Proposed application of the IUPAC FAIRSpec Finding Aid for standardized repository data introduction and delivery of metadata via an API

ACS National Meeting, Mar. 26, 2023

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

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.

Objectives of the FAIRSpec Project

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 a generalized “IUPAC FAIRData Collection"
- **A standard for describing properties and relationships of digital objects** within the metadata records of an “IUPAC FAIRSpec Finding Aid.”
- **A standard for the serialization** of an IUPAC FAIRSpec Finding Aid.
- **A proposal for methods of data and metadata extraction** from “IUPAC FAIRSpec-Ready” aggregations.

Products to Date – Guiding Principles

IUPAC specification for the FAIR management of spectroscopic data in chemistry (IUPAC FAIRSpec) – guiding principles

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

From the journal Pure and Applied Chemistry
https://doi.org/10.1515/pac-2021-2009

https://doi.org/10.1515/pac-2021-2009

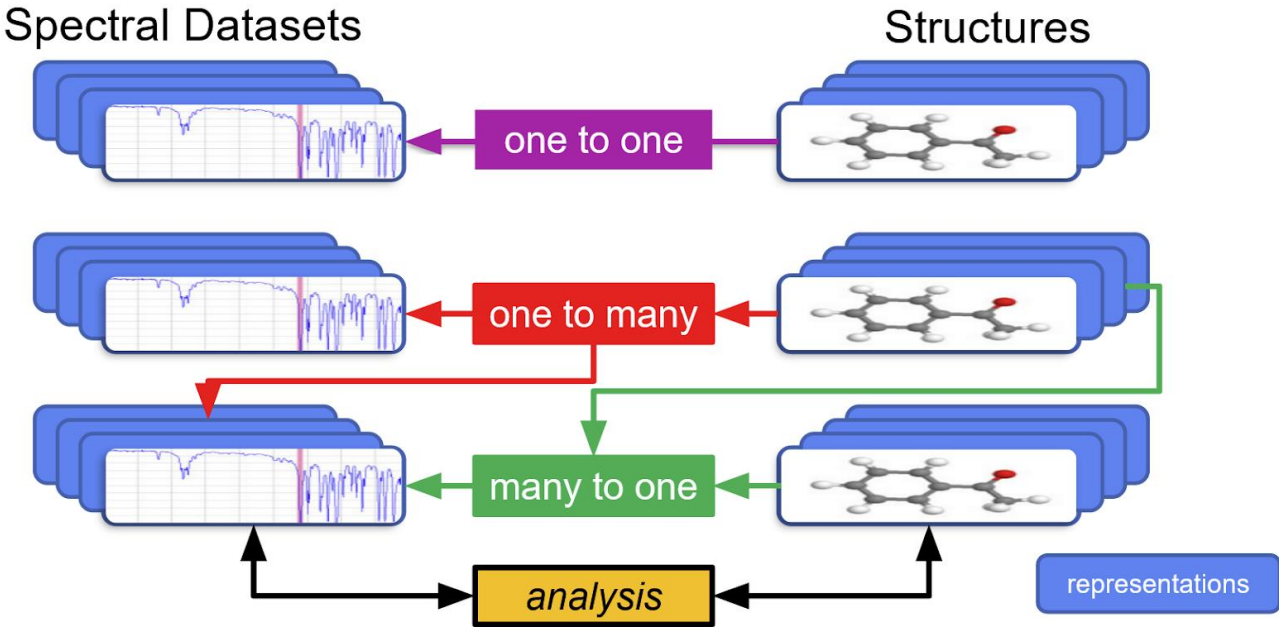
Products to Date – Guiding Principles

- **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.

https://doi.org/10.1515/pac-2021-2009

Key Concept: Associations – Relational Metadata

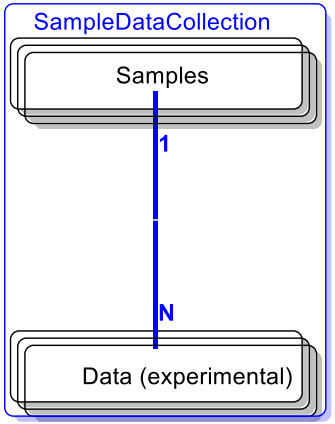
One to One and One to Many FAIR Relationships

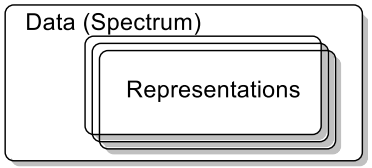


The IUPAC FAIRData Collection – Early On

An IUPAC FAIRData Collection can start off as simply a set of spectra with sample identifiers.

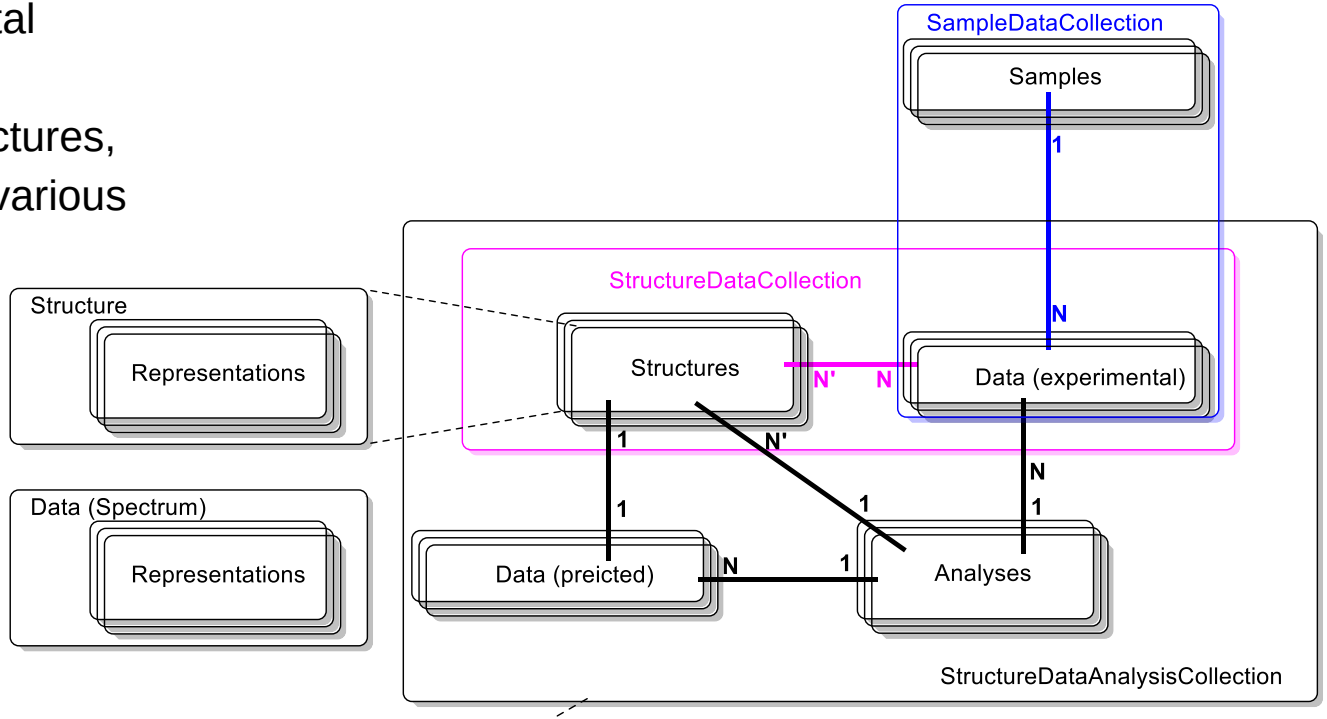
Representations include raw instrument datasets, peak listings, images, etc.





The IUPAC FAIRData Collection

They can be as complex as a collection of experimental spectra with associated samples, chemical structures, predicted spectra, and various analysis objects.



Products to Date – GitHub Project

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

- Repository for project digital outputs
- Collection of 13 ACS pilot study spectroscopic “supporting data” aggregations
- Eclipse open-source Java project
  - IUPAC FAIRData object model programmatic description and implementation development
  - IUPAC FAIRSpec Finding Aid serialization development and testing
  - metadata/data extractor prototyping

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

“Finding Aid” Inspiration – EAD

Finding Aids

Encoded Archival Description (EAD) at the Library of Congress

What are Finding Aids?

Handwritten poems by Walt Whitman ... Leonard Bernstein's scrapbooks ... Thomas Edison's patents ... photographs and memoranda from the NAACP ... Margaret Mead's field notes ... The collections of the Library of Congress offer researchers rich and deep access to primary source material of unparalleled interest and significance.

What is EAD?

LC finding aids are XML documents created using the Encoded Archival Description (EAD), an international standard maintained by the Library of Congress in partnership with the Society of American Archivists.

https://www.loc.gov/rr/ead/

Products to Date – Demo Site

https://chemapps.stolaf.edu/iupac/site/ifd4/

- Collection of 13 IUPAC FAIRSpec Finding Aids (JSON) and their associated IUPAC FAIRSpec Collections generated by the extractor prototype from the ACS pilot SI data packages.

- Small JavaScript library demonstrating (minimal) processing and rendering of the finding aids.

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 12 supporting information data sets from the ACS FAIRData pilot.

Select an ACS article

FAIRSpecFindingAid acs.orglett.0c00571

Title	Synthesis of Novel Heterocycles by Amide Activation and Umpolung Cyclization
Authors	Haoqi Zhang, Margaux Riomet, Alexander Roller, Nuno Maulide
Publication	http://pubs.acs.org/doi/pdf/10.1021/acs.orglett.0c00571
Data Origin	https://ndownloader.figshare.com/files/21975525 (189.9 MB)
Collections	Compounds(30) Spectra(114) Structures(30)

FAIRSpecFindingAid acs.orglett.0c00624

Title	Intermolecular Vicinal Diaminative Assembly of Tetrahydroquinoxalines via Metal-free Oxidative [4 + 2] Cycloaddition Strategy
Authors	Dangui Wang, Huaibin Yu, Shaohan Sun, Fangrui Zhong
Publication	https://pubs.acs.org/doi/pdf/10.1021/acs.orglett.0c00624
Data Origin	https://ndownloader.figshare.com/files/21947274 (15.2 MB)
Collections	Compounds(42) Spectra(80) Structures(42)

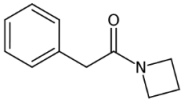
FAIRSpecFindingAid for acs.orglett.0c00571

Title	Synthesis of Novel Heterocycles by Amide Activation and Umpolung Cyclization
Authors	Haoqi Zhang, Margaux Riomet, Alexander Roller, Nuno Maulide
Publication	http://pubs.acs.org/doi/pdf/10.1021/acs.orglett.0c00571
Data Origin	https://ndownloader.figshare.com/files/21975525 (189.9 MB)
Collections	Compounds(30) Spectra(114) Structures(30)

Compound 1c

from SMILES:

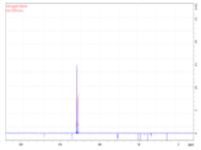
inchikey HXFKEAUPENVJFI-UHFFFAOYSA-N
molecular_formula H 13 C 11 N 1 O 1



mol_2d 1c.mol (1.3 KB)
smiles c1cccc2c1.C2C(=O)N1CCC1
standard_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
fixedh_inchi InChI=1/C11H13NO/c13-11(12-7-4-8-12)9-10-5-2-1-3-6-10/h1-3,5-6H,4,7-9H2

Spectra 1c/13C-NMR

spectrum_document 1.pdf (117.4 KB)

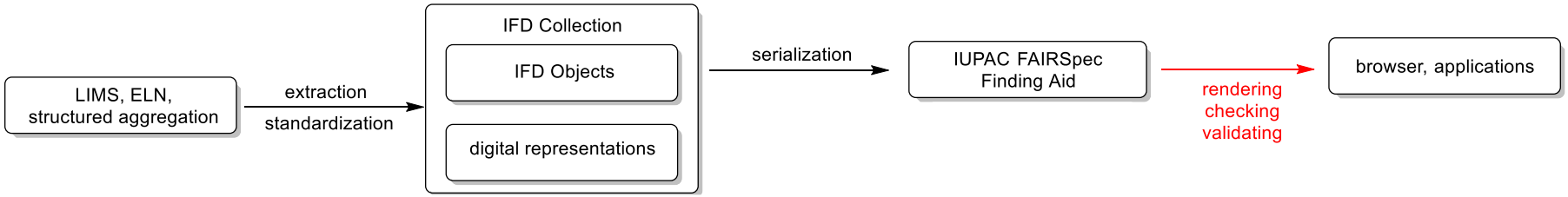
spectrum_image 

vendor_dataset 13C-NMR.zip (1.2 MB)

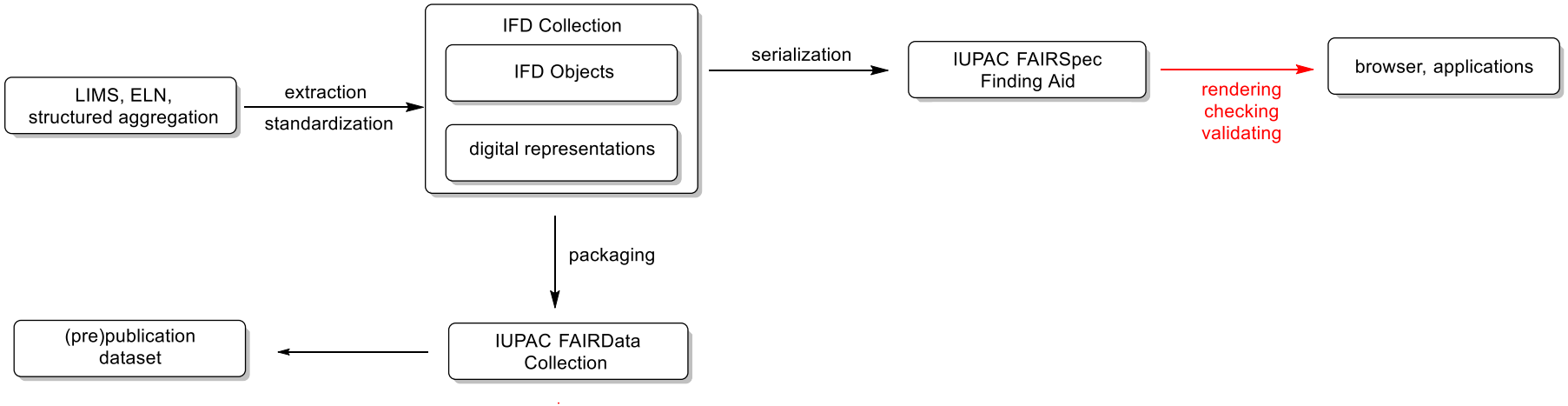
IFD Properties

expt_absolute_temperature	298.1525
expt_dimension	1D
expt_nucl1	13C
expt_nucl2	1H
expt_pulse_prog	deptqgpsp
expt_solvent	CDCl3
expt_title	Auftraggeber Maulide mari-0099-carac

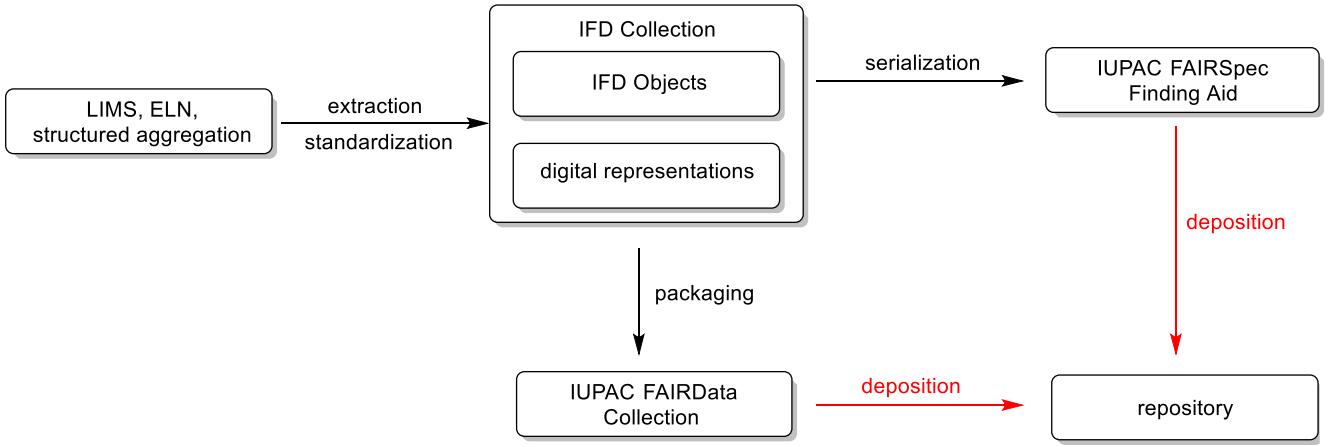
The IUPAC FAIRSpec Finding Aid – Research Mode



The IUPAC FAIRSpec Finding Aid – Publication Mode

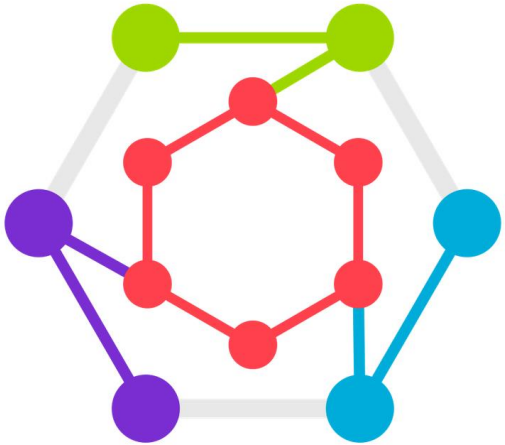


The IUPAC FAIRSpec Finding Aid – Deposition Mode



- Would provide standardized ingest mechanism for a repository
- Would allow pre-ingest validation and error correction
- Would allow automated workflow for bulk deposition of datasets
- Would interface with systems generating and preserving instrument data

API Inspiration – OPTIMADE

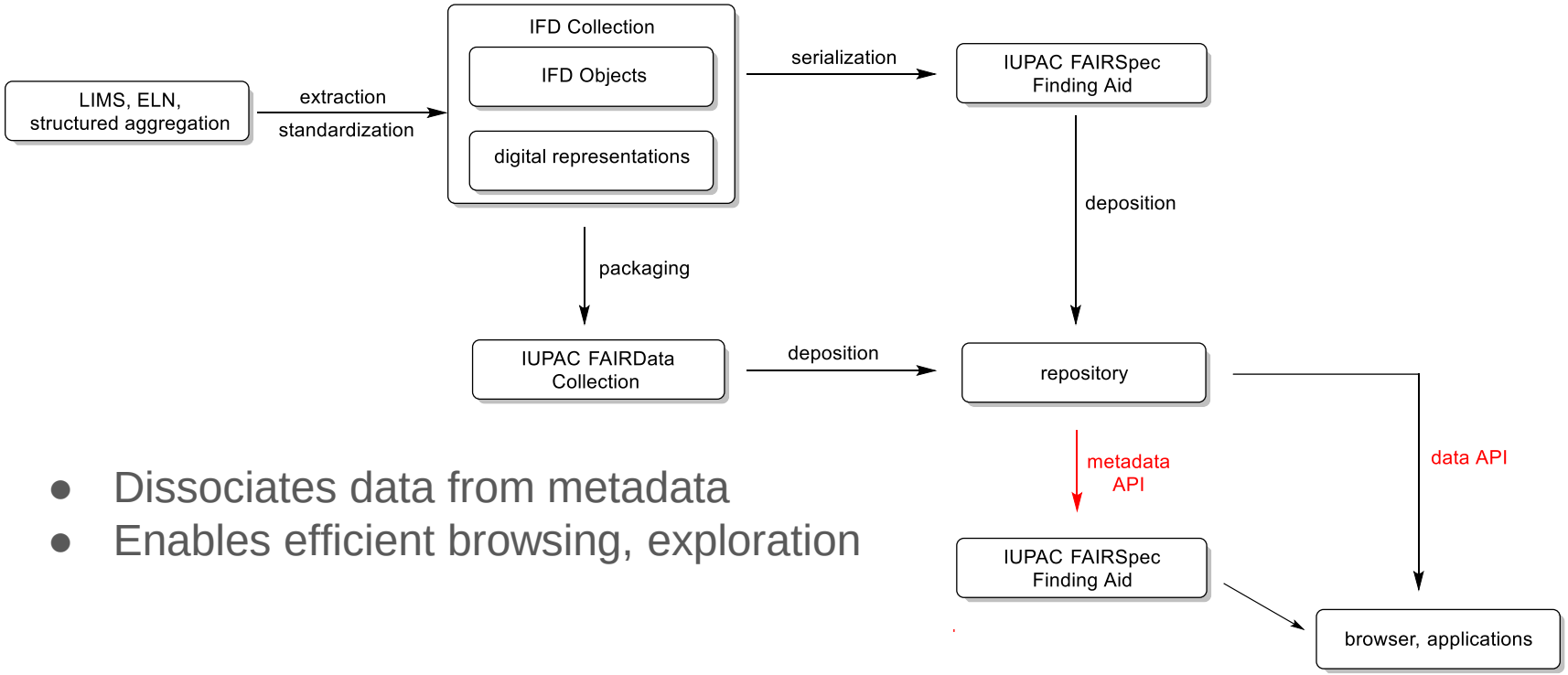


OPTIMADE

Open Databases Integration for Materials Design

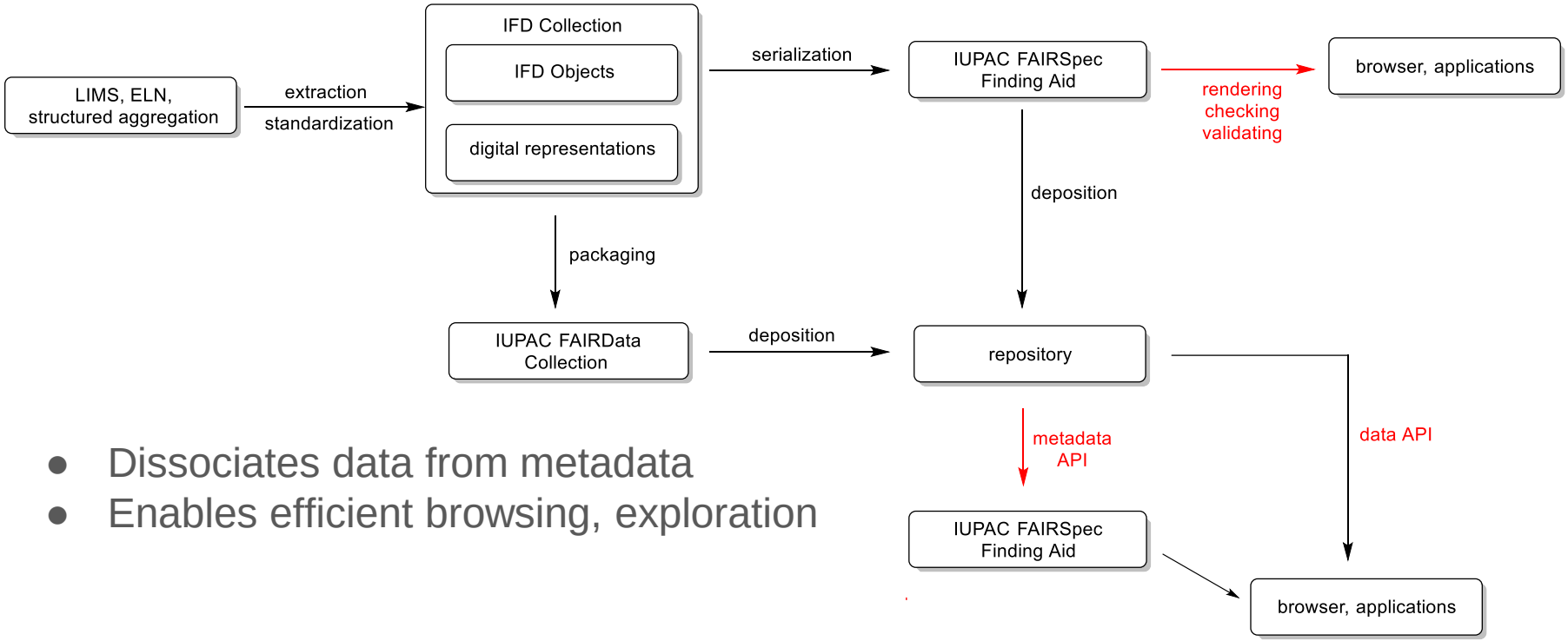
The **Open Databases Integration for Materials Design** (OPTIMADE) consortium aims to make materials databases interoperable by developing a specification for a common REST API.

The IUPAC FAIRSpec Finding Aid – API Mode



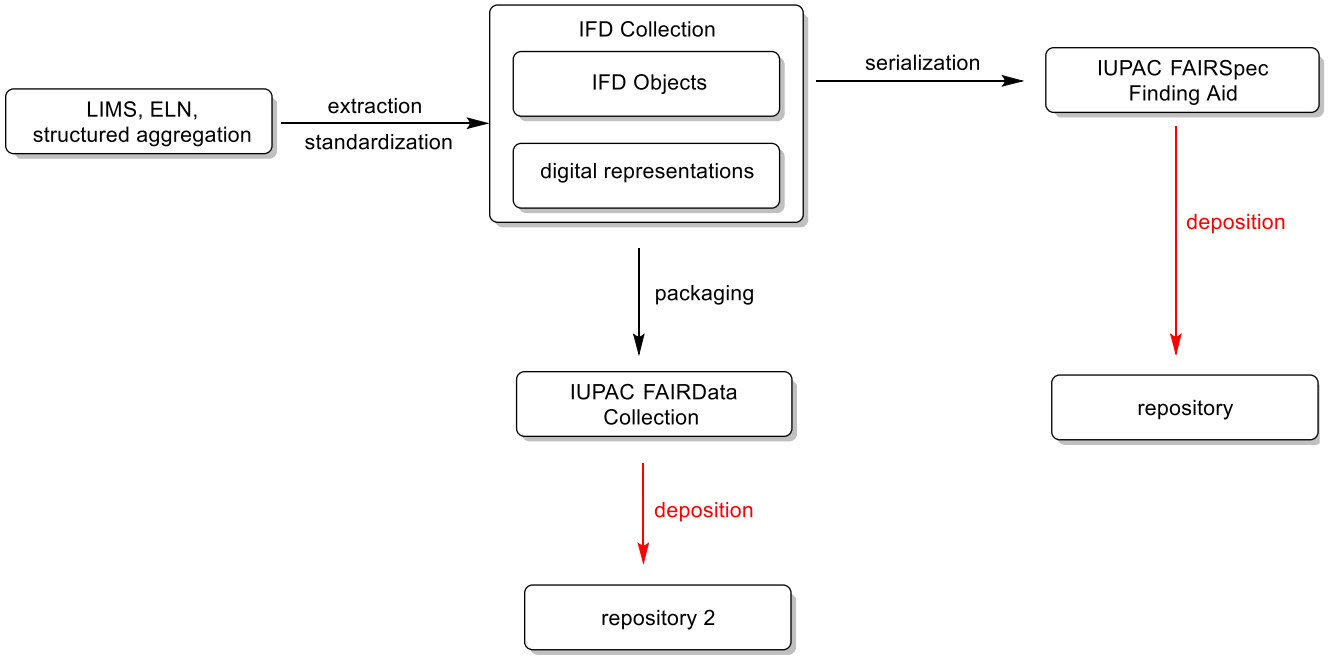
- Dissociates data from metadata
- Enables efficient browsing, exploration

The IUPAC FAIRSpec Finding Aid – API Mode

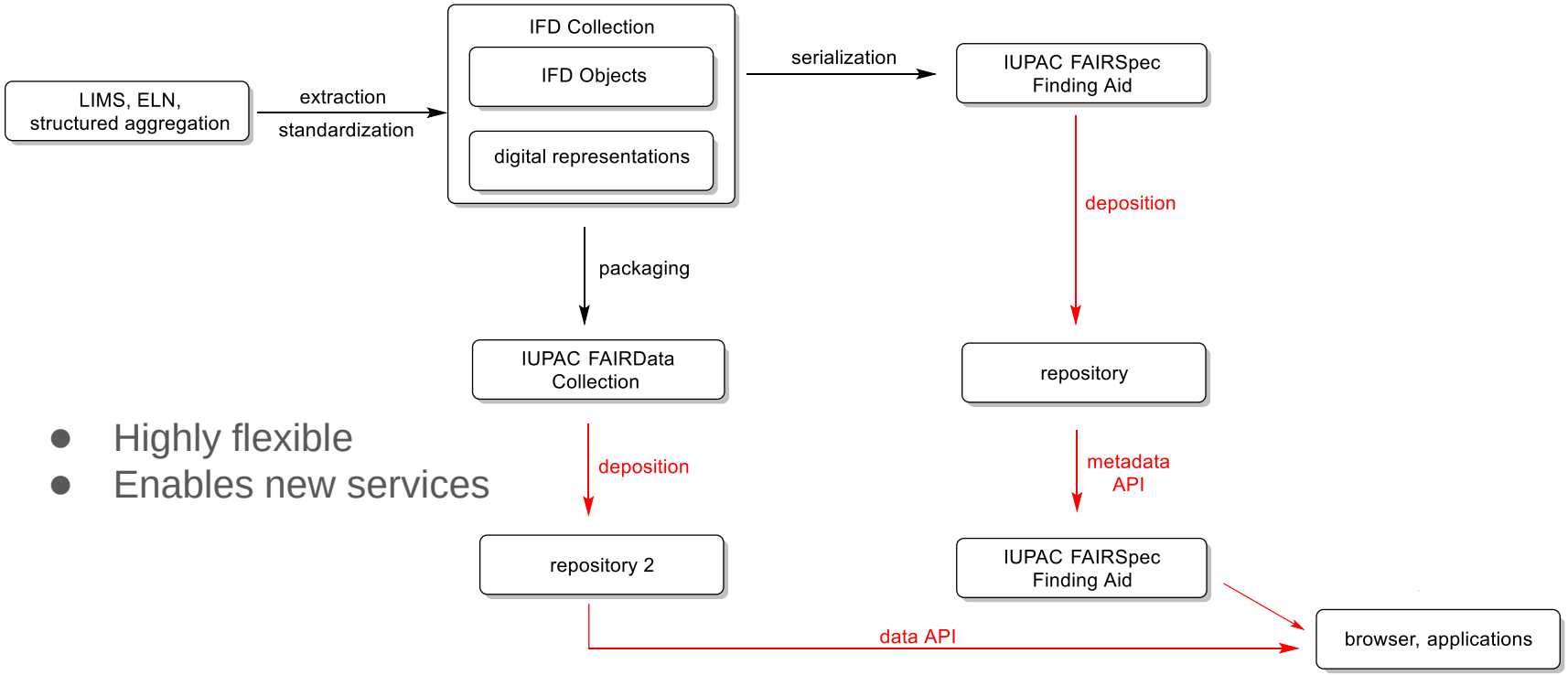


- Dissociates data from metadata
- Enables efficient browsing, exploration

The IUPAC FAIRSpec Finding Aid – Distributed Mode

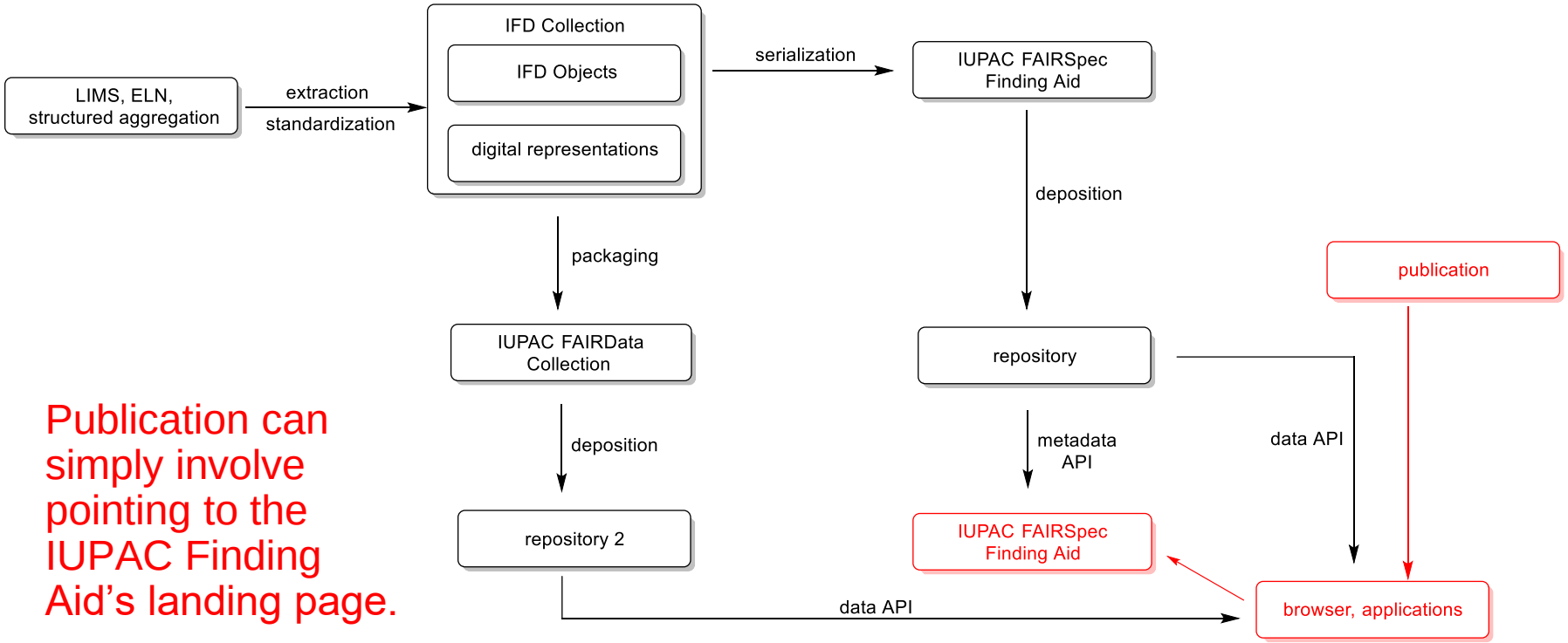


The IUPAC FAIRSpec Finding Aid – Distributed Mode



- Highly flexible
- Enables new services

The IUPAC FAIRSpec Finding Aid – Publication Mode



Publication can simply involve pointing to the IUPAC Finding Aid's landing page.

acs.orglett.0c00571

Show Finding Aid

Collection Folder

FAIRSpecFindingAidacs.orglett.0c00571

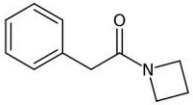
Title	Synthesis of Novel Heterocycles by Amide Activation and Umpolung Cyclization
Authors	Haoqi Zhang, Margaux Riomet, Alexander Roller, Nuno Maulide
Publication	http://pubs.acs.org/doi/pdf/10.1021/acs.orglett.0c00571
Data Origin	https://ndownloader.figshare.com/files/21975525 (189.9 MB)
Collections	Compounds(30) Spectra(114) Structures(30)

Compound 1c

from SMILES:

inchikey HXFKEAUPENVJFI-UHFFFAOYSA-N

molecular_formula H 13 C 11 N 1 O 1



mol_2d 1c.mol (1.3 KB)

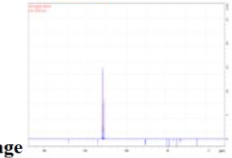
smiles c1cccc2c1.C2C(=O)N1CCC1

standard_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

fixedh_inchi InChI=1/C11H13NO/c13-11(12-7-4-8-12)9-10-5-2-1-3-6-10/h1-3,5-6H,4,7-9H2

Spectra 1c/13C-NMR

spectrum_document 1.pdf (117.4 KB)

spectrum_image 

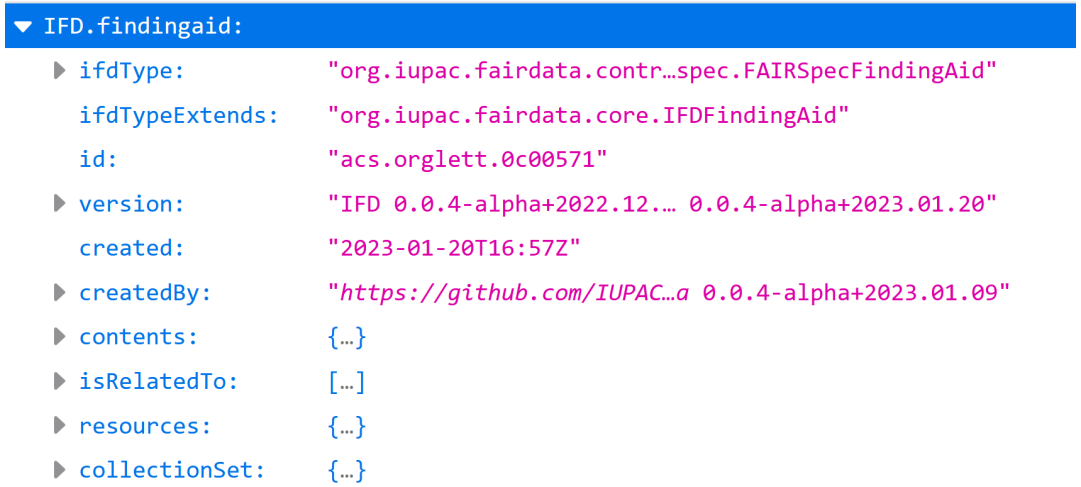
vendor_dataset 13C-NMR.zip (1.2 MB)

IFD Properties

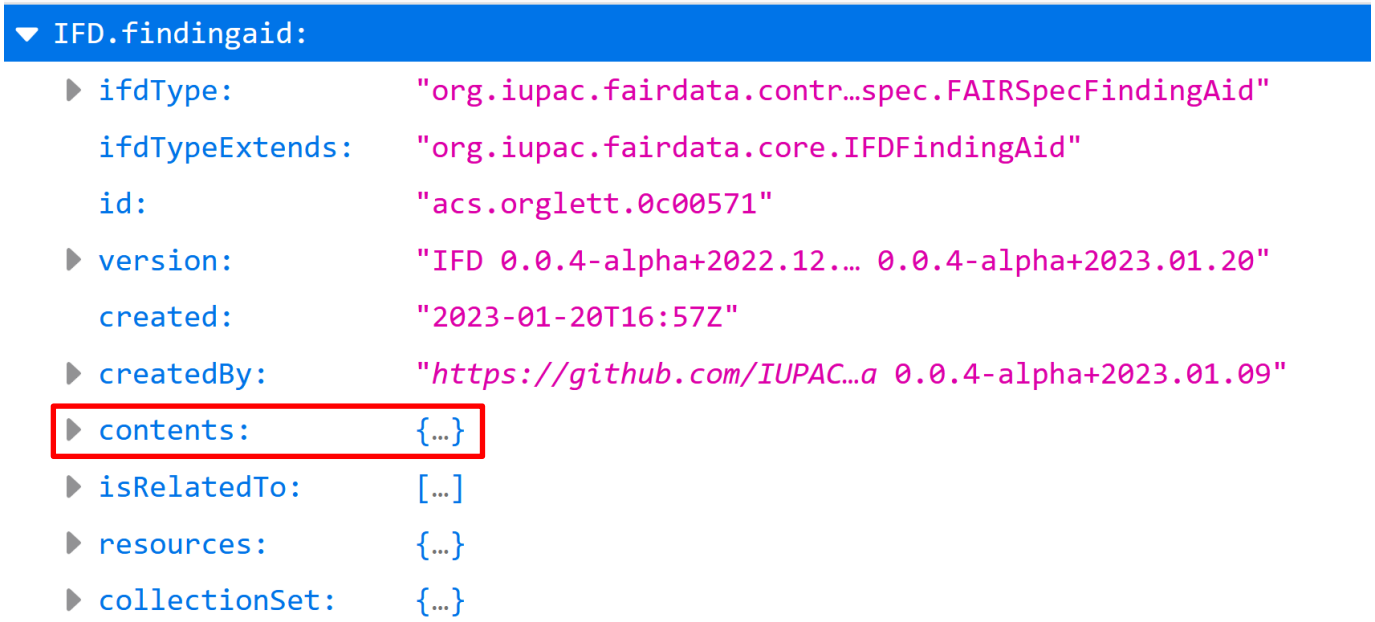
expt_absolute_temperature	298.1525
expt_dimension	1D
expt_nucl1	13C
expt_nucl2	1H
expt_pulse_prog	deptqgpsp

The IUPAC FAIRSpec Finding Aid – A Closer Look

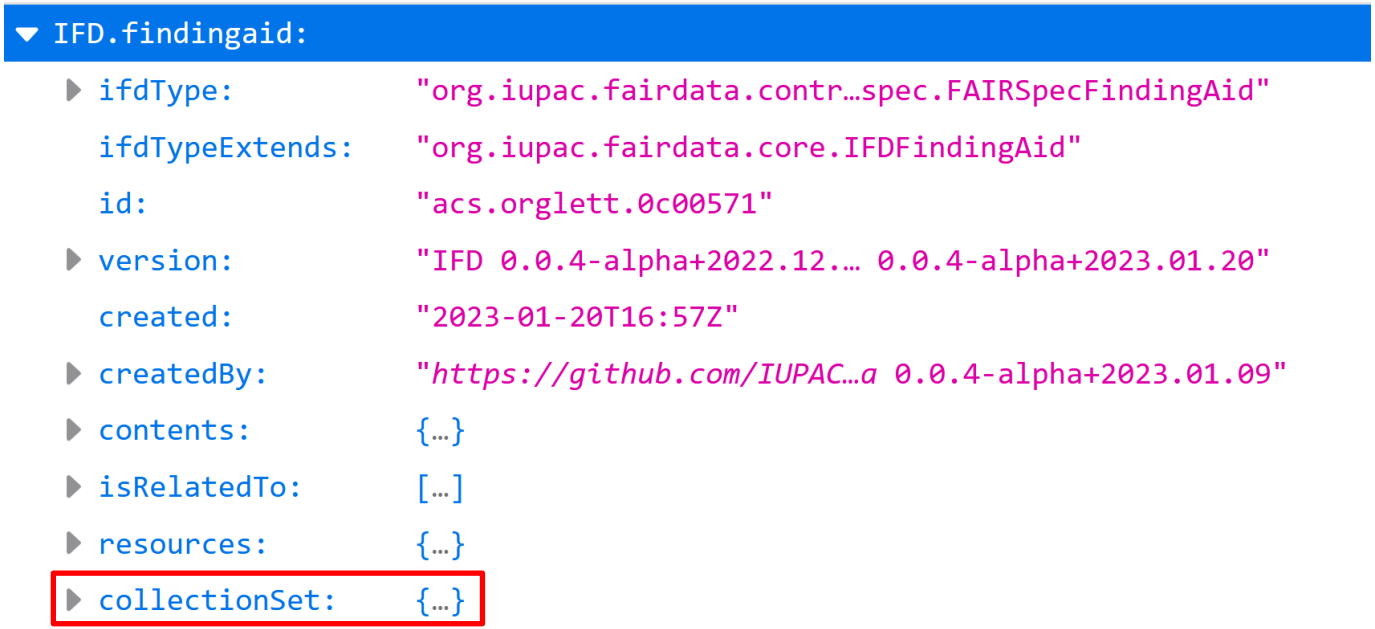
- Serializable as JSON
  (or XML, in principle)
- Highly structured
- Object-oriented
- Domain/Subdomain-specific
- Broadly extensible
- Applicable throughout the full data life cycle (and beyond!)
- Easily implemented

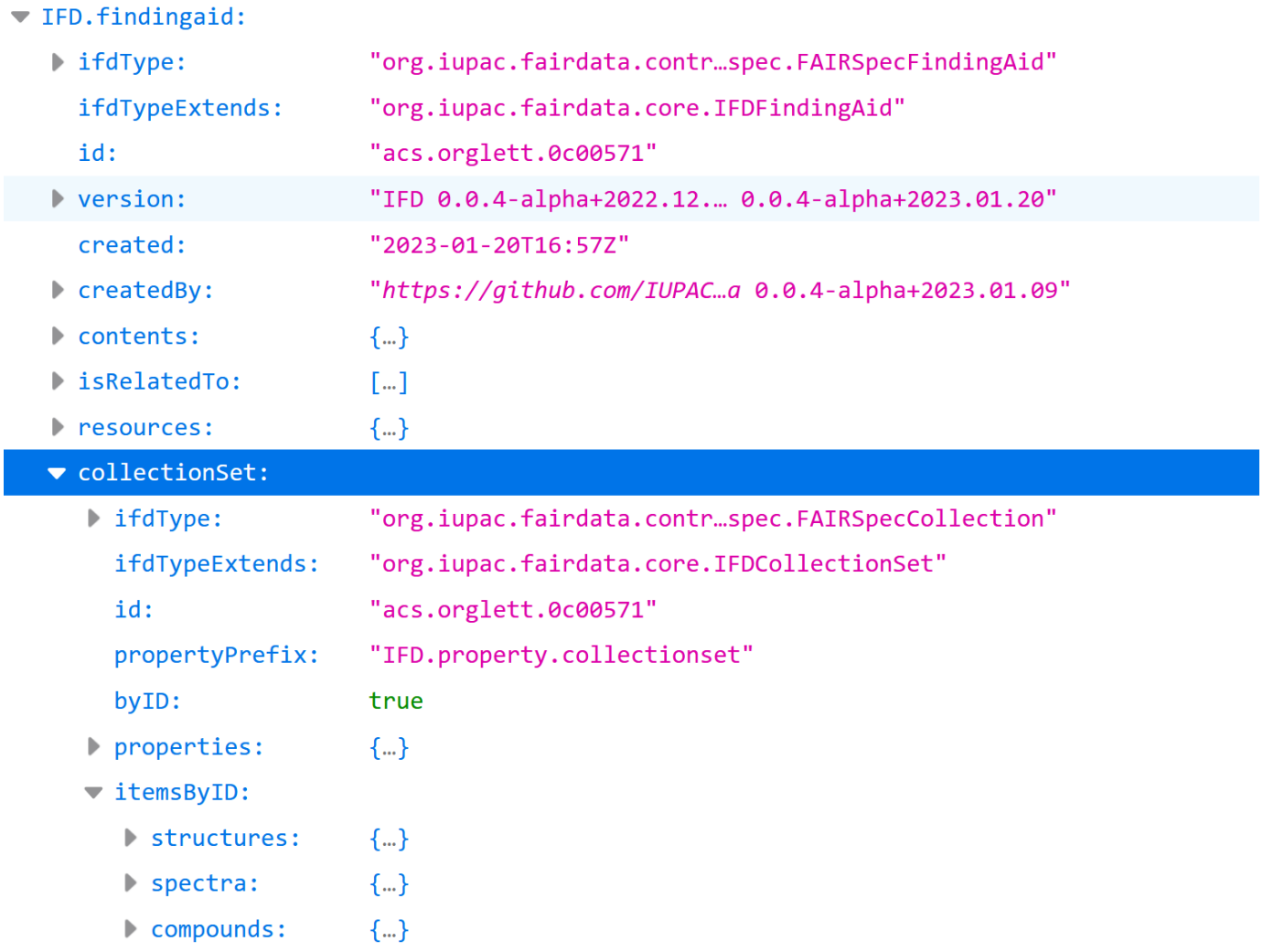


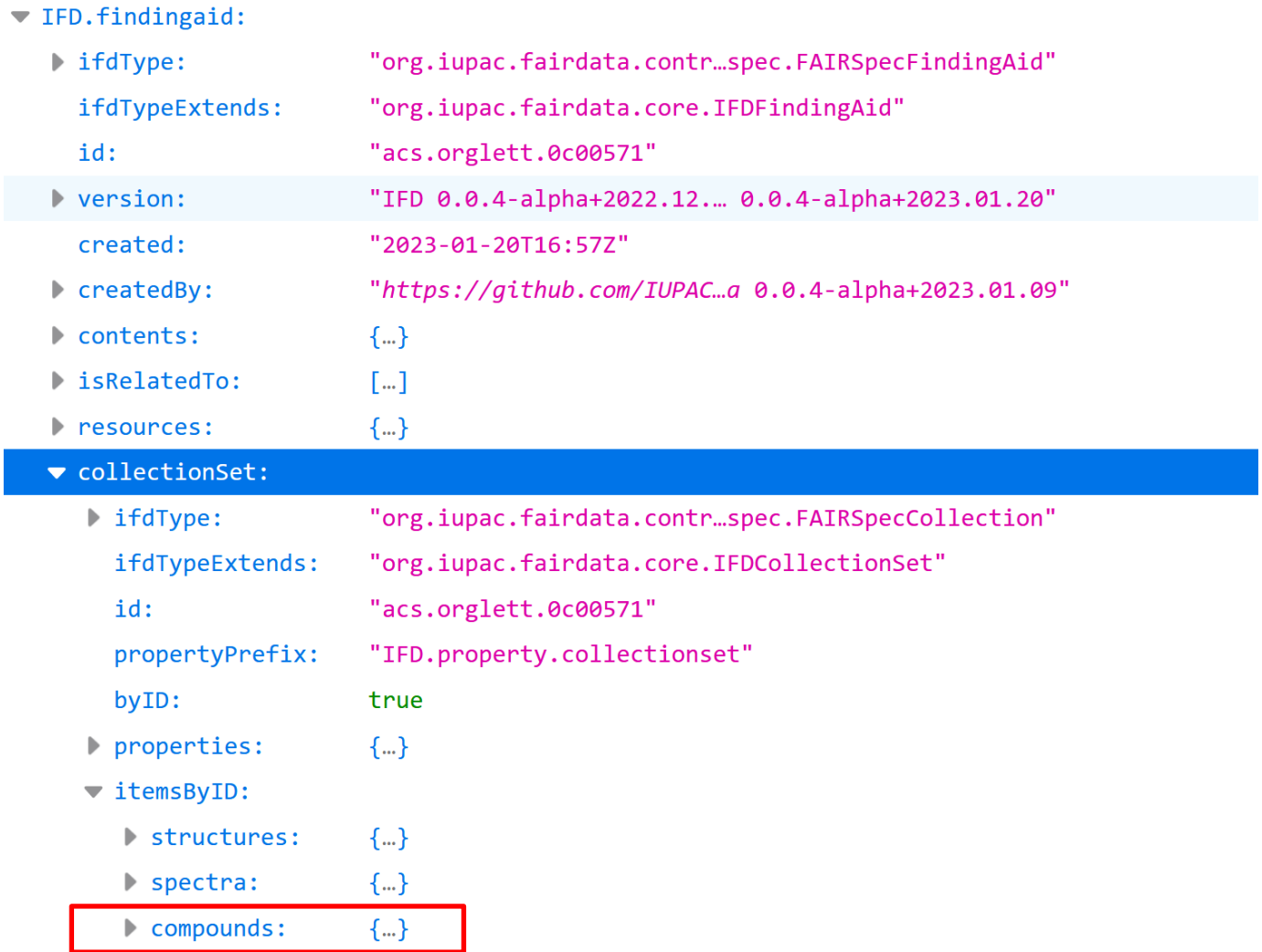
https://www.loc.gov/rr/ead


IFD.findingaid:
ifdType: "org.iupac.fairdata.contr…spec.FAIRSpecFindingAid"
ifdTypeExtends: "org.iupac.fairdata.core.IFDFindingAid"
id: "acs.orglett.0c00571"
version: "IFD 0.0.4-alpha+2022.12.… 0.0.4-alpha+2023.01.20"
created: "2023-01-20T16:57Z"
createdBy: "https://github.com/IUPAC…a 0.0.4-alpha+2023.01.09"
contents: {…}
isRelatedTo: […]
resources: {…}
collectionSet: {…}

```
contents:
  collections:
    0:
      count:        30
      id:           "structures"
      type:         "org.iupac.fairdata.struc….IFDStructureCollection"
      typeExtends:  "org.iupac.fairdata.core.IFDCollection"
    1:
      count:        114
      id:           "spectra"
      type:         "org.iupac.fairdata.datao…IFDDataObjectCollection"
      typeExtends:  "org.iupac.fairdata.core.IFDCollection"
    2:
      count:        30
      id:           "compounds"
      type:         "org.iupac.fairdata.contr…RSpecCompoundCollection"
      typeExtends:  "org.iupac.fairdata.deriv…FDAssociationCollection"
```


IFD.findingaid:
ifdType: "org.iupac.fairdata.contr…spec.FAIRSpecFindingAid"
ifdTypeExtends: "org.iupac.fairdata.core.IFDFindingAid"
id: "acs.orglett.0c00571"
version: "IFD 0.0.4-alpha+2022.12.… 0.0.4-alpha+2023.01.20"
created: "2023-01-20T16:57Z"
createdBy: "https://github.com/IUPAC…a 0.0.4-alpha+2023.01.09"
contents: {…}
isRelatedTo: […]
resources: {…}
collectionSet: {…}

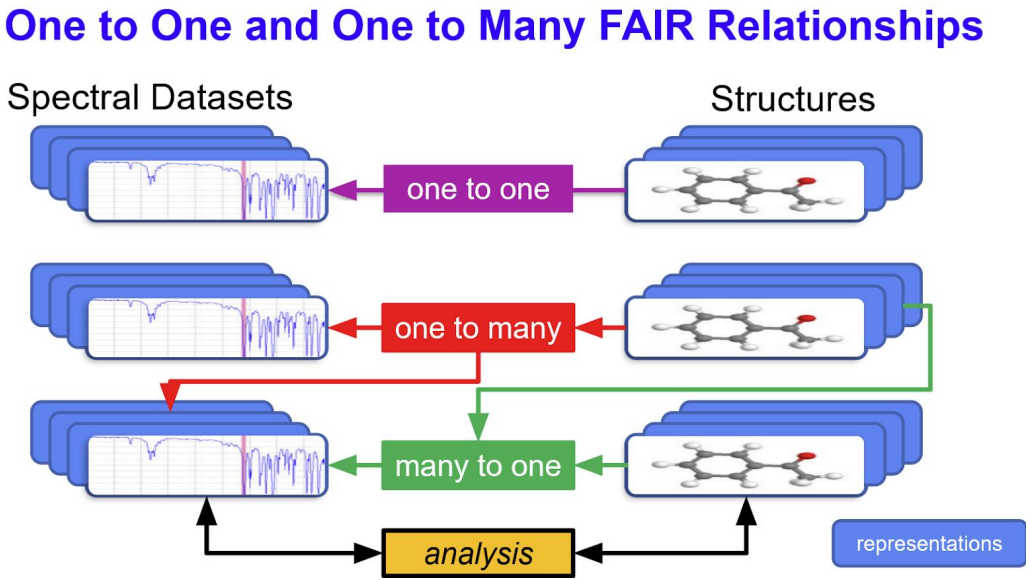

IFD.findingaid:
ifdType: "org.iupac.fairdata.contr…spec.FAIRSpecFindingAid"
ifdTypeExtends: "org.iupac.fairdata.core.IFDFindingAid"
id: "acs.orglett.0c00571"
version: "IFD 0.0.4-alpha+2022.12.… 0.0.4-alpha+2023.01.20"
created: "2023-01-20T16:57Z"
createdBy: "https://github.com/IUPAC…a 0.0.4-alpha+2023.01.09"
contents: {…}
isRelatedTo: […]
resources: {…}
collectionSet:
ifdType: "org.iupac.fairdata.contr…spec.FAIRSpecCollection"
ifdTypeExtends: "org.iupac.fairdata.core.IFDCollectionSet"
id: "acs.orglett.0c00571"
propertyPrefix: "IFD.property.collectionset"
byID: true
properties: {…}
itemsByID:
structures: {…}
spectra: {…}
compounds: {…}


IFD.findingaid:
ifdType: "org.iupac.fairdata.contr…spec.FAIRSpecFindingAid"
ifdTypeExtends: "org.iupac.fairdata.core.IFDFindingAid"
id: "acs.orglett.0c00571"
version: "IFD 0.0.4-alpha+2022.12.… 0.0.4-alpha+2023.01.20"
created: "2023-01-20T16:57Z"
createdBy: "https://github.com/IUPAC…a 0.0.4-alpha+2023.01.09"
contents: {…}
isRelatedTo: […]
resources: {…}
collectionSet:
ifdType: "org.iupac.fairdata.contr…spec.FAIRSpecCollection"
ifdTypeExtends: "org.iupac.fairdata.core.IFDCollectionSet"
id: "acs.orglett.0c00571"
propertyPrefix: "IFD.property.collectionset"
byID: true
properties: {…}
itemsByID:
structures: {…}
spectra: {…}
compounds: {…}


compounds:
ifdType: "org.iupac.fairdata.contr…RSpecCompoundCollection"
ifdTypeExtends: "org.iupac.fairdata.deriv…FDAssociationCollection"
id: "compounds"
itemType: "org.iupac.fairdata.contr…SpecCompoundAssociation"
itemTypeExtends: "org.iupac.fairdata.deriv…ata.core.IFDAssociation"
itemsByID:
1c:
id: "1c"
itemsByID:
structures:
0: "1c"
spectra:
0: "1c/13C-NMR"
1: "1c/1H-NMR"
2: "1c/HRMS"

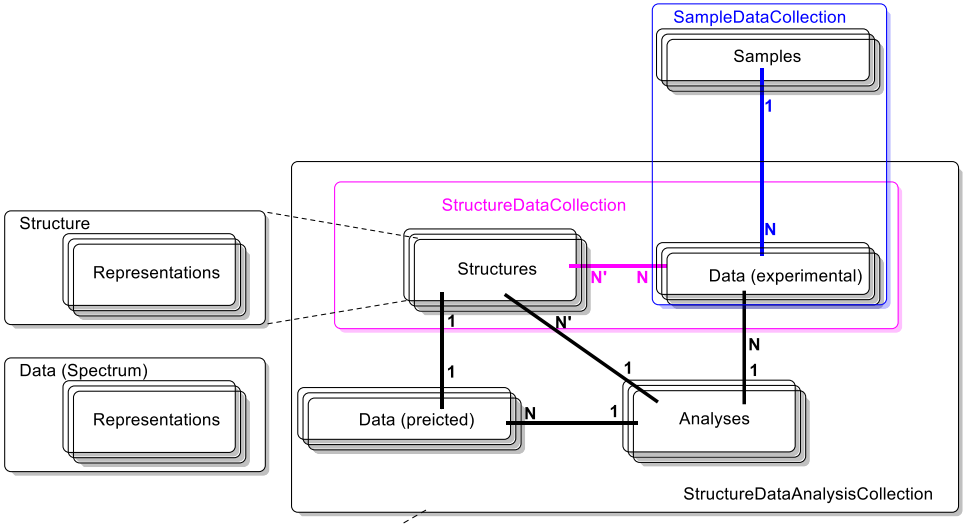
Summary

Spectral data is nothing without a sample identifier or an association with one or more chemical structures. Collections should be valued, not dismembered.



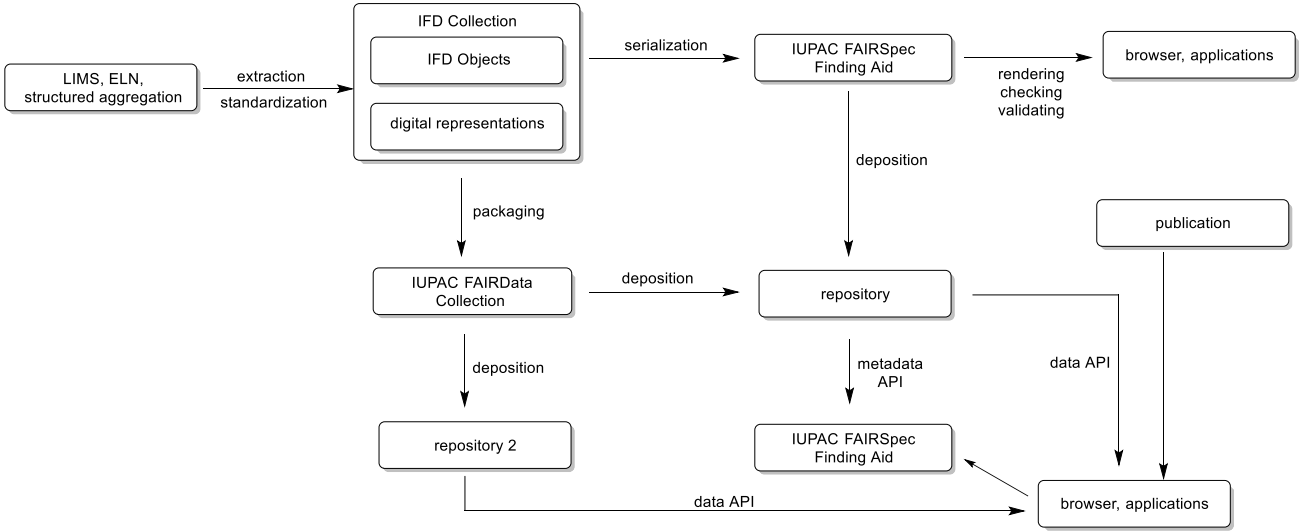
Summary

The developing **IUPAC FAIRData Collection Model** is capable of describing complex domain-specific relationships between a wide variety data objects.



Summary

The developing **IUPAC FAIRSpec Finding Aid Standard** is positioned to be the basis for a **common protocol** for the management and distribution of spectral data *and their associated chemical structures* throughout the whole data cycle.



Thank you!

Please join us tomorrow (Monday) afternoon for a discussion of what it means to be “FAIRSpec-Ready”

IUPAC FAIRSpec-ready collections: Recommendations for researchers, authors, and publishers

SESSION: Advancing FAIR Chemistry: Developing New Services for Sharing Chemical Data

2:00 PM - 5:55 PM (in person)
Room 112 - Indiana Convention Center

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