Open databases: what do we have, where are we going

Saulius Gražulis

Vilnius, 2023

Vilnius University Institute of Biotechnology

Id: slides.tex 2050 2023-02-16 14:24:22Z saulius February 16, 2023

Data in Crystallography

- Numbers of published protein structures;
- Numbers of published “small molecule” structures;
- Number of crystallographic and chemical papers
- Incidentally, number of chemical entities in chemical databases

Available data records

Experimental databases:

Database	Nr. rec.[1]	License	Web Ref.
PDB	**201 515**	Open	wwpdb.org, rcsb.org
COD	**497 457**	Open	crystallography.net
MAGNDATA	**2 034**	Open	Bilbao MAGNDATA
B-IncStrDB	**256**	Open	Bilbao B-IncStrDB

[1] As of 2023-02-15

Crystallographic databases

Open Access:

- Protein Data Bank;
- Crystallography Open Database (and its “sisters”);
- Bilbao Magnetic Structure Database

Proprietary:

- CCDC
- ICSD
- PDF
- Pauling File
- ...

Crystallographic databases

Open Access:

- Protein Data Bank;
- Crystallography Open Database (and its “sisters”);
- Bilbao Magnetic Structure Database

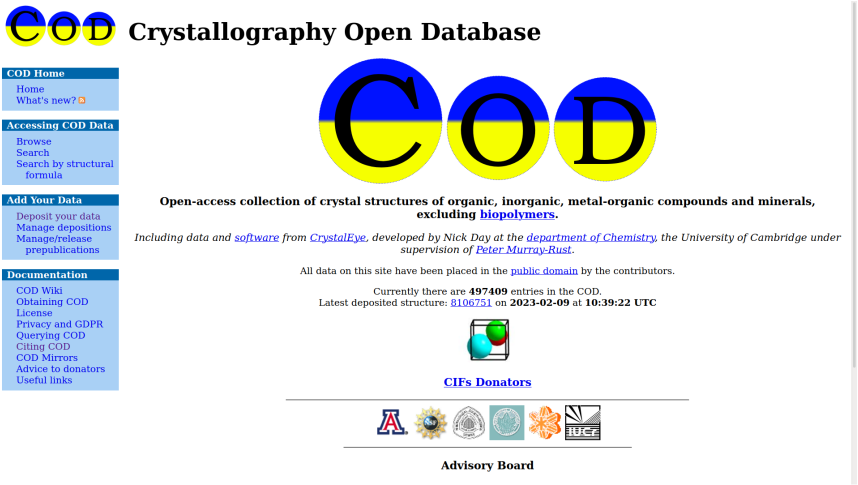
Proprietary:

- CCDC
- ICSD
- PDF
- Pauling File
- ...

About $\mathbf{10^6}$ – $\mathbf{10^7}$ crystallographic records are available.

The Crystallography Open Database

https://www.crystallography.net/cod



Crystallography Open Database

Open-access collection of crystal structures of organic, inorganic, metal-organic compounds and minerals, excluding biopolymers.

Including data and software from CrystalEye, developed by Nick Day at the department of Chemistry, the University of Cambridge under supervision of Peter Murray-Rust.

All data on this site have been placed in the public domain by the contributors.

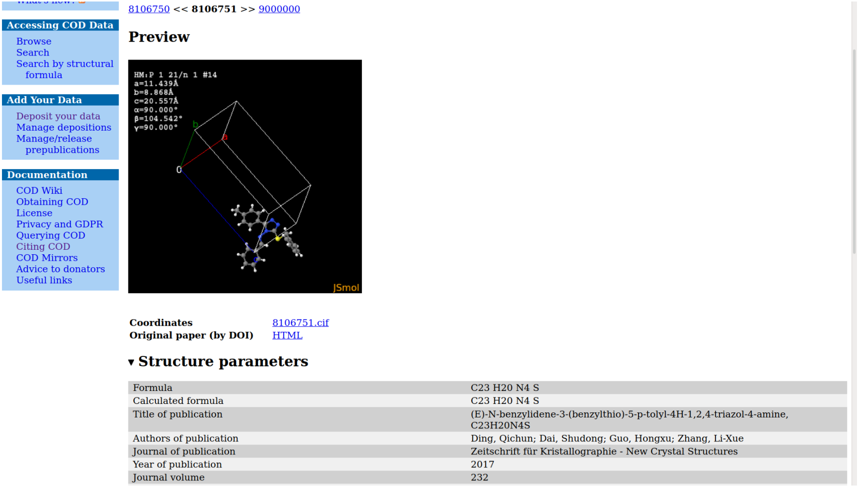
Currently there are **497409** entries in the COD.
Latest deposited structure: 8106751 on **2023-02-09** at **10:39:22 UTC**

CIFs Donators

Advisory Board

The Crystallography Open Database

https://www.crystallography.net/cod



What's new?

Accessing COD Data

Browse
Search
Search by structural formula

Add Your Data

Deposit your data
Manage depositions
Manage/release prepublications

Documentation

COD Wiki
Obtaining COD
License
Privacy and GDPR
Querying COD
Citing COD
COD Mirrors
Advice to donators
Useful links

8106750 << **8106751** >> 9000000

Preview

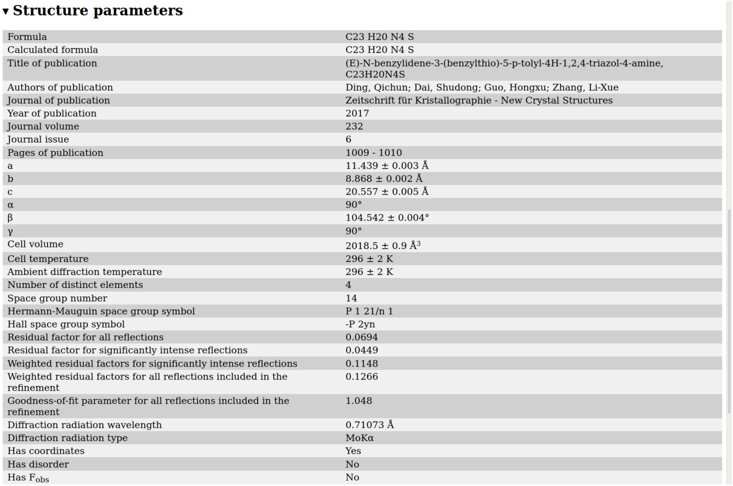
Coordinates 8106751.cif
Original paper (by DOI) HTML

▾ Structure parameters

Formula	C23 H20 N4 S
Calculated formula	C23 H20 N4 S
Title of publication	(E)-N-benzylidene-3-(benzylthio)-5-p-tolyl-4H-1,2,4-triazol-4-amine, C23H20N4S
Authors of publication	Ding, Qichun; Dai, Shudong; Guo, Hongxu; Zhang, Li-Xue
Journal of publication	Zeitschrift für Kristallographie - New Crystal Structures
Year of publication	2017
Journal volume	232

The Crystallography Open Database

https://www.crystallography.net/cod



▾ **Structure parameters**

Formula	C23 H20 N4 S
Calculated formula	C23 H20 N4 S
Title of publication	(E)-N-benzylidene-3-(benzylthio)-5-p-tolyl-4H-1,2,4-triazol-4-amine, C23H20N4S
Authors of publication	Ding, Qichun; Dai, Shudong; Guo, Hongxu; Zhang, Li-Xue
Journal of publication	Zeitschrift für Kristallographie - New Crystal Structures
Year of publication	2017
Journal volume	232
Journal issue	6
Pages of publication	1009 - 1010
a	11.439 ± 0.003 Å
b	8.868 ± 0.002 Å
c	20.557 ± 0.005 Å
α	90°
β	104.542 ± 0.004°
γ	90°
Cell volume	2018.5 ± 0.9 Å^3
Cell temperature	296 ± 2 K
Ambient diffraction temperature	296 ± 2 K
Number of distinct elements	4
Space group number	14
Hermann-Mauguin space group symbol	P 1 21/n 1
Hall space group symbol	-P 2yn
Residual factor for all reflections	0.0694
Residual factor for significantly intense reflections	0.0449
Weighted residual factors for significantly intense reflections	0.1148
Weighted residual factors for all reflections included in the refinement	0.1266
Goodness-of-fit parameter for all reflections included in the refinement	1.048
Diffraction radiation wavelength	0.71073 Å
Diffraction radiation type	MoKα
Has coordinates	Yes
Has disorder	No
Has F_{obs}	No

Quality criteria for data

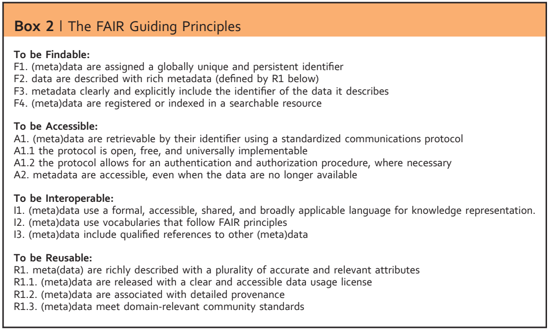
- data should be FAIR;
- data should be machine readable;
- data should support scientific conclusions;
- data should be open;

Quality criteria for data

- data should be FAIR;
- data should be machine readable;
- data should support scientific conclusions;
- data should be open;

"As open as possible, as closed as necessary"
[Landi et al., 2020]

COD FAIRness



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

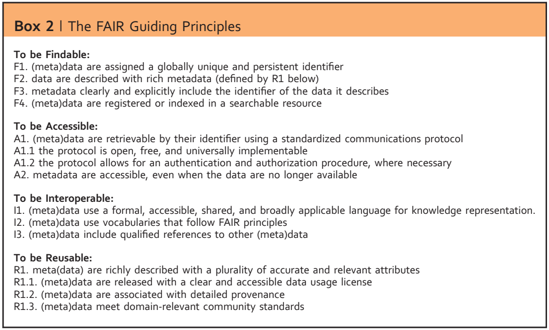
To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

✓	F1.	COD IDs
✓	F2.	CIF `_journal_...`, etc.
✓	F3.	CIF `_cod_database_code`
✓	F4.	crystallography.net
✓	A1.	HTTP(S), SVN, Rsync
✓	A1.1.	HTTP(S), SVN, Rsync
✓	A1.2.	HTTP(S), SVN, Rsync
✓	A2.	COD retraction policy
✓	I1.	CIF syntax
✓	I2.	CIF dictionaries
✓	I3.	COD cross-references
✓	R1.	CIF `_journal_...`, etc.
✓	R1.1	COD: CC0
✓	R1.2	COD SVN repository
✓	R1.3	IUCr criteria

[Wilkinson et al., 2016]

COD FAIRness



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

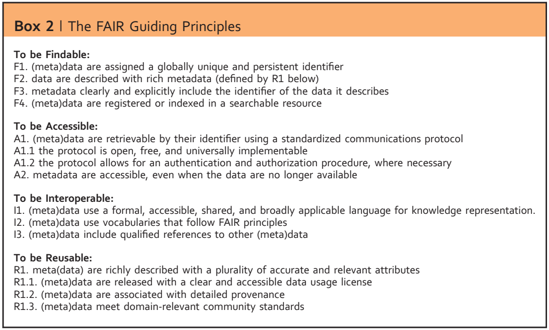
To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

[Wilkinson et al., 2016]

✓	F1.	COD IDs
✓	F2.	CIF `_journal_...`, etc.
✓	F3.	CIF `_cod_database_code`
✓	F4.	crystallography.net
✓	A1.	HTTP(S), SVN, Rsync
✓	A1.1.	HTTP(S), SVN, Rsync
✓	A1.2.	HTTP(S), SVN, Rsync
✓	A2.	COD retraction policy
✓	I1.	CIF syntax
✓	I2.	CIF dictionaries
✓	I3.	COD cross-references
✓	R1.	CIF `_journal_...`, etc.
✓	R1.1	COD: CC0
✓	R1.2	COD SVN repository
✓	R1.3	IUCr criteria

COD FAIRness



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

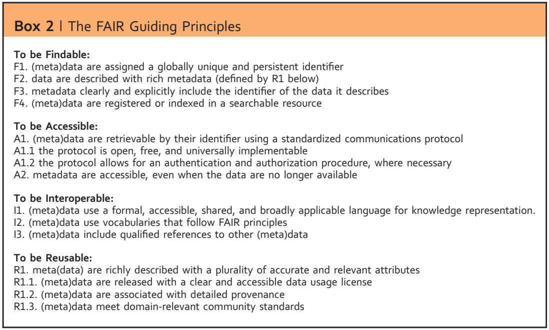
To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

✓ F1.	COD IDs
✓ F2.	CIF `_journal_...`, etc.
✓ F3.	CIF `_cod_database_code`
✓ F4.	crystallography.net
✓ A1.	HTTP(S), SVN, Rsync
✓ A1.1.	HTTP(S), SVN, Rsync
✓ A1.2.	HTTP(S), SVN, Rsync
✓ A2.	COD retraction policy
✓ I1.	CIF syntax
✓ I2.	CIF dictionaries
✓ I3.	COD cross-references
✓ R1.	CIF `_journal_...`, etc.
✓ R1.1	COD: CC0
✓ R1.2	COD SVN repository
✓ R1.3	IUCr criteria

[Wilkinson et al., 2016]

COD FAIRness



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

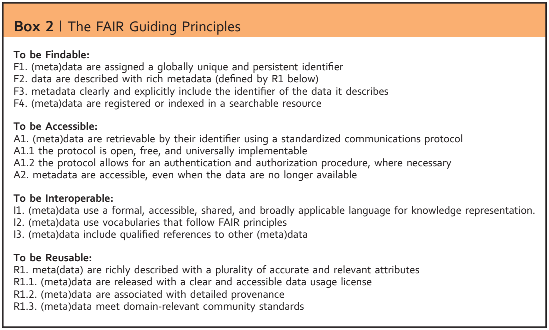
To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

✓	F1.	COD IDs
✓	F2.	CIF `_journal_...`, etc.
✓	F3.	CIF `_cod_database_code`
✓	F4.	crystallography.net
✓	A1.	HTTP(S), SVN, Rsync
✓	A1.1.	HTTP(S), SVN, Rsync
✓	A1.2.	HTTP(S), SVN, Rsync
✓	A2.	COD retraction policy
✓	I1.	CIF syntax
✓	I2.	CIF dictionaries
✓	I3.	COD cross-references
✓	R1.	CIF `_journal_...`, etc.
✓	R1.1	COD: CC0
✓	R1.2	COD SVN repository
✓	R1.3	IUCr criteria

[Wilkinson et al., 2016]

COD FAIRness



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

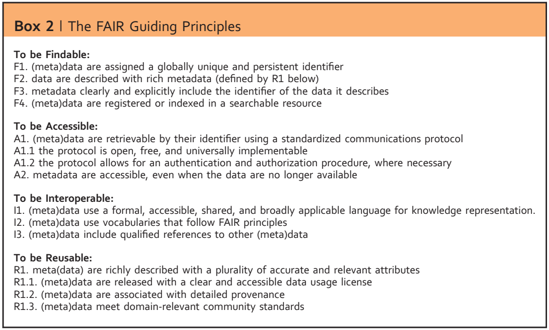
To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

✓	F1.	COD IDs
✓	F2.	CIF `_journal_...`, etc.
✓	F3.	CIF `_cod_database_code`
✓	F4.	crystallography.net
✓	A1.	HTTP(S), SVN, Rsync
✓	A1.1.	HTTP(S), SVN, Rsync
✓	A1.2.	HTTP(S), SVN, Rsync
✓	A2.	COD retraction policy
✓	I1.	CIF syntax
✓	I2.	CIF dictionaries
✓	I3.	COD cross-references
✓	R1.	CIF `_journal_...`, etc.
✓	R1.1	COD: CC0
✓	R1.2	COD SVN repository
✓	R1.3	IUCr criteria

[Wilkinson et al., 2016]

COD FAIRness



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

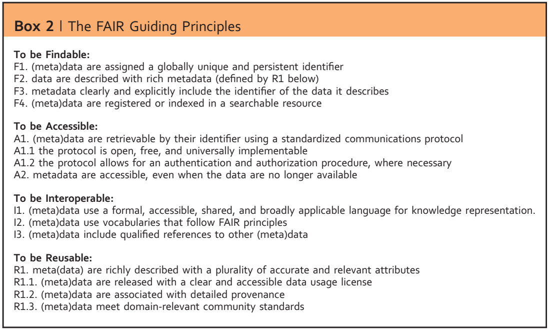
To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

✓	F1.	COD IDs
✓	F2.	CIF `_journal_...`, etc.
✓	F3.	CIF `_cod_database_code`
✓	F4.	crystallography.net
✓	A1.	HTTP(S), SVN, Rsync
✓	A1.1.	HTTP(S), SVN, Rsync
✓	A1.2.	HTTP(S), SVN, Rsync
✓	A2.	COD retraction policy
✓	I1.	CIF syntax
✓	I2.	CIF dictionaries
✓	I3.	COD cross-references
✓	R1.	CIF `_journal_...`, etc.
✓	R1.1	COD: CC0
✓	R1.2	COD SVN repository
✓	R1.3	IUCr criteria

[Wilkinson et al., 2016]

COD FAIRness



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

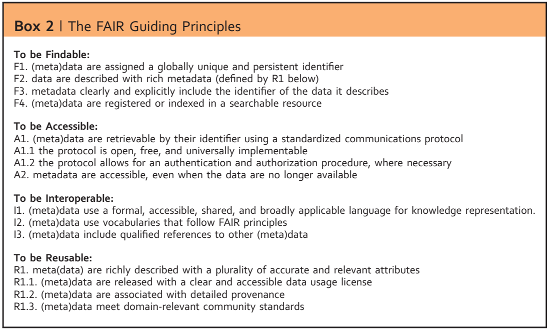
To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

[Wilkinson et al., 2016]

✓	F1.	COD IDs
✓	F2.	CIF `_journal_...`, etc.
✓	F3.	CIF `_cod_database_code`
✓	F4.	crystallography.net
✓	A1.	HTTP(S), SVN, Rsync
✓	A1.1.	HTTP(S), SVN, Rsync
✓	A1.2.	HTTP(S), SVN, Rsync
✓	A2.	COD retraction policy
✓	I1.	CIF syntax
✓	I2.	CIF dictionaries
✓	I3.	COD cross-references
✓	R1.	CIF `_journal_...`, etc.
✓	R1.1	COD: CC0
✓	R1.2	COD SVN repository
✓	R1.3	IUCr criteria

COD FAIRness



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

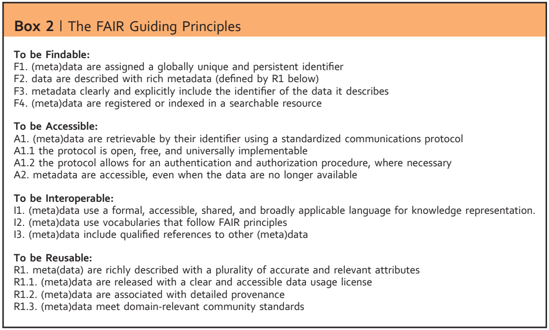
To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

✓	F1.	COD IDs
✓	F2.	CIF `_journal_...`, etc.
✓	F3.	CIF `_cod_database_code`
✓	F4.	crystallography.net
✓	A1.	HTTP(S), SVN, Rsync
✓	A1.1.	HTTP(S), SVN, Rsync
✓	A1.2.	HTTP(S), SVN, Rsync
✓	A2.	COD retraction policy
✓	I1.	CIF syntax
✓	I2.	CIF dictionaries
✓	I3.	COD cross-references
✓	R1.	CIF `_journal_...`, etc.
✓	R1.1	COD: CC0
✓	R1.2	COD SVN repository
✓	R1.3	IUCr criteria

[Wilkinson et al., 2016]

COD FAIRness



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

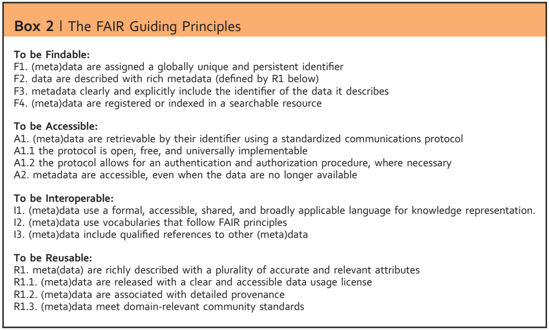
To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

✓	F1.	COD IDs
✓	F2.	CIF `_journal_...`, etc.
✓	F3.	CIF `_cod_database_code`
✓	F4.	crystallography.net
✓	A1.	HTTP(S), SVN, Rsync
✓	A1.1.	HTTP(S), SVN, Rsync
✓	A1.2.	HTTP(S), SVN, Rsync
✓	A2.	COD retraction policy
✓	I1.	CIF syntax
✓	I2.	CIF dictionaries
✓	I3.	COD cross-references
✓	R1.	CIF `_journal_...`, etc.
✓	R1.1	COD: CC0
✓	R1.2	COD SVN repository
✓	R1.3	IUCr criteria

[Wilkinson et al., 2016]

COD FAIRness



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

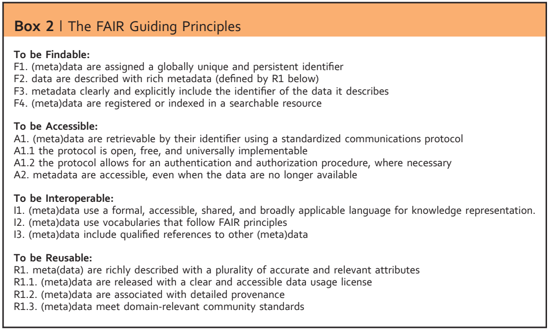
To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

✓	F1.	COD IDs
✓	F2.	CIF `_journal_...`, etc.
✓	F3.	CIF `_cod_database_code`
✓	F4.	crystallography.net
✓	A1.	HTTP(S), SVN, Rsync
✓	A1.1.	HTTP(S), SVN, Rsync
✓	A1.2.	HTTP(S), SVN, Rsync
✓	A2.	COD retraction policy
✓	I1.	CIF syntax
✓	I2.	CIF dictionaries
✓	I3.	COD cross-references
✓	R1.	CIF `_journal_...`, etc.
✓	R1.1	COD: CC0
✓	R1.2	COD SVN repository
✓	R1.3	IUCr criteria

[Wilkinson et al., 2016]

COD FAIRness



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

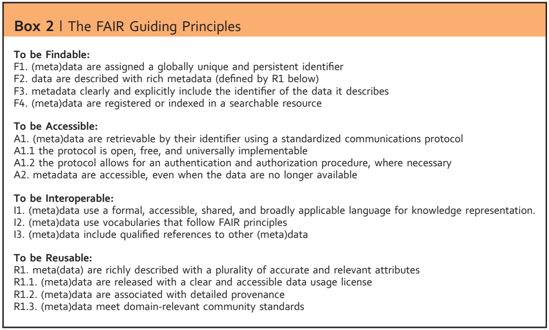
To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

✓	F1.	COD IDs
✓	F2.	CIF `_journal_...`, etc.
✓	F3.	CIF `_cod_database_code`
✓	F4.	crystallography.net
✓	A1.	HTTP(S), SVN, Rsync
✓	A1.1.	HTTP(S), SVN, Rsync
✓	A1.2.	HTTP(S), SVN, Rsync
✓	A2.	COD retraction policy
✓	I1.	CIF syntax
✓	I2.	CIF dictionaries
✓	I3.	COD cross-references
✓	R1.	CIF `_journal_...`, etc.
✓	R1.1	COD: CC0
✓	R1.2	COD SVN repository
✓	R1.3	IUCr criteria

[Wilkinson et al., 2016]

COD FAIRness



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

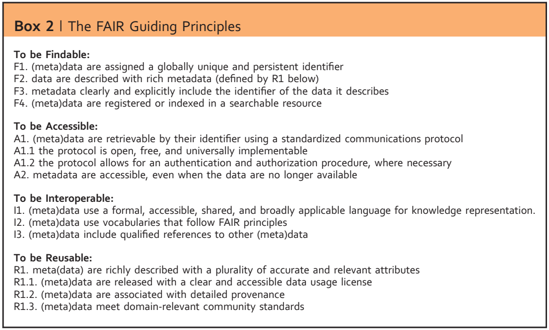
To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

[Wilkinson et al., 2016]

✓	F1.	COD IDs
✓	F2.	CIF `_journal_...`, etc.
✓	F3.	CIF `_cod_database_code`
✓	F4.	crystallography.net
✓	A1.	HTTP(S), SVN, Rsync
✓	A1.1.	HTTP(S), SVN, Rsync
✓	A1.2.	HTTP(S), SVN, Rsync
✓	A2.	COD retraction policy
✓	I1.	CIF syntax
✓	I2.	CIF dictionaries
✓	I3.	COD cross-references
✓	R1.	CIF `_journal_...`, etc.
✓	R1.1	COD: CC0
✓	R1.2	COD SVN repository
✓	R1.3	IUCr criteria

COD FAIRness



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

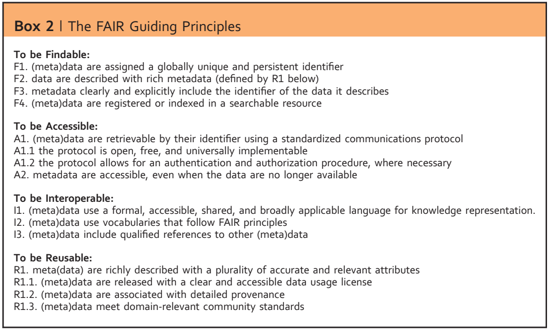
To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

✓	F1.	COD IDs
✓	F2.	CIF `_journal_...`, etc.
✓	F3.	CIF `_cod_database_code`
✓	F4.	crystallography.net
✓	A1.	HTTP(S), SVN, Rsync
✓	A1.1.	HTTP(S), SVN, Rsync
✓	A1.2.	HTTP(S), SVN, Rsync
✓	A2.	COD retraction policy
✓	I1.	CIF syntax
✓	I2.	CIF dictionaries
✓	I3.	COD cross-references
✓	R1.	CIF `_journal_...`, etc.
✓	R1.1	COD: CC0
✓	R1.2	COD SVN repository
✓	R1.3	IUCr criteria

[Wilkinson et al., 2016]

COD FAIRness



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

✓	F1.	COD IDs
✓	F2.	CIF `_journal_...`, etc.
✓	F3.	CIF `_cod_database_code`
✓	F4.	crystallography.net
✓	A1.	HTTP(S), SVN, Rsync
✓	A1.1.	HTTP(S), SVN, Rsync
✓	A1.2.	HTTP(S), SVN, Rsync
✓	A2.	COD retraction policy
✓	I1.	CIF syntax
✓	I2.	CIF dictionaries
✓	I3.	COD cross-references
✓	R1.	CIF `_journal_...`, etc.
✓	R1.1	COD: CC0
✓	R1.2	COD SVN repository
✓	R1.3	IUCr criteria

[Wilkinson et al., 2016]

COD data purposes

- Find exact structure of the crystal;
- Determine material structure-property relations;
- Demonstrate that the synthesised compound is the one we expected;

COD data purposes

- Find exact structure of the crystal;
- Determine material structure-property relations;
- Demonstrate that the synthesised compound is the one we expected;

We must be prepared for unexpected data reuse

IUCr quality criteria

- CIF framework
  - CIF syntax (CIF 1.1, CIF 2);
  - CIF Dictionaries;
- IUCr publication requirements (Platon Alerts);

COD data curation principles

- Strictly stick to IUCr standards (CIF syntax, dictionaries);
- Do not invent data;
- Better to have no data than wrong data;
- Consult original papers or authors themselves if in doubt;
- Document: record and explain (justify) all changes;
- Keep track of all changes in a version control system;
- Keep data provenance (original file names);

Three levels of data validation

- Check of file syntax;
- Validation against dictionaries;
- Domain-specific checks:
    - internal consistency;
    - coherence with raw data;
    - scientific plausibility;

COD data validation

COD data validation policies:

1. Syntactic checks:

```
$ cifparse 7234818.cif
```

2. Semantic validation (against dictionaries)

```
$ cif_validate -D cif_core.dic 7234818.cif
```

3. Database-specific checks

```
$ cif_cod_check 7234818.cif
```

Syntax errors in *published* CIFS

Among 3 most prolific publishers in 2021–2022:

- ≈ 12 000 files harvested,
- ≈ 43 000 structures deposited to the COD,
- **52** correctable syntax errors detected in **14** files.

E.g.:

```
cifparse: example1.cif(15,39) data_block_1: ERROR, incorrect CIF syntax:
 _exptl_crystal_description  structure obtained
                                       ^
```

Syntax errors in *published* CIFS

Among 3 most prolific publishers in 2021–2022:

- ≈ 12 000 files harvested,
- ≈ 43 000 structures deposited to the COD,
- **52** correctable syntax errors detected in **14** files.

E.g.:

```
cifparse: example1.cif(15,39) data_block_1: ERROR, incorrect CIF syntax:
 _exptl_crystal_description  structure obtained
                                       ^
```

Most of these errors are fixed automatically by the COD CIF parser [Merkys et al., 2016], but ...

Syntax errors in *published* CIFS

Among 3 most prolific publishers in 2021–2022:

- ≈ 12 000 files harvested,
- ≈ 43 000 structures deposited to the COD,
- **52** correctable syntax errors detected in **14** files.

E.g.:

```
cifparse: example1.cif(15,39) data_block_1: ERROR, incorrect CIF syntax:
 _exptl_crystal_description  structure obtained
                                       ^
```

Most of these errors are fixed automatically by the COD CIF parser [Merkys et al., 2016], but ...

Data do not get the same attention from reviewers as the main text.

Syntax formally right, but ...

```
_publ_contact_author
;
    Name, Surname
    Department of Chemistry
    University of ...
    ;
_publ_contact_letter This is the CIF file for ...
_publ_contact_author_phone           ;
;
_publ_section_title
;
  The correct title follows ...
;
```

[Boerrigter 2023, pers. comm.]

Syntax formally right, but ...

```
_publ_contact_author
;
    Name, Surname
    Department of Chemistry
    University of ...
    ;
_publ_contact_letter This is the CIF file for ...
_publ_contact_author_phone            ;
;
_publ_section_title
;
  The correct title follows ...
;
```

[Boerrigter 2023, pers. comm.]

Syntax formally right, but ...

```
_publ_contact_author
;
    Name, Surname
    Department of Chemistry
    University of ...
    ;
_publ_contact_letter This is the CIF file for ...
_publ_contact_author_phone            ;
;
_publ_section_title
;
  The correct title follows ...
;
```

[Boerrigter 2023, pers. comm.]

Data review and the use of proper authoring tools could help...

Description of semantics

CIF dictionaries

```
data_cell_length_
    loop_ _name                '_cell_length_a'
                               '_cell_length_b'
                               '_cell_length_c'
    _category                    cell
    _type                        numb
    _type_conditions             esd
    _enumeration_range           0.0:
    _units                       A
    _units_detail              'angstroms'
    _definition
;              Unit-cell lengths in angstroms corresponding to the structure
               reported. The values of _refln_index_h, *_k, *_l must
               correspond to the cell defined by these values and _cell_angle_
               values. The values of _diffrn_refln_index_h, *_k, *_l may not
               correspond to these values if a cell transformation took place
               following the measurement of the diffraction intensities. See
               also _diffrn_reflns_transf_matrix_.
;
```

COD data curation – validation against dictionaries

- Several types of dictionaries (DDL1, DDL2, DDLm);
- COD validation tools in CIF1 and CIF2 frameworks (`cif_validate`, `ddlm_validate`[2]);

[Vaitkus et al., 2021]

[2] Available in the `cod-tools` package on Debian and Ubuntu systems.

[3] https://sql.crystallography.net/db/cod_validation/validation_issue

COD data curation – validation against dictionaries

- Several types of dictionaries (DDL1, DDL2, DDLm);
- COD validation tools in CIF1 and CIF2 frameworks (`cif_validate`, `ddlm_validate`[2]);

[Vaitkus et al., 2021]

Running validation on all COD yields over **11 mln.** validation messages...[3]

[2] Available in the `cod-tools` package on Debian and Ubuntu systems.

[3] https://sql.crystallography.net/db/cod_validation/validation_issue

COD validation examples

```
/usr/bin/cif_validate: 1506432.cif data_1506432:
NOTE, data item '_atom_site_aniso_label' contains value 'F40'
that was not found among the values of the parent data item
'_atom_site_label'.
```

COD validation examples

```
/usr/bin/cif_validate: 1506432.cif data_1506432:
NOTE, data item '_atom_site_aniso_label' contains value 'F40'
that was not found among the values of the parent data item
'_atom_site_label'.
```

```
loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_U_iso_or_equiv
# ... some data names omitted for brevity
>F40 F 0.21810(11) -1.5061(4) 0.7984(2) 0.0684(9) # ...
F41 F 0.29902(11) -1.4446(4) 0.8587(2) 0.0724(9)  # ...
```

COD validation examples

```
/usr/bin/cif_validate: 1506432.cif data_1506432:
NOTE, data item '_atom_site_aniso_label' contains value 'F40'
that was not found among the values of the parent data item
'_atom_site_label'.
```

```
loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_U_iso_or_equiv
# ... some data names omitted for brevity
>F40 F 0.21810(11) -1.5061(4) 0.7984(2) 0.0684(9) # ...
F41 F 0.29902(11) -1.4446(4) 0.8587(2) 0.0724(9)  # ...
```

Validation *might* help to catch data errors if applied consistently during the publication.

COD entry validation examples

- Example: wrong coordinates;
- Example: missing/wrong keys;
- Example: mistyped enumerator values;
- Example: typos in data/OCR errors?

COD entry validation examples

- Example: wrong coordinates;
- Example: missing/wrong keys;
- Example: mistyped enumerator values;
- Example: typos in data/OCR errors?

Ideally, validation should be applied during the data peer review process

Corrupted data in a text field

```
_iucr_refine_reflections_details
;
   0   0   2   -0.20    0.30   99-0.77969 0.78029 0.62494-0.62494 0.03182 0.03182
   0   0   2   -0.30    0.30  209 0.78190-0.78130-0.62292 0.62292 0.03182 0.03182

# ... lines omitted for brevity

 -15  -3  -5   -4.60    8.40  316-0.62905-0.26313 0.12897-0.27170 0.76065-0.92814
  15  -3   5   -7.40    8.00  166 0.27655 0.61563-0.16429 0.02155 0$1 0$1$0$1(0(2?
"10 0$0(0$0(0$0 0(0 0(0$0"4 0"2 0(2%0(6%0"2 0"0 0 0?   ?    4    0
0$0$0$5$0&7 0&0 0&8??  ?    4    0                  00 0(4 0"2 0"2 0(2%0(4$0" ...
0                 "10 0$4 0 4 0 4$0(0$0(0$0&4 0&2 0"0%0"5(0(4 0(0 0 0??  ?   ..."

# ... lines omitted for brevity
;
```

[Boerrigter 2023, pers. comm.]

Corrupted data in a text field

```
_iucr_refine_reflections_details
;
   0   0   2   -0.20    0.30  99-0.77969 0.78029 0.62494-0.62494 0.03182 0.03182
   0   0   2   -0.30    0.30 209 0.78190-0.78130-0.62292 0.62292 0.03182 0.03182

# ... lines omitted for brevity

 -15  -3  -5   -4.60    8.40 316-0.62905-0.26313 0.12897-0.27170 0.76065-0.92814
  15  -3   5   -7.40    8.00 166 0.27655 0.61563-0.16429 0.02155 0$1 0$1$0$1(0(2?
"10 0$0(0$0(0$0 0(0 0(0$0"4 0"2 0(2%0(6%0"2 0"0 0 0?   ?    4   0
0$0$0$5$0&7 0&0 0&8??  ?    4   0                  00 0(4 0"2 0"2 0(2%0(4$0"  ...
0                 "10 0$4 0 4 0 4$0(0$0(0$0&4 0&2 0"0%0"5(0(4 0(0 0 0??  ?    ..."

# ... lines omitted for brevity
;
```

[Boerrigter 2023, pers. comm.]

It would be better to use CIF `loop_` constructs and *avoid* text fields with internal structure.

Corrupted numeric tables

```
/usr/bin/cif_validate: 2009384.cif data_2009384:
NOTE, data item '_atom_site_aniso_U_11' value
'H91' violates type constraints -- the value
should be a numerically interpretable string,
e.g. '42', '42.00', '4200E-2'.
```

Corrupted numeric tables

```
/usr/bin/cif_validate: 2009384.cif data_2009384:
NOTE, data item '_atom_site_aniso_U_11' value
'H91' violates type constraints -- the value
should be a numerically interpretable string,
e.g. '42', '42.00', '4200E-2'.
```

```
loop_
_atom_site_aniso_label
_atom_site_aniso_U_11
_atom_site_aniso_U_22
_atom_site_aniso_U_33
_atom_site_aniso_U_12
_atom_site_aniso_U_13
_atom_site_aniso_U_23
# ... some atoms omitted for brevity
C9 0.086(10) 0.061(8) 0.053(8) -0.003(7) -0.025(7) 0.008(7)
H5 0.062 H81 0.111 H82 0.111 H83
0.111 H91 0.081 H92 0.081 H93 0.081
```

COD entry checks – IUCr criteria checks

- Checks on prepublications and Personal communications;
- Checks on published structures;
- *Statistics of structures in the database*

COD internal consistency – checks against Fobs; QM

- Checks of/against deposited F_{obs} data;

[Henn, 2019]

COD has over **58 000** Fobs files; most recent COD files contain SHELX HKL data as a text field...

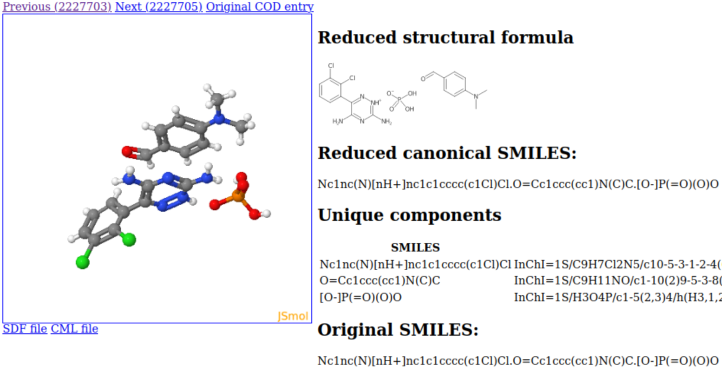
- Checks using QM relaxation with F/LOSS DFT and QM codes; work in progress ...

COD internal consistency – chemistry checks

- Perception of chemical formulae and checks of chemical plausibility; work in progress – publication submitted;
  - *example of a corrected publication entry*;
- Overlay of chemical graphs obtained from different sources (CIF coordinates, supplementary CML files, chemical names); A. Merkys, CODCHEM, publication accepted;

COD Molecules

http://molecules.crystallography.net/~saulius/cod-molecules/cod/2227704.html



Previous (2227703) Next (2227705) Original COD entry

SDF file CML file

Reduced structural formula

Reduced canonical SMILES:

Nc1nc(N)[nH+]nc1c1cccc(c1Cl)Cl.O=Cc1ccc(cc1)N(C)C.[O-]P(=O)(O)O

Unique components

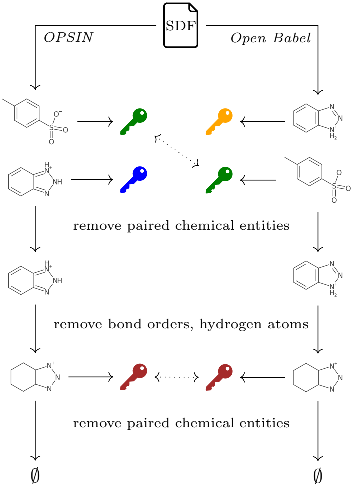
SMILES	
Nc1nc(N)[nH+]nc1c1cccc(c1Cl)Cl	InChI=1S/C9H7Cl2N5/c10-5-3-1-2-4(
O=Cc1ccc(cc1)N(C)C	InChI=1S/C9H11NO/c1-10(2)9-5-3-8(
[O-]P(=O)(O)O	InChI=1S/H3O4P/c1-5(2,3)4/h(H3,1,2

Original SMILES:

Nc1nc(N)[nH+]nc1c1cccc(c1Cl)Cl.O=Cc1ccc(cc1)N(C)C.[O-]P(=O)(O)O

[Vaitkus 2023, in preparation]

Matching the chemical structure graphs


SDF
OPSIN
Open Babel
remove paired chemical entities
remove bond orders, hydrogen atoms
remove paired chemical entities
∅
∅

[Merkys 2023, in press]

Fraudulent structures...

- more than 100 published structures were falsified;
- looked "OK" based on usual criteria;
- detected by crystallographers in the IUCr-led effort; based on implausible chemistry

Can we limit data fraud and honest mistakes?

- data *must* be reviewed as the main text, and possibly even more thoroughly;
- collaborative tools are necessary (a-la GitLab or GitHub); work in progress;
- reviewers for data as well as reviewers for paper text?

What is the role and capabilities of reviewers?

- Discussions in “Science” (2006):
  - *“The reporting of scientific results is based on trust”*; *“journals are not designed to catch fraud”* [Couzin, 2006];

    on the other hand,

  - *“It recommended “substantially stricter” requirements for reporting primary data and a risk assessment for accepted papers”* [Couzin, 2006];
- Errors are errors no matter of they are honest or deliberate – same approaches to detect them should work;

Recommendations for data publication

For scientists and educators

- Invest into preparing your data – make sure that you data are well documented, have complete metadata; measurements, models and computations are reproducible;
- Educate researchers students:
    - importance of syntax – files *must* be machine readable;
    - importance of metadata;
    - importance of validation;
    - importance of data consistency checks, curation and review;

Recommendations for data publication

For publishers

Improve data publication procedures:

- recommend publishers to use more formal checks, e.g. dictionary validation;
- recommend publishers to use more quality criteria;
- recommend publishers to conduct data peer-review, not just the paper text peer review;
- ensure correct cross-references between data;
- use appropriate tools for data review;

Acknowledgements

VU Institute of Biotechnology (KICIS)

Andrius Merkys
Antanas Vaitkus
Algirdas Grybauskas

QM community

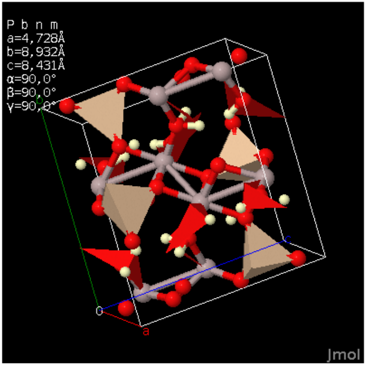
Audrius Alkauskas
Vytautas Žalandauskas
Lukas Razinkovas
Björkman Torbjörn
Stefaan Cottenier
Nicola Marzari
Giovanni Pizzi
Lubomir Smrcok
Linas Vilčiauskas
Chris Wolverton

COD Advisory board

Daniel Chateigner
Robert T. Downs
Werner Kaminsky
Armel Le Bail
Luca Lutterotti
Peter Moeck
Peter Murray-Rust
Miguel Quirós

Thank you!

http://en.wikipedia.org/wiki/Topaz



Coordinates	2207377.cif
Original IUCr paper	HTML

http://www.crystallography.net/2207377.html

References I

Couzin, J. (2006).
Scientific fraud.
Science, 314(5807):1853–1853.

Henn, J. (2019).
Metrics for crystallographic diffraction- and fit-data: a review of existing ones and the need for new ones.
Crystallography Reviews, 25(2):83–156.

Landi, A., Thompson, M., Giannuzzi, V., Bonifazi, F., Labastida, I., da Silva Santos, L. O. B., and Roos, M. (2020).
”A” of FAIR – as open as possible, as closed as necessary.
Data Intelligence, 2(1-2):47–55.

Merkys, A., Vaitkus, A., Butkus, J., Okulič-Kazarinas, M., Kairys, V., and Gražulis, S. (2016).
COD::CIF::Parser: an error-correcting CIF parser for the Perl language.
Journal of Applied Crystallography, 49(1):292–301.

Vaitkus, A., Merkys, A., and Gražulis, S. (2021).
Validation of the Crystallography Open Database using the Crystallographic Information Framework.
Journal of Applied Crystallography, 54(2):1–12.

References II

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1).