IUPAC specification for the FAIR management of spectroscopic data in chemistry (IUPAC FAIRSpec) - preliminary (meta)data object model

ACS National Meeting, Mar. 21, 2022

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

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.

The proposed standards involve several aspects:

- **A set of principles** underlying what we mean by "FAIR" in relation to spectroscopic data.
- **A detailed object model** for describing the contents and relationships within an “IUPAC FAIRData Collection" in terms of objects and relationships of objects.
- **A recommendation for the organization of digital objects** within a collection,
- **A standard for describing properties of digital objects** within the metadata records of the finding aid,
- **A proposal for methods of data and metadata extraction** and the generation of IUPAC FAIRData Finding Aids.
- **A standard for the serialization of the finding aid** for an IUPAC FAIRData Collection.

The proposed standards involve several aspects:

- **A set of principles** underlying what we mean by "FAIR" in relation to spectroscopic data.
- **A detailed object model** for describing the contents and relationships within an “IUPAC FAIRData Collection" in terms of objects and relationships of objects.
- **A recommendation for the organization of digital objects** within a collection,
- **A standard for describing properties of digital objects** within the metadata records of the finding aid,
- **A proposal for methods of data and metadata extraction** and the generation of IUPAC FAIRData Finding Aids.
- **A standard for the serialization of the finding aid** for an IUPAC FAIRData Collection.

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, in press

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.

The proposed standards involve several aspects:

- **A set of principles** underlying what we mean by "FAIR" in relation to spectroscopic data.
- **A detailed object model** for describing the contents and relationships within an “IUPAC FAIRData Collection" in terms of objects and relationships of objects.
- **A recommendation for the organization of digital objects** within a collection,
- **A standard for describing properties of digital objects** within the metadata records of the finding aid,
- **A proposal for methods of data and metadata extraction** and the generation of IUPAC FAIRData Finding Aids.
- **A standard for the serialization of the finding aid** for an IUPAC FAIRData Collection.

Today’s presentation – the object model

1. Representations and Properties
2. Aggregations, Associations, and Collections
3. The IUPAC FAIRData Collection
4. The Pieces of the Puzzle
5. The Full Enterprise

Today’s presentation – the object model

1. Representations and Properties
2. Aggregations, Associations, and Collections
3. The IUPAC FAIRData Collection
4. The Pieces of the Puzzle
5. The Full Enterprise

Principles for Today

FAIRSpec Principle 3D. *Value is in the eye of the reuser.*
FAIRSpec Principle 5C. *Respect format and implementation diversity.*
FAIRSpec Principle 5D. *All data formats should be valued.*

Digital Entity

digital entity

Anything that can be represented by a bitstream.

Digital Object

digital object

A digital entity composed of **a structured sequence of bits** that has a name and can be identified with attributes that describe its properties.

Representations

representation

One of a set of **digital objects** that may take any one of a number of forms that allow for various levels of data reuse.

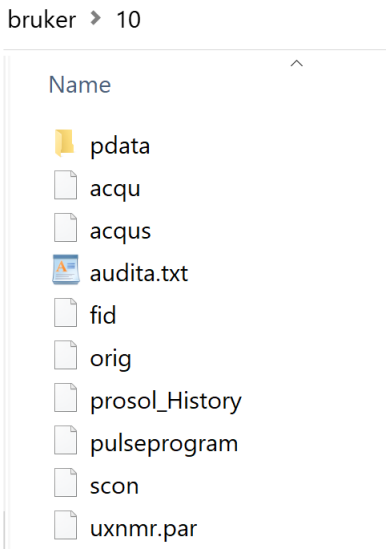
Properties

property

A **key:value pair** that describes a characteristic of a digital object.

Examples of Spectroscopic Representations

… an instrument dataset



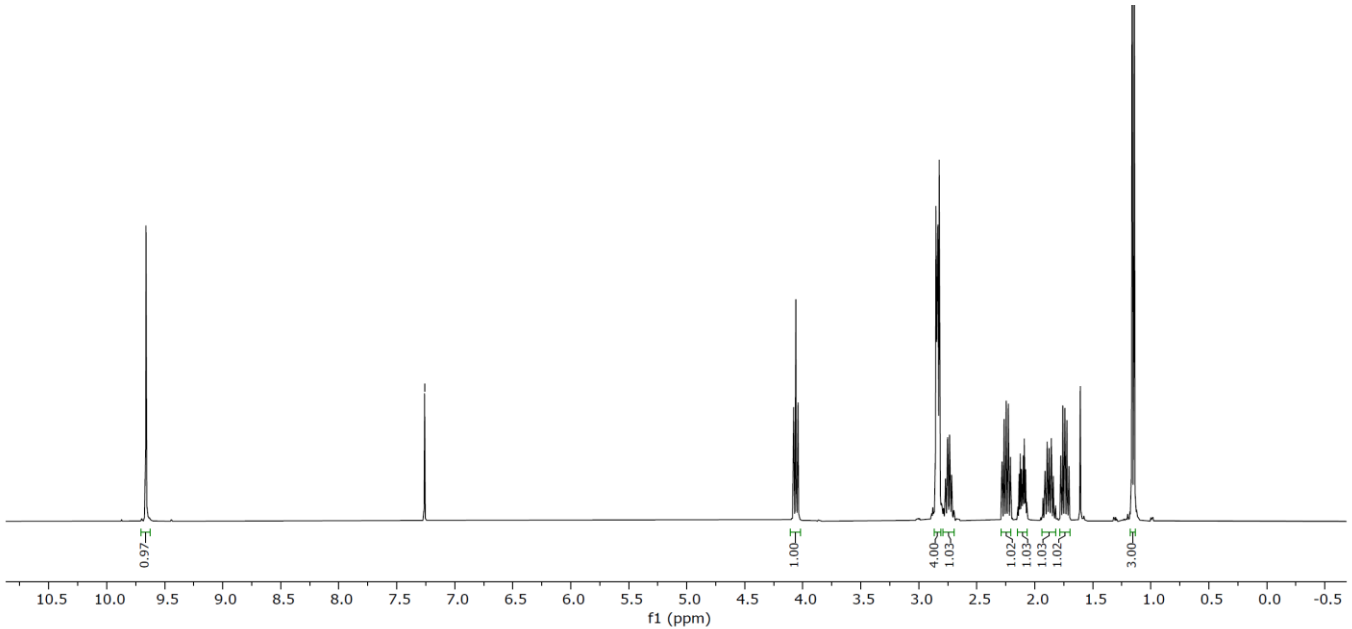
Examples of Spectroscopic Representations

… a JCAMP-DX file

```
##TITLE= Beta_Pinene
##JCAMP-DX= 6.0 $$ MestReNova 14.0.1-23559
##DATA TYPE= NMR SPECTRUM
##DATA CLASS= XYDATA
##ORIGIN= Mestrelab Research S.L.
##OWNER= skim592
…
```

Examples of Spectroscopic Representations

… an image



Examples of Spectroscopic Representations

… a linear description

^{1}H NMR (400 MHz, $CDCl_3$) δ 5.45 (ddq, *J* = 4.3, 2.9, 1.4 Hz, 1H), 4.09 (t, *J* = 5.94 Hz, 1H), 3.13 – 3.02 (m, 1H), 2.98 (s, 1H), 2.59 (ddtd, *J* = 16.1, 5.2, 2.4, 1.3 Hz, 1H), 2.34 (ddd, *J* = 11.5, 5.4, 1.9 Hz, 1H), 1.87 (tq, *J* = 6.1, 4.0 Hz, 1H), 1.79 (ddd, *J* = 14.4, 8.5, 4.9 Hz, 1H), 1.72 – 1.64 (m, 4H), 1.63 – 1.58 (m, 1H), 1.57 – 1.49 (m, 1H), 1.37 (dtd, *J* = 12.0, 5.6, 0.6 Hz, 1H), 1.05 (d, *J* = 6.5 Hz, 3H), 1.02 (s, 3H), 0.99 – 0.94 (m, 12H), 0.94 (s, 3H), 0.65 – 0.56 (m, 7H), 0.52 (td, *J* = 9.3, 5.0 Hz, 1H) ppm;

Examples of Structure Representations

… a 3D MOL file

```
C8H10N4O2
APtclcactv03202207183D 0   0.00000     0.00000

 24 25  0  0  0  0  0  0  0  0999 V2000
    1.3120   -1.0479    0.0025 N   0  0  0  0  0  0  0  0  0  0  0  0
    2.2465   -2.1762    0.0031 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.7906    0.2081    0.0010 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.9938    0.3838    0.0002 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.9714    1.2767   -0.0001 N   0  0  0  0  0  0  0  0  0  0  0  0
    1.5339    2.6294   -0.0017 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.4026    1.0989   -0.0001 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.4446    1.9342   -0.0010 N   0  0  0  0  0  0  0  0  0  0  0  0
   -2.5608    1.2510   -0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.2862   -0.0680    0.0015 N   0  0  0  0  0  0  0  0  0  0  0  0
   -3.2614   -1.1612    0.0029 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.9114   -0.1939    0.0014 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.0163   -1.2853   -0.0022 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.4380   -2.4279   -0.0068 O   0  0  0  0  0  0  0  0  0  0  0  0
    3.2697   -1.8004    0.0022 H   0  0  0  0  0  0  0  0  0  0  0  0
```

Examples of Structure Representations

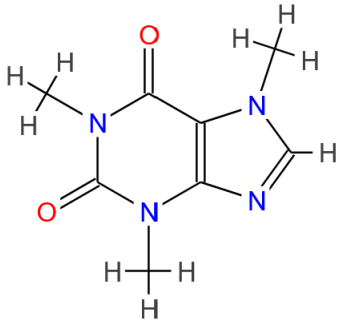
… a 2D MOL file

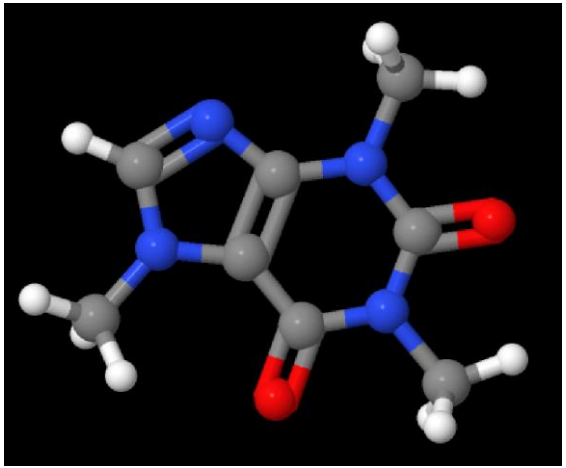
```
JME 2015-12-06 Sun Mar 20 11:19:51 GMT-500 2022

 24 25  0  0  0  0  0  0  0  0999 V2000
    2.3188    5.1567    0.0000 N   0  0  0  0  0  0  0  0  0  0
    0.8874    5.9831    0.0000 C   0  0  0  0  0  0  0  0  0  0
    3.7502    5.9831    0.0000 C   0  0  0  0  0  0  0  0  0  0
    3.7502    7.6359    0.0000 O   0  0  0  0  0  0  0  0  0  0
    5.1815    5.1567    0.0000 C   0  0  0  0  0  0  0  0  0  0
    5.1815    3.5039    0.0000 C   0  0  0  0  0  0  0  0  0  0
    3.7502    2.6775    0.0000 N   0  0  0  0  0  0  0  0  0  0
    3.7502    1.0247    0.0000 C   0  0  0  0  0  0  0  0  0  0
    2.3188    3.5039    0.0000 C   0  0  0  0  0  0  0  0  0  0
    0.8874    2.6775    0.0000 O   0  0  0  0  0  0  0  0  0  0
    6.7454    5.6604    0.0000 N   0  0  0  0  0  0  0  0  0  0
    7.2589    7.2314    0.0000 C   0  0  0  0  0  0  0  0  0  0
    7.7100    4.3303    0.0000 C   0  0  0  0  0  0  0  0  0  0
    6.7454    3.0003    0.0000 N   0  0  0  0  0  0  0  0  0  0
    1.3998    6.8705    0.0000 H   0  0  0  0  0  0  0  0  0  0
    0.0000    6.4955    0.0000 H   0  0  0  0  0  0  0  0  0  0
    0.3751    5.0957    0.0000 H   0  0  0  0  0  0  0  0  0  0
```

Examples of Structure Representations

… an image





Examples of Properties

```
IFD.property.spec.nmr.expt.label:                  "1c/13C-NMR"
IFD.property.spec.nmr.expt.nucl.1:                 "13C"
IFD.property.spec.nmr.expt.nucl.2:                 "1H"
IFD.property.spec.nmr.expt.pulse.prog:             "deptqgpsp"
IFD.property.spec.nmr.expt.temperature.absolute:   298.1525
IFD.property.spec.nmr.instr.freq.nominal:          600
IFD.property.spec.nmr.instr.manufacturer.name:     "Bruker"
IFD.property.spec.nmr.instr.probe.type:            "Z126545_0016 (CPP BBO 600S3 BB-H&F-D-05 Z)"

IFD.property.struc.compound.label:   "1c"
IFD.property.struc.inchi:            "InChI=1S/C11H13NO/c13-11(12-7-4-8-12)9-10-5-2-1-3-6-10/h1-3,5-6H,4,7-9H2"
IFD.property.struc.inchikey:         "HXFKEAUPENVJFI-UHFFFAOYSA-N"
IFD.property.struc.smiles:           "c1cccc2c1.C2C(=O)N1CCC1"
```

Today’s presentation – the object model

1. Representations and Properties
2. Aggregations, Associations, and Collections
3. The IUPAC FAIRData Collection
4. The Pieces of the Puzzle
5. The Full Enterprise

Digital Aggregations

digital aggregation

A **bundle of digital entities** lacking the metadata required to provide context and to describe its contents in a machine-actionable manner.

Examples of Digital Aggregations

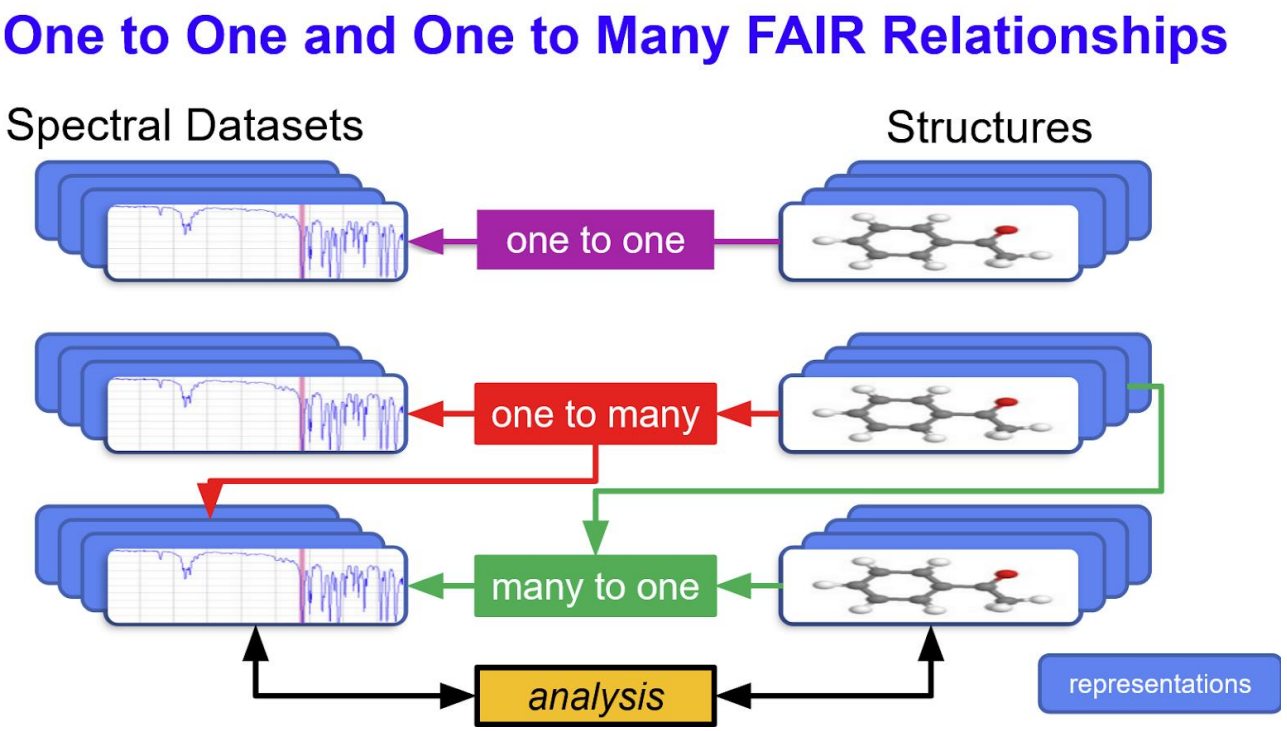
ACS Aggregation	Size (MB)		digital entities	
	(zip)	(raw)	files	type
joc.0c00770	25	37	720	11 cmpd dirs; 24 Bruker datasets & 12 mnova files
orglett.0c00874	27	40	1616	36 cmpd dirs; 76 Bruker datasets
orglett.0c00967	29	41	1354	33 cmpd dirs; 62 Bruker datasets
orglett.0c01022	15	52	66	2 dirs; 64 mnova 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

Association

association

A meaningful **context-dependent connection** made between two or more objects.

Associations



Digital Collections

digital collection

A **bundle of digital objects** with associated metadata that provide context and characteristics of its digital objects and associations in a machine-actionable manner.

Today's presentation – the object model

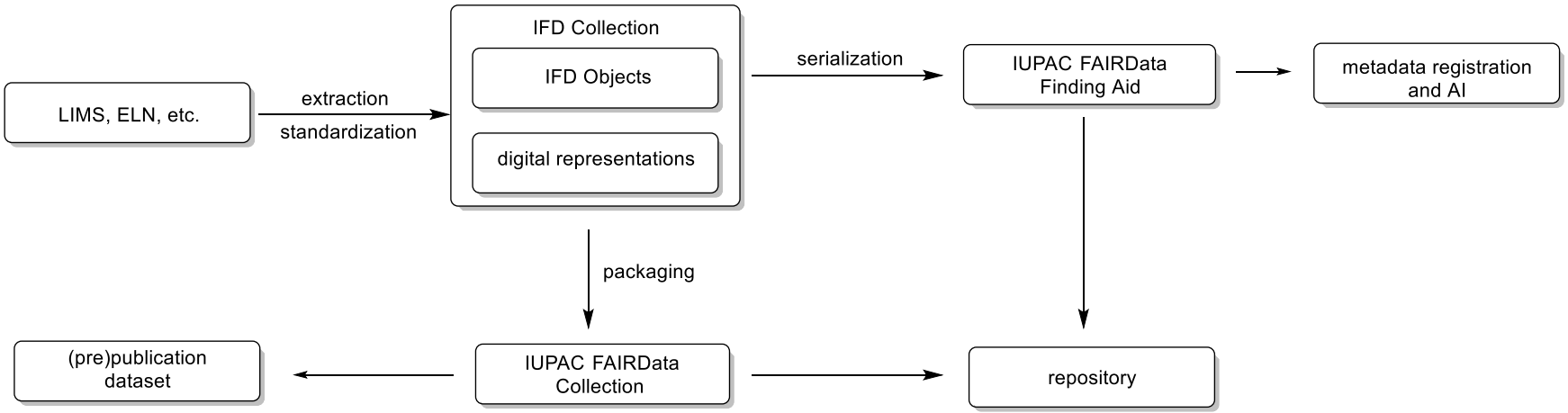
1. Representations and Properties
2. Aggregations, Associations, and Collections
3. The IUPAC FAIRData Collection
4. The Pieces of the Puzzle
5. The Full Enterprise

The IUPAC FAIRData Collection

IUPAC FAIRData Collection

A digital collection organized in concordance with the IUPAC FAIRData Recommendations, with an associated **IUPAC FAIRData Finding Aid**.

The IUPAC FAIRData Collection

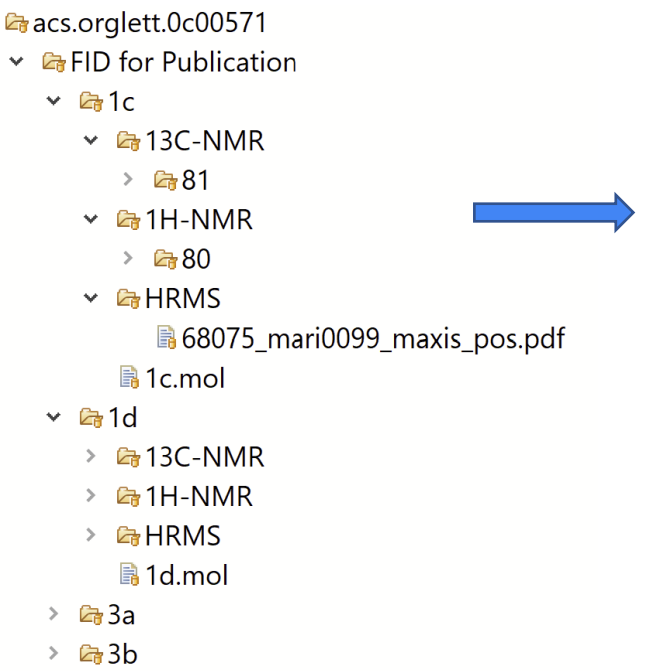


The IUPAC FAIRData Finding Aid

IUPAC FAIRData Finding Aid

A digital object that describes the collection's representations in a machine-actionable manner, including their properties and their associations.

The IUPAC FAIRData Finding Aid



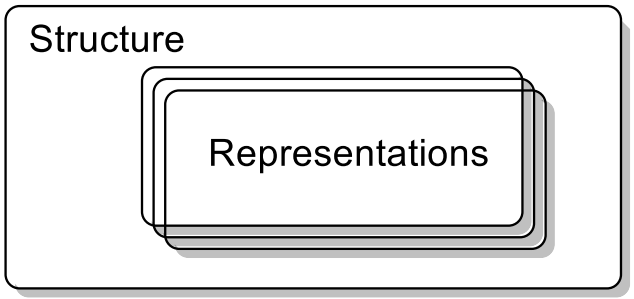
https://chemapps.stolaf.edu/iupac/demo/demo.htm?pub=571

Today’s presentation – the object model

1. Representations and Properties
2. Aggregations, Associations, and Collections
3. The IUPAC FAIRData Collection
4. The Pieces of the Puzzle
5. The Full Enterprise

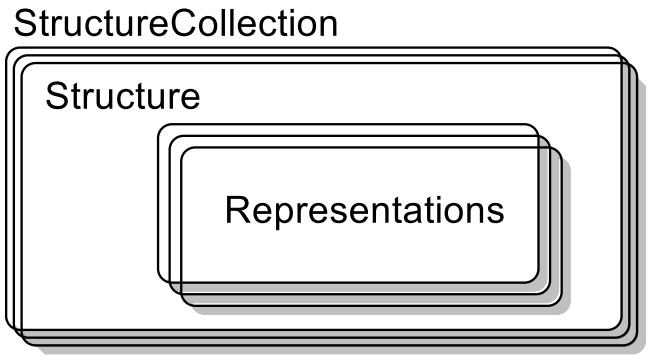
The Pieces of the Puzzle

a structure with its associated representations



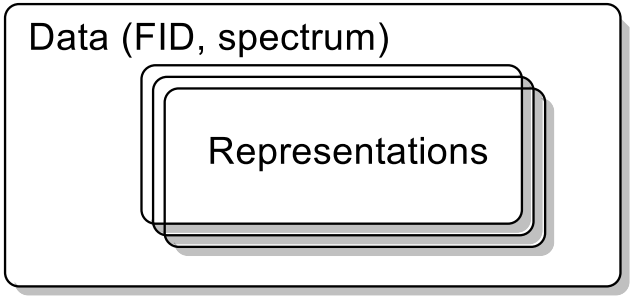
The Pieces of the Puzzle

a collection of structures



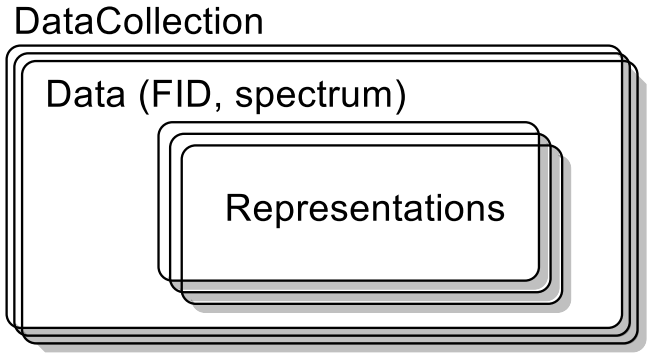
The Pieces of the Puzzle

spectroscopic data



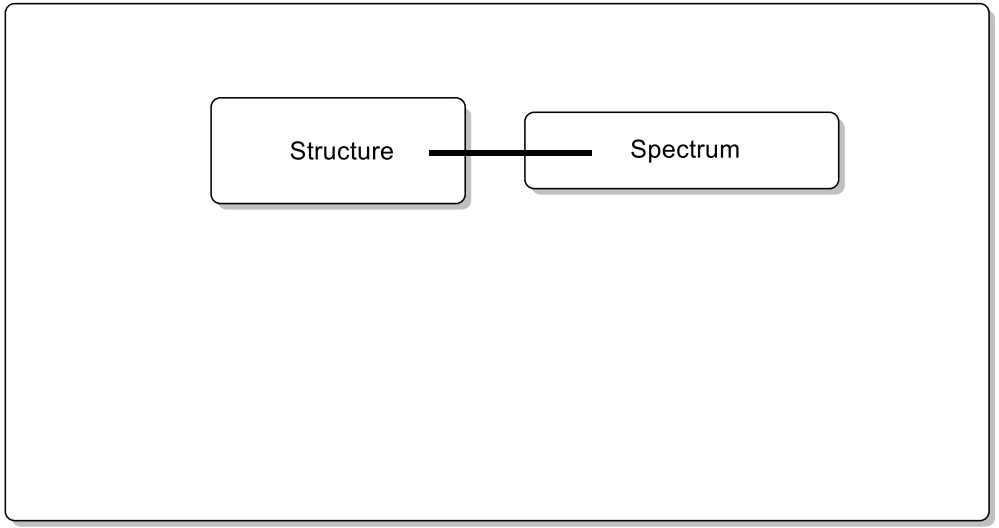
The Pieces of the Puzzle

a collection of spectra



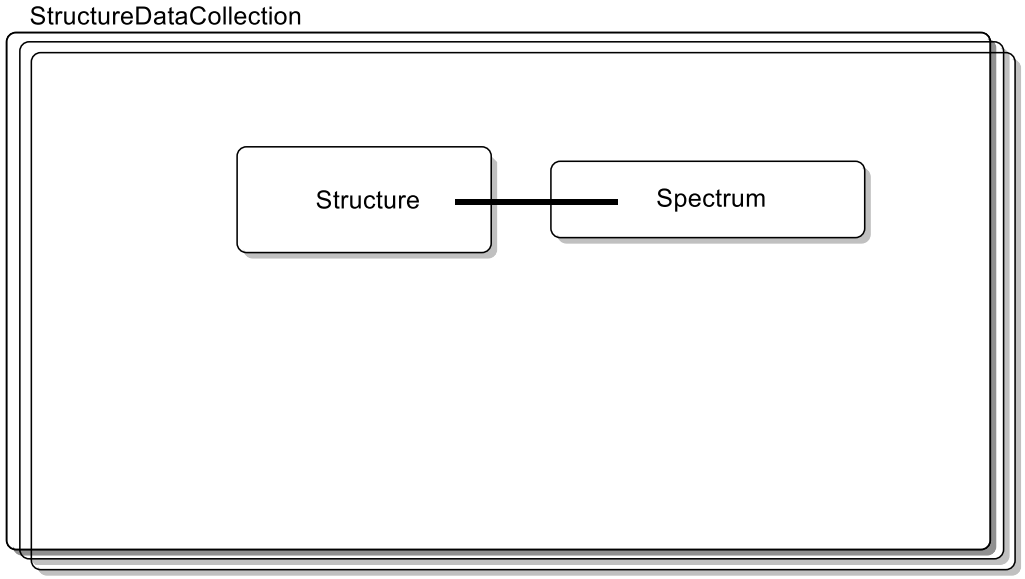
The Pieces of the Puzzle

a simple structure – spectrum association



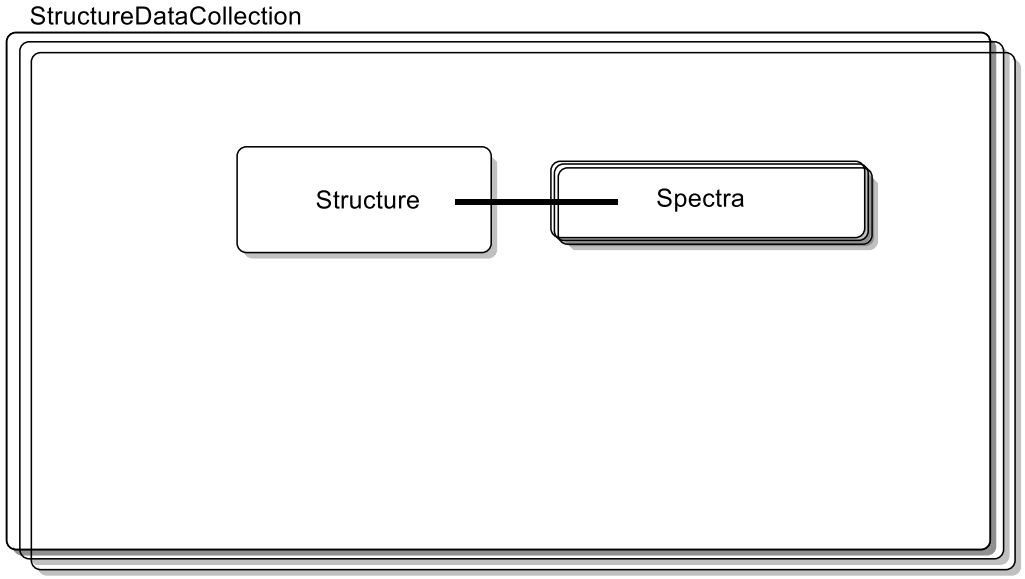
The Pieces of the Puzzle

a collection of simple structure – spectrum associations



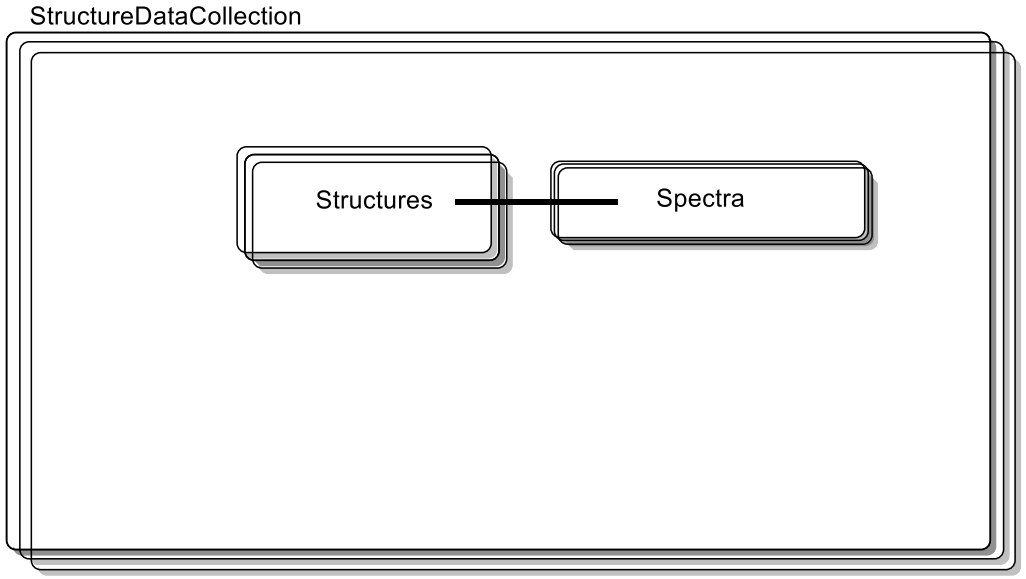
The Pieces of the Puzzle

a more typical collection of structure – spectra associations



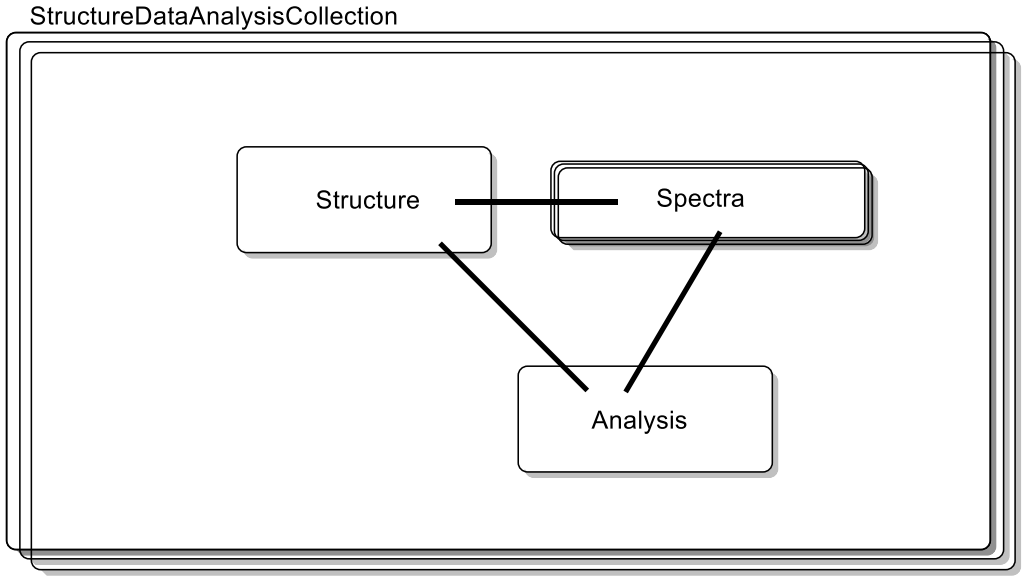
The Pieces of the Puzzle

allowing for mixtures



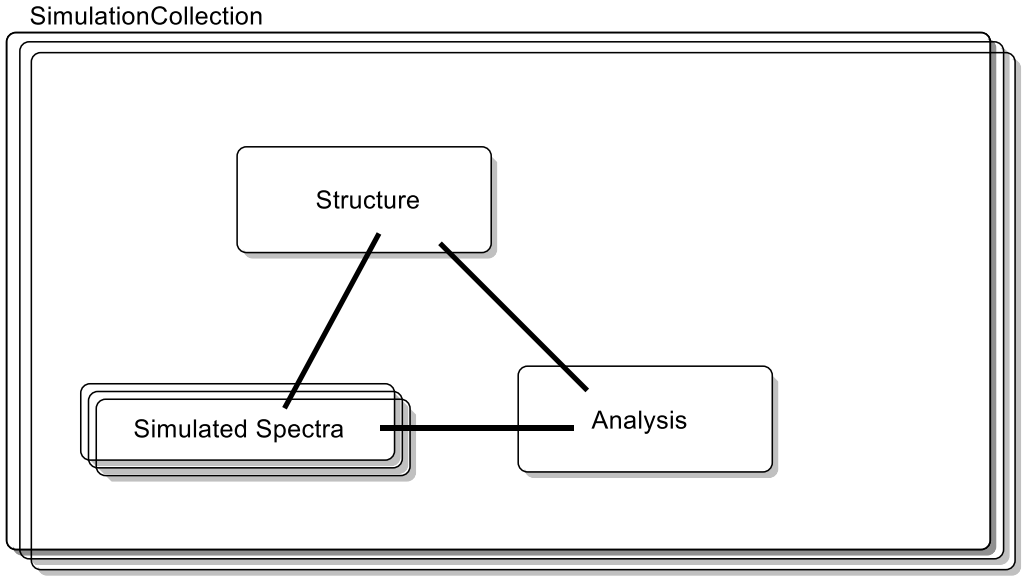
The Pieces of the Puzzle

adding analysis



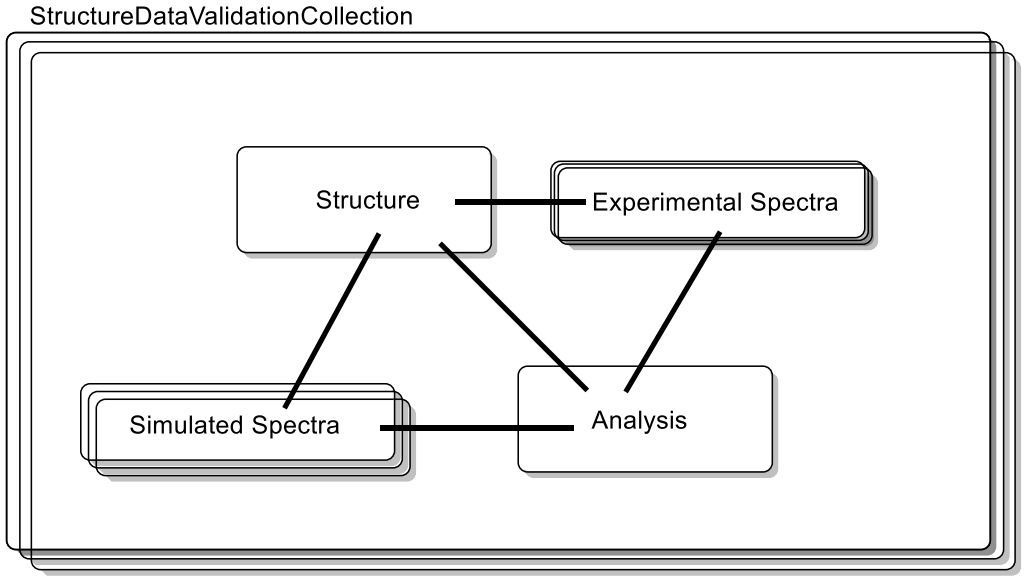
More Pieces of the Puzzle

a simulation



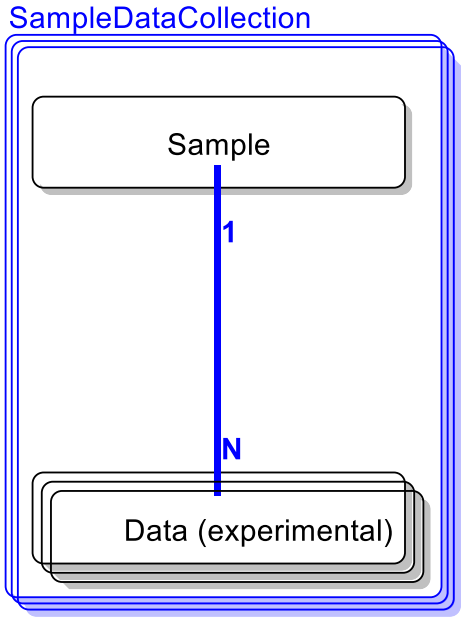
The Pieces of the Puzzle

adding simulation



More Pieces of the Puzzle

a collection of samples and
their associated spectra

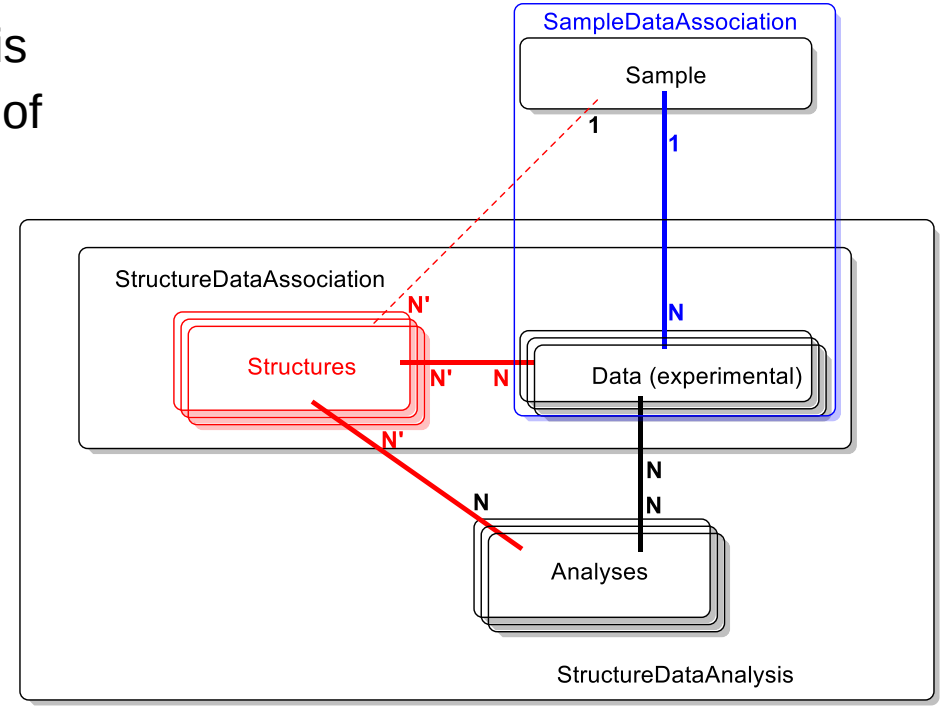


More Pieces of the Puzzle

The goal of spectroscopic data analysis is generally to make a 1:1 association of a sample with a chemical structure.

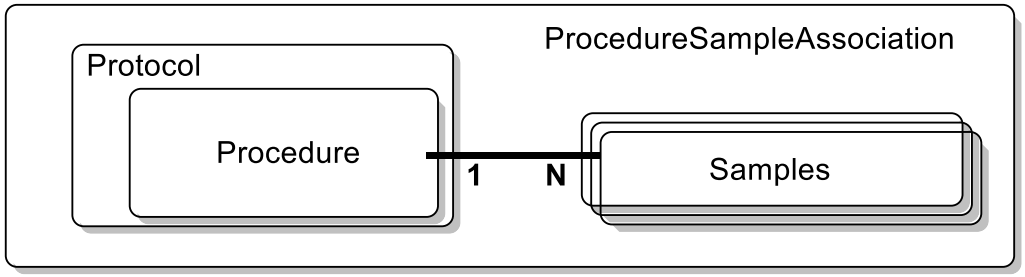
The inference that a given sample is a compound with a given structure is a product of this analysis.

The result may not always be 1:1.



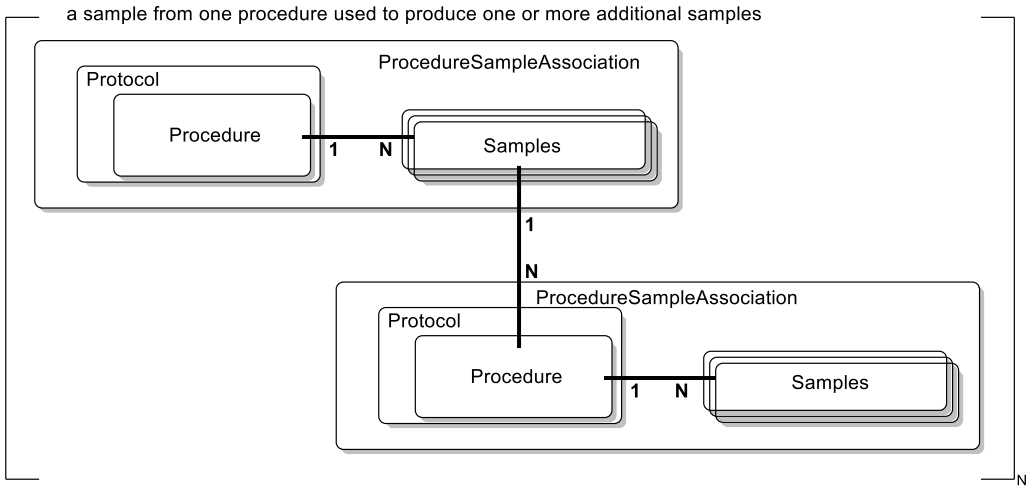
More Pieces of the Puzzle

a procedure based on a protocol producing one or more samples



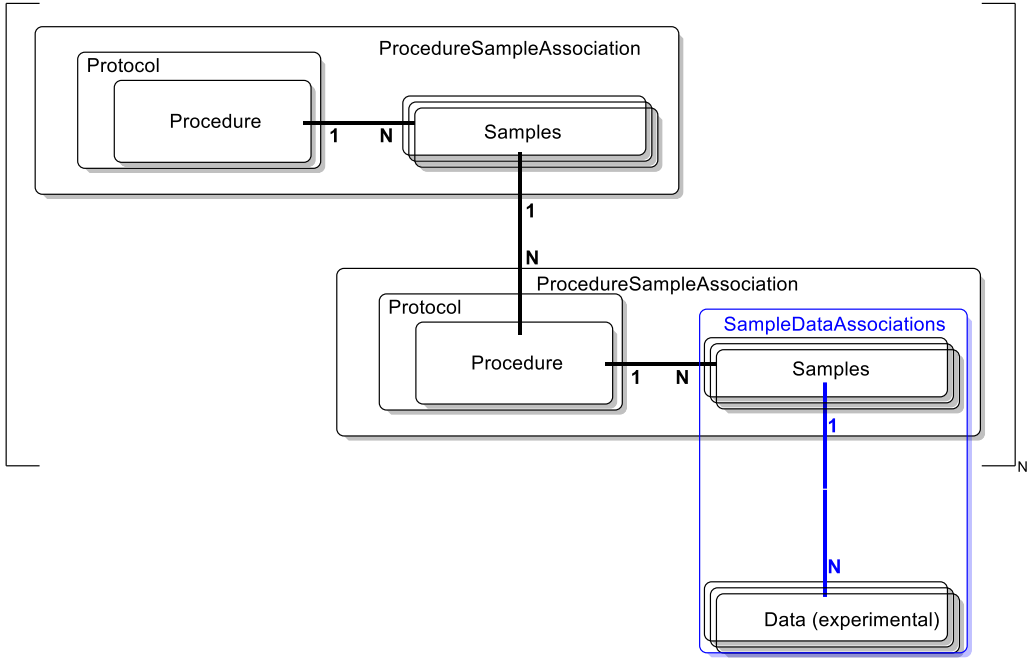
More Pieces of the Puzzle

a sample from one procedure used to produce one or more additional samples



The ELN Piece

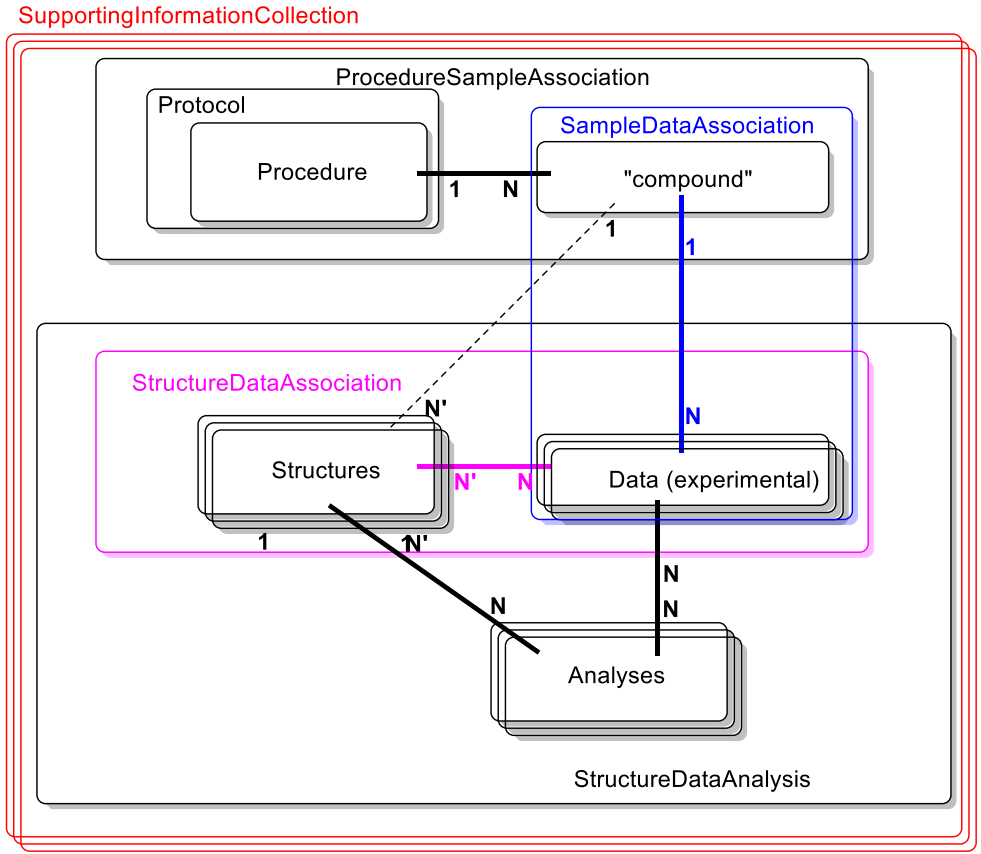
Electronic laboratory notebooks implementing IUPAC FAIRSpec Recommendations could provide the needed sample-data association.



The Publication Piece

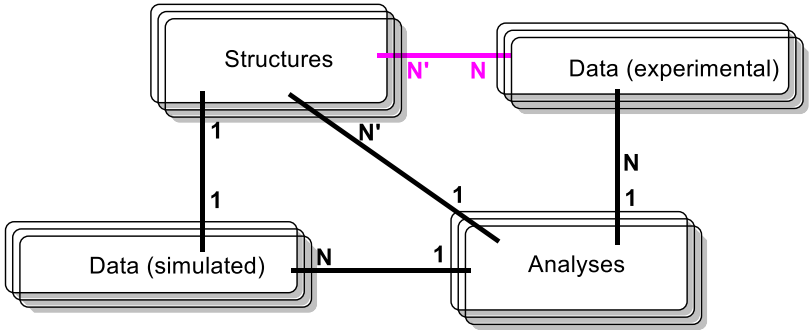
The “supporting information” for a publication in chemistry could be one possible representation of an IUPAC FAIRData Collection.

Note that there is not necessarily a 1:1 connection between structure and “compound”



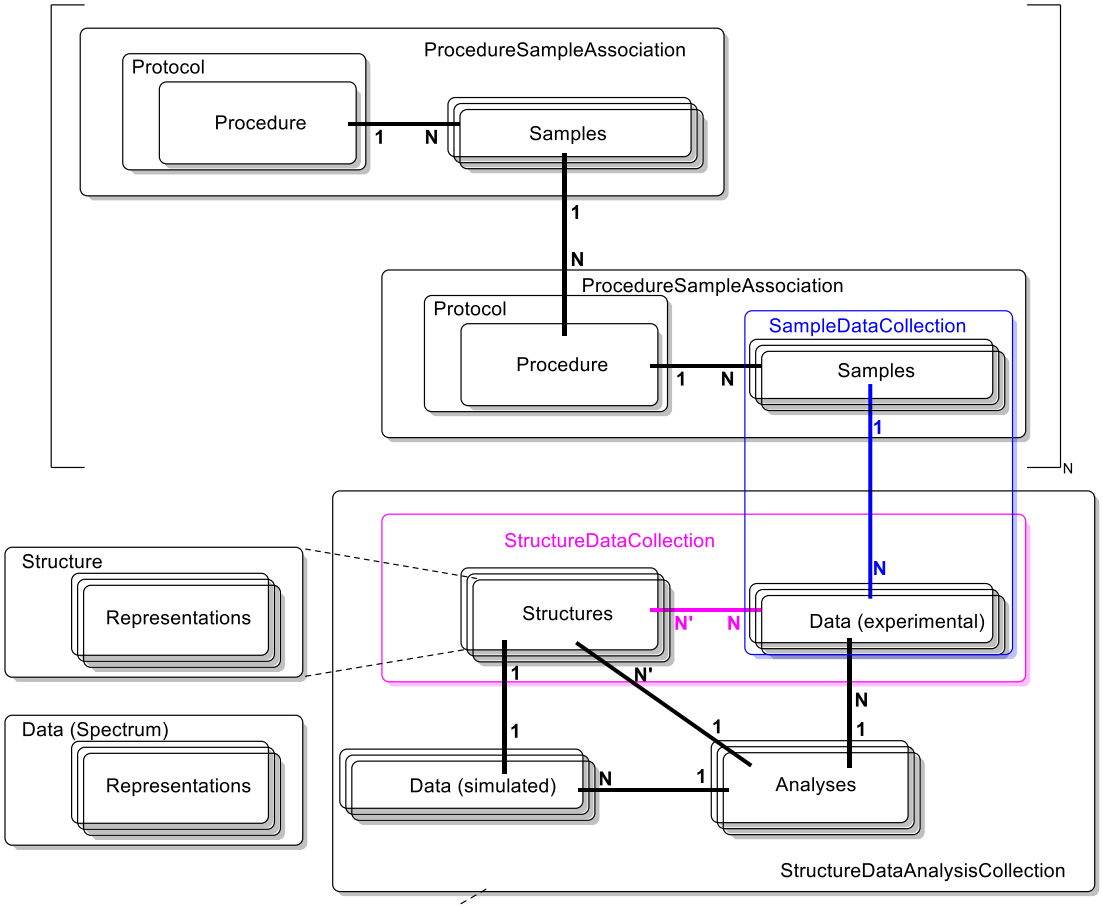
The Validation Piece

Based on IUPAC FAIRData Collections, emergent services could offer value-added pre-publication validation services.



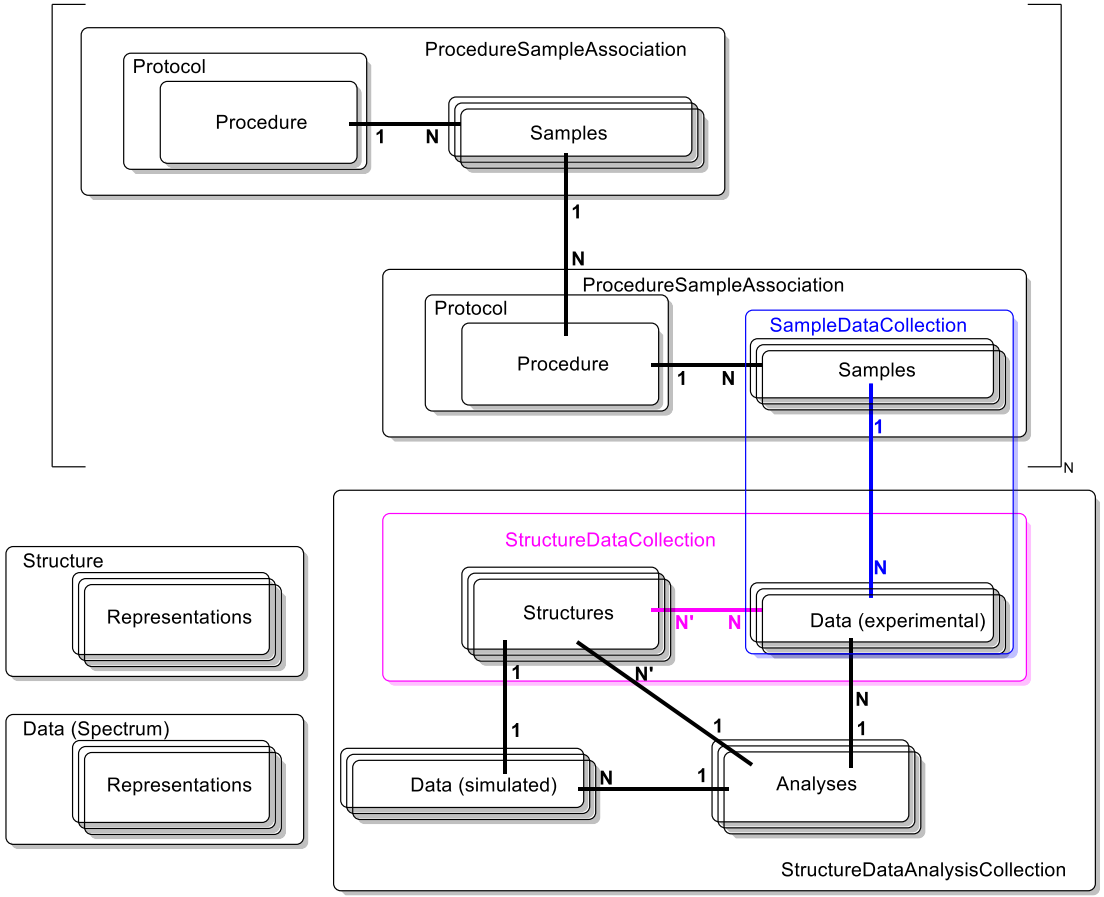
The Repository Piece

A repository could implement a query structure that could return any or all of these associations as IUPAC FAIRData Collections of whatever representations are desired by the (re)user.



The IUPAC FAIRData Digital Object Model

The full object model, all of which (or any part of which) could be described using an IUPAC FAIRData Finding Aid.



Stay tuned!

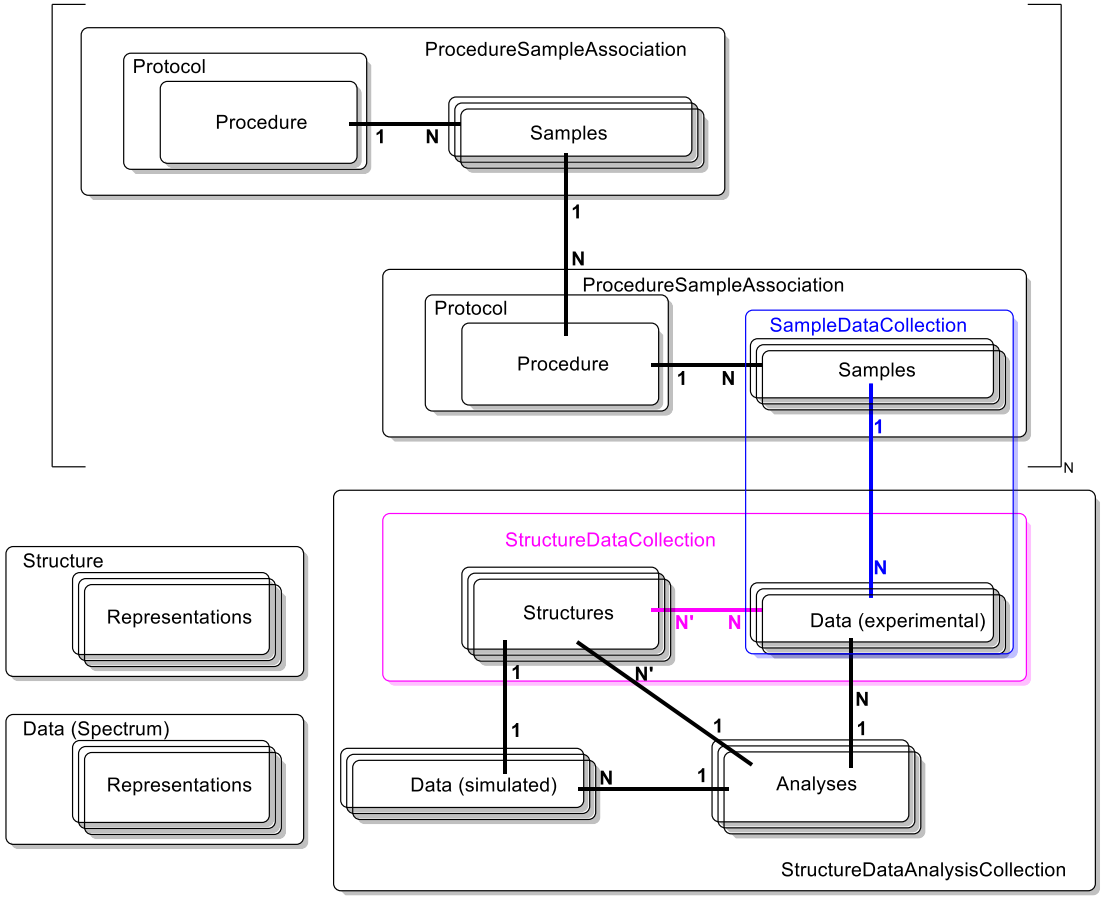
The proposed standards involve several aspects:

- **A set of principles** underlying what we mean by "FAIR" in relation to spectroscopic data.
- **A detailed object model** for describing the contents and relationships within an “IUPAC FAIRData Collection" in terms of objects and relationships of objects.
- **A recommendation for the organization of digital objects** within a collection,
- **A standard for describing properties of digital objects** within the metadata records of the finding aid,
- **A proposal for methods of data and metadata extraction** and the generation of IUPAC FAIRData Finding Aids.
- **A standard for the serialization of the finding aid** for an IUPAC FAIRData Collection.

In Summary

We have presented an object model that is based on the IUPAC FAIRSpec Guiding Principles.

The model defines a comprehensive set of objects that can be associated, represented, and collected in a variety of ways.

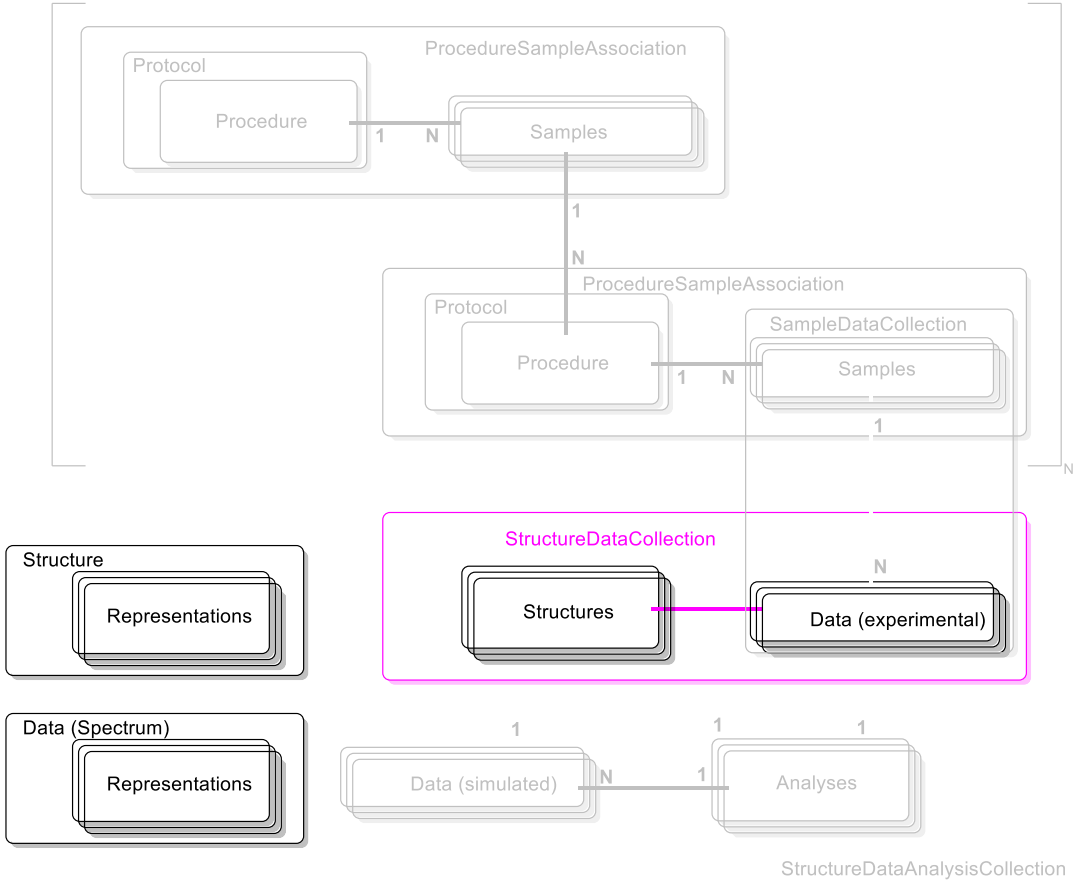


In Summary

The object model is *modular, extensible,* and *flexible.*

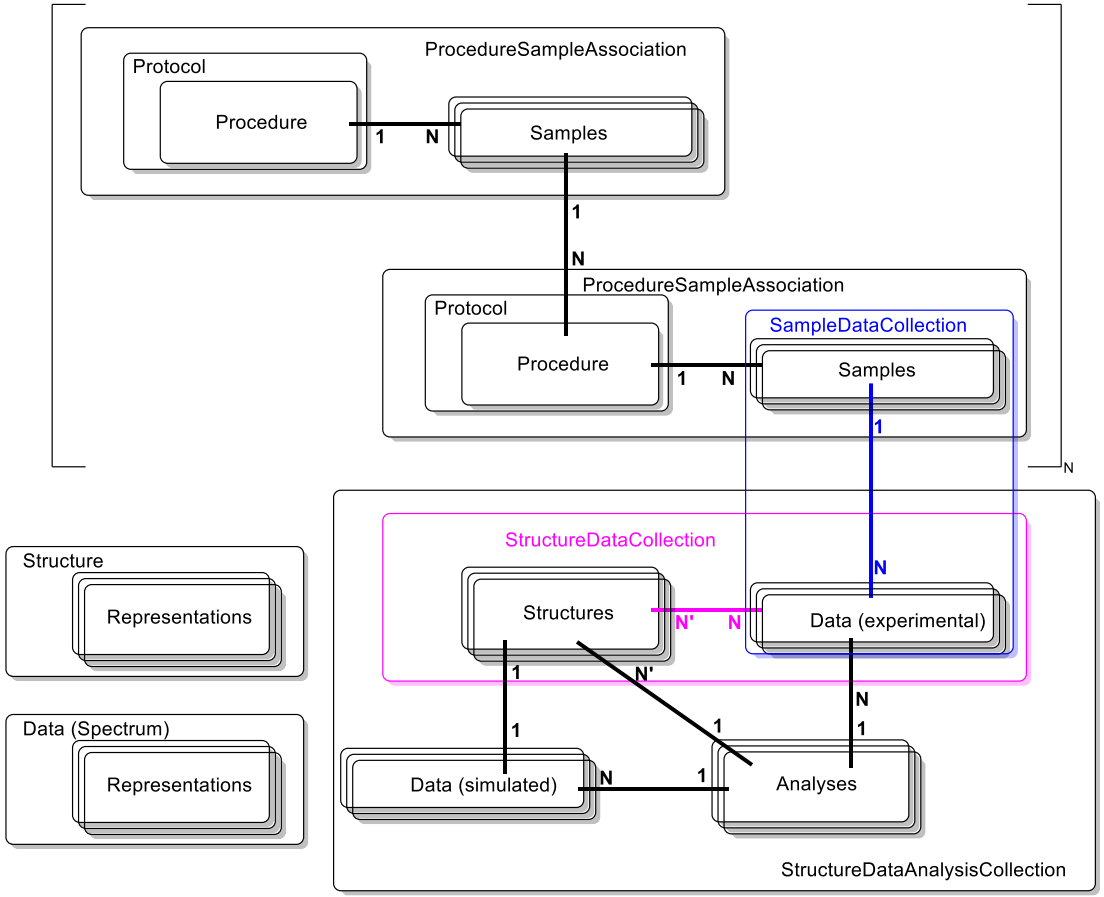
Our project scope and expertise is in the area of structure-spectra collections.

These are the pieces we will develop.



In Summary

We hope that others with other expertise and perspectives will join us in this endeavor to complete the puzzle and revolutionize the world of chemistry.



Guiding Principles for the FAIR Management of Spectroscopic Data

Additional resources

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

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.

INTERNATIONAL UNION OF
PURE AND APPLIED CHEMISTRY

Bob Hanson

Damien Jeannerat

Thank you!

hansonr@stolaf.edu

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

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