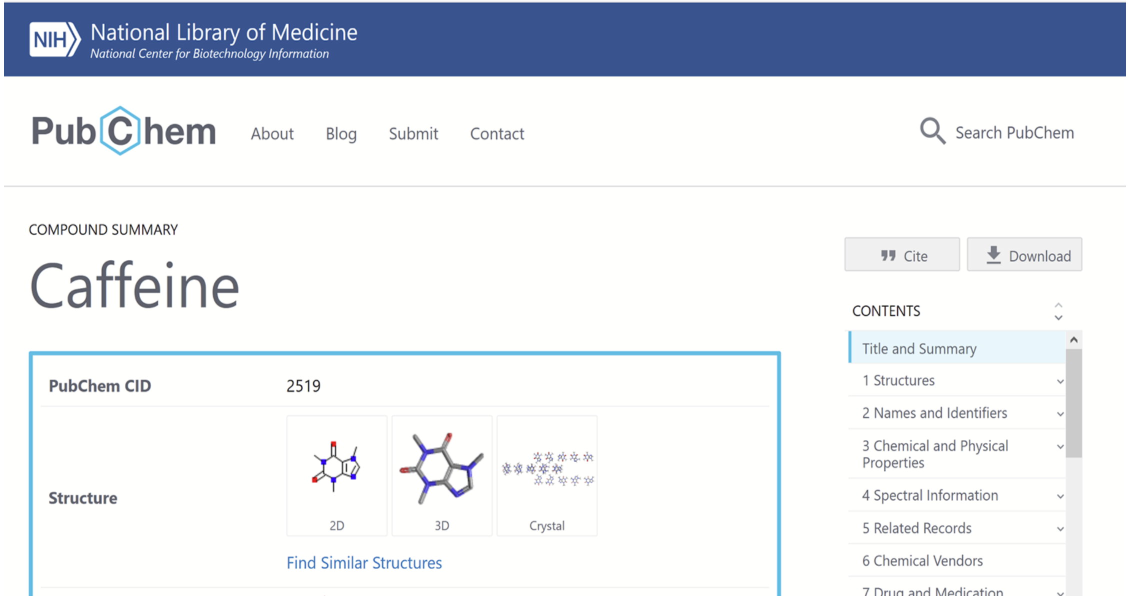
> Metadata is everything other than the raw data that we will analyse.

This definition runs the risk of being too broad, but there is a greater danger from too little metadata being stored and shared rather than too much, so anything that might accidentally lead to an over abundance of metadata is a risk worth taking.

Examples – Current Practice
PubChem

A PubChem landing page can be thought of as a sort of digital finding aid.

Extracts key metadata of (presumed) interest and points to more resources.



COMPOUND SUMMARY

Caffeine

PubChem CID	2519
Structure	

https://pubchem.ncbi.nlm.nih.gov/compound/2519

Examples – Current Practice

PubChem

A PubChem landing page can be thought of as a sort of digital finding aid.

Extracts key metadata of (presumed) interest and points to more resources.

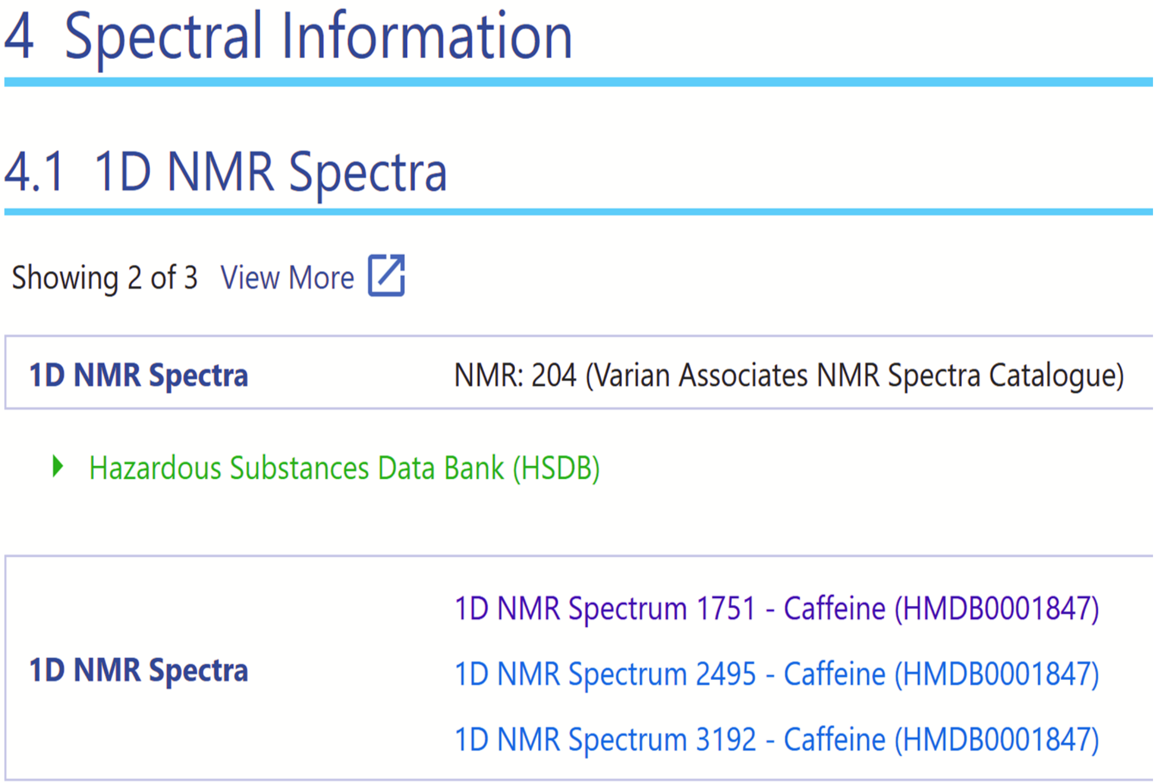
1. Structures	**11. Identification**
2. Names and Identifiers	**12. Safety and Hazards**
3. Chemical and Physical Properties	**13. Toxicity**
4. Spectral Information	**14. Associated Disorders and Diseases**
5. Related Records	**15. Literature**
6. Chemical Vendors	**16. Patents**
7. Drug and Medication Information	**17. Biomolecular Interactions and Pathways**
8. Food Additives and Ingredients	**18. Biological Test Results**
9. Pharmacology and Biochemistry	**19. Classification**
10. Use and Manufacturing	**20. Information Sources**

Examples – Current Practice

PubChem

…and points to more resources…

and portals…



4 Spectral Information

4.1 1D NMR Spectra

Showing 2 of 3 View More

1D NMR Spectra	NMR: 204 (Varian Associates NMR Spectra Catalogue)

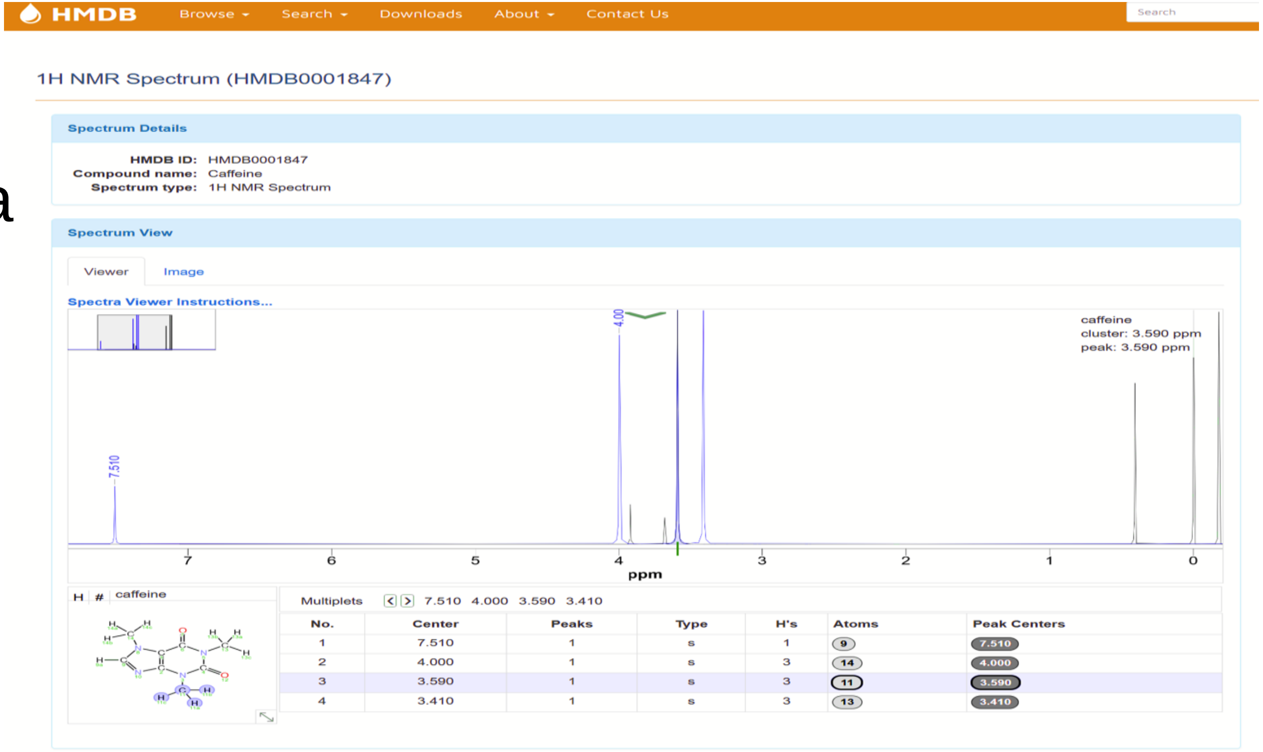
▸ Hazardous Substances Data Bank (HSDB)

1D NMR Spectra	1D NMR Spectrum 1751 - Caffeine (HMDB0001847) 1D NMR Spectrum 2495 - Caffeine (HMDB0001847) 1D NMR Spectrum 3192 - Caffeine (HMDB0001847)

Examples – Current Practice

Human Metabolome Database

...where we can explore the data visually...



1H NMR Spectrum (HMDB0001847)

Spectrum Details

HMDB ID: HMDB0001847
Compound name: Caffeine
Spectrum type: 1H NMR Spectrum

Spectrum View

Spectra Viewer Instructions...

caffeine
cluster: 3.590 ppm
peak: 3.590 ppm

Multiplets 7.510 4.000 3.590 3.410

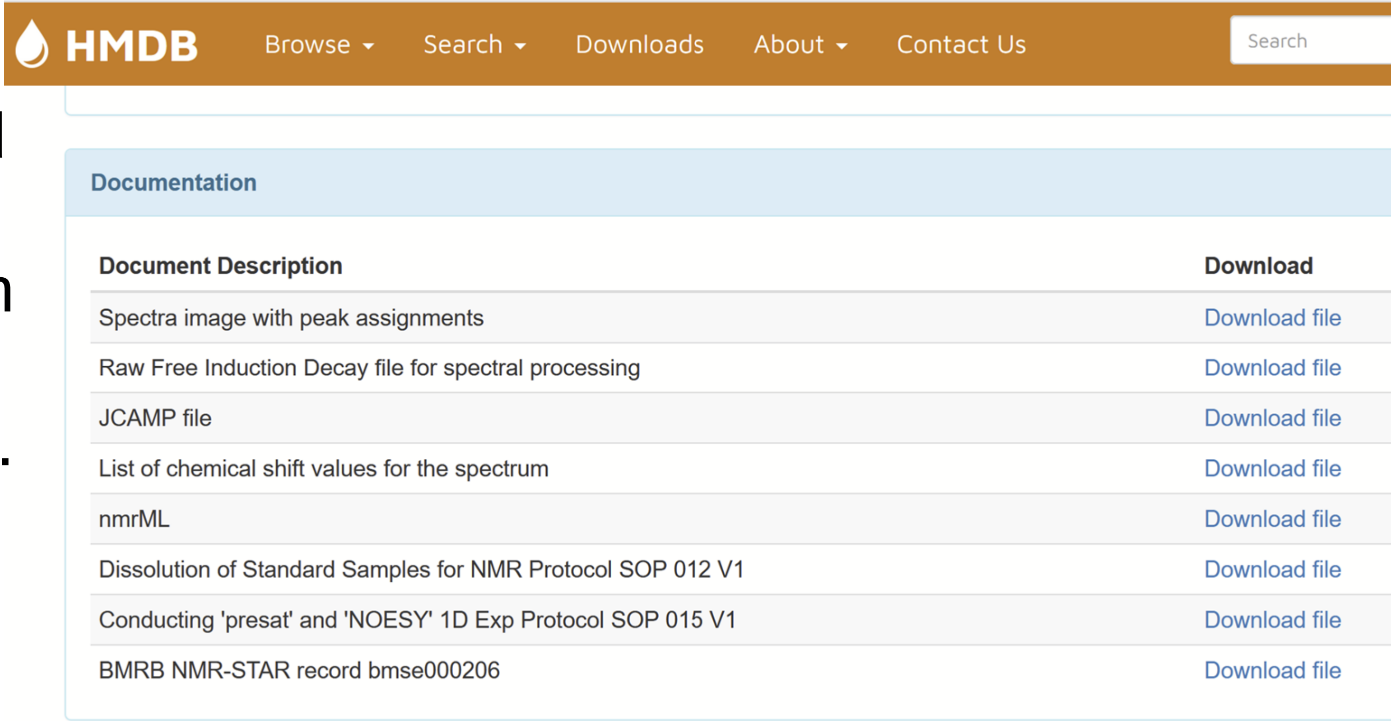
No.	Center	Peaks	Type	H's	Atoms	Peak Centers
1	7.510	1	s	1	9	7.510
2	4.000	1	s	3	14	4.000
3	3.590	1	s	3	11	3.590
4	3.410	1	s	3	13	3.410

https://hmdb.ca/spectra/nmr_one_d/1751

Examples – Current Practice

Human Metabolome Database

...and download the machine-readable data in a number of representations.



Documentation

Document Description	Download
Spectra image with peak assignments	Download file
Raw Free Induction Decay file for spectral processing	Download file
JCAMP file	Download file
List of chemical shift values for the spectrum	Download file
nmrML	Download file
Dissolution of Standard Samples for NMR Protocol SOP 012 V1	Download file
Conducting 'presat' and 'NOESY' 1D Exp Protocol SOP 015 V1	Download file
BMRB NMR-STAR record bmse000206	Download file

https://hmdb.ca/spectra/nmr_one_d/1751