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

CECAM-MADICES, Feb 7, 2022

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

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.

Project Timeline

- 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
- Nov 2022 – Dec 2022
  - Finalize recommendations
- Jan 2023 – Dec 2023
  - Collaborate with implementers

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

Submitted for publication Oct 2021

- 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.***

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 registered and standardized.

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. Design to adapt to future needs.
C. Respect digital diversity.
D. All data formats should be valued.

Glossary of about 30 terms

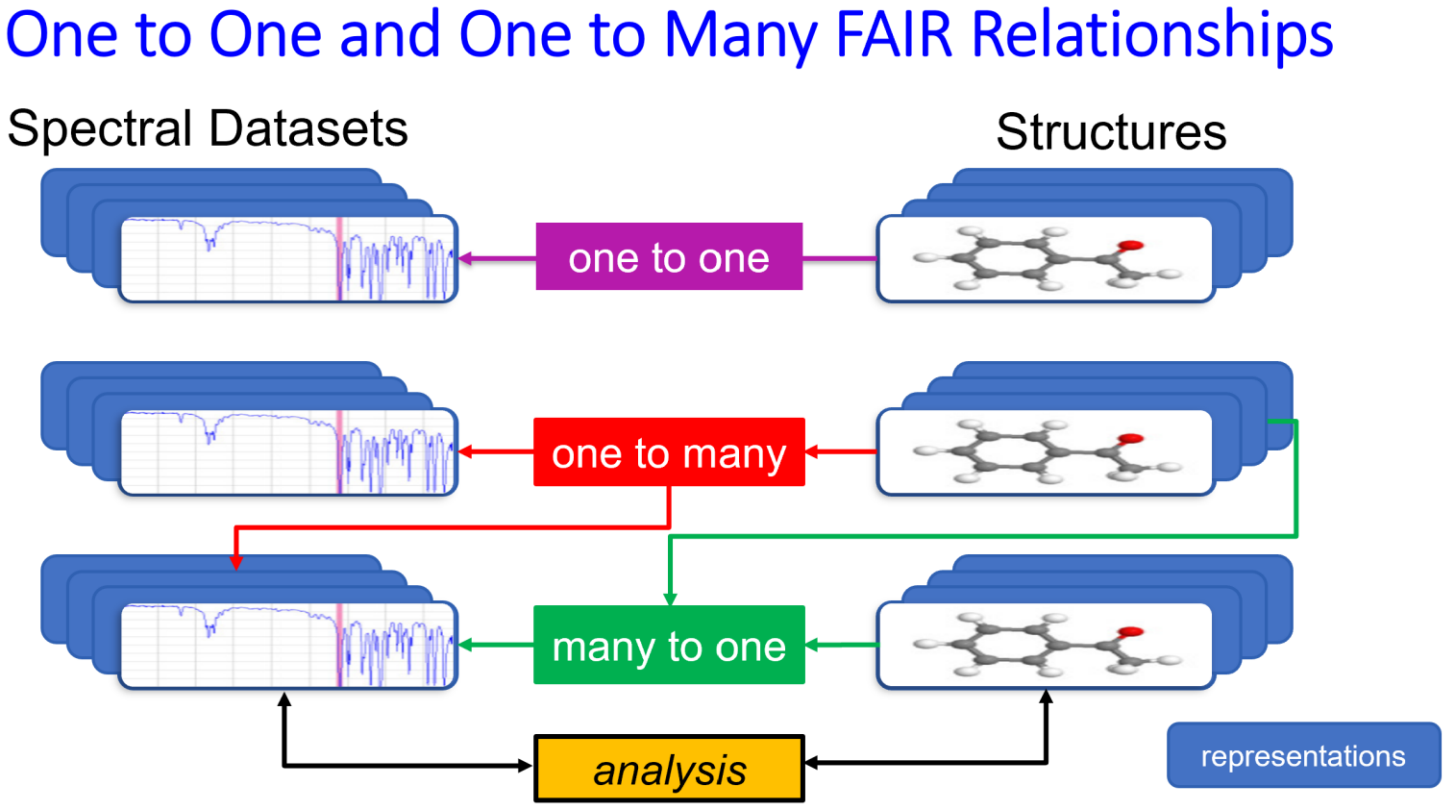
chemical structure identifier
curation
data and metadata extraction
data management plan
data model
data provenance
data repository
data representation
dataset (spectroscopic)
digital aggregation
digital collection
digital entity
digital finding aid
digital object
Digital Object Identifier (DOI)
IUPAC 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)

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)
reuser
PID graph
serialization (of a finding aid)

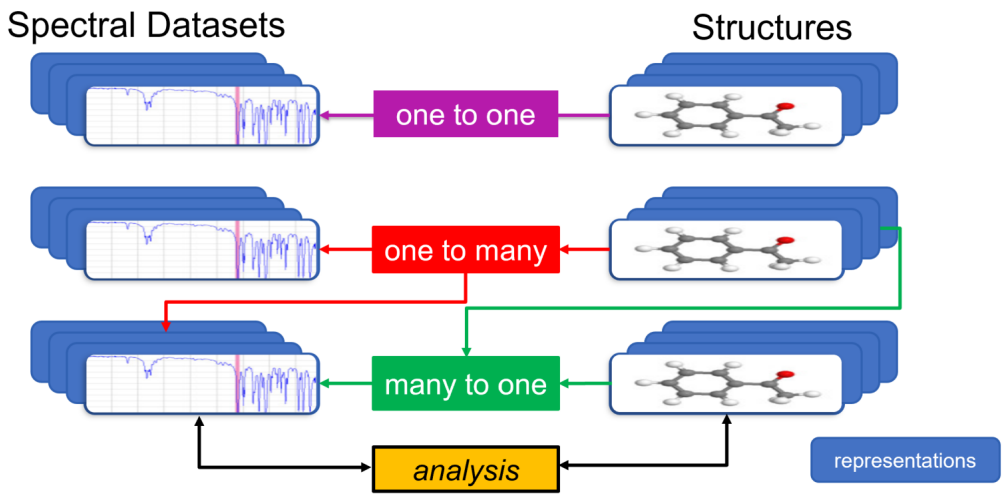
IUPAC FAIRSpec Principles

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*.



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.

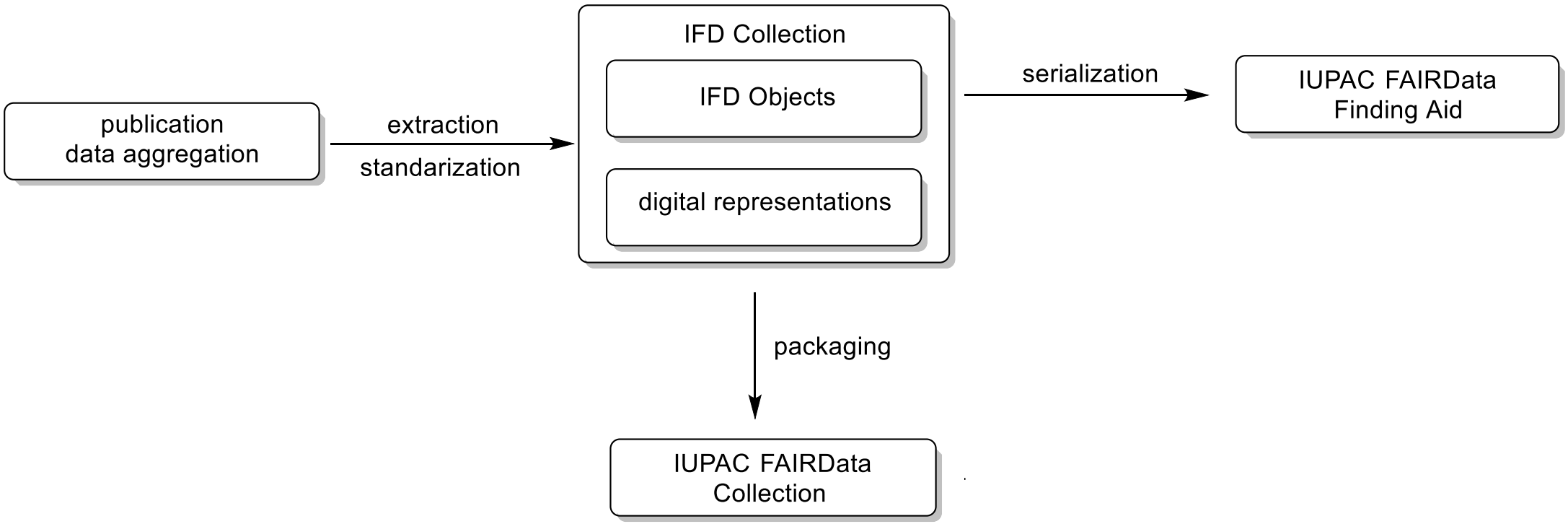


+

```
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:       {…}
```

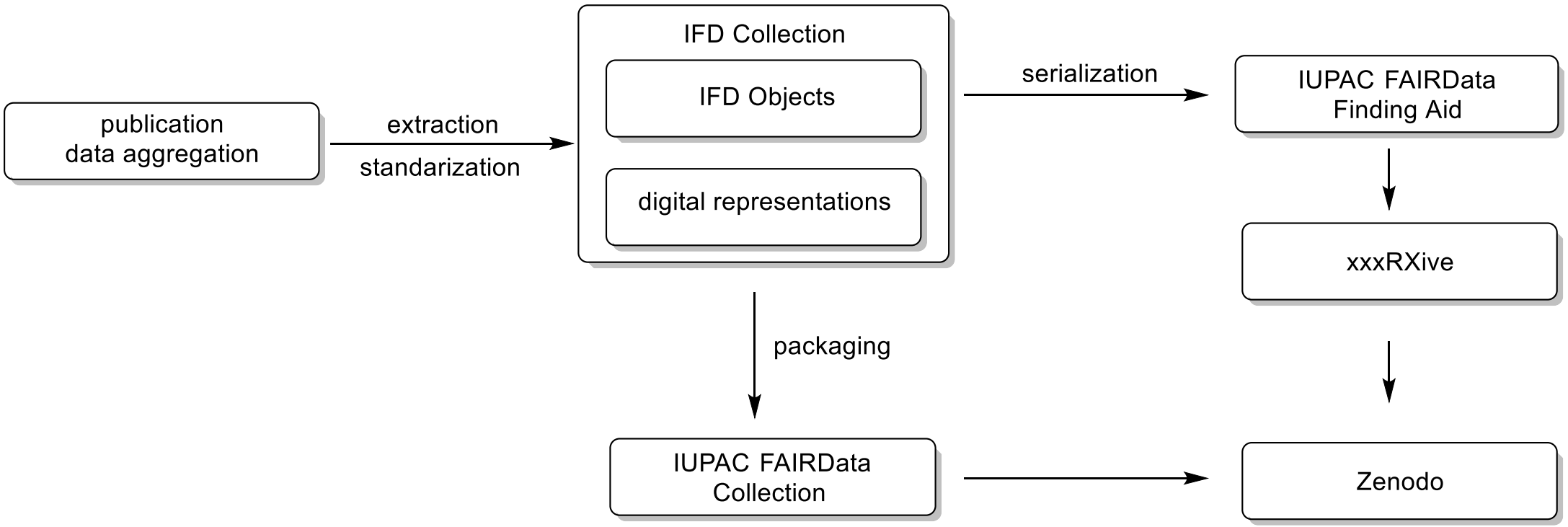
IUPAC FAIRSpec Principles

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.



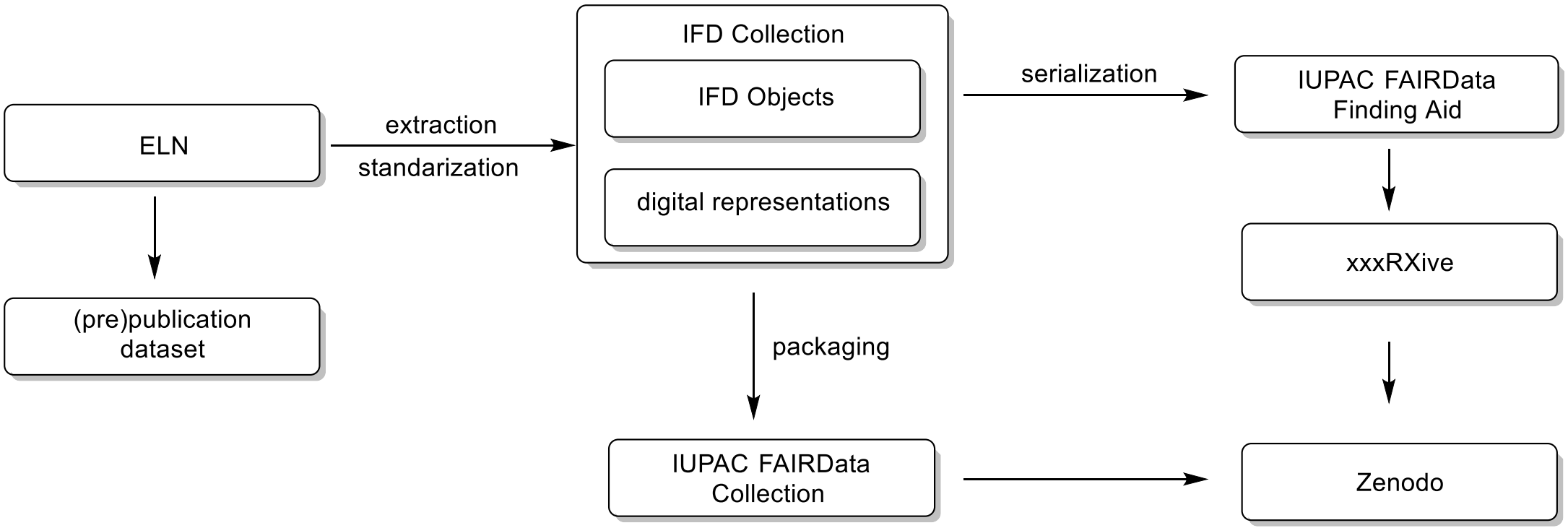
IUPAC FAIRSpec Principles

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



IUPAC FAIRSpec Principles

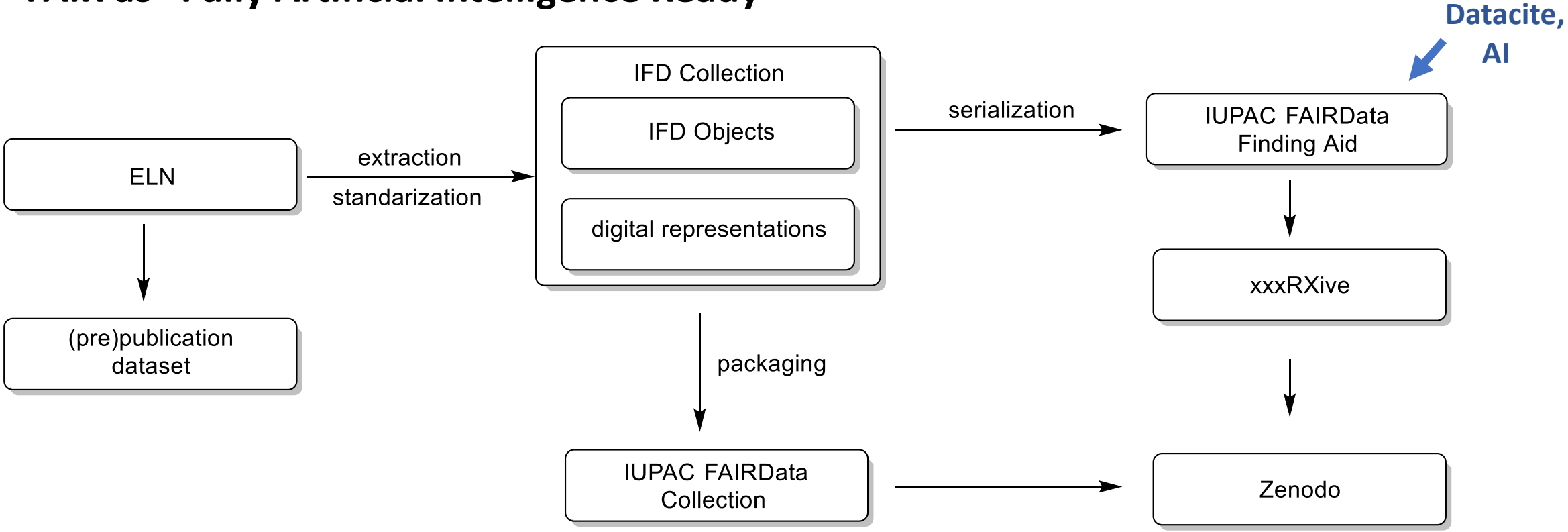
The standard emphasizes the **importance of management throughout the lifecycle of the data (and beyond)**



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 FAIRSpec Metadata Model

Collection Properties

```
IFD.property.collection.data.license.name:          "cc-by-nc-4.0"
IFD.property.collection.data.license.uri:           "https://creativecommons.org/licenses/by-nc/4.0"
IFD.property.collection.id:                         "acs.orglett.0c00571"
IFD.property.collection.len:                        199412175
IFD.property.collection.ref:                        "acs.orglett.0c00571._IFD_collection.zip"
IFD.property.collection.source.publication.uri:     "https://doi.org/10.1021/acs.orglett.0c00571"
```

The IUPAC FAIRSpec Metadata Model

Spectroscopic Data Properties

```
IFD.property.spec.nmr.expt.label:                "1d/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: 298.1511
IFD.property.spec.nmr.instr.freq.nominal:        700
IFD.property.spec.nmr.instr.manufacturer.name:   "Bruker"
IFD.property.spec.nmr.instr.probe.type:          "Z122896_0005 (CP QCI 700S3 H/F-C/N-D-05 Z)"
```

The IUPAC FAIRSpec Metadata Model

Spectroscopic Data Representations

```
IFD_REP_SPEC_NMR_VENDOR_DATASET        = "IFD.representation.spec.nmr.vendor.dataset";
IFD_REP_SPEC_NMR_SPECTRUM_PDF          = "IFD.representation.spec.nmr.spectrum.pdf";
IFD_REP_SPEC_NMR_SPECTRUM_IMAGE        = "IFD.representation.spec.nmr.spectrum.image";
IFD_REP_SPEC_NMR_SPECTRUM_DESCRIPTION = "IFD.representation.spec.nmr.spectrum.description";
IFD_REP_SPEC_NMR_PEAKLIST              = "IFD.representation.spec.nmr.peaklist";
IFD_REP_SPEC_NMR_JCAMP_FID_1D          = "IFD.representation.spec.nmr.jcamp.fid.1d";
IFD_REP_SPEC_NMR_JCAMP_FID_2D          = "IFD.representation.spec.nmr.jcamp.fid.2d";
IFD_REP_SPEC_NMR_JCAMP_SPEC_1r_1D      = "IFD.representation.spec.nmr.jcamp.spec.1r.1d";
IFD_REP_SPEC_NMR_JCAMP_SPEC_1i1r_1D    = "IFD.representation.spec.nmr.jcamp.spec.1i1r.1d";
IFD_REP_SPEC_NMR_JCAMP_SPEC_2D         = "IFD.representation.spec.nmr.jcamp.spec.2d";
```

The IUPAC FAIRSpec Metadata Model

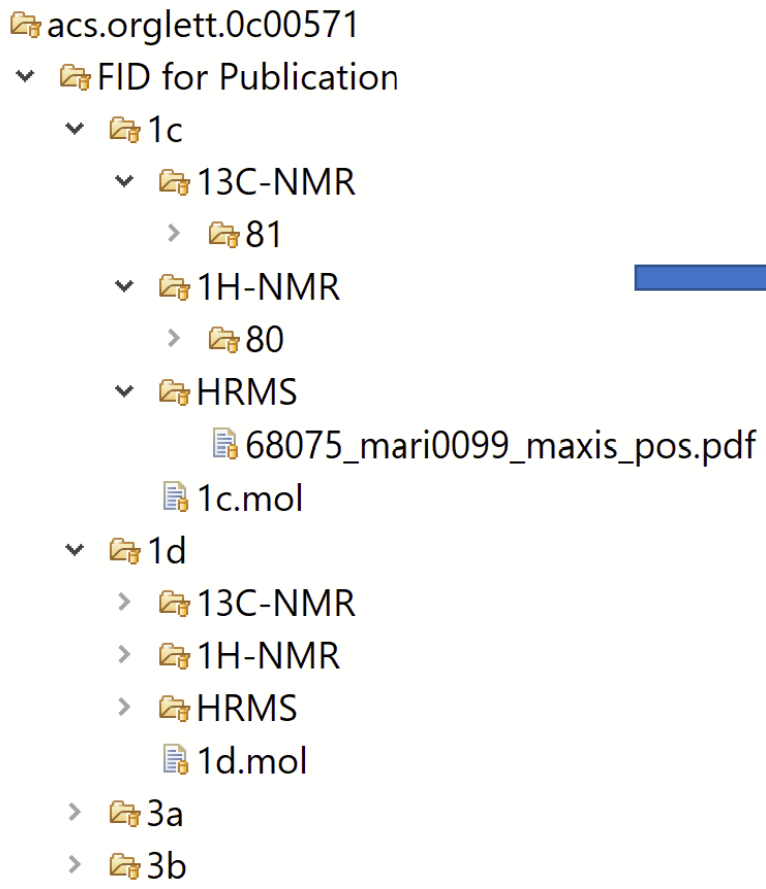
Chemical Structure Properties

```
IFD.property.struc.compound.label:    "3a"
IFD.property.struc.inchi:             "InChI=1S/C20H22N2.CHF3O3…H,6-7,12-15H2;(H,5,6,7)"
IFD.property.struc.inchikey:          "KZHKHOYBVCYSSO-UHFFFAOYSA-N"
IFD.property.struc.smiles:            "c1cccc2c1.C32c4c5CC[N+1]=C3N6CCCC6.c5ccc4"
```

Chemical Sample Properties (TO DO)

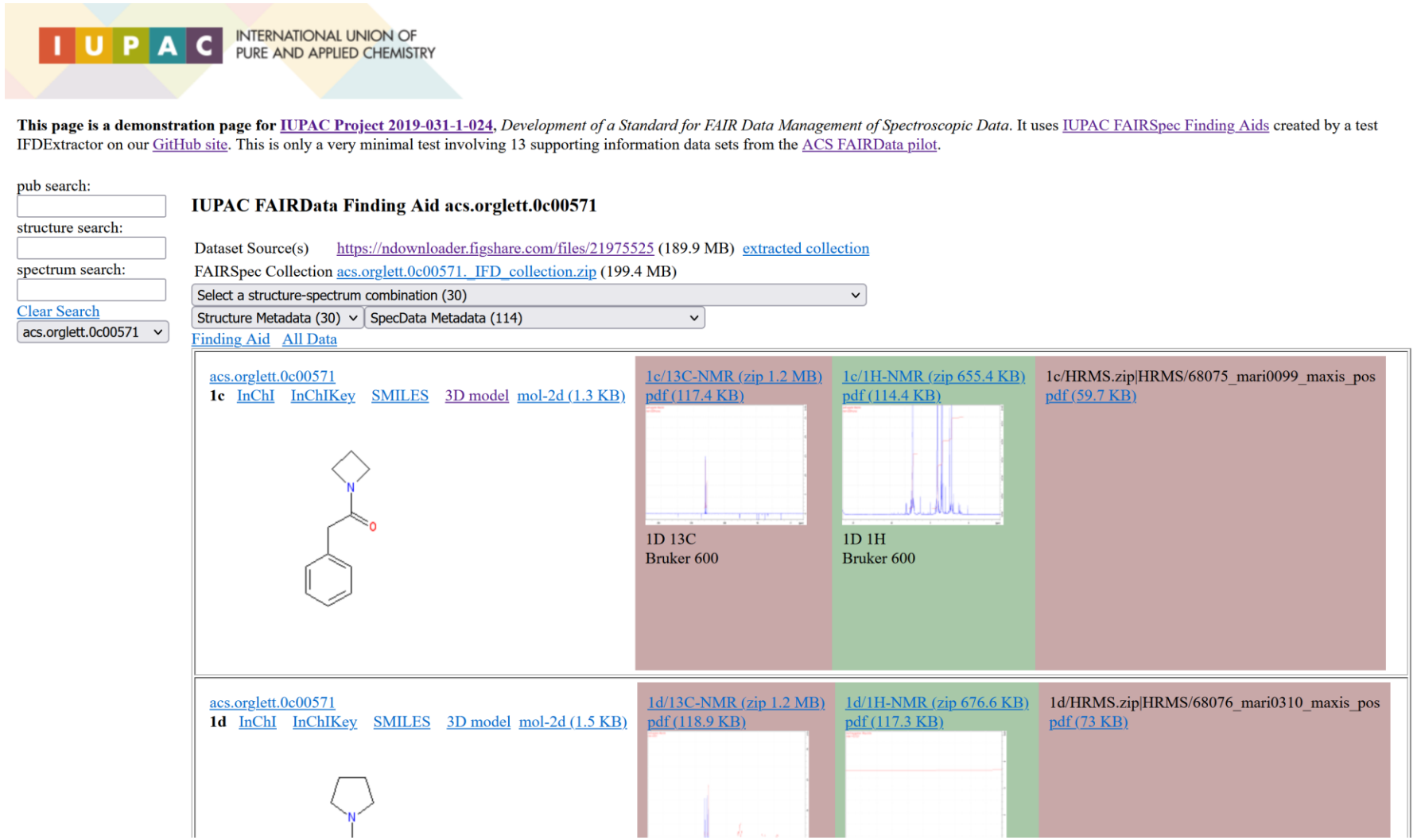
Chemical Analysis Properties (TO DO)

Prototype Extractor and JSON-based Finding Aid



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

Preliminary Data Model -- the “IUPAC FAIRData Finding Aid”



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