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
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

FAIRSpec PROJECT TEAM
IUPAC Project: 2019-031-1-024
Development of a Standard for FAIR Data Management of Spectroscopic Data

Bob Hanson
Damien Jeannerat

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**

<table>
<thead>
<tr>
<th>NMR Media Type</th>
<th>Boolean</th>
<th>Metadata</th>
<th>Search for:</th>
</tr>
</thead>
<tbody>
<tr>
<td>mnova</td>
<td>AND</td>
<td>General</td>
<td>obtusallene</td>
</tr>
<tr>
<td></td>
<td>AND</td>
<td>InChIKey</td>
<td>BMXMTENBOUZINM-XSOLQRZSA-N</td>
</tr>
</tbody>
</table>

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-Epoxypotusallene IV, 12-Epoxypotusallene 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

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

**Member of collection / collaboration**

<table>
<thead>
<tr>
<th>DOI</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.14469/hpc/1267</td>
<td>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 G and Marilzabicycloallene D</td>
</tr>
</tbody>
</table>
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.

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

- 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

<table>
<thead>
<tr>
<th>ACS Collection</th>
<th>Size (MB)</th>
<th>digital entities</th>
<th>digital entities type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(zip)</td>
<td>(raw)</td>
<td>files</td>
</tr>
<tr>
<td><strong>joc.0c00770</strong></td>
<td>25</td>
<td>37</td>
<td>720</td>
</tr>
<tr>
<td><strong>orglett.0c00874</strong></td>
<td>27</td>
<td>40</td>
<td>1616</td>
</tr>
<tr>
<td><strong>orglett.0c00967</strong></td>
<td>29</td>
<td>41</td>
<td>1354</td>
</tr>
<tr>
<td><strong>orglett.0c01022</strong></td>
<td>15</td>
<td>52</td>
<td>66</td>
</tr>
<tr>
<td><strong>orglett.0c01197</strong></td>
<td>79</td>
<td>101</td>
<td>61</td>
</tr>
<tr>
<td><strong>orglett.0c01277</strong></td>
<td>52</td>
<td>74</td>
<td>2463</td>
</tr>
<tr>
<td><strong>orglett.0c01297</strong></td>
<td>57</td>
<td>73</td>
<td>1544</td>
</tr>
</tbody>
</table>
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?)
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

Spectral Datasets

Structures

one to one

one to many

many to one

analysis

representations
# Levels of NMR data reusability

<table>
<thead>
<tr>
<th>data representations and reuse level</th>
<th>possible processing</th>
<th>viewing and analysis facilitated (* with additional processing)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>full</td>
<td>near-full</td>
</tr>
<tr>
<td>raw data (FID + parameters)</td>
<td>10</td>
<td>yes</td>
</tr>
<tr>
<td>minimally processed data, (r+i spectra)</td>
<td>9</td>
<td>yes</td>
</tr>
<tr>
<td>fully processed data (real spectrum)</td>
<td>8</td>
<td>yes</td>
</tr>
</tbody>
</table>
## Levels of NMR data reusability

<table>
<thead>
<tr>
<th>Data representations and reuse level</th>
<th>Possible processing</th>
<th>Viewing and analysis facilitated (* with additional processing)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>full</td>
<td>near-full</td>
</tr>
<tr>
<td>peak tables, shifts, integration, and splitting</td>
<td>7</td>
<td>yes*</td>
</tr>
<tr>
<td>PDF</td>
<td>6</td>
<td>yes</td>
</tr>
<tr>
<td>journal-style description</td>
<td>5</td>
<td>yes*</td>
</tr>
<tr>
<td>image (PNG)</td>
<td>4</td>
<td>yes</td>
</tr>
<tr>
<td>peak table -- shifts only</td>
<td>3</td>
<td>yes</td>
</tr>
</tbody>
</table>

# 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
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 metadata? (Is that important?)
2. Are there examples of any of this already out there?
3. What about predicted vs. experimental vs. simulated spectral “data”?
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://bmrbi.org/
4. Human Metabolome Database (HMDB) https://hmdb.ca
5. Library of Congress Encoded Archival Description (EAD) https://www.loc.gov/ead
7. Natural Products Magnetic Resonance Database (NP-MRD) http://www.npmrd-project.org
8. NMRDB https://www.nmrd.org/
9. NMReData https://nmredata.org/
10. NMR Markup Language/Controlled Vocabulary (nmrML/nmrCV) http://nrmml.org/
11. NMRShiftDB https://nmrshiftdb.nmr.uni-koeln.de/
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…

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 DEs.

Digital Collection – An aggregation which contains DOs and DEs identified by a PID and described by metadata.

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?
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.

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.

<table>
<thead>
<tr>
<th>1. Structures</th>
<th>11. Identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Names and Identifiers</td>
<td>12. Safety and Hazards</td>
</tr>
<tr>
<td>3. Chemical and Physical Properties</td>
<td>13. Toxicity</td>
</tr>
<tr>
<td>5. Related Records</td>
<td>15. Literature</td>
</tr>
<tr>
<td>7. Drug and Medication Information</td>
<td>17. Biomolecular Interactions and Pathways</td>
</tr>
<tr>
<td>8. Food Additives and Ingredients</td>
<td>18. Biological Test Results</td>
</tr>
<tr>
<td>10. Use and Manufacturing</td>
<td>20. Information Sources</td>
</tr>
</tbody>
</table>
**Examples – Current Practice**

**PubChem**

...and points to more resources... *and portals*...
Examples – Current Practice
Human Metabolome Database

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

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.

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