INTERNATIONAL UNION OF
PURE AND APPLIED CHEMISTRY

IUPAC General Assembly August 2021
Committee on Publications and Cheminformatics Data Standards

Progress Report: IUPAC FAIR standard for spectroscopic data description & management

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

INTERNATIONAL UNION OF
PURE AND APPLIED CHEMISTRY

Bob Hanson

Damien Jeannerat

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

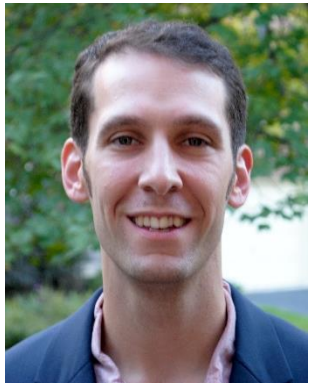
Mark Archibald

Ian Bruno

Stuart Chalk

Tony Davies

Robert Lancashire



Jeff Lang

Henry Rzepa

March 2020 Project Initiation – Two-year time frame

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

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

March 2020 Project Initiation – Two-year time frame

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

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

first year **design**; second year **build**

Design Decisions

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

1. This project is not about the creation of any new file formats
2. We will not limit ourselves to one specific spectroscopic technique.
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.)
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 Decisions

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

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
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)

The IUPAC FAIRSpec Specification, Version 0.0.1

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

[draft version 2021.08.06]

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

https://docs.google.com/document/d/1WYB3f04dFdVzlvf7aEwdVNwEwLpQ7YBAA00pGbc8Jp0/edit?usp=sharing

1. Basic IUPAC FAIRSpec Principles
2. Preferred Collection Organization
3. The IUPAC FAIRSpec Metadata Model
4. Data and Metadata Extraction and Serialization
5. The IUPAC FAIRSpec Data Model

1. Basic IUPAC FAIRSpec Principles

1.1 FAIR spectroscopic data management should be an ongoing concern

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

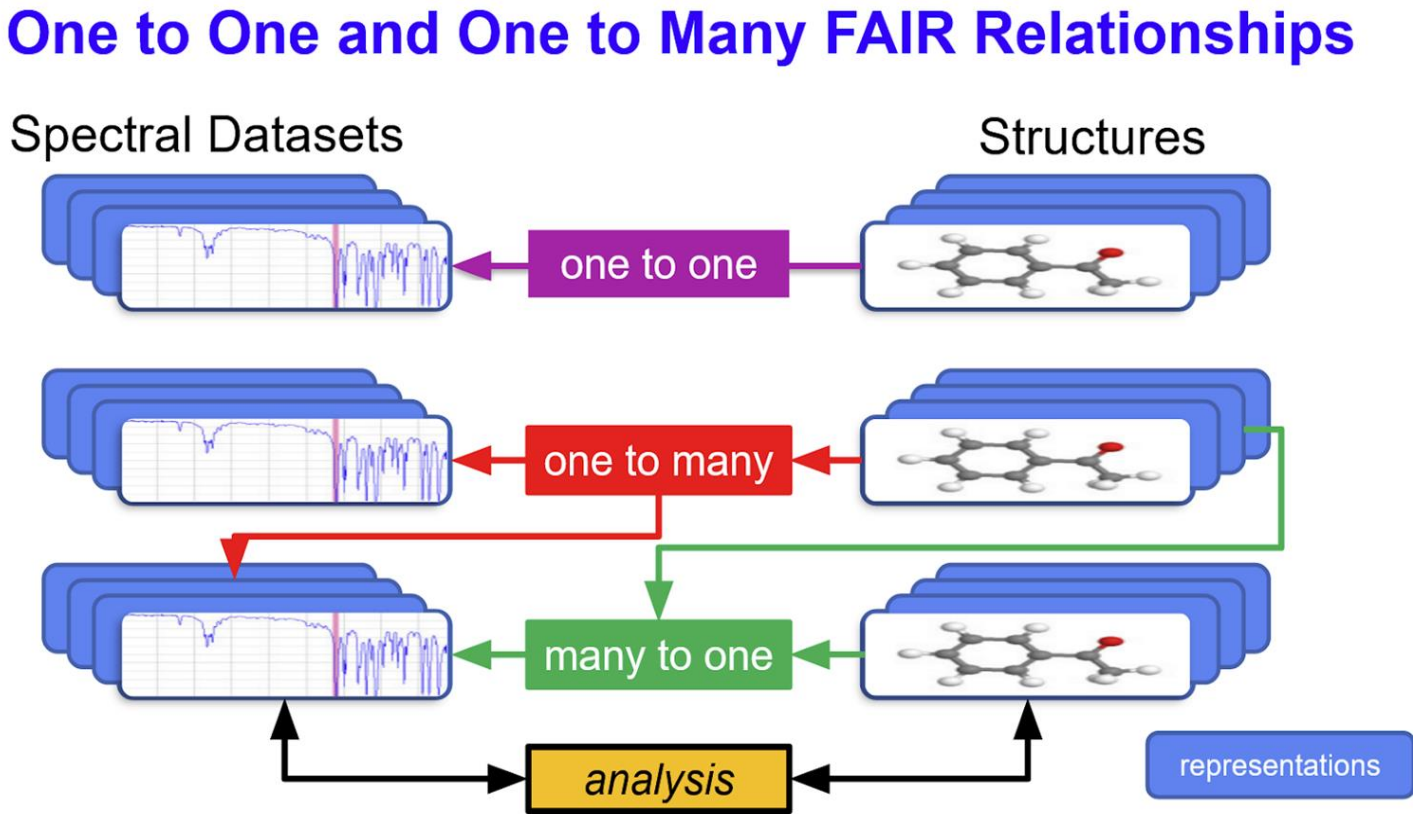
- part of the design of an **ongoing scientific endeavor**
- **intrinsic value** in real time to the originating research group(s)
- should **allow for cycles of data *generation*, (re)*processing*, and (re)*analysis***
- requires **distributed curation**

1. Basic IUPAC FAIRSpec Principles

1.2 Context is important

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

- emphasize the **value of a *collection***
- emphasize the **connection between chemical structure and chemical properties**
- allow for the fact that **spectroscopic relationships develop over time**



IFS Finding Aid Demonstrator

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

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

IUPAC 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 IFSExtractor 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

IFS Finding Aid acs.orglett.0c00571

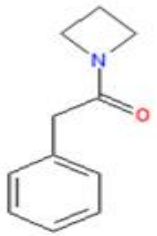
Title	Synthesis of Novel Heterocycles by Amide Activation and Umpolung Cyclization
Authors	Haoqi Zhang, Margaux Riomet, Alexander Roller, Nuno Maulide
url	https://pubs.acs.org/doi/10.1021/acs.orglett.0c00571
Dataset Source(s)	https://ndownloader.figshare.com/files/21975525 (189.9 MB) extracted collection
FAIRSpec Collection	acs.orglett.0c00571._IFS_collection.zip (199.4 MB)

Select a structure-spectrum combination (30)

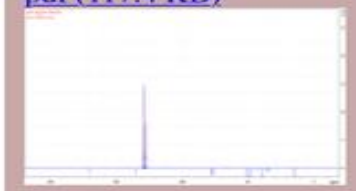
Structure Metadata (30) SpecData Metadata (114)

Finding Aid All Data

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

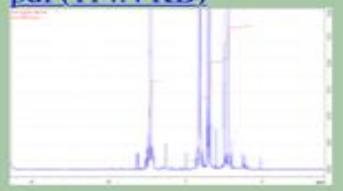


1c/13C-NMR (zip 1.2 MB)
pdf (117.4 KB)



1D 13C
Bruker 600
298.1525 K

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

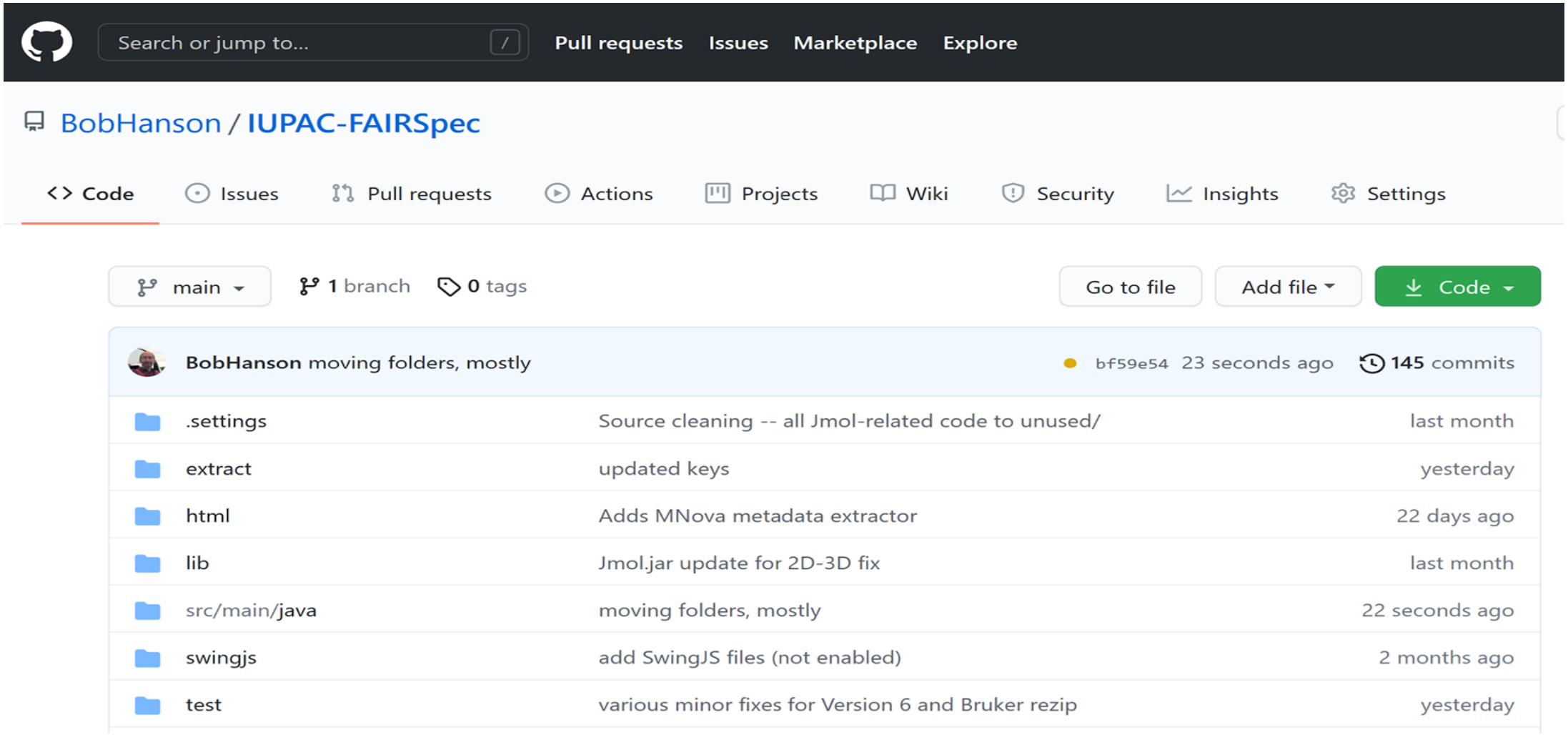


1D 1H
Bruker 600
298.1512 K

IUPAC-FAIRSpec GitHub Repository

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

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY



Communications

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

- ACS National Meeting, April 14, 2021
  - **Division of Chemical Information:**
    Framing FAIR: Scientific Research Data Sharing Policies, Frameworks and Principles – Presentation
- Are you taking your Metadata seriously, Spectroscopy Europe, 31(2), 17-23, 2019 https://doi.org/10.1255/sew.2019.a1
- Are Your Spectroscopic Data FAIR?, Spectroscopy Europe, 30(4), 21-24, 2018 https://doi.org/10.1255/sew.2018.a2
- FAIR enough? Spectroscopy Europe, 33(2), 25-31, 2021 https://doi.org/10.1255/sew.2021.a9
- FAIR Practice, Spectroscopy Europe, 33(3), 18-21, 2021 https://doi.org/10.1255/sew.2021.a14

Technical Appendix

1. Basic IUPAC FAIRSpec Principles

1.3 FAIR data management standards should be modular, extensible, and flexible

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

- *modular*, allowing for core standards to be **developed in different subdisciplines**
- *extensible*, expressing clear versioning and allowing for **future needs**
- *flexible,* **respecting variety** and should not require one particular data format

BRUKER

JEOL

jcamp-dx

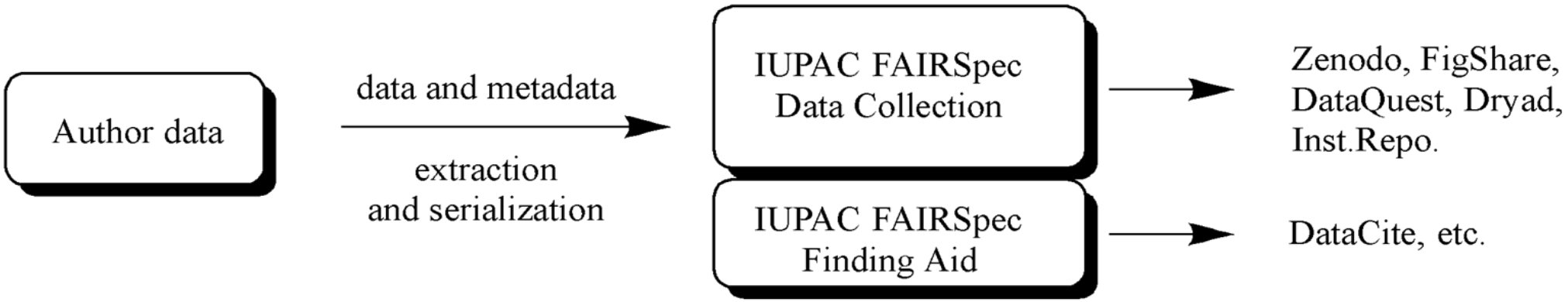
NOVA

1. Basic IUPAC FAIRSpec Principles

1.4 Findability, Accessibility, Interoperability, and Reusability does not stop at the ZIP file

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

- the standard should describe a **digital *finding aid*** that allows a reuser to quickly ascertain whether additional scrutiny of the data collection is warranted
- the standard should **allow for repackaging** or "extraction" of metadata and other digital objects from an original dataset in order to provide a better reuser experience

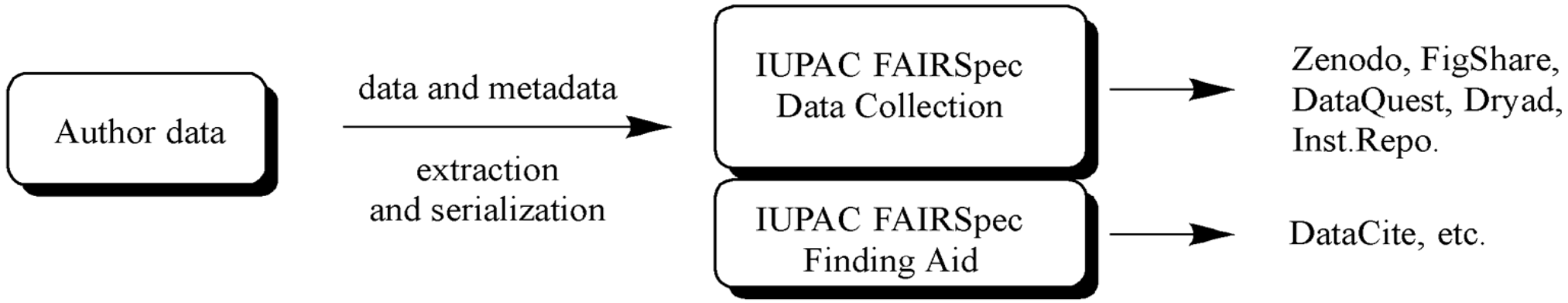


1. Basic IUPAC FAIRSpec Principles

1.4 Findability, Accessibility, Interoperability, and Reusability do not stop at the ZIP file

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

- standard must 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”
- standards should respect the fact that data can have **multiple *representations***, as reuse of data relies upon data being in a form that is meaningful *for the reuser*.



2. The IUPAC FAIRSpec Metadata Model

2.1 IFS Properties

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

2.1.1 Collection Properties

```
IFS.property.collection.data.license.name:          "cc-by-nc-4.0"
IFS.property.collection.data.license.uri:           "https://creativecommons.org/licenses/by-nc/4.0"
IFS.property.collection.id:                         "acs.orglett.0c00571"
IFS.property.collection.len:                        199413401
IFS.property.collection.ref:                        "acs.orglett.0c00571._IFS_collection.zip"
IFS.property.collection.source.publication.uri:     "https://doi.org/10.1021/acs.orglett.0c00571"
```

2. The IUPAC FAIRSpec Metadata Model

2.1 IFS Properties

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

2.1.2 Spectroscopic Data Properties

```
IFS.property.spec.nmr.expt.label:                "3c/1H-NMR"
IFS.property.spec.nmr.expt.nucl.1:               "1H"
IFS.property.spec.nmr.expt.pulse.prog:           "zg30"
IFS.property.spec.nmr.expt.temperature.K:        298.1486
IFS.property.spec.nmr.instr.freq.nominal:        600
IFS.property.spec.nmr.instr.manufacturer.name:   "Bruker"
IFS.property.spec.nmr.instr.probe.type:          "Z126545_0016 (CPP BBO 600S3 BB-H&F-D-05 Z)"
```

2. The IUPAC FAIRSpec Metadata Model

2.1 IFS Properties

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

2.1.3 Chemical Structure Properties

```
IFS.property.struc.compound.label:      "3a"
IFS.property.struc.inchi:               "InChI=1S/C20H22N2.CHF3O3S
                                        /c1-2-9-17(10-3-1)19-18-11-5-4-8-16(18)12-13-21-20(19)22-14-6-7-15-22;
                                        2-1(3,4)8(5,6)7/h1-5,8-11,19H,6-7,12-15H2;(H,5,6,7)"
IFS.property.struc.inchikey:            "KZHKHOYBVCYSSO-UHFFFAOYSA-N"
IFS.property.struc.smiles:              "c1cccc2c1.C32c4c5CC[N+1]=C3N6CCCC6.c5ccc4"
```

2.1.4 Chemical Sample Properties (todo)

2. The IUPAC FAIRSpec Metadata Model

2.2 IFS Representations

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

2.2.1 Spectroscopic Data Representations

IFS.representation.spec.nmr.jcamp.fid.1d
IFS.representation.spec.nmr.jcamp.fid.2d
IFS.representation.spec.nmr.jcamp.spec.1i1r.1d
IFS.representation.spec.nmr.jcamp.spec.1r.1d
IFS.representation.spec.nmr.jcamp.spec.2d
IFS.representation.spec.nmr.peaklist
IFS.representation.spec.nmr.spectrum.description
IFS.representation.spec.nmr.spectrum.image
IFS.representation.spec.nmr.spectrum.pdf
IFS.representation.spec.nmr.vendor.dataset

2. The IUPAC FAIRSpec Metadata Model

2.2 IFS Representations

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

2.2.2 Chemical Structure Representations

IFS.representation.struc.cdx
IFS.representation.struc.cdxml
IFS.representation.struc.mol
IFS.representation.struc.mol.2d
IFS.representation.struc.mol.3d
IFS.representation.struc.png
IFS.representation.struc.sdf
IFS.representation.struc.sdf.2d
IFS.representation.struc.sdf.3d
IFS.representation.struc.unknown

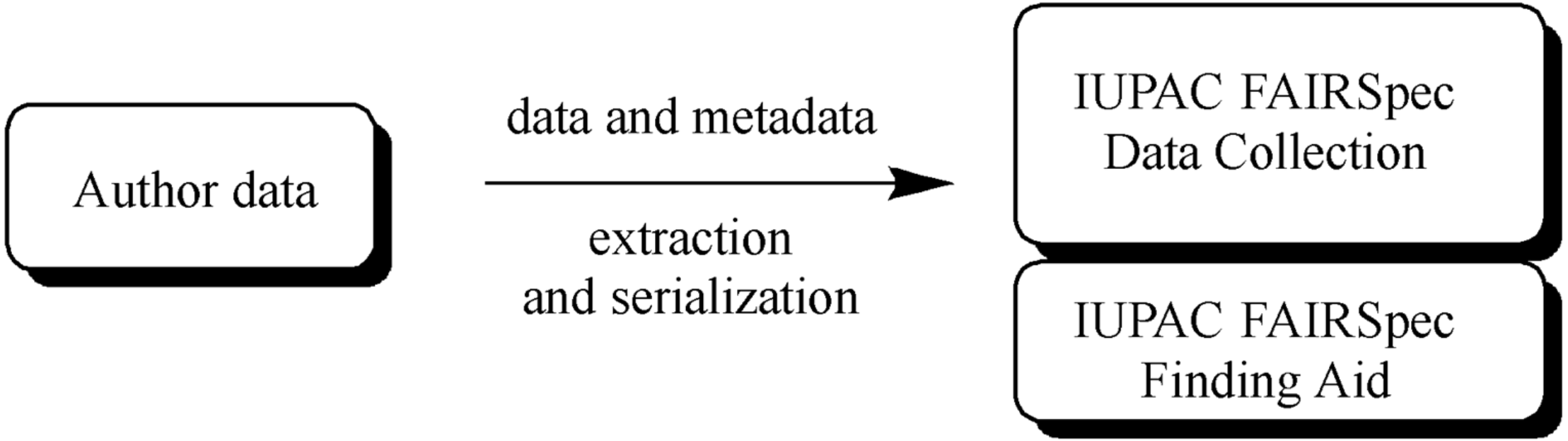
2.2.3 Chemical Sample Representations (todo)

3. Preferred Collection Organization

3.1 Structured data collections

3.2 Describing a collection - The IFS Finding Aid

3.3 Data and metadata extraction

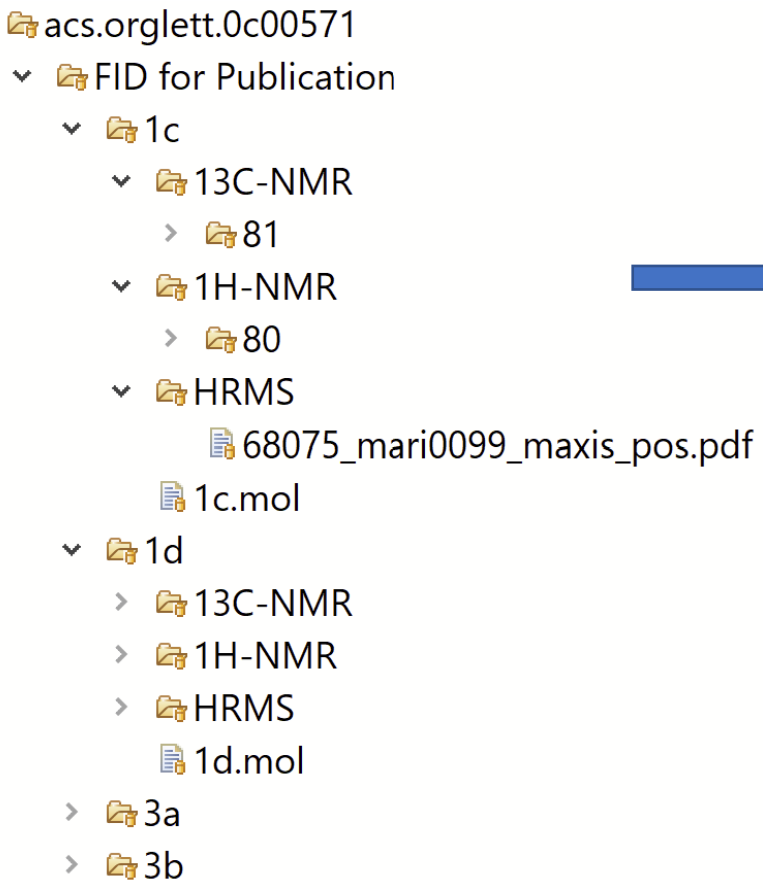


3. Preferred Collection Organization

3.1 Structured data collections

3.2 Describing a collection - The IFS Finding Aid

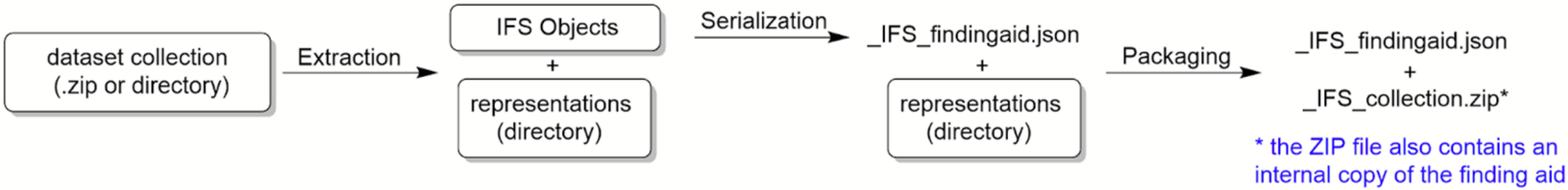
3.3 Data and metadata extraction



4. Data / Metadata Extraction & Serialization

4.1 Workflow
4.2 Data and Metadata Extraction
4.3 IFSSpecDataFindingAid Serialization

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY



INTERNATIONAL UNION OF
PURE AND APPLIED CHEMISTRY

- 5. The IUPAC FAIRSpec Data Model
  - 5.1 Definitions
  - 5.2 Common Core Classes
  - 5.3 Common Associative Classes
  - 5.4 Spectroscopy-Specific Classes