The IUPAC FAIRData Metadata Object Model v. 0.0.2

NFDI4Chem/IUPAC Joint Meeting

Karlsruhe, Germany
June 7-8, 2022

Mark Archibald, Ian Bruno, Stuart J. Chalk, Antony N. Davies,
Damien Jeannerat, **Robert M. Hanson**, Robert J. Lancashire,
Chandu Nainala, Henry S. Rzepa

IUPAC Project 2019-031-1-024

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

Note!

Some of the figures in this hastily pieced together presentation are old.

Names of classes and properties are not up to date.

See https://github.com/IUPAC/IUPAC-FAIRSpec and the public for-comment Google Doc for the latest details.

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

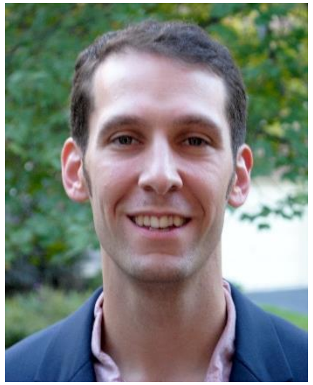
Mark
Archibald

Ian
Bruno

Stuart Chalk

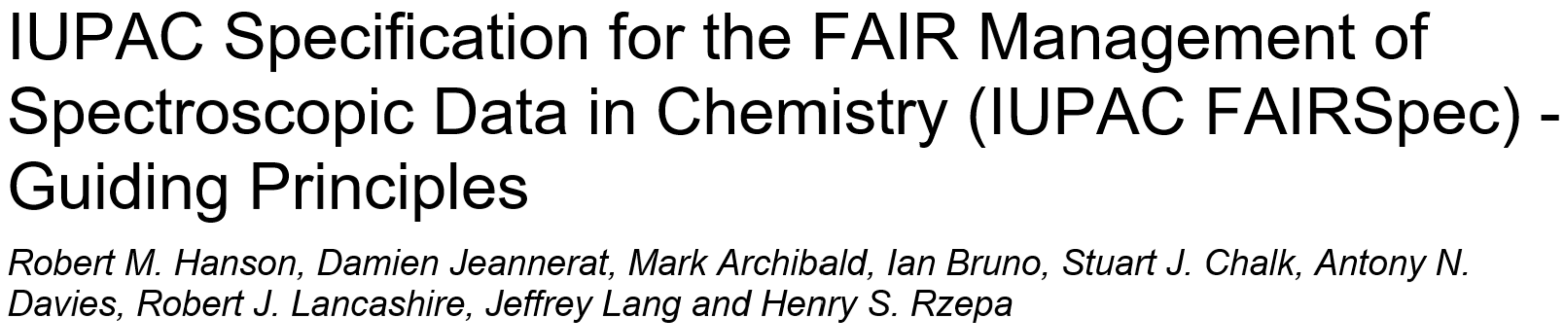
Tony Davies

Robert
Lancashire



Jeff
Lang

Henry Rzepa



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

https://chemrxiv.org/engage/chemrxiv/article-details/626671c388636c48051e91dd
https://www.degruyter.com/document/doi/10.1515/pac-2021-2009/html

Presents 20 principles in five areas:

1. FAIR Management of data should be an ongoing concern

2. Context is important.

3. FAIR management of data requires curation.

4. Metadata must be standardized and registered.

5. FAIR data management standards should be *modular, extensible, and flexible.*

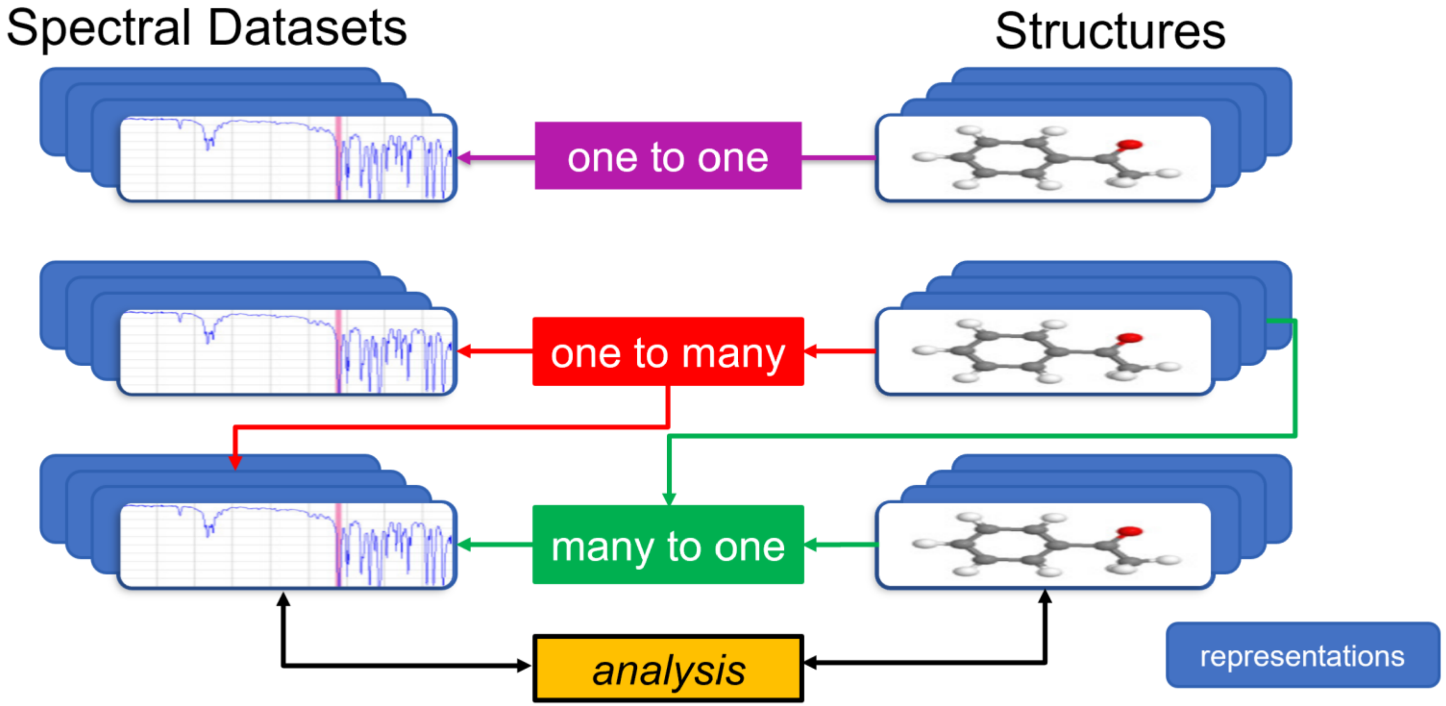
Glossary of about 30 terms

chemical structure identifier
curation
data and metadata extraction
data management plan
data model
data provenance
data repository
data representation
dataset (spectroscopic)
digital aggregation
digital collection
digital entity
digital finding aid
digital object
Digital Object Identifier (DOI)
IUPAC FAIRSpec Data Collection
IUPAC FAIRSpec Data Model
metadata
metadata crosswalk
metadata element
metadata harvesting
metadata registration
metadata registration agency
metadata schema
metadata store
open data
persistent identifier (PID)
PID graph
serialization (of a finding aid)

IUPAC FAIRSpec Principles

The standard respects the reality that data can have **multiple *representations***, and that reuse of data relies upon data being in a form that is meaningful *for the reuser*.

One to One and One to Many FAIR Relationships



Spectral Datasets

Structures

one to one

one to many

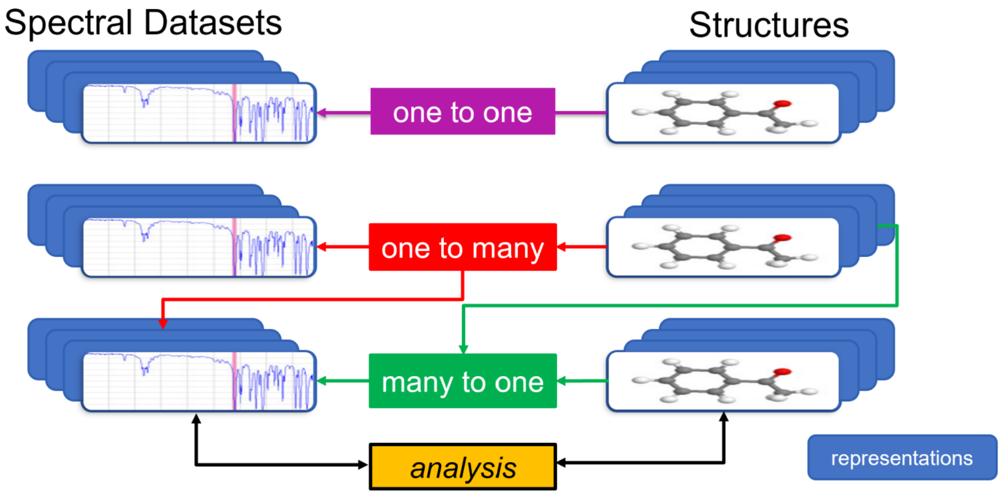
many to one

analysis

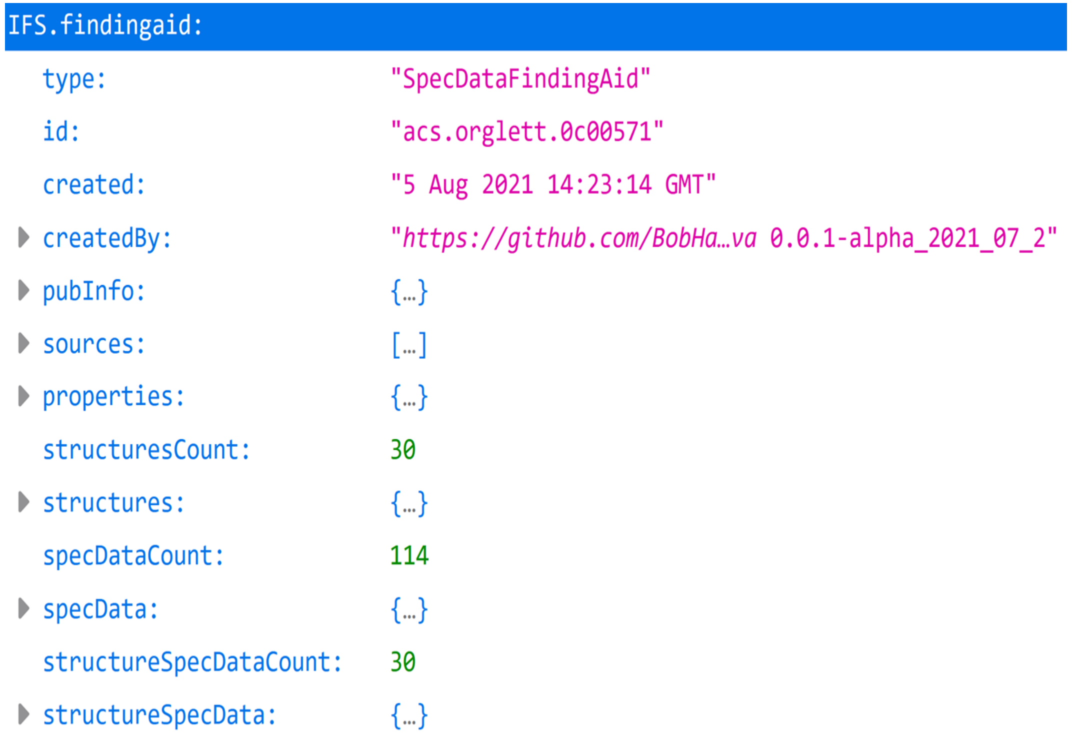
representations

IUPAC FAIRSpec Principles

The standard describes a ***digital collection*** with associated ***digital finding aid*** that allows a reuser to quickly ascertain whether additional scrutiny of the data collection is warranted.



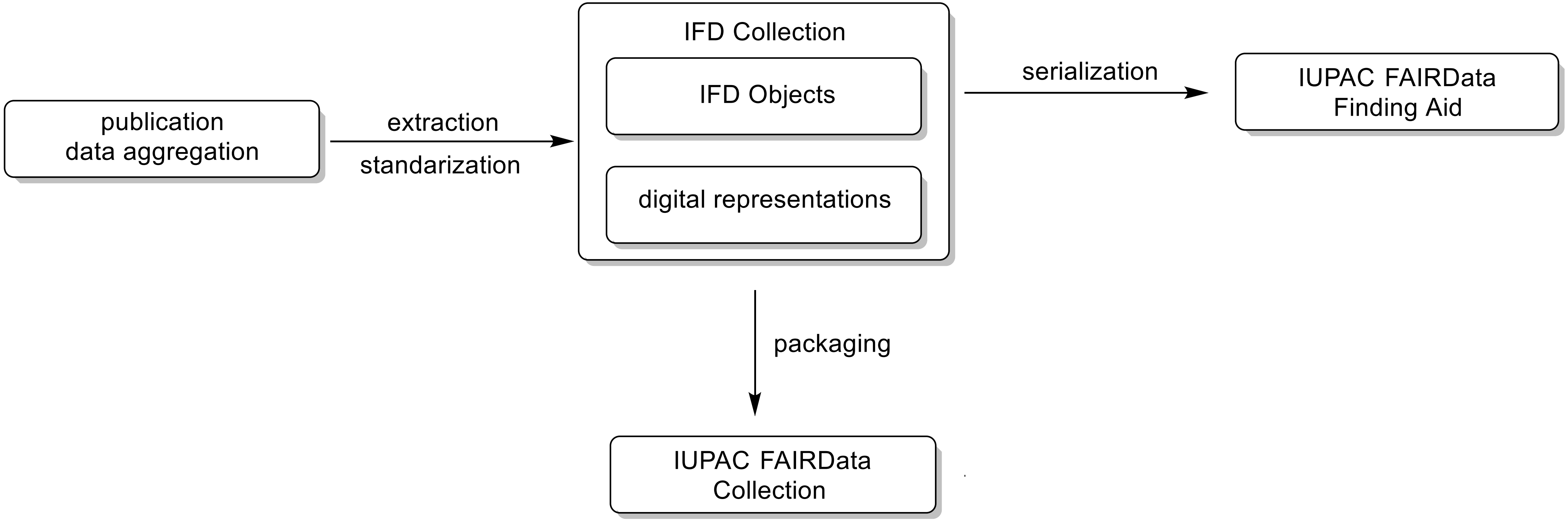
+



```
IFS.findingaid:
  type:                    "SpecDataFindingAid"
  id:                      "acs.orglett.0c00571"
  created:                 "5 Aug 2021 14:23:14 GMT"
▸ createdBy:               "https://github.com/BobHa…va 0.0.1-alpha_2021_07_2"
▸ pubInfo:                 {…}
▸ sources:                 […]
▸ properties:              {…}
  structuresCount:         30
▸ structures:              {…}
  specDataCount:           114
▸ specData:                {…}
  structureSpecDataCount:  30
▸ structureSpecData:       {…}
```

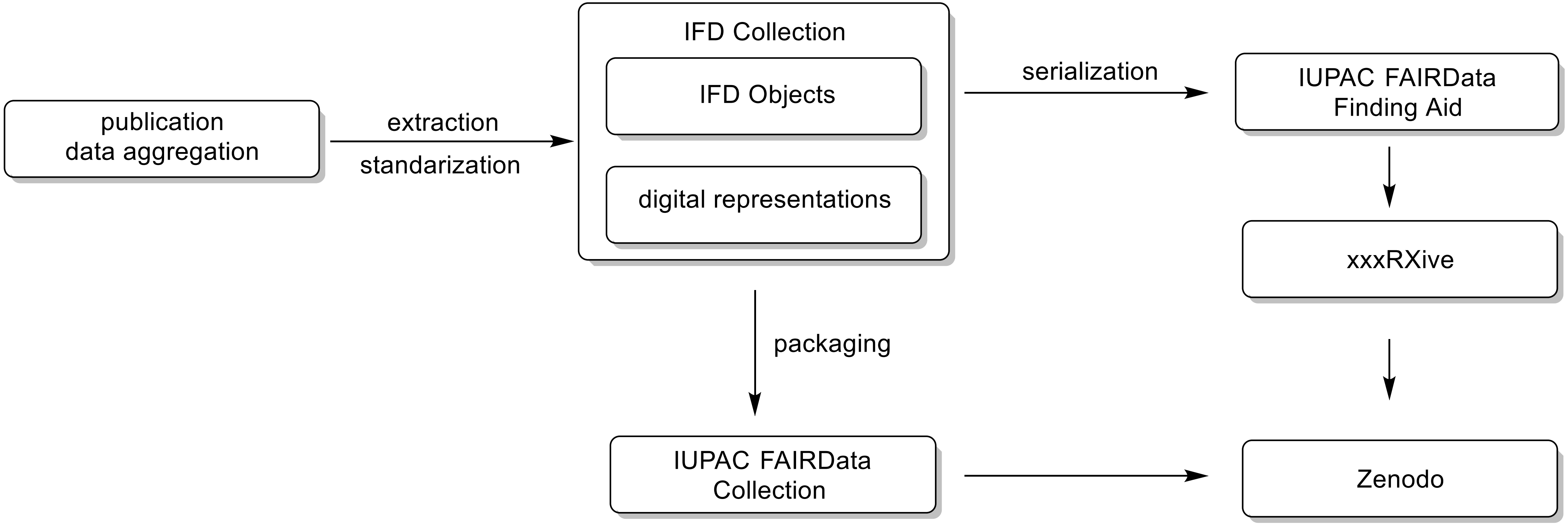
IUPAC FAIRSpec Principles

The standard **allows for repackaging** or "extraction" of metadata and other digital objects from an original dataset in order to provide a better reuser experience.



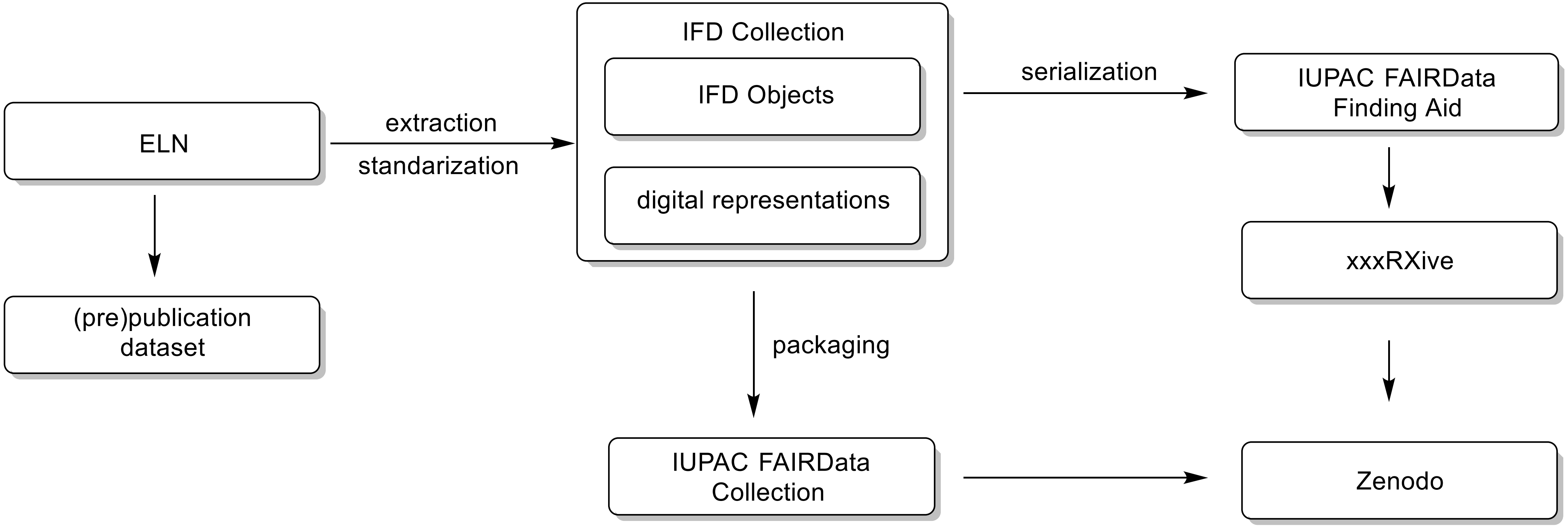
IUPAC FAIRSpec Principles

The standard allows for **distributed data storage**.



IUPAC FAIRSpec Principles

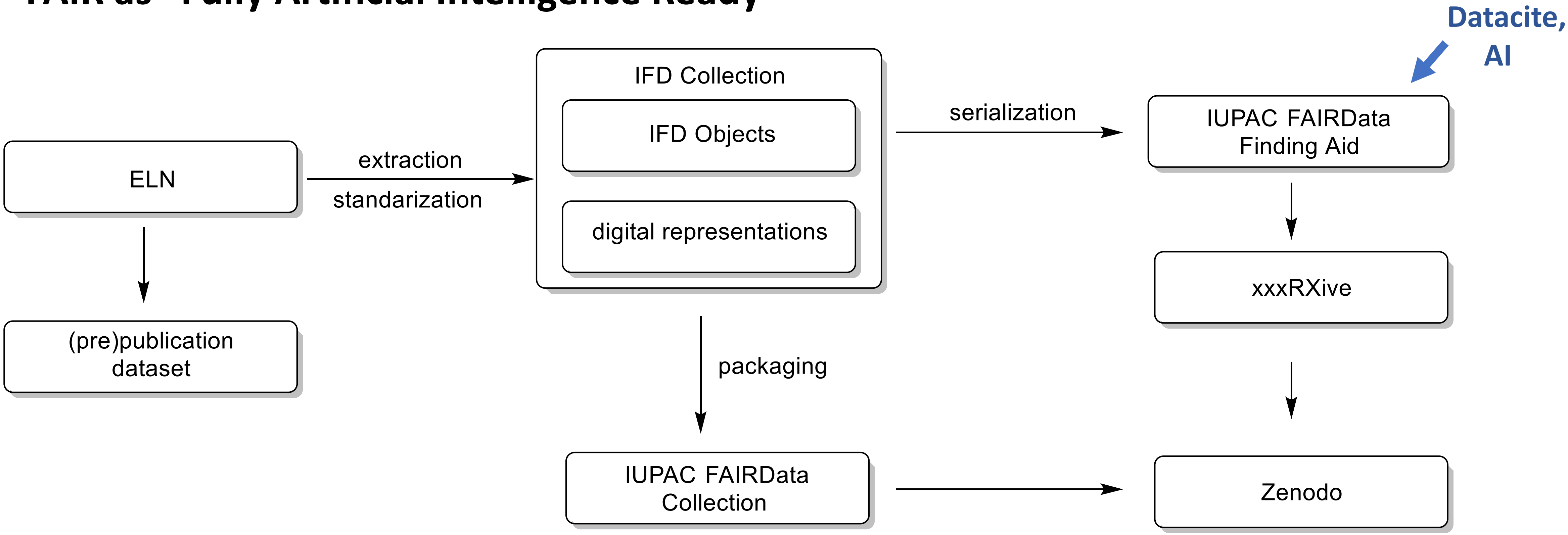
The standard emphasizes the **importance of management throughout the lifecycle of the data (and beyond)**



IUPAC FAIRSpec Principles

The standard will be clearly defined and, as much as possible, mappable onto other metadata standards that are in use or will be in future –

FAIR as "Fully Artificial Intelligence Ready"



The IUPAC FAIRData Metadata Object Model

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

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.

The IUPAC FAIRData Metadata Object Model

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

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.

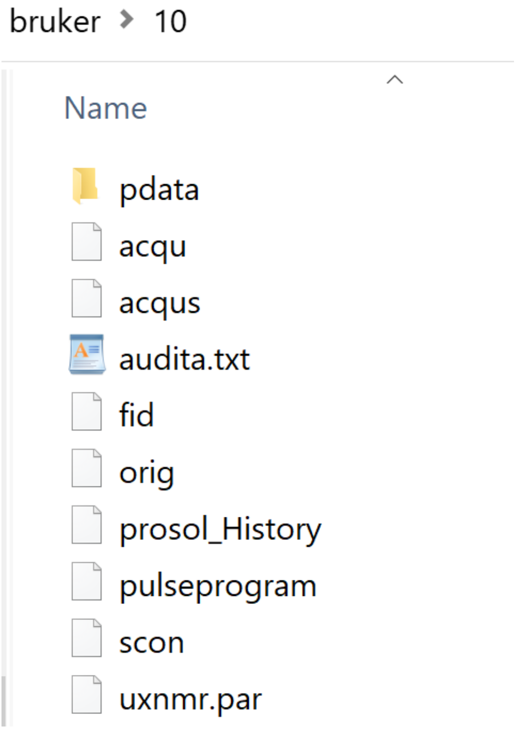
Properties

property

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

Examples of Spectroscopic Representations

… an instrument dataset



bruker › 10

Name

pdata
acqu
acqus
audita.txt
fid
orig
prosol_History
pulseprogram
scon
uxnmr.par

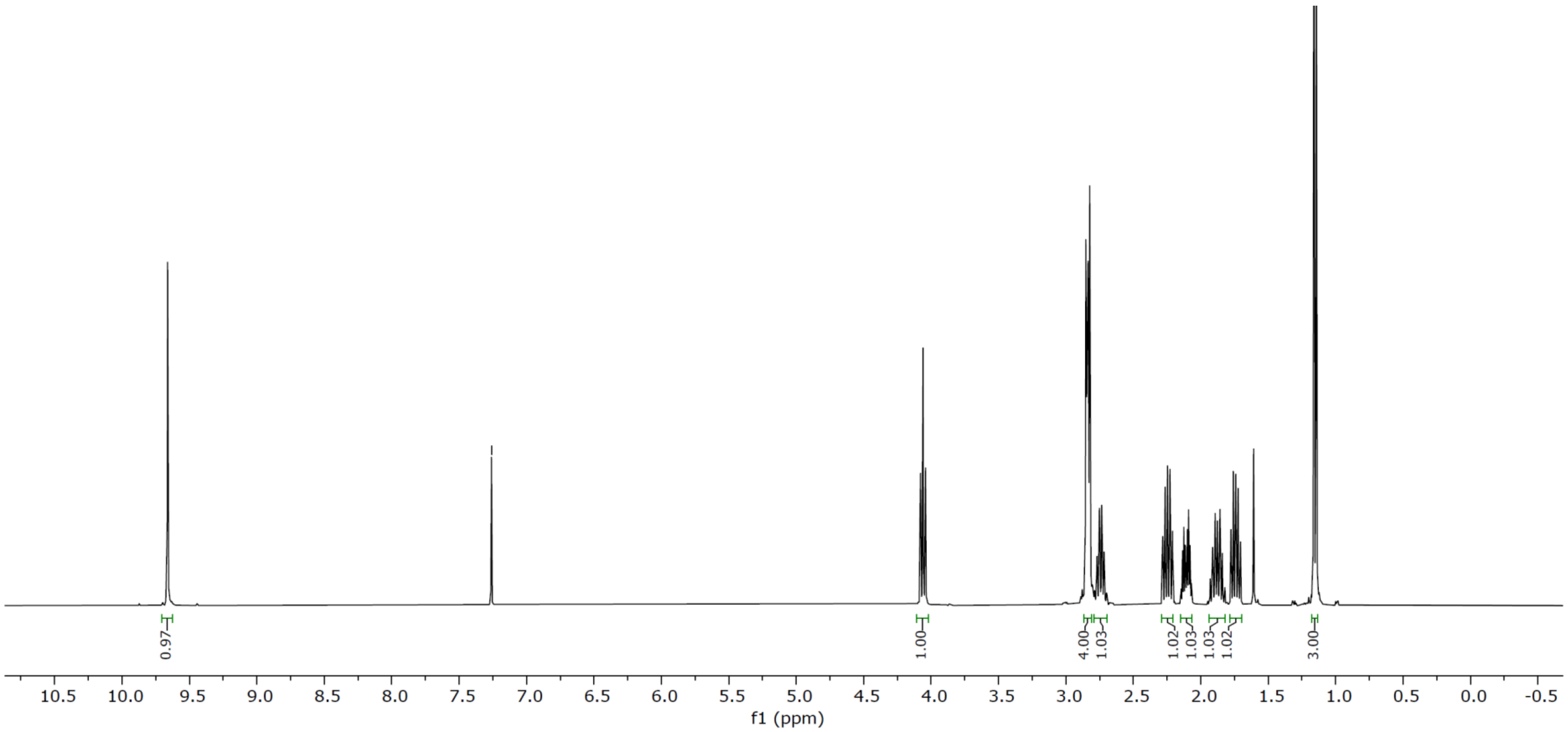
Examples of Spectroscopic Representations

… 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
…
```

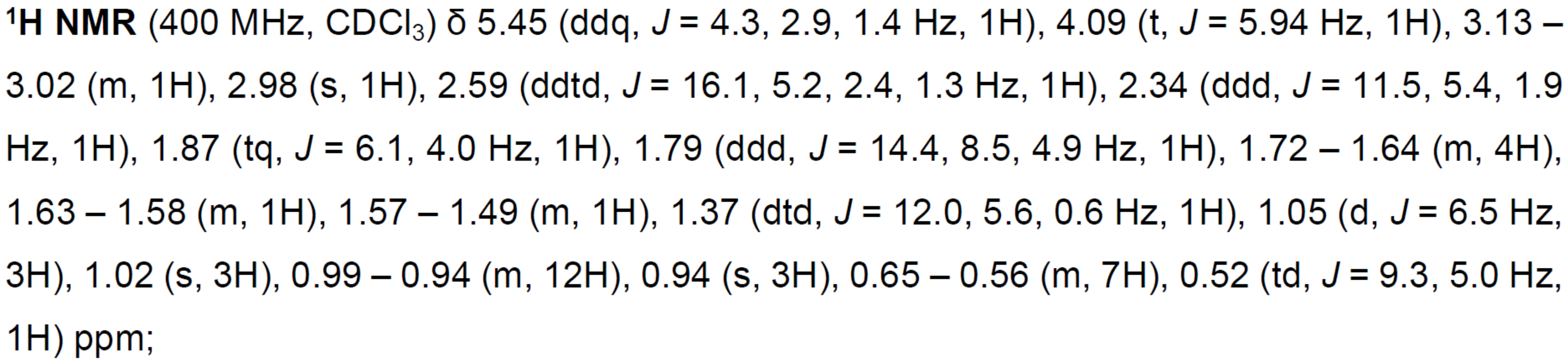
Examples of Spectroscopic Representations

… 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

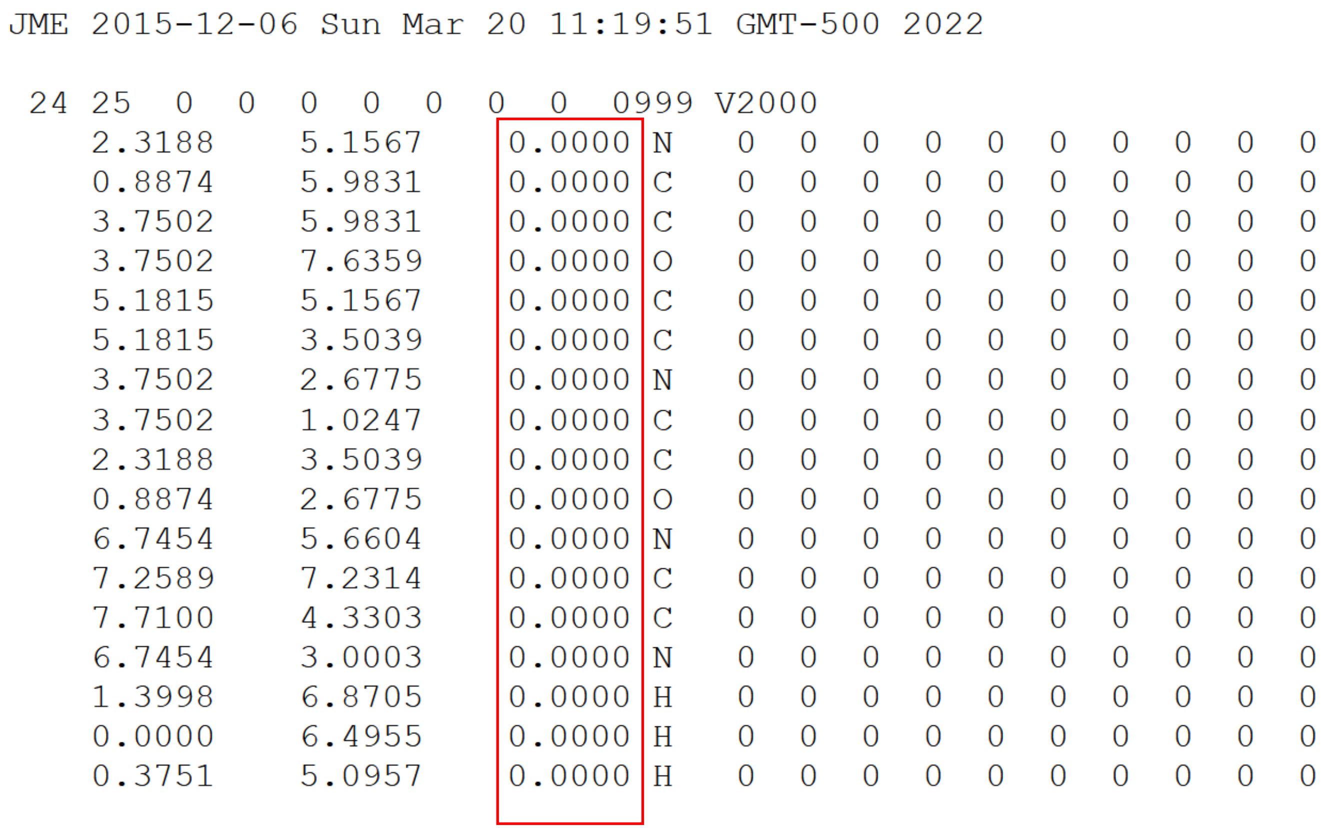
… a 3D MOL file

```
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

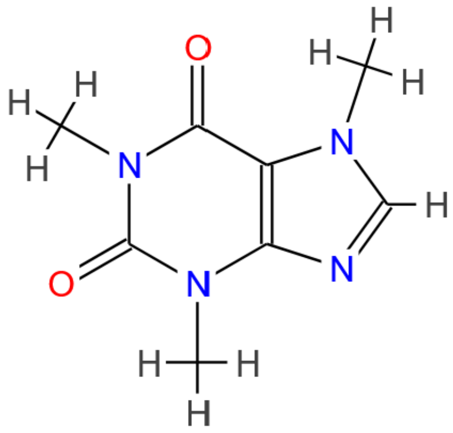


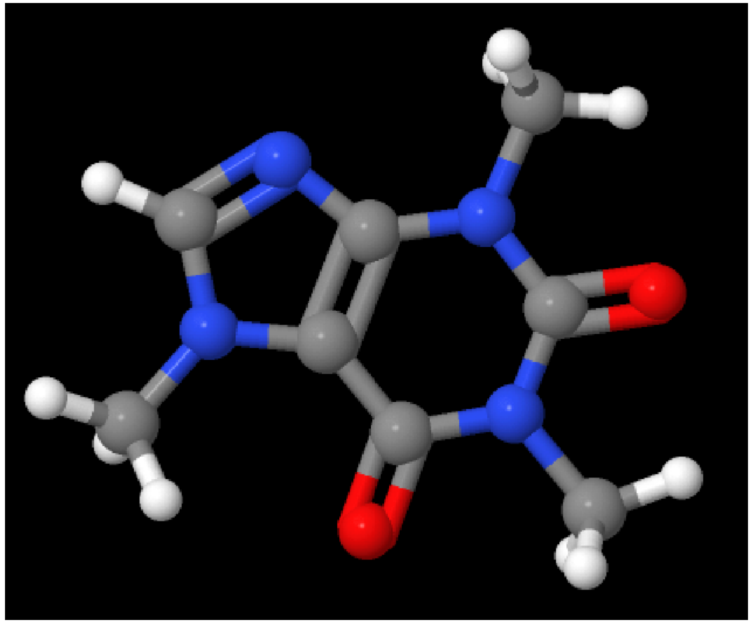
```
JME 2015-12-06 Sun Mar 20 11:19:51 GMT-500 2022

 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"
```

The IUPAC FAIRData Metadata Object Model

1. Digital Entities and Digital Objects
2. Representations and Properties
3. Aggregations, Associations, and Collections
4. The IUPAC FAIRData Collection
5. The Pieces of the Puzzle
6. 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

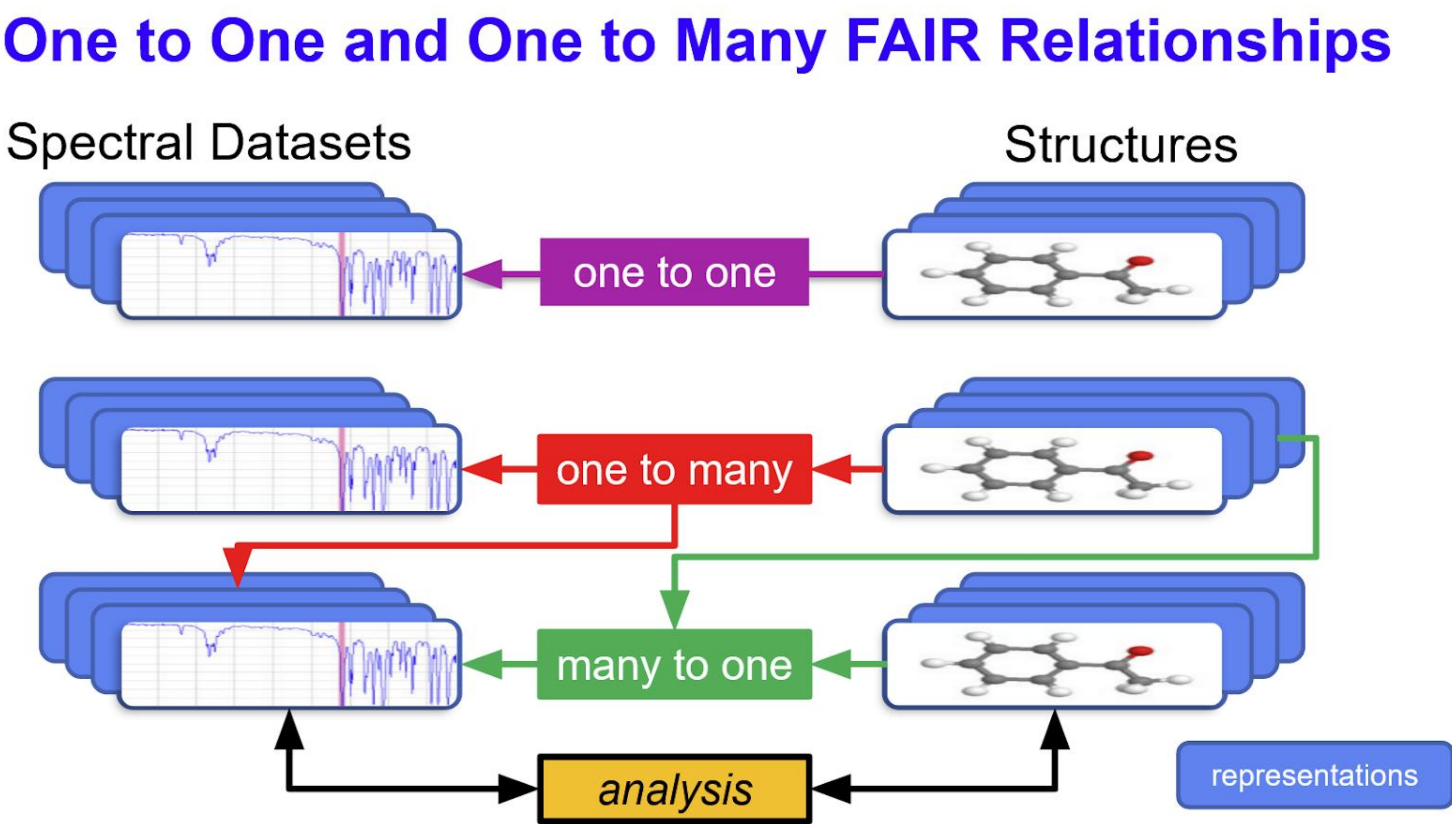
ACS Aggregation	Size (MB)		digital entities	
	(zip)	(raw)	files	type
joc.0c00770	25	37	720	11 cmpd dirs; 24 Bruker datasets & 12 mnova files
orglett.0c00874	27	40	1616	36 cmpd dirs; 76 Bruker datasets
orglett.0c00967	29	41	1354	33 cmpd dirs; 62 Bruker datasets
orglett.0c01022	15	52	66	2 dirs; 64 mnova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

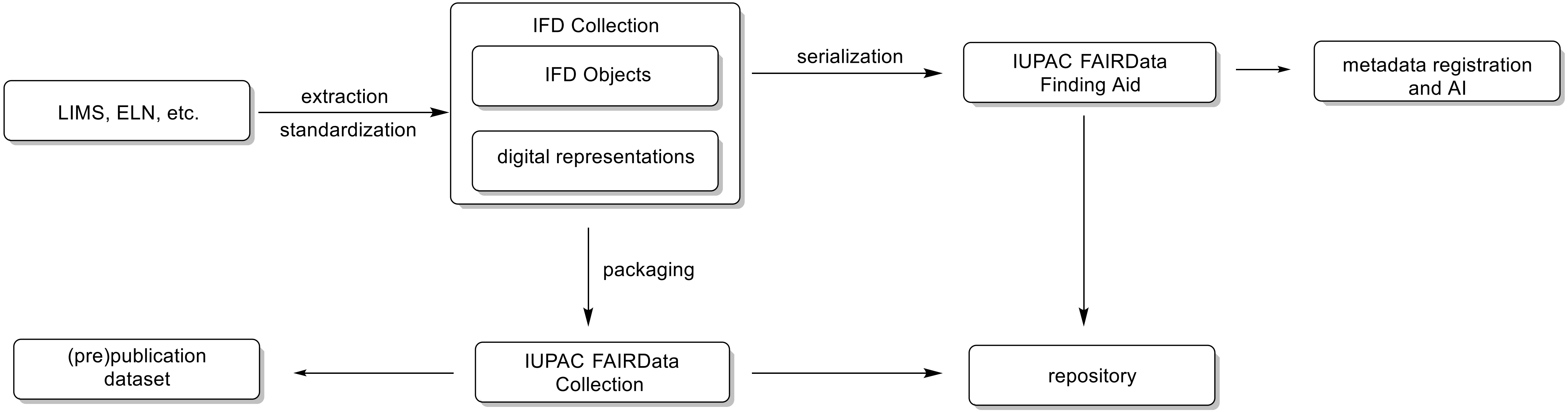
1. Digital Entities and Digital Objects
2. Representations and Properties
3. Aggregations, Associations, and Collections
4. The IUPAC FAIRData Collection
5. The Pieces of the Puzzle
6. The Full Enterprise

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

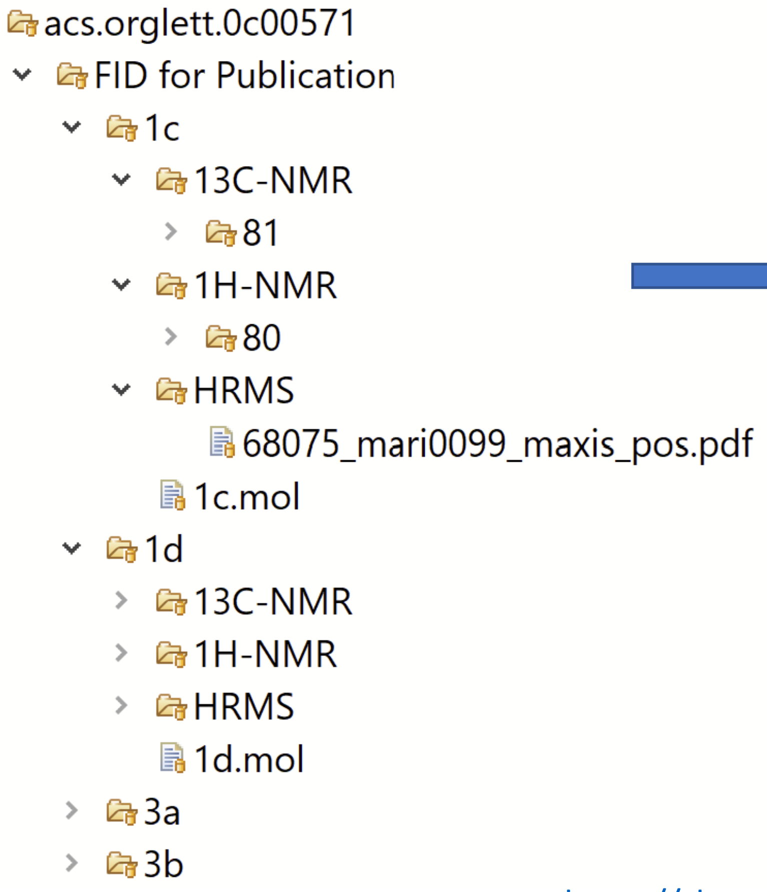


The IUPAC FAIRData Finding Aid

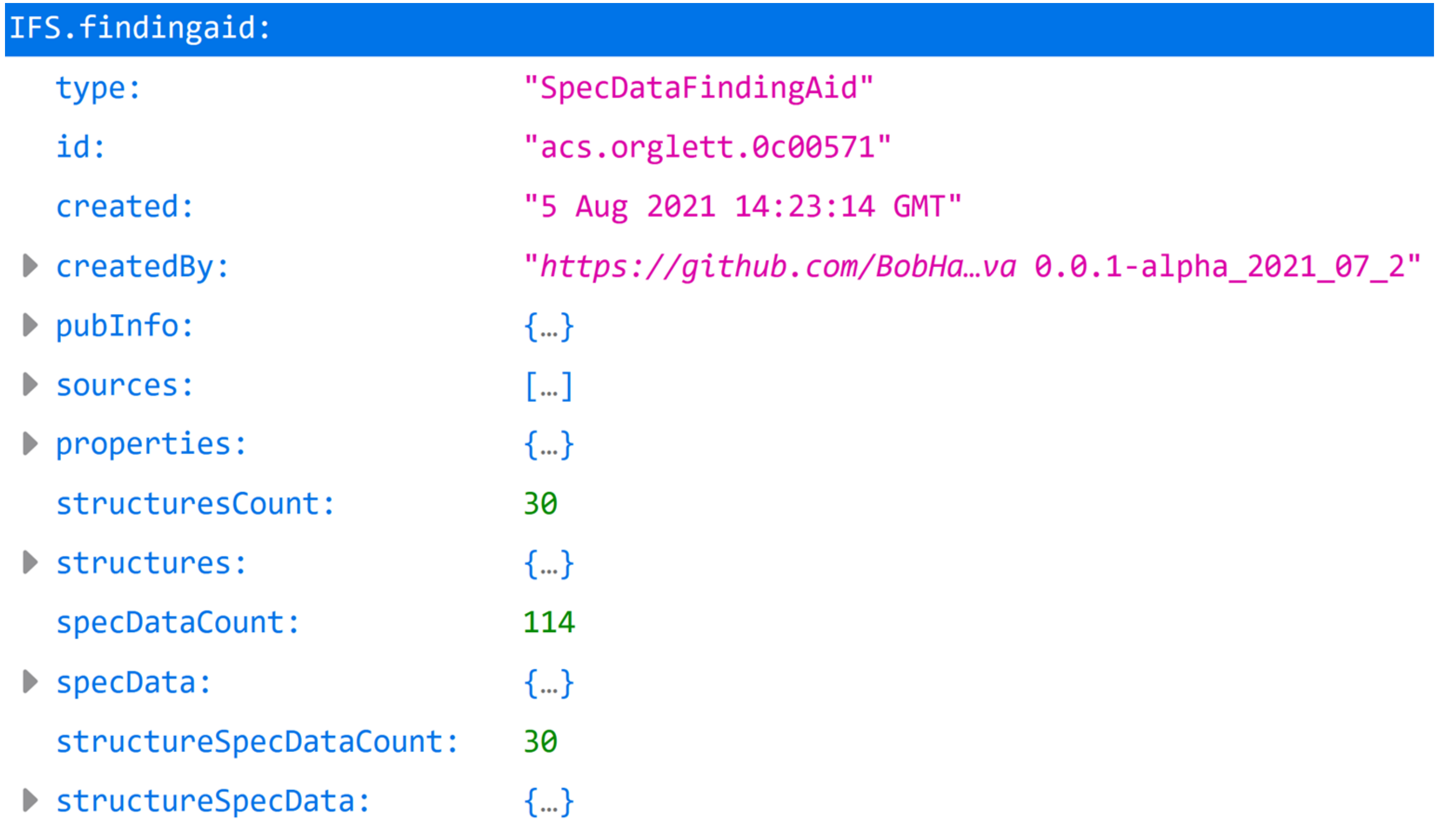
IUPAC FAIRData Finding Aid

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



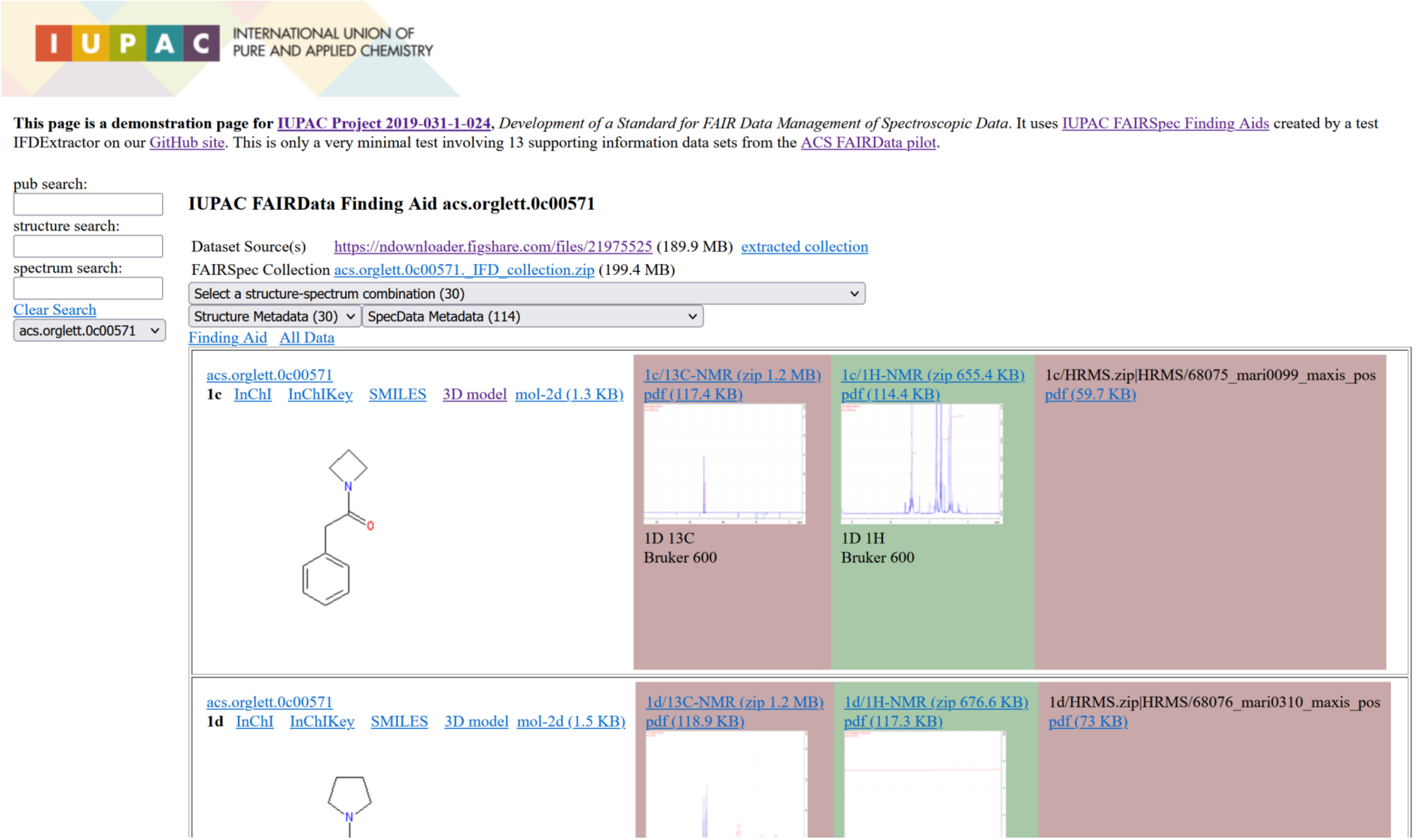
```
acs.orglett.0c00571
  ⌄ FID for Publication
      ⌄ 1c
          ⌄ 13C-NMR
              › 81
          ⌄ 1H-NMR
              › 80
          ⌄ HRMS
                68075_mari0099_maxis_pos.pdf
            1c.mol
      ⌄ 1d
          › 13C-NMR
          › 1H-NMR
          › HRMS
            1d.mol
      › 3a
      › 3b
```



```
IFS.findingaid:
    type:                      "SpecDataFindingAid"
    id:                        "acs.orglett.0c00571"
    created:                   "5 Aug 2021 14:23:14 GMT"
  ▶ createdBy:                 "https://github.com/BobHa…va 0.0.1-alpha_2021_07_2"
  ▶ pubInfo:                   {…}
  ▶ sources:                   […]
  ▶ properties:                {…}
    structuresCount:           30
  ▶ structures:                {…}
    specDataCount:             114
  ▶ specData:                  {…}
    structureSpecDataCount:    30
  ▶ structureSpecData:         {…}
```

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

(early) Preliminary Data Model -- the "IUPAC FAIRData Finding Aid"



INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

This page is a demonstration page for IUPAC Project 2019-031-1-024, *Development of a Standard for FAIR Data Management of Spectroscopic Data*. It uses IUPAC FAIRSpec Finding Aids created by a test IFDExtractor on our GitHub site. This is only a very minimal test involving 13 supporting information data sets from the ACS FAIRData pilot.

pub search:

structure search:

spectrum search:

Clear Search

acs.orglett.0c00571

IUPAC FAIRData Finding Aid acs.orglett.0c00571

Dataset Source(s) https://ndownloader.figshare.com/files/21975525 (189.9 MB) extracted collection

FAIRSpec Collection acs.orglett.0c00571._IFD_collection.zip (199.4 MB)

Select a structure-spectrum combination (30)

Structure Metadata (30) SpecData Metadata (114)

Finding Aid All Data

acs.orglett.0c00571
1c InChI InChIKey SMILES 3D model mol-2d (1.3 KB)

1c/13C-NMR (zip 1.2 MB)
pdf (117.4 KB)
1D 13C
Bruker 600

1c/1H-NMR (zip 655.4 KB)
pdf (114.4 KB)
1D 1H
Bruker 600

1c/HRMS.zip|HRMS/68075_mari0099_maxis_pos
pdf (59.7 KB)

acs.orglett.0c00571
1d InChI InChIKey SMILES 3D model mol-2d (1.5 KB)

1d/13C-NMR (zip 1.2 MB)
pdf (118.9 KB)

1d/1H-NMR (zip 676.6 KB)
pdf (117.3 KB)

1d/HRMS.zip|HRMS/68076_mari0310_maxis_pos
pdf (73 KB)

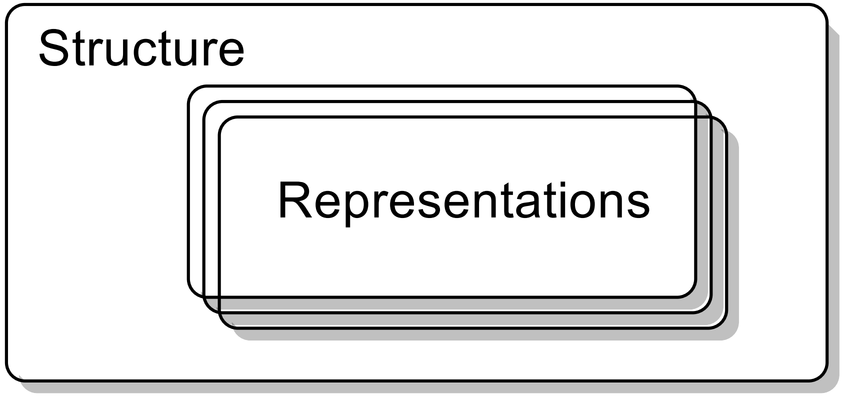
https://chemapps.stolaf.edu/iupac/demo/demo.htm?pub=571

The IUPAC FAIRData Metadata 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

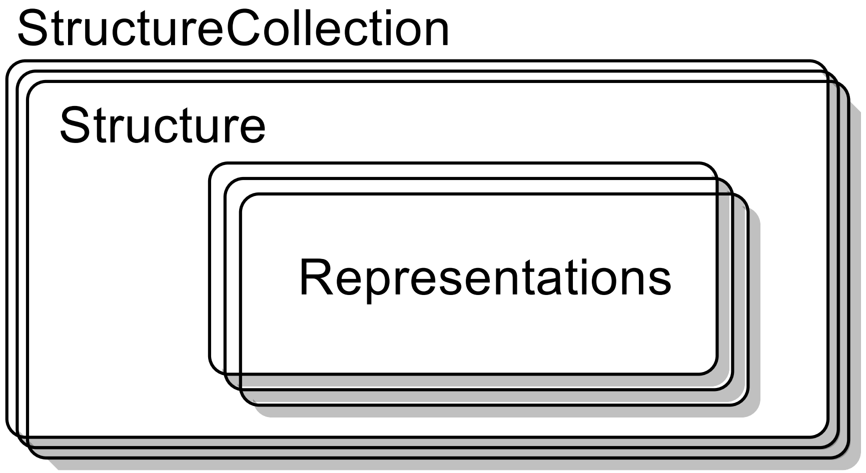
The Pieces of the Puzzle

a structure with its associated representations



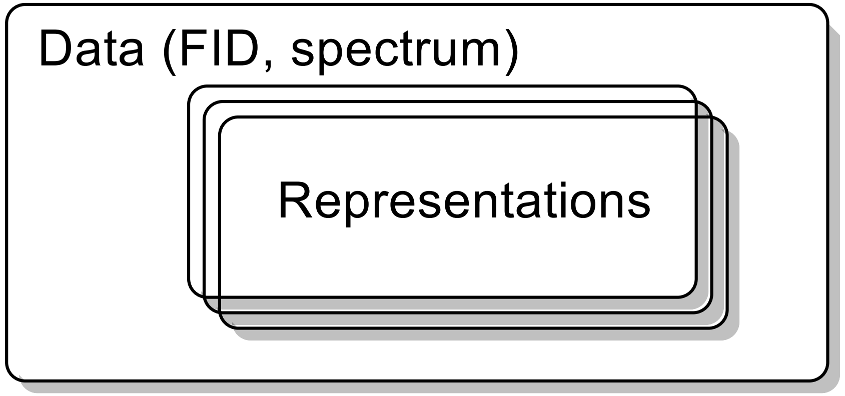
The Pieces of the Puzzle

a collection of structures



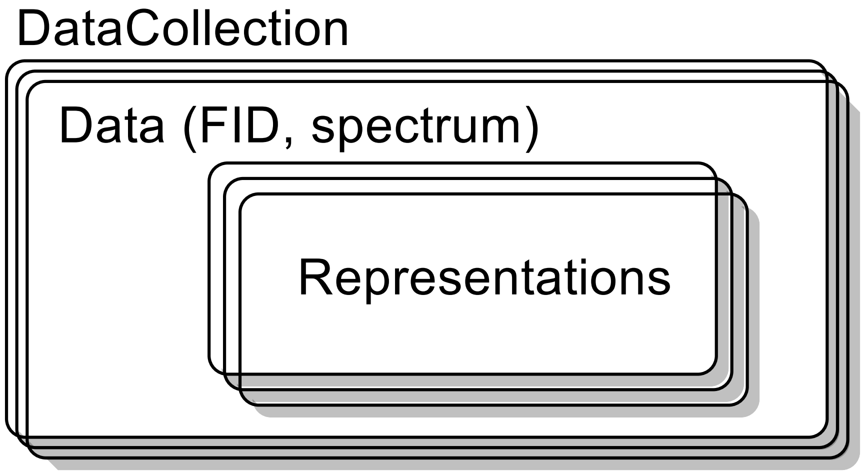
The Pieces of the Puzzle

spectroscopic data



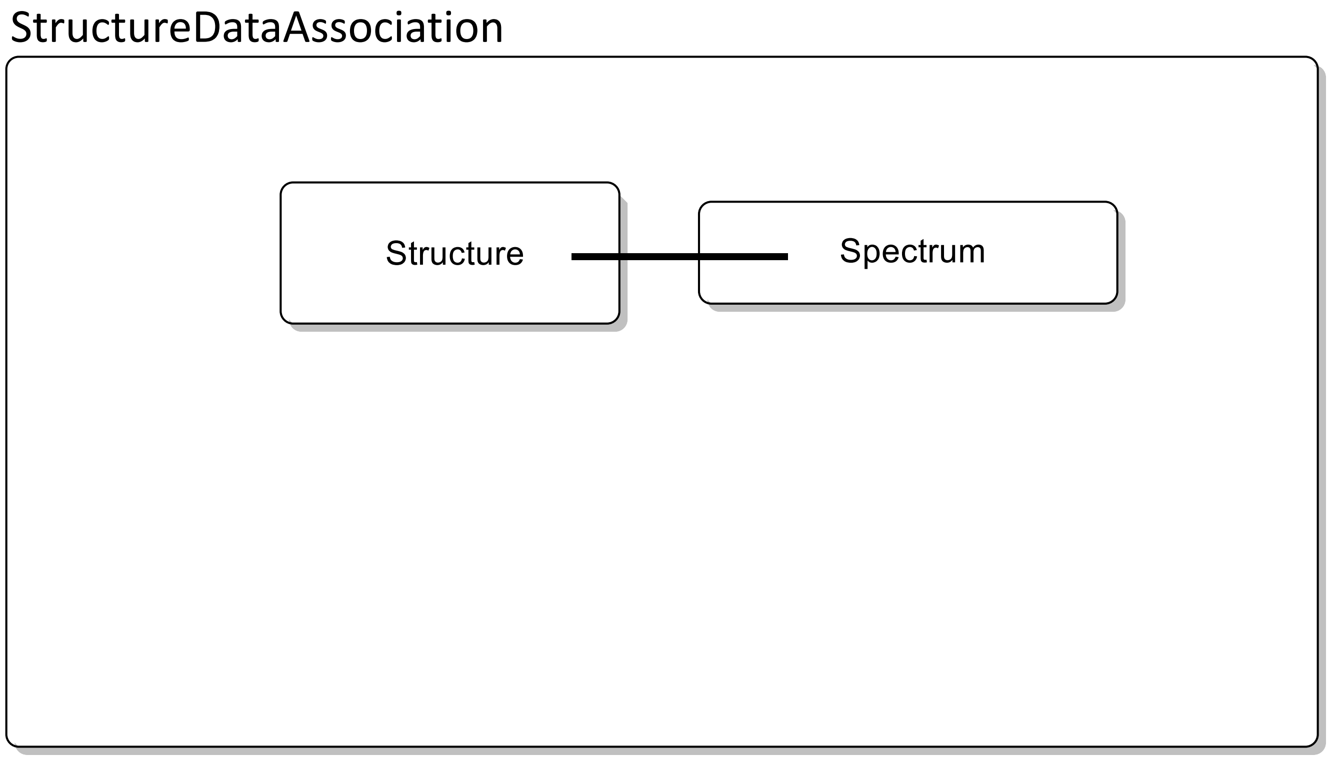
The Pieces of the Puzzle

a collection of spectra



The Pieces of the Puzzle

a simple structure – spectrum association



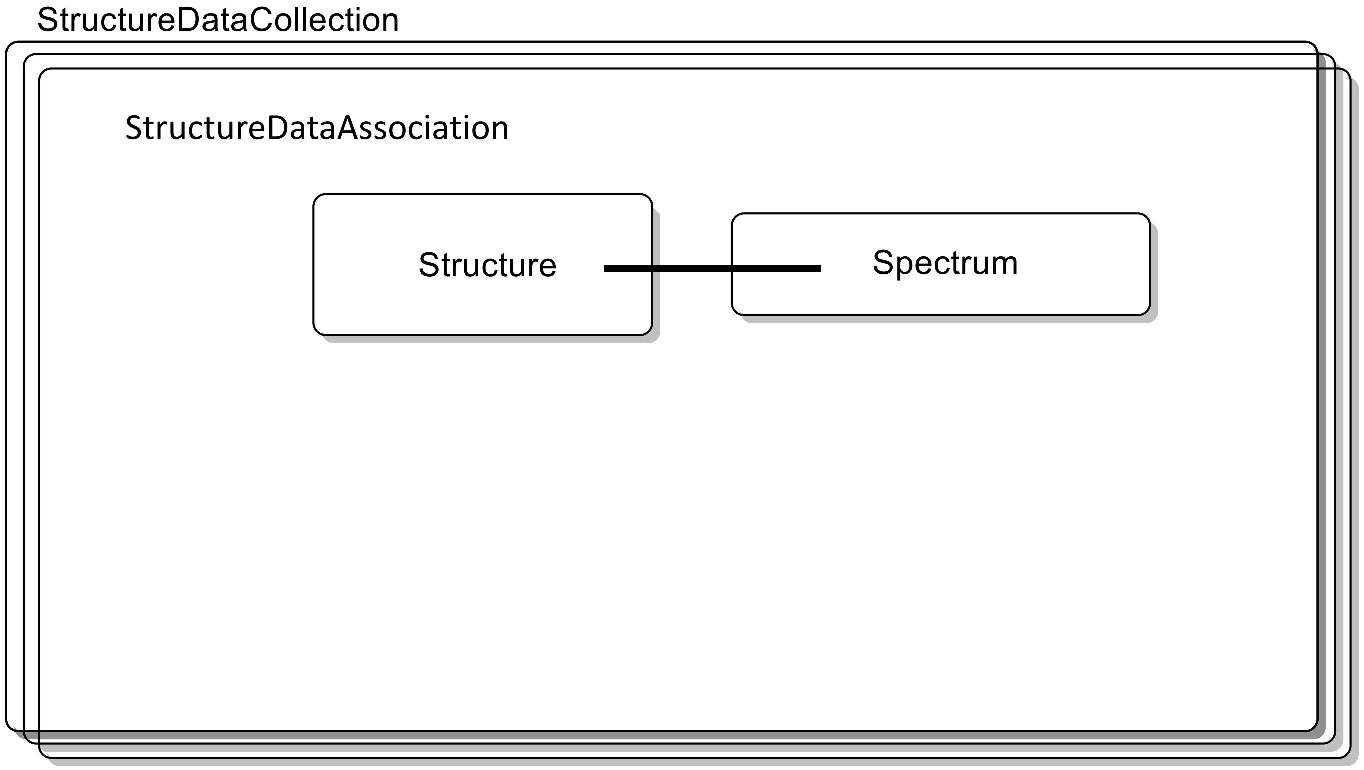
StructureDataAssociation

Structure

Spectrum

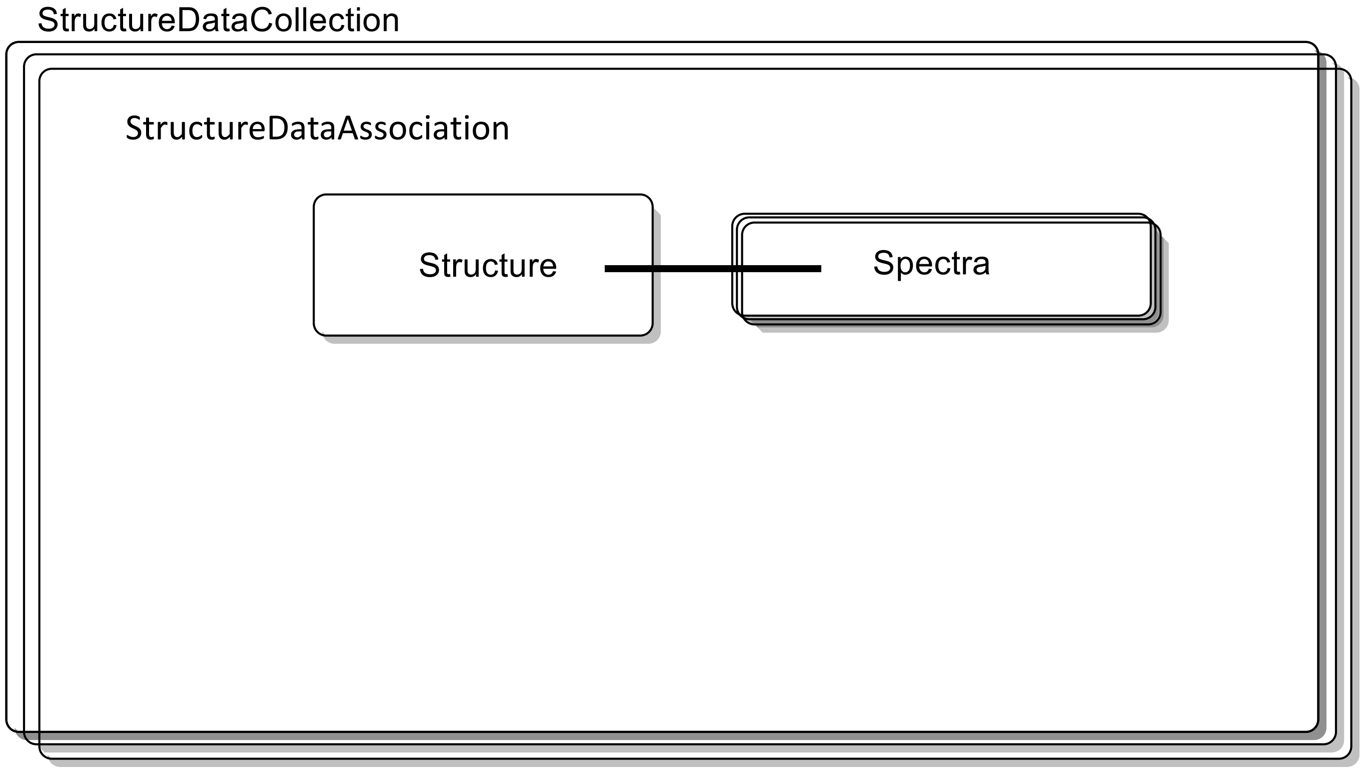
The Pieces of the Puzzle

a collection of simple structure – spectrum associations



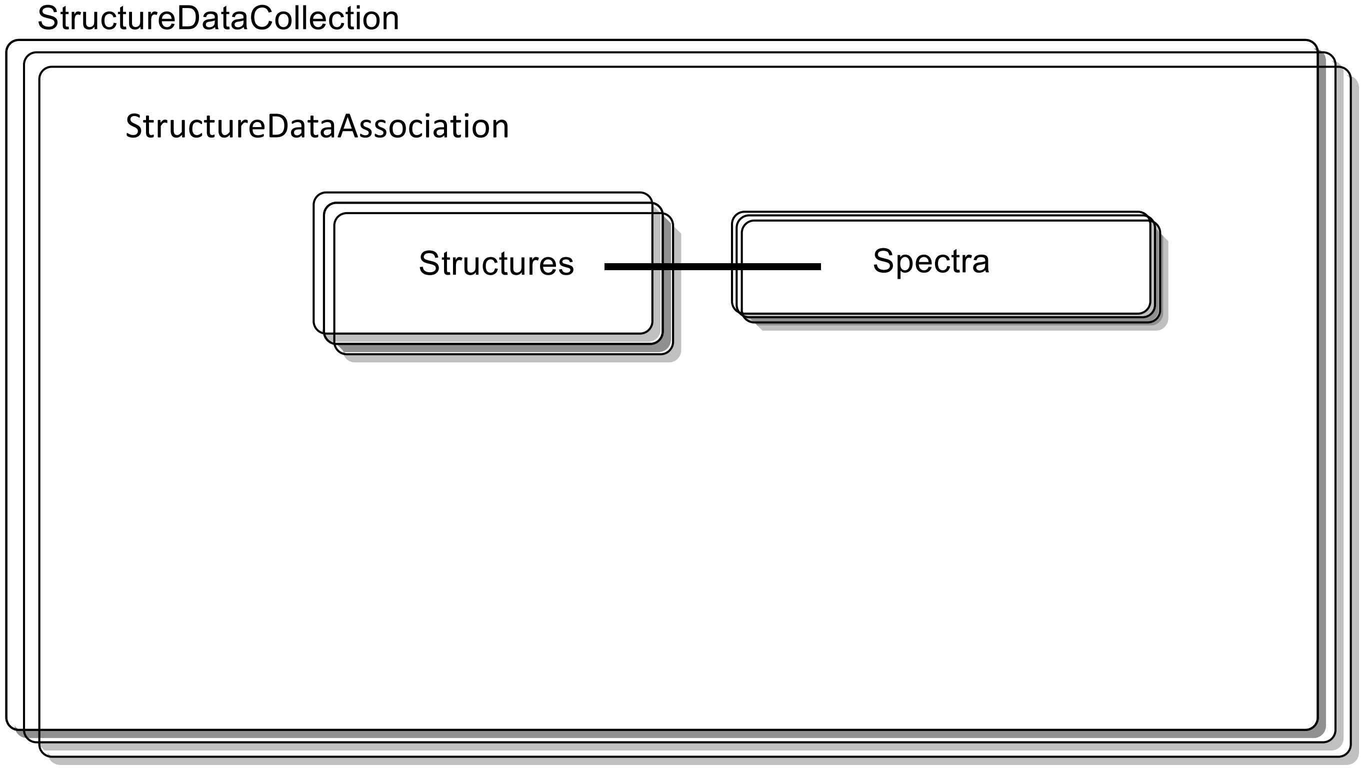
The Pieces of the Puzzle

a more typical collection of structure – spectra associations



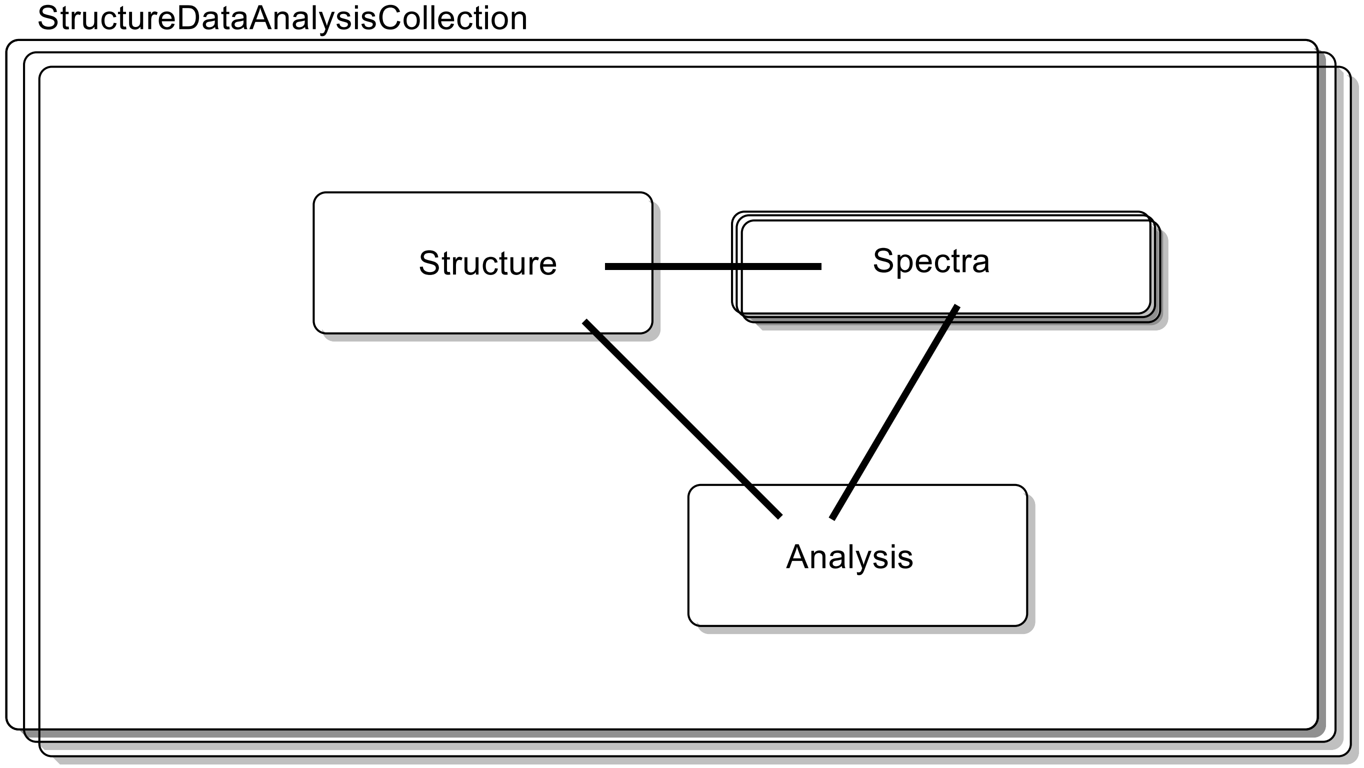
The Pieces of the Puzzle

allowing for mixtures



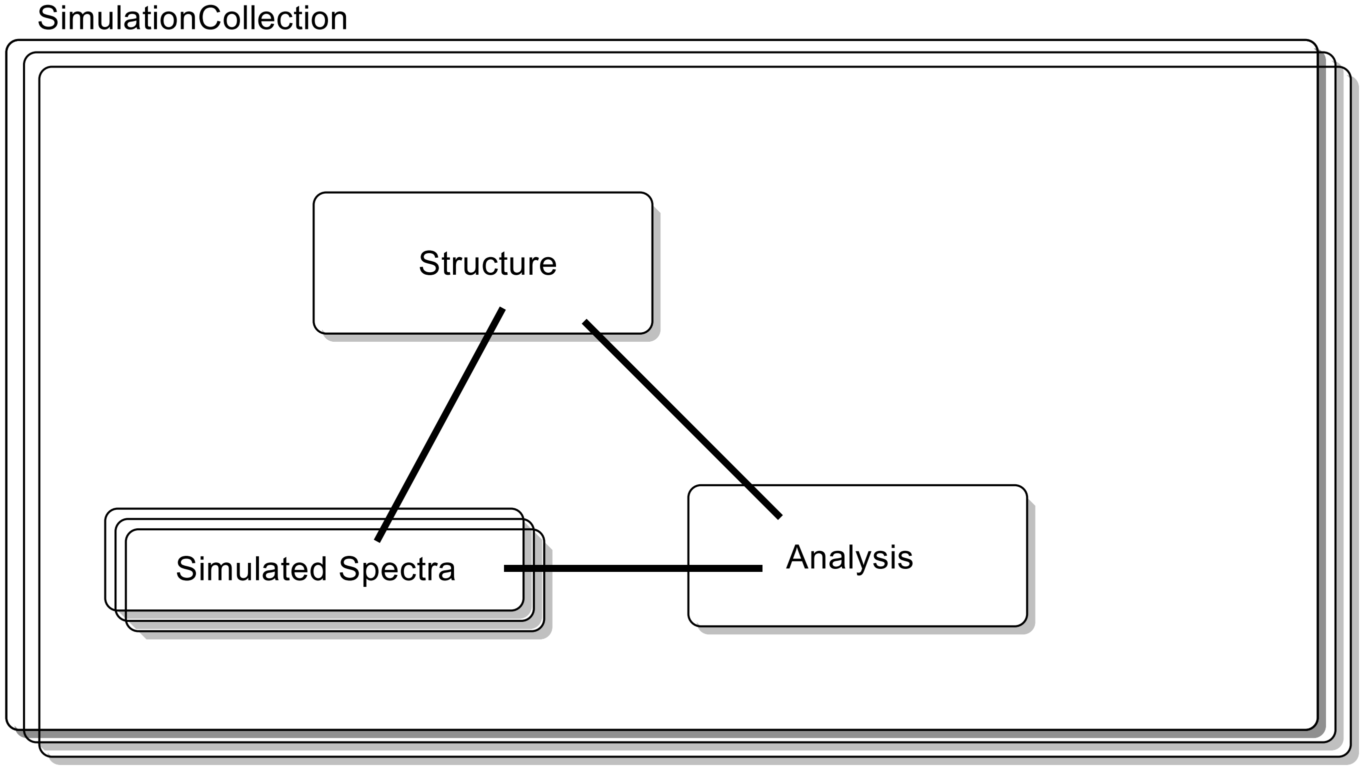
The Pieces of the Puzzle

adding analysis



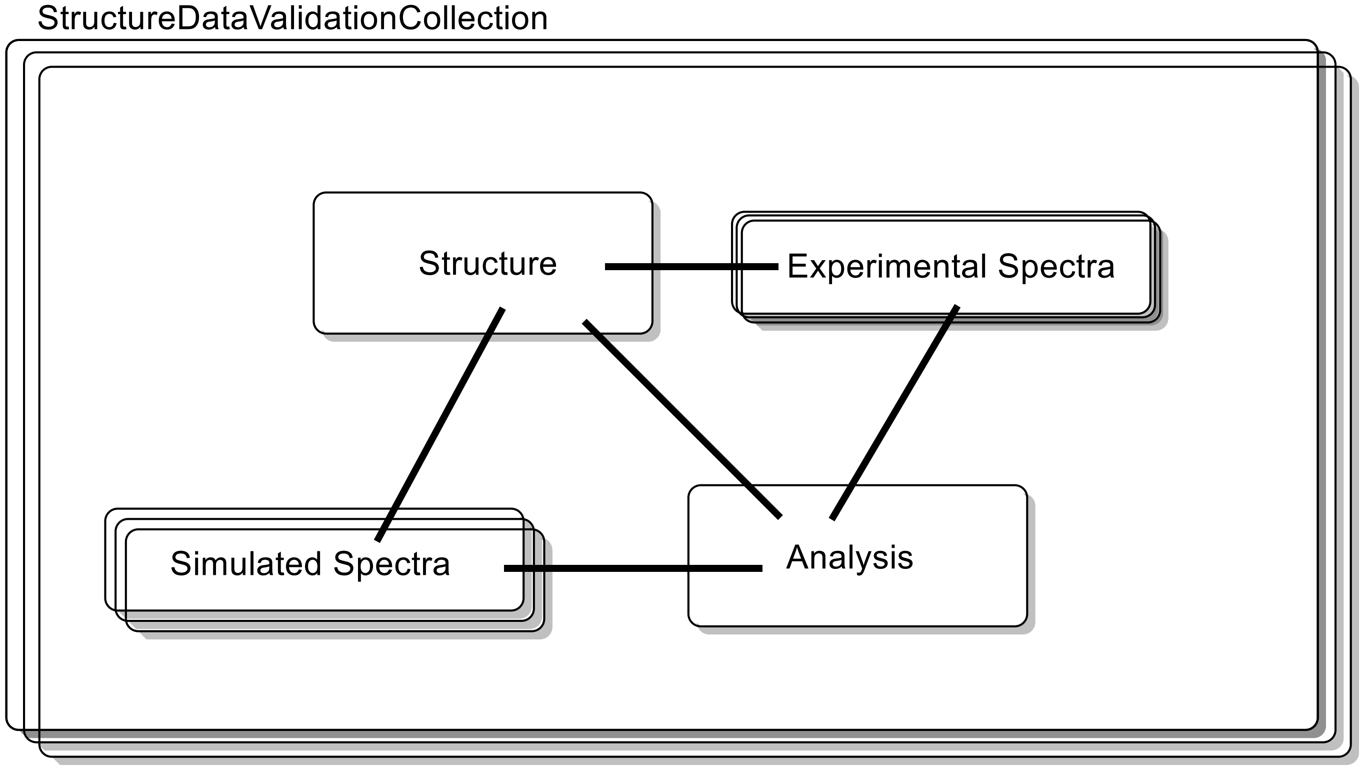
More Pieces of the Puzzle

a simulation



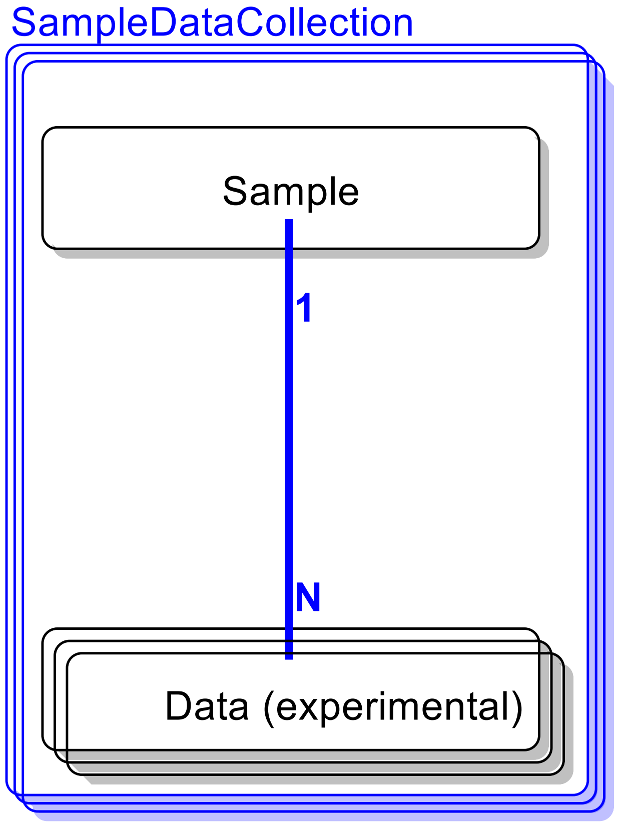
The Pieces of the Puzzle

adding simulation



More Pieces of the Puzzle

a collection of samples and their associated spectra

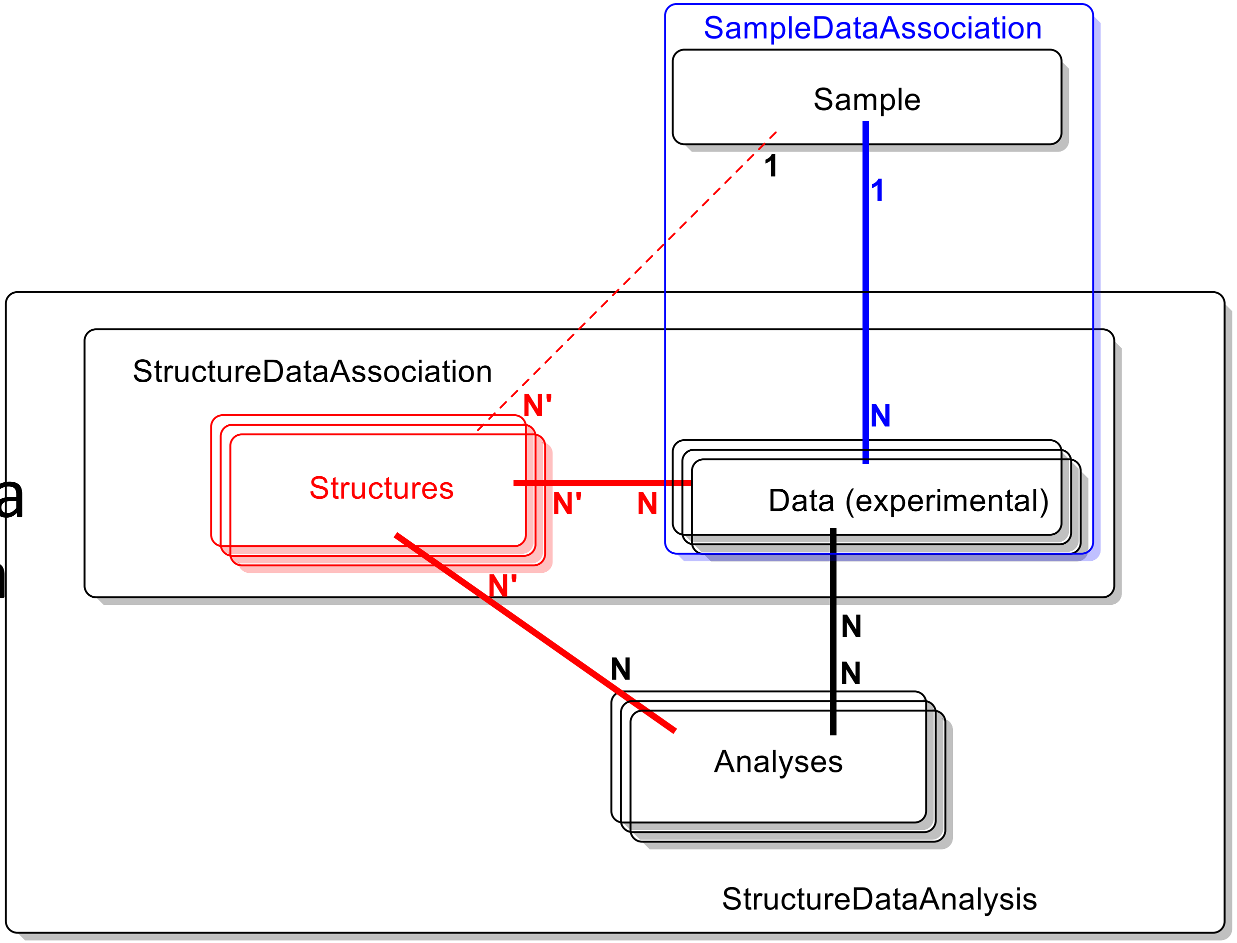


More Pieces of the Puzzle

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

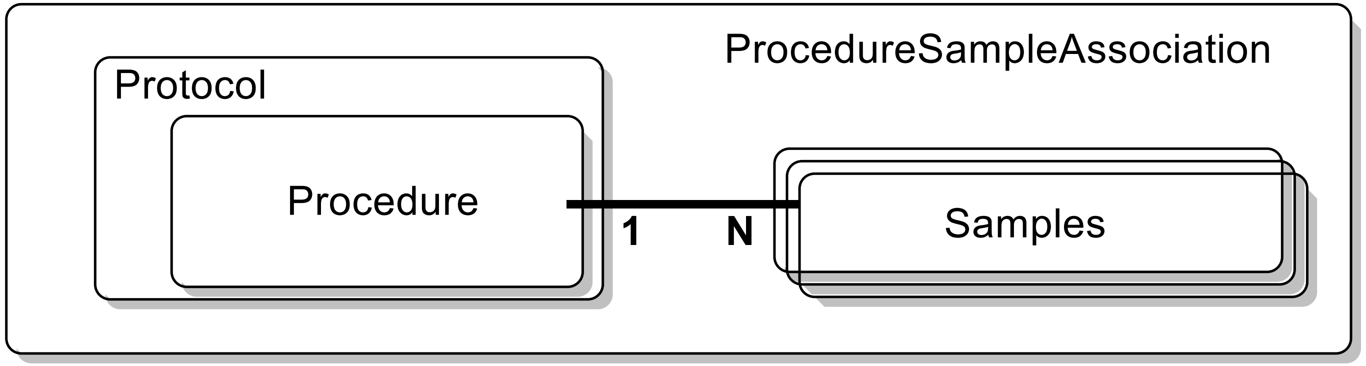
The inference that a given sample is a compound with a given structure is a product of this analysis.

The result may not always be 1:1.



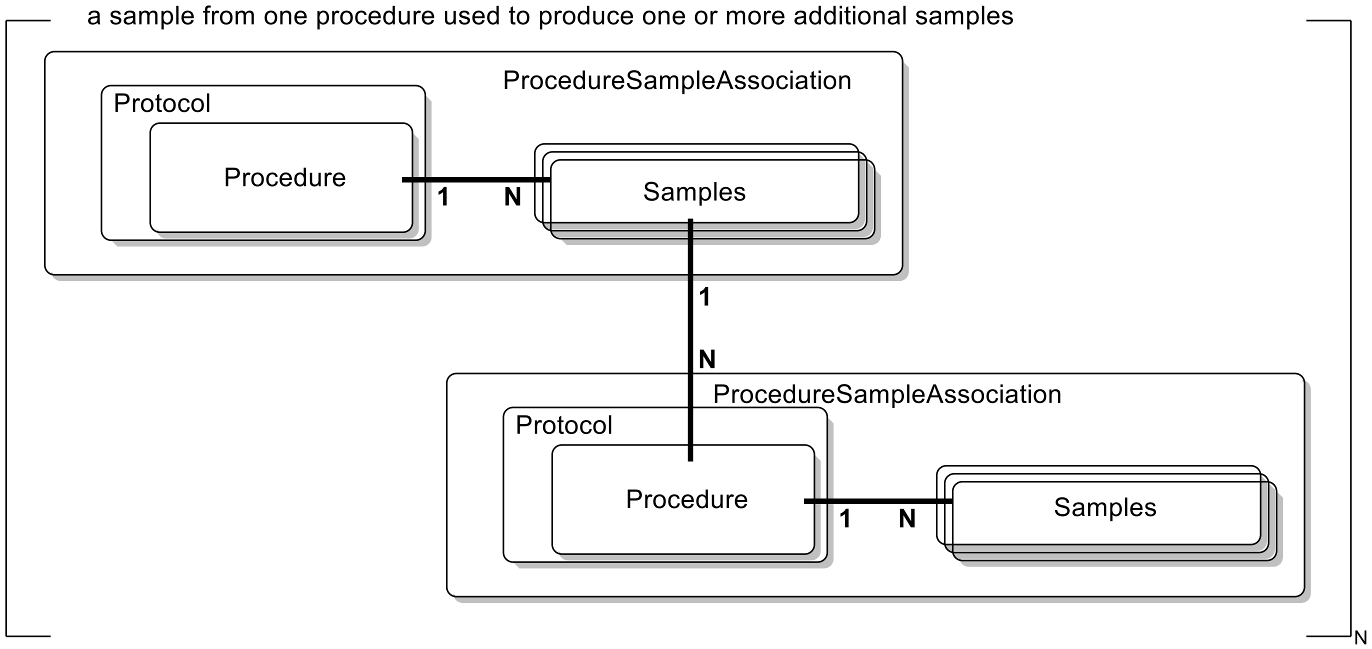
More Pieces of the Puzzle

a procedure based on a protocol producing one or more samples



More Pieces of the Puzzle

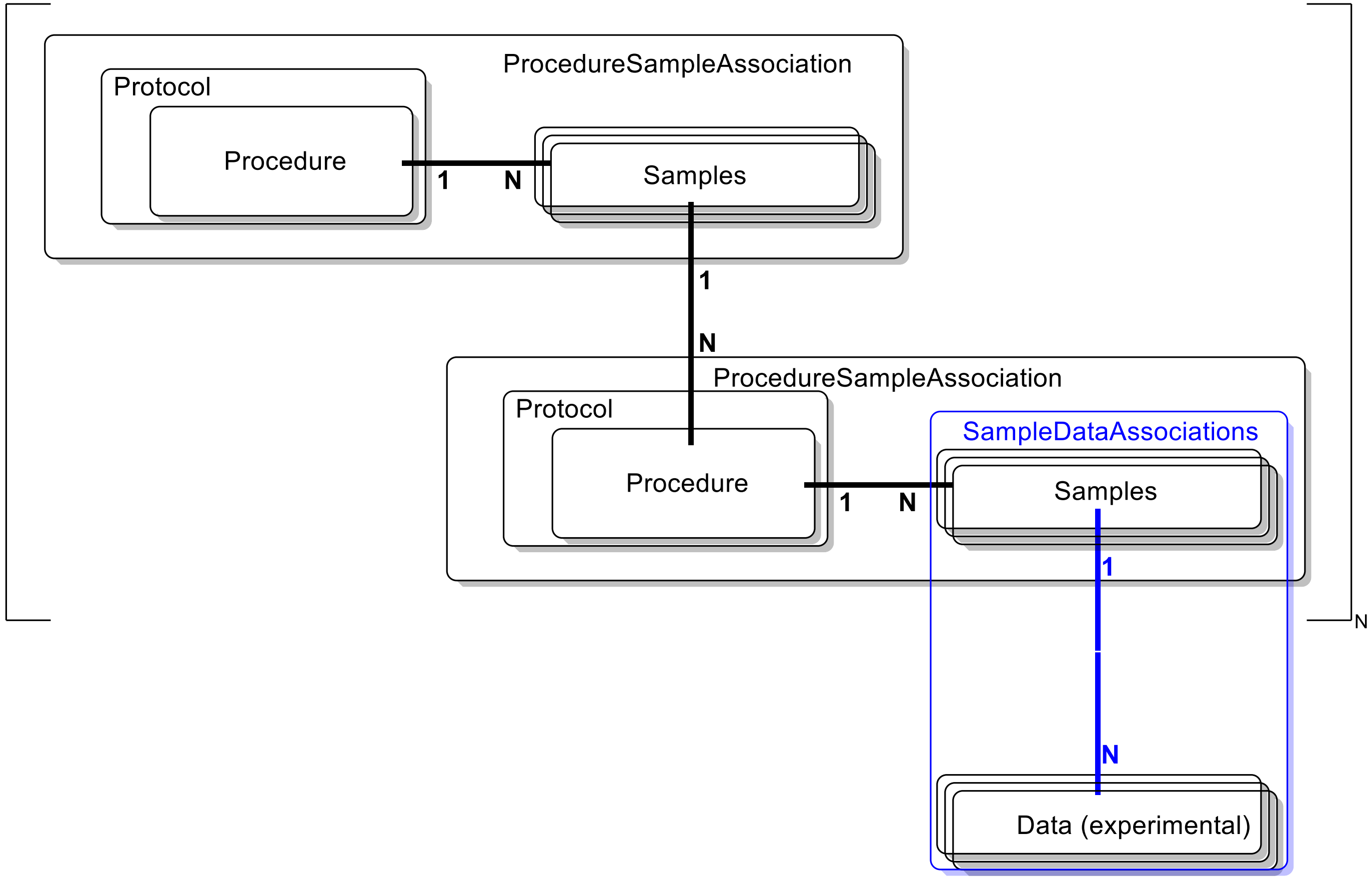
a sample from one procedure used to produce one or more additional samples



a sample from one procedure used to produce one or more additional samples
ProcedureSampleAssociation
Protocol
Procedure
1
N
Samples
1
N
ProcedureSampleAssociation
Protocol
Procedure
1
N
Samples
N

The ELN Piece

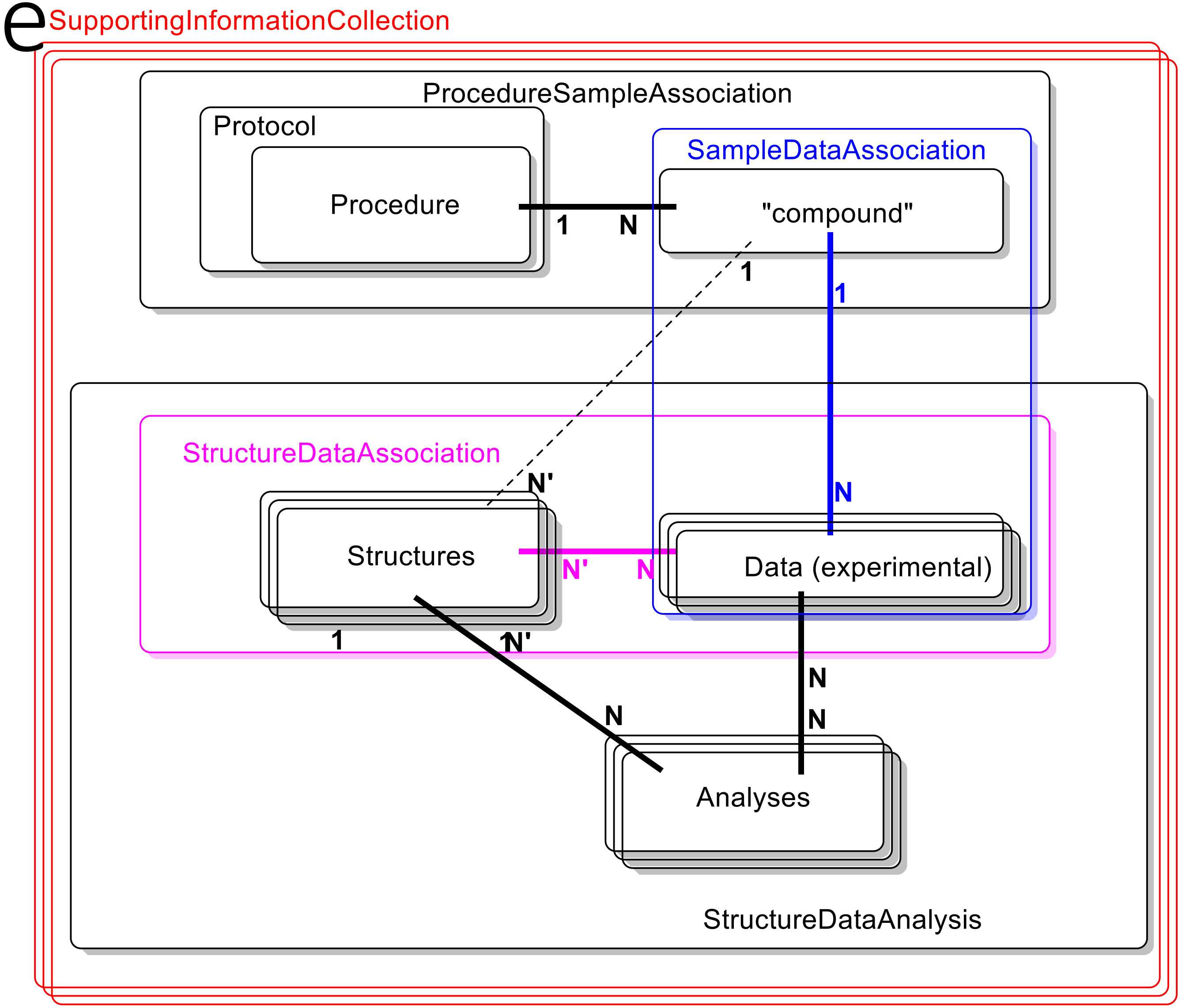
Electronic laboratory notebooks implementing IUPAC FAIRSpec Recommendations could provide the needed sample-data association.



The Publication Piece

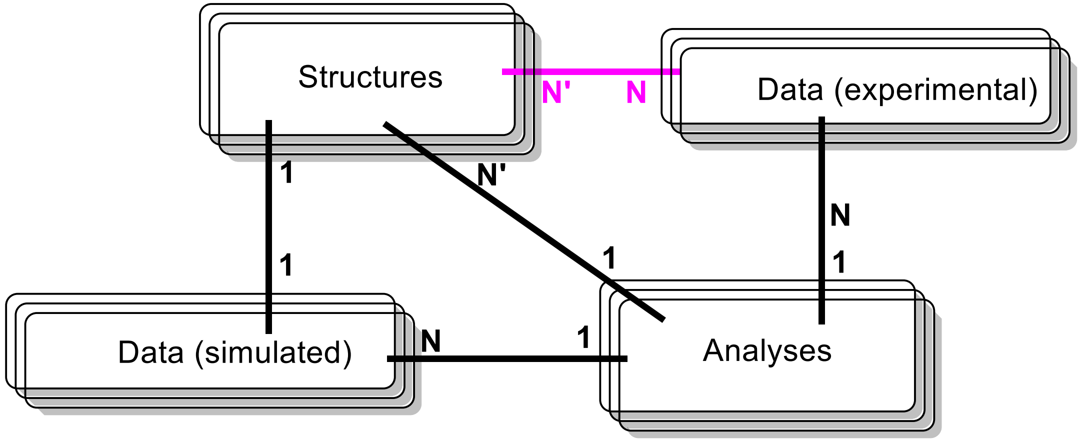
The "supporting information" for a publication in chemistry could be one possible representation of an IUPAC FAIRData Collection.

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



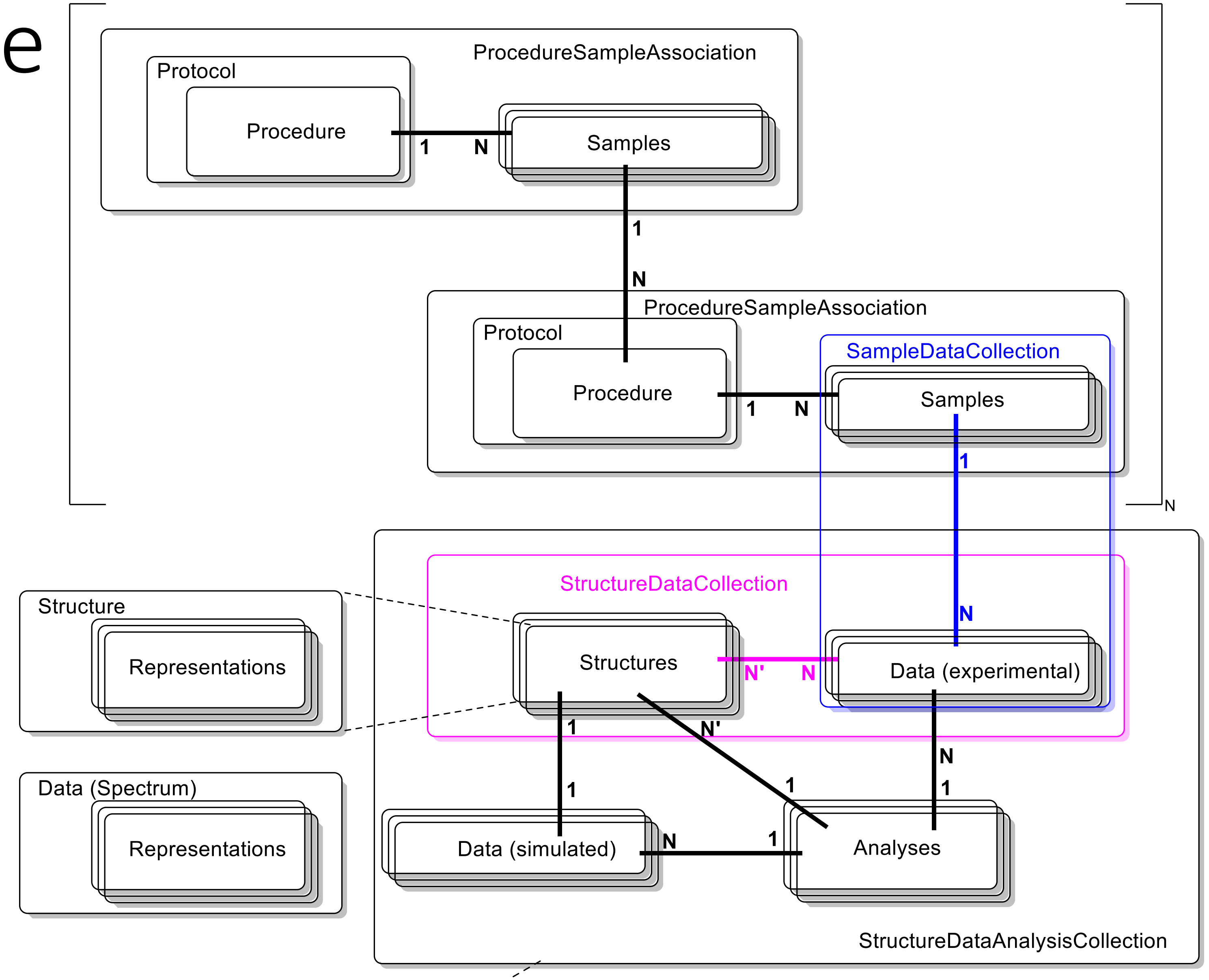
The Validation Piece

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



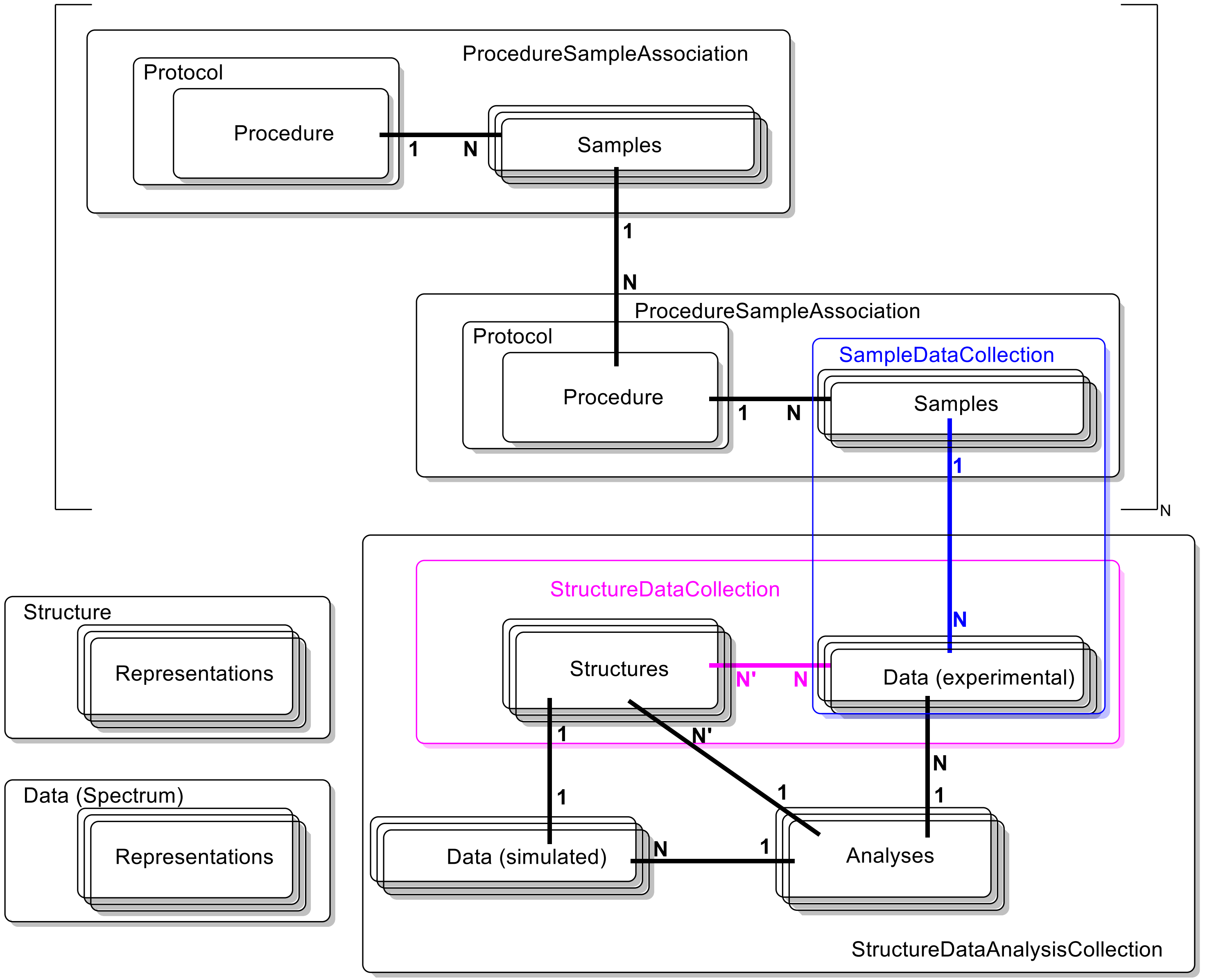
The Repository Piece

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 Metadata Object Model

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



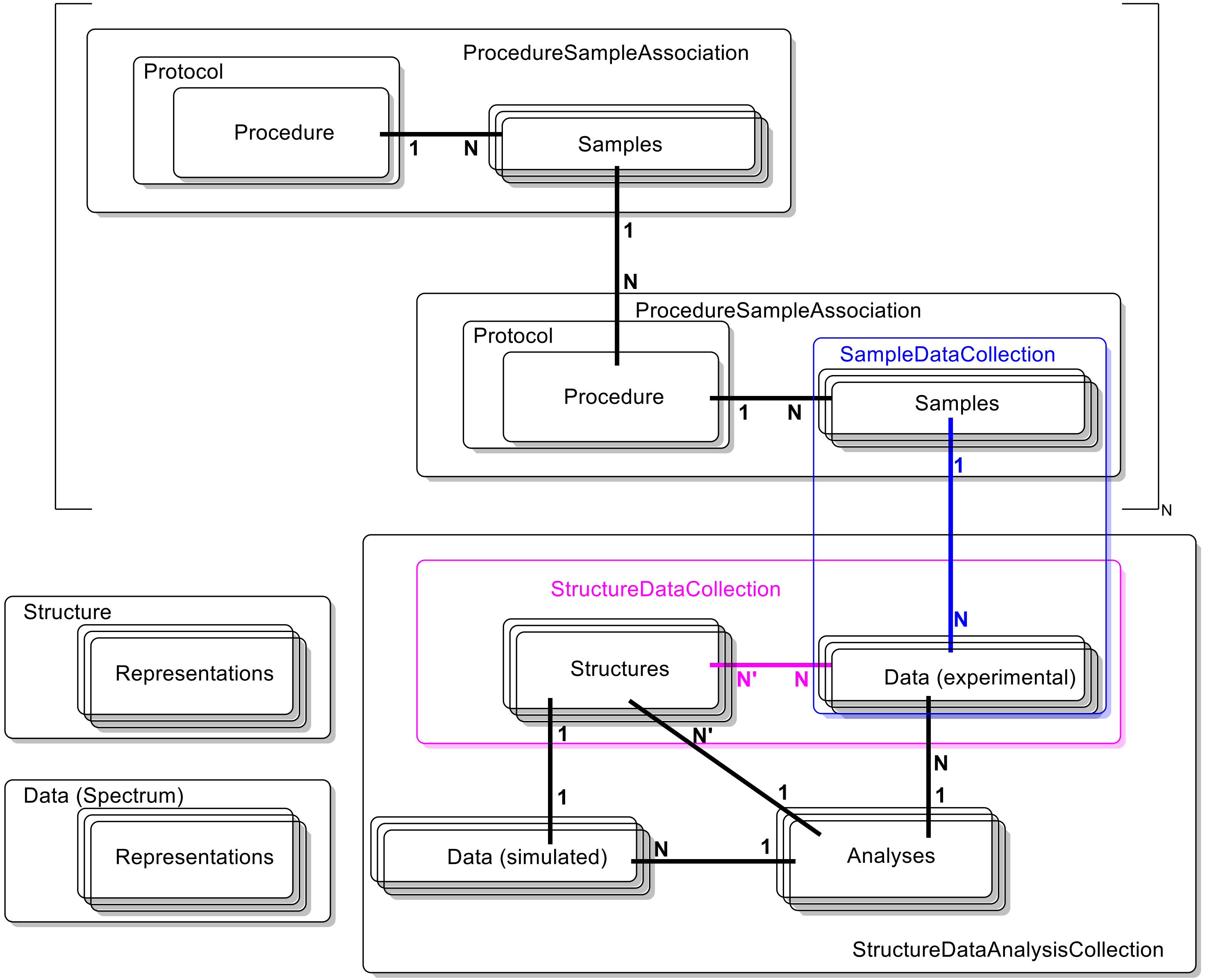
The proposed standards involve several aspects:

- **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 standard for describing properties of digital objects** within the metadata records of the finding aid
- **A standard for the serialization of the finding aid** for an IUPAC FAIRData Collection
- **A proposal for methods of data and metadata extraction** and the generation of IUPAC FAIRData Finding Aids
- **A recommendation for the organization of digital objects** within a 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.

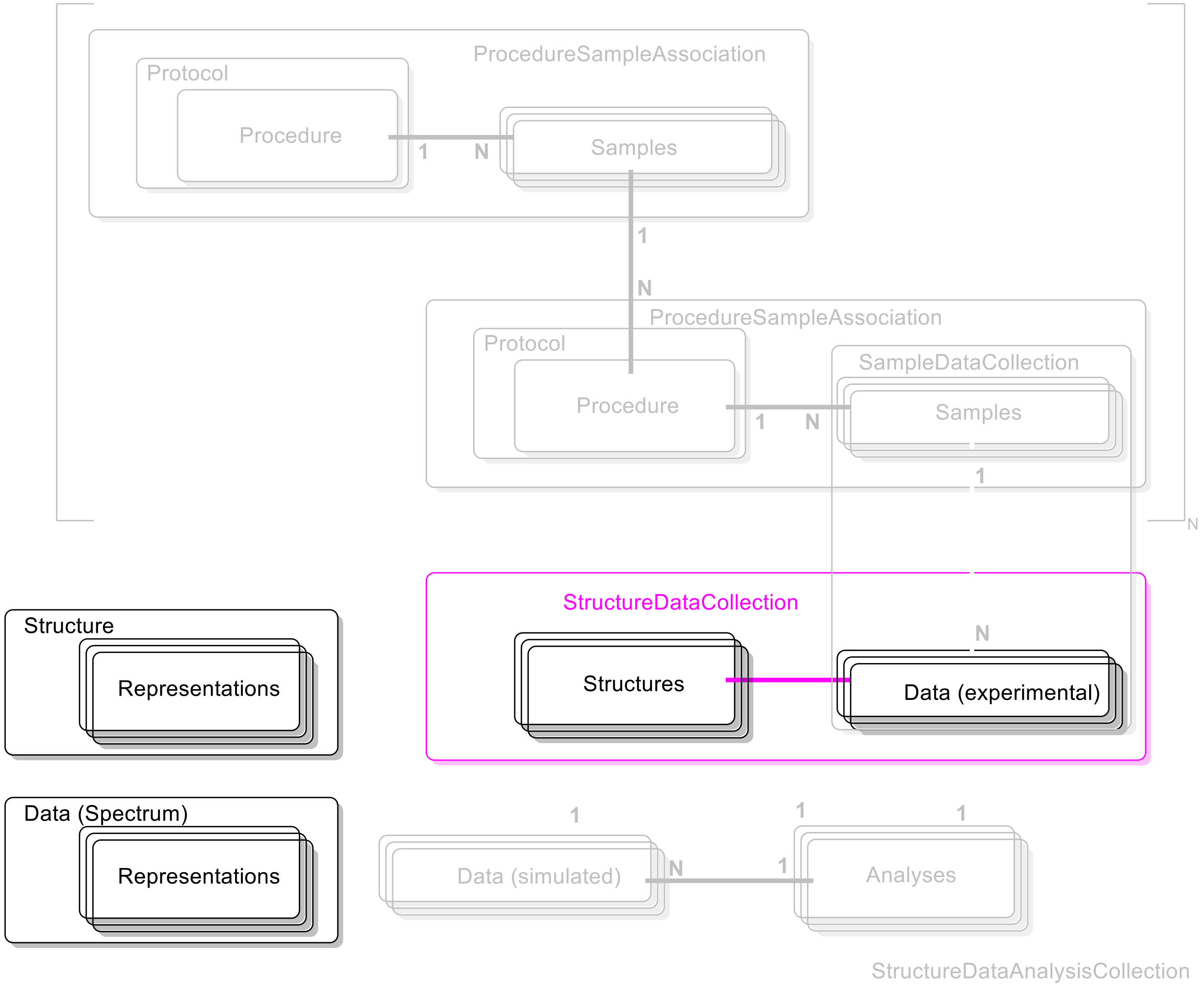


In Summary

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

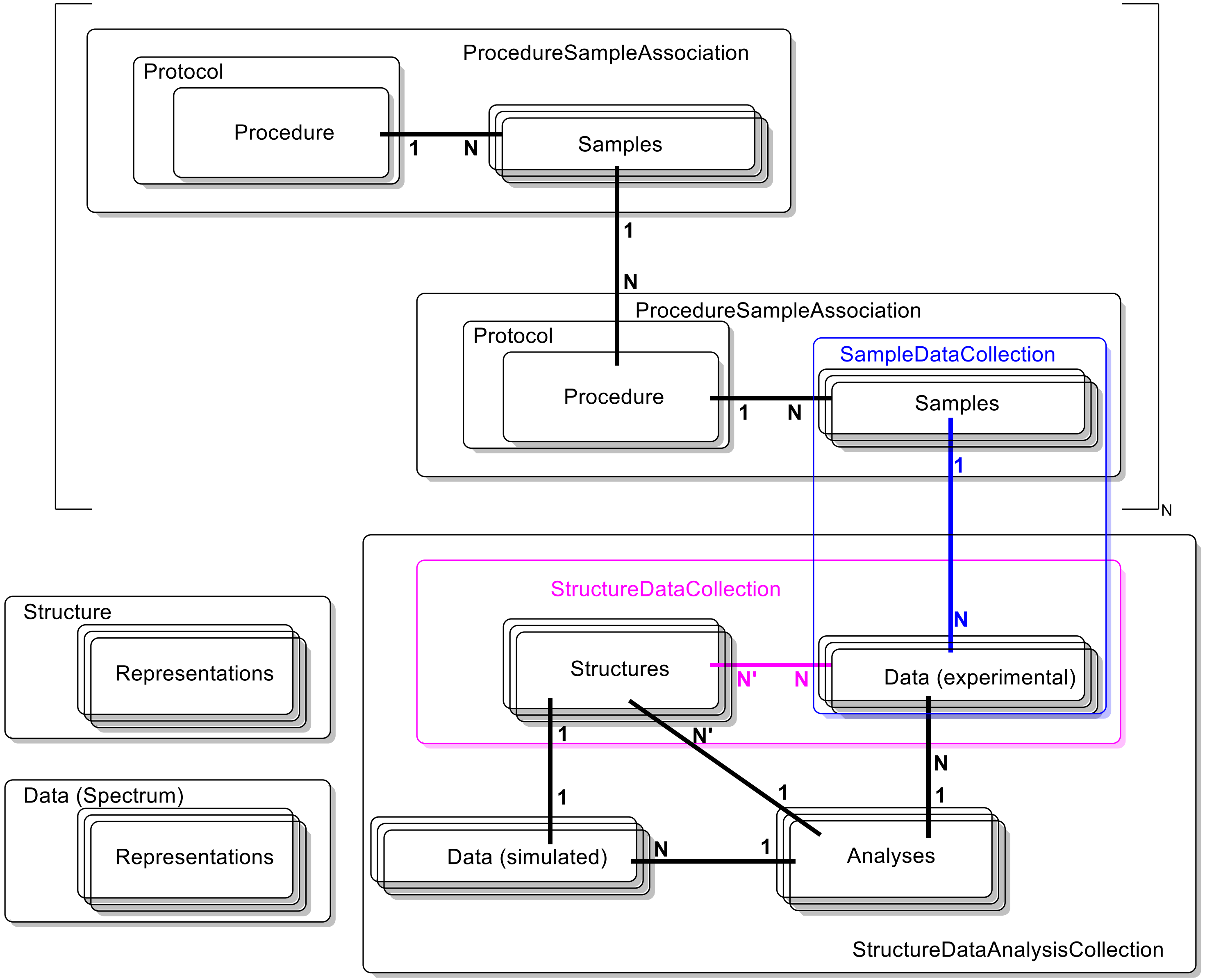
Our project scope and expertise is in the area of structure-spectra collections.

These are the pieces we will develop.



In Summary

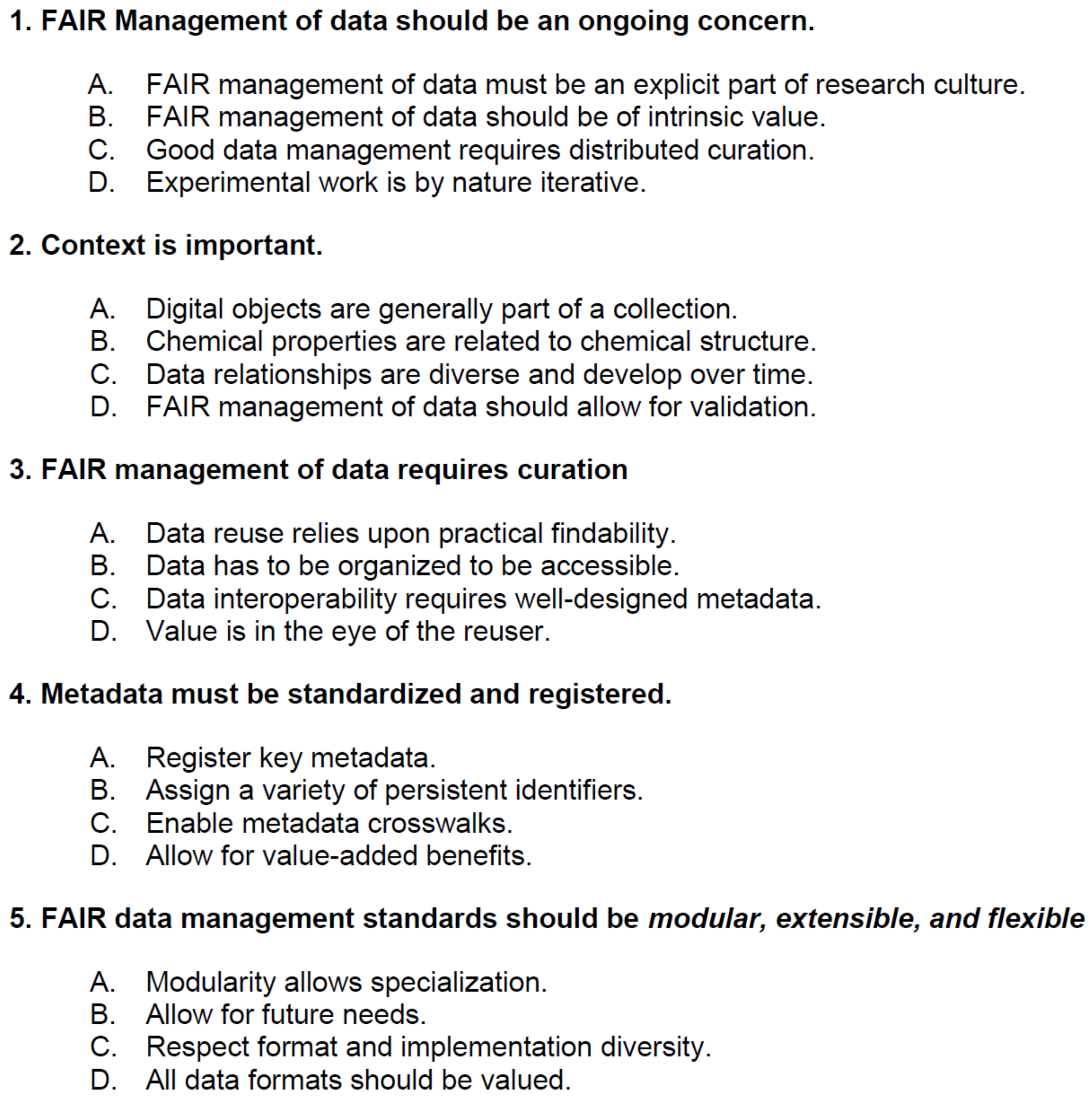
We hope that others with other expertise and perspectives will join us in this endeavor to complete the puzzle and revolutionize the world of chemistry.



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.

Project Timeline (ambitious version)

- Mar 2020 – Dec 2021
    - COVID!
    - Vision development
    - Develop partnerships
    - FAIRSpec Principles development
    - Request and analyze author-submitted datasets
- Jan 2022 – Jun 2022
    - Work with partners on details of recommendations
- Jun 2022 – Oct 2022
    - Preliminary recommendations for comment
    - Work with potential implementers
- Nov 2022 – Dec 2022
    - Finalize recommendations
- Jan 2023 – Dec 2023
    - Continue to collaborate closely with implementers
    - Continue to refine recommendations (V. 2)