The relational underpinnings of CIF

James Hester

ANSTO, Sydney

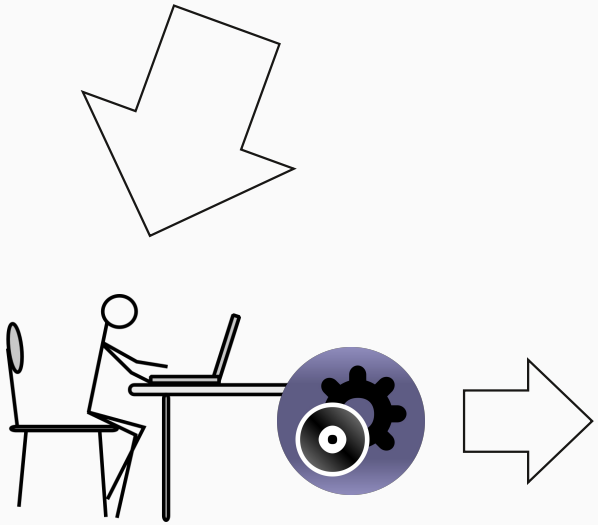
How CIF transfers meaning

CIF needs humans:

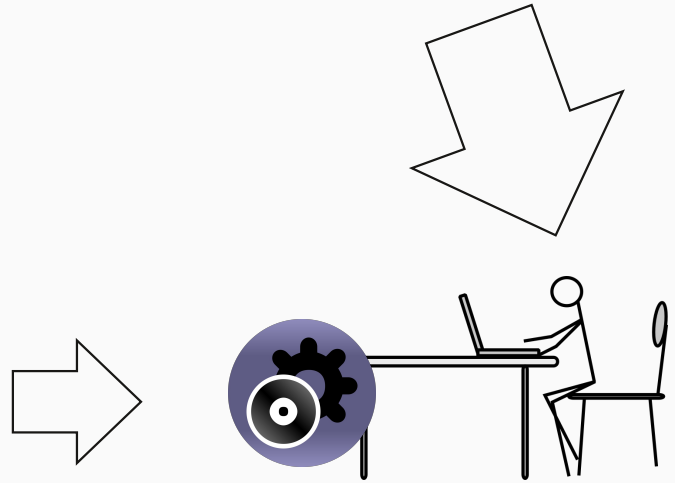
```
_definition.id       '_pd_proc.intensity_total'
_description.text
;
   Intensity values for the processed diffractogram
   at each data point where background, normalization,
   or other corrections have not been applied.
;
   _name.category_id          pd_proc
   _name.object_id            intensity_total
   _type.contents             Real
   _enumeration.range         0.0:
```

Whitespace-separated columns of ASCII-formatted numbers. Comments on lines with first character "#".

_pd_proc.intensity_total is in column 2.



```
# Data from file WBT0069064.nx.hdf
#     Angle  Intensity     Error
    15.08374   3297.29   96.79420
    15.14624   2749.89   34.75352
    15.20874   3046.08   55.95003
    15.27124   3053.61   50.83487
    15.33374   3028.87   57.34422
    15.39624   3036.24   54.20394
    15.45874   2966.64   53.41039
    15.52124   3058.82   62.81126
    15.58374   2824.74   57.53835
    15.64624   2900.90   54.75753
    15.70874   2837.56   56.48932
    15.77125   2927.63   53.42696
    15.83375   2866.92   54.19454
    15.89625   2852.62   50.36857
    15.95875   2911.23   51.58636
```



Why machine-readable dictionaries?

```
save_topol_atom.node_id

    _definition.id                '_topol_atom.node_id'
    _description.text
;
    The node associated with this atom, if applicable.
    It must match a value provided for _topol_node.id.
;
    _name.linked_item_id          '_topol_node.id'
    _type.purpose                 Link
    _type.contents                Integer
    _enumeration.range            1:

save_

save_topol_atom.symop_id

    _definition.id                '_topol_atom.symop_id'
    _description.text
;
    The identifier of the symmetry operation that is to be
    applied to the coordinates of the atom given by
    _topol_atom.atom_label before addition of
    the translations given by _topol_atom.translation.
    The value must match a value of _space_group_symop.id.
    If this item is omitted or assigned to '.', the identity
    operation is assumed.
;
    _name.linked_item_id          '_space_group_symop.id'
    _type.contents                Integer
    _enumeration.range            1:192
    _enumeration.default          1

save_
```

- For validation
- For transformation to other formats
- For precision of description
- For automated database construction!

`_topol_atom.node_id` (*Integer*)

The node associated with this atom, if applicable. It must match a value provided for **`_topol_node.id`**.

*Values must match those for the following item(s): **`_topol_node.id`**.*

The permitted range is $1 \rightarrow \infty$.

`_topol_atom.symop_id` (*Integer*)

The identifier of the symmetry operation that is to be applied to the coordinates of the atom given by **`_topol_atom.atom_label`** before addition of the translations given by **`_topol_atom.translation`**. The value must match a value of **`_space_group_symop.id`**. If this item is omitted or assigned to '.', the identity operation is assumed.

*Values must match those for the following item(s): **`_space_group_symop.id`**.*

The permitted range is $1 \rightarrow 192$. Where no value is given, the assumed value is '1'.

Outline

1. Introduce relational model
2. Introduce category theory
3. Link the two together
4. CIF is relational
5. Why is this useful
6. Multiple data blocks

The Relational Model

"Relation" = Table

Key: column(s) whose values always select unique row

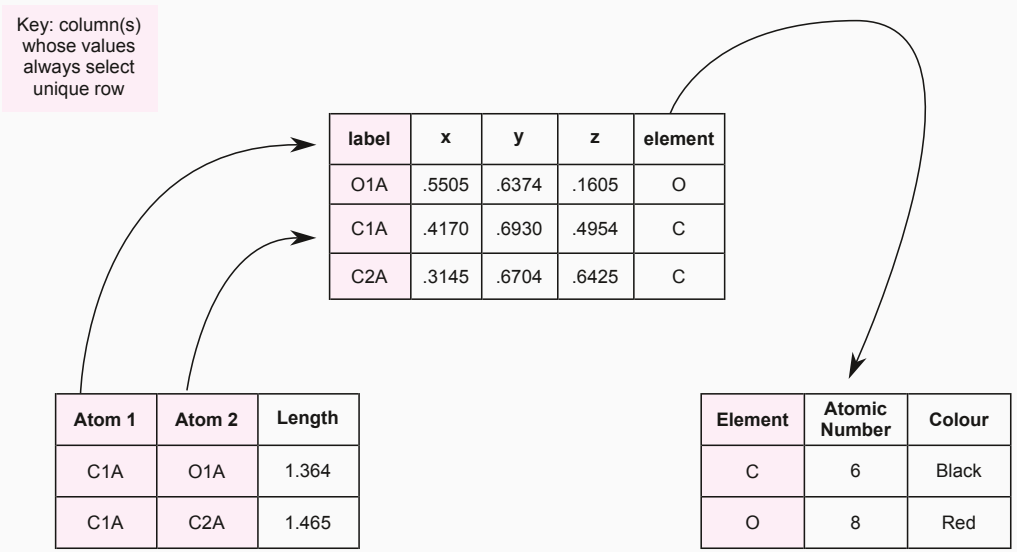
label	x	y	z	element
O1A	.5505	.6374	.1605	O
C1A	.4170	.6930	.4954	C
C2A	.3145	.6704	.6425	C

"Attribute" = Column Header

"Tuple" = Row

- A relation has no repeated column names, or row contents
- The order of columns and rows has no meaning
- Each column has an associated domain from which its values are drawn
- "Key" (or candidate key): values can be used to select unique row

The Relational Model



Relations refer to particular rows in other relations by referencing the values of their key data names.

Functional view of relational model

label	x	y	z	element
O1A	.5505	.6374	.1605	O
C1A	.4170	.6930	.4954	C
C2A	.3145	.6704	.6425	C

Atom 1	Atom 2	Length
C1A	O1A	1.364
C1A	C2A	1.465

Each column in a relation is a function mapping the key data name(s) to the set from which the column values are drawn.

$x(\mathrm{O1A}) = 0.5505$

$\mathit{Length}(\mathrm{C1A}, \mathrm{O1A}) = 1.364$

$\mathit{element}(\mathrm{O1A}) = \mathrm{O}$

"Normal forms": efficient relations

- Focused on improving database performance
- Reduce duplication of information
- Improve orthogonality of information
- Allow parallel operations on databases

We just want:

- To reduce information duplication
- Minimum disruption for future expansion: maximally non-committal

Can I update a single row in my loop and maintain consistency/truth?

Third normal form

- All non-key data names depend only on the values of the keys
- Key data names do not depend on one another
- No transitive dependencies allowed!
- "every non-key column must provide a fact about the key, the whole key, and nothing but the key" (William Kent)[1].

label	x	y	z	element	atomic number
O1A	.5505	.6374	.1605	O	8
C1A	.4170	.6930	.4954	C	6
C2A	.3145	.6704	.6425	C	6

Example of transitive dependency

[1] William Kent, "A Simple Guide to Five Normal Forms in Relational Database Theory", *Communications of the ACM* (1983)**26**(2), 120-125

CIF dictionaries describe relations

```
save_ATOM_SITE

    _definition.id              ATOM_SITE
    _definition.scope           Category
    _definition.class           Loop
    _definition.update          2023-02-03
    _description.text
;
    The CATEGORY of data items used to describe
    atom site information
;
    _name.category_id           ATOM
    _name.object_id             ATOM_SITE
    _category_key.name          '_atom_site.label'
```

```
save_atom_site.type_symbol

  _definition.id              '_atom_site.type_symbol'
  _definition.update          2021-10-27
  _description.text
;
    A code to identify the atom specie(s)
    occupying this site.
;
  _name.category_id           atom_site
  _name.object_id             type_symbol
  _name.linked_item_id        '_atom_type.symbol'
  _type.purpose               Link
  _type.source                Related
  _type.contents              Word
```

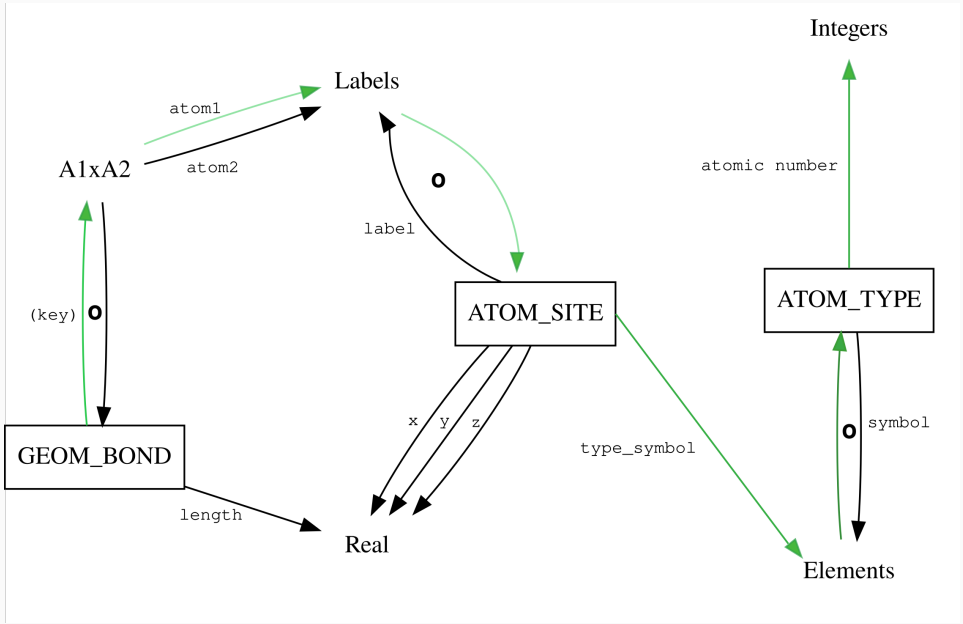
```
 loop_
_atom_site.label
_atom_site.fract_x
_atom_site.fract_y
_atom_site.fract_z
_atom_site.type_symbol
  o1   .550(5)   .637(5)   .160(1)   .035(3) O
  o2   .029(3)   .031(3)   .040(3) -.008(3) O
  c1   .028(4)   .036(5)   .023(4)   .000(4) C
```

- Table (category) is given a name: **ATOM_SITE**
- Key data name(s) identified using **_category_key.name**
- Data names to be tabulated in this table assigned via **_name.category_id**
- Data name that links to another category's data name given in **_name.linked_item_id**

(Mathematical) Category theory

- Objects connected by arrows ("morphisms")
- Arrows can be composed ($f \circ g$ is also a morphism)
- All objects have an identity morphism ($f \circ i = f$)
- Many, many theorems.

Relations are categories over sets and functions



Key: column(s) whose values always select unique row

label	x	y	z	element
O1A	.5505	.6374	.1605	O
C1A	.4170	.6930	.4954	C
C2A	.3145	.6704	.6425	C

Atom 1	Atom 2	Length
C1A	O1A	1.364
C1A	C2A	1.465

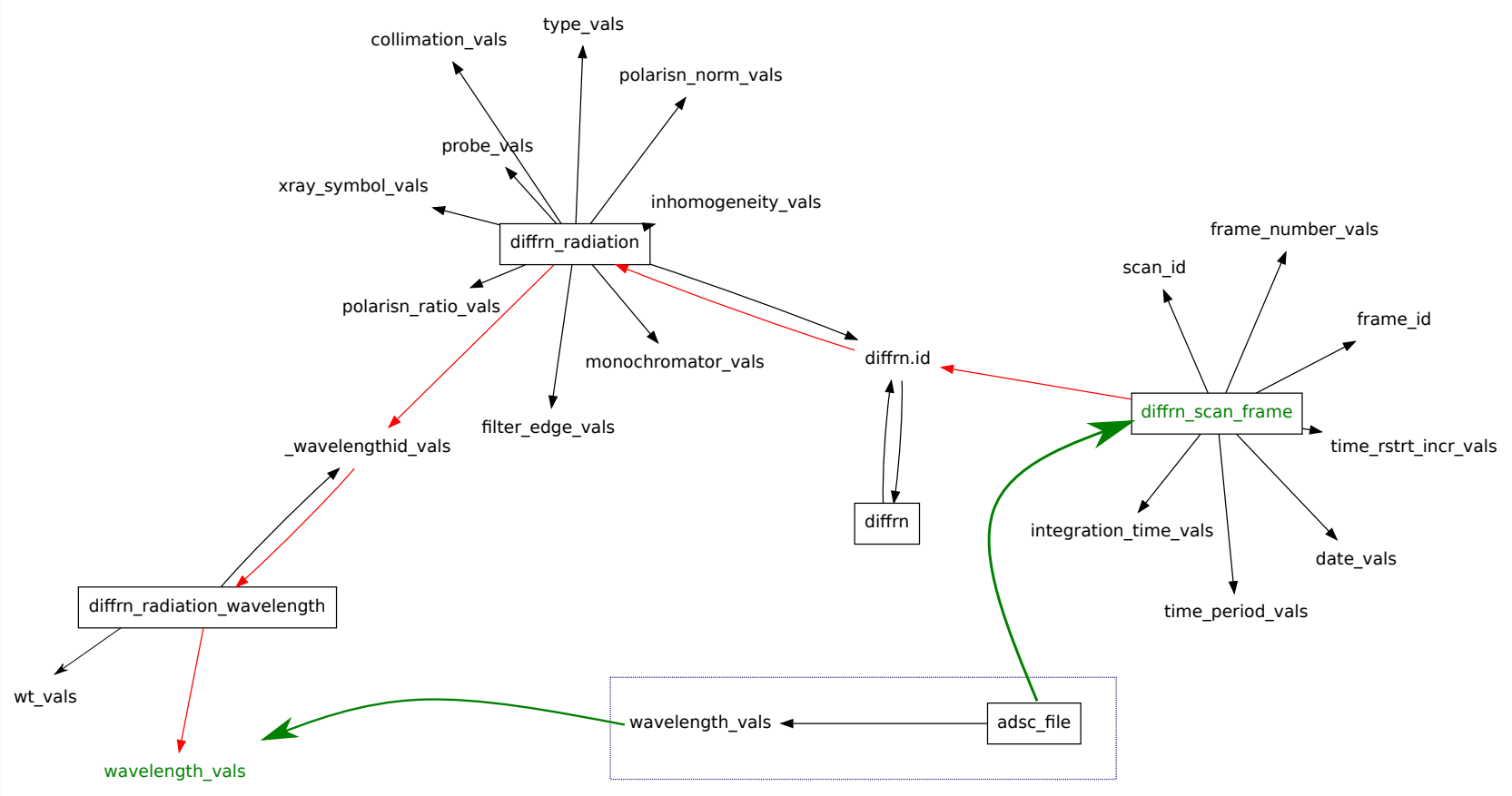
Element	Atomic Number	Colour
C	6	Black
O	8	Red

- Category theory "objects" $\rightarrow$ Sets
- Category theory "morphisms" $\rightarrow$ Functions
- A relation tabulates values for the functions
- Symbol "o": composition gets back to original value
- Green arrows: function to obtain atomic number from bond.

...CIF dictionaries describe categories!

- A CIF category $\approx$ category theory object
- A CIF data name $\approx$ morphism

Aside: Category theory is useful



Relations can be automatically transformed if corresponding objects are identified.

Spivak, D. I. "Functorial Data Migration" *Information and Computation* **217** (2012) 31-51

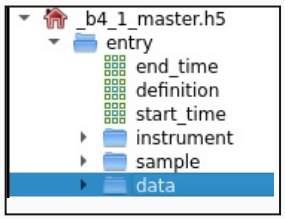
`www.categoricaldata.net`

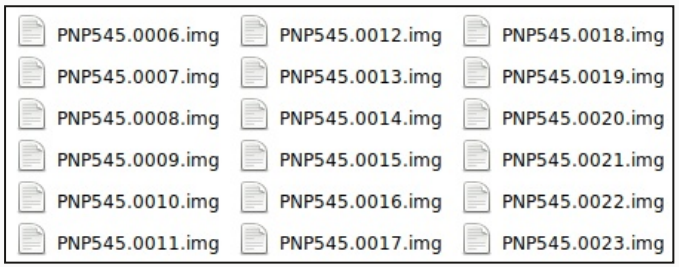
Relevance to writing CIF dictionaries

1. (Functional picture) Given which key data names can I unambiguously determine a value for my data name?
2. Do I need a data name in this category or is there a path from the keys already? (composition of functions)
3. Is the dependence of a data name on the key via another data name in the same category? (remove transitivity)
4. There is not one perfect way to express the data relationships (functors) - choose a way that suits the context e.g. widespread convention

Data Containers

```
#\#CIF_2.0
#
data_general
...
data_structure_1
...
data_structure_2
...
data_calibration
...
```





Our relational data sit inside containers (blocks, files, directories, ...)

There are no containers in the relational model

How do we combine container contents and retain the relational model?

Projection and Scoping

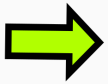
A	Item1	Item2
1	20.34	151.4
2	17.4	132.1

A	B	Data
1	x	1.1
1	y	3.2
2	x	-1.1
1	z	2.2
2	y	1.0
1	q	4.1

...

A	C	Data
1	RR	Perth
2	EE	Perth
2	RR	Perth
1	FF	Hobart
2	SS	Perth
1	EE	Darwin

...



A=1

Item1 = 20.34

Item2 = 151.4

B	Data
x	1.1
y	3.2
z	2.2
q	4.1

...

C	Data
RR	Perth
FF	Hobart
EE	Darwin

...

A=2

Item1 = 17.4

Item2 = 132.1

B	Data
x	-1.1
y	1.0

C	Data
EE	Perth
RR	Perth
SS	Perth

...

1. Collect rows that have a particular value of a chosen key data name (projection)
2. Drop columns for which the value can be inferred (scoping)

Reverting Projection and Scoping

A	Item1	Item2
1	20.34	151.4
2	17.4	132.1

A	B	Data
1	x	1.1
1	y	3.2
2	x	-1.1
1	z	2.2
2	y	1.0
1	q	4.1

...

A	C	Data
1	RR	Perth
2	EE	Perth
2	RR	Perth
1	FF	Hobart
2	SS	Perth
1	EE	Darwin

...

A=1

Item1 = 20.34

Item2 = 151.4

B	Data
x	1.1
y	3.2
z	2.2
q	4.1

...

C	Data
RR	Perth
FF	Hobart
EE	Darwin

...

A=2

Item1 = 17.4

Item2 = 132.1

B	Data
x	-1.1
y	1.0

C	Data
EE	Perth
RR	Perth
SS	Perth

...

1. Add back the missing columns (undo scoping)
2. Merge separate tables together (undo projection)

In CIF terms

_cat1.A 1

_cat1.Item1 20.34

_cat1.Item2 151.4

_cat2.B	_cat2.Data	
x	1.1	
y	3.2	
z	2.2	...
q	4.1	

_cat3.C	_cat3.Data	
RR	Perth	
FF	Hobart	...
EE	Darwin	

_cat1.A 2

_cat1.Item1 17.4

_cat1.Item2 132.1

_cat2.B	_cat2.Data	
x	-1.1	
y	1.0	

_cat3.C	_cat3.Data	
EE	Perth	
RR	Perth	...
SS	Perth	

- A "Set" category is one that has been projected and therefore takes single values for data names
- Child data names of "Set" category key data names are elided
- If key data names are provided for "Set" categories, information in that category from separate data blocks can be combined (multiple crystals, diffraction conditions, etc.)

`cat1` is a `Set` category
`_cat2.A` and `_cat3.A` are elided
`_name.linked_item_id` for `_cat2.A` and `_cat3.A` is `_cat1.A`
The key for `cat1` is `_cat1.A`