For discussion – version of 27th January 2020, revised by IUCr 4 February 2020 (Previous version(s): 10 January 2019)

Enhancements to CIF content and practice

A set of recommendations and proposals arising from a meeting between IUCr and CCDC staff at IUCr, 5 Abbey Square, Chester CH1 2HU, 7-9 January 2019

Present: Gillian Holmes (IUCr), Mike Hoyland (IUCr), Natalie Johnson (CCDC), Brian McMahon (IUCr), Peter Strickland (IUCr), Simon Westrip (IUCr), Seth Wiggin (CCDC)

Updated:

- **06/01/2020**, Additional information added for proposal four concerning reporting of detector information in CIFs in CSD, Natalie Johnson (CCDC)
- **27/01/2020**, Additional comments added to draft recommendation 1 and 2.
- **04/02/2020,** IUCr responses and comments, most significantly addition of *_id_iucr items to proposal 2, redrafting of definitions in proposal 3, extraction of section 5 as a separate discussion document.
- **18/02/2020,** Additional CCDC comments to proposals 1 and 3.

.

1. Draft Recommendation: Recognition of scientist(s) responsible for data collection

It is established practice that an article submitted to a journal in CIF format may record the authors of the article using data names in the PUBL and PUBL_AUTHOR categories. These data items should appear once only in the submitted CIF. (Very often the CIF contains a separate data block, often called `data_global`, where these items, as well as any others relating to the publication as a whole, are recorded.) Figure 1 is an example of how these might be used.

```
data_global

_publ_section_title
; The crystal structures of benzylammonium phenylacetate and its hydrate
;
_publ_contact_author_name 'He\&s, David'
_publ_contact_author_email hessd@ill.fr

_publ_contact_author_address
'Institut Laue-Langevin, 71 Avenue des Martyrs, 38000 Grenoble, France'
_publ_requested_journal 'Acta Crystallographica Section E'

loop_
    _publ_author_name
    _publ_author_address
    _publ_author_email
'He\&s, David'
; Institut Laue-Langevin, 71 Avenue des Martyrs, 38000 Grenoble, France
;
hessd@ill.fr
'Mayer, Peter'
; Ludwig-Maximilians-Universit\"at, Department Chemie, Butenandtstrasse
  5--13, 81377 M\"unchen, Germany
;
p.mayer@lmu.de
```

Figure 1. Example of publication metadata, including names and affiliations of authors, and details of the contact author (i.e. the author who assumes responsibility for the submission and any associated correspondence with the journal).

There is an increasing requirement to provide specific credit to the crystallographer or other scientists responsible for the data collection and handling of each crystal structure reported. This is usually, but not always, one or more of the publication authors. It is recommended that the CIF data block for each structure convey this information using similar data names found in the AUDIT_AUTHOR and AUDIT_CONTACT_AUTHOR. Figure 2 shows an example.

```
data_1
_audit_contact_author_name 'Mayer, Peter'
_audit_contact_author_email p.mayer@lmu.de
_audit_contact_author_address
; Ludwig-Maximilians-Universit\"at, Department Chemie, Butenandtstrasse
  5--13, 81377 M\"unchen, Germany
;

_audit_author_name    'Mayer, Peter'
_audit_author_address
; Ludwig-Maximilians-Universit\"at, Department Chemie, Butenandtstrasse
  5--13, 81377 M\"unchen, Germany
;

data_2
_audit_contact_author_name 'Mayer, Peter'
_audit_contact_author_email p.mayer@lmu.de
_audit_contact_author_address
; Ludwig-Maximilians-Universit\"at, Department Chemie, Butenandtstrasse
  5--13, 81377 M\"unchen, Germany
;

loop_
    _audit_author_name
    _audit_author_address
'Mayer, Peter'
; Ludwig-Maximilians-Universit\"at, Department Chemie, Butenandtstrasse
  5--13, 81377 M\"unchen, Germany
;
'He\&s, David'
; Institut Laue-Langevin, 71 Avenue des Martyrs, 38000 Grenoble, France
;
```

Figure 2. Example of metadata associated with the collection and analysis of data in individual data blocks.

Implementation notes:

[1] From (*January 2021**) articles published in IUCr journals will contain a statement attributing responsibility for each reported structure based on the AUDIT_AUTHOR and AUDIT_CONTACT_AUTHOR records in each data block.

[2] From (*January 2021**) the CCDC will associate any person and associated affiliation named in the AUDIT_CONTACT_AUTHOR fields of a submitted CIF as the responsible

'crystallographer'. Where this information is absent, the CCDC deposition procedure will continue to request the name and address of the responsible crystallographer.

* Dates are notional, but it is suggested that IUCr and CCDC actions occur at the same time, and that these proposals be given to the community at least several months before implementation begins, to allow software developers to update their packages accordingly.

CCDC comment:

We are aware of discussion to add information about each authors contribution into the CIF. Perhaps a way of future-proofing this field would be the addition of another field, such as _audit_contact_author_contribution where information such as 'crystallographer', or more specific information such as 'data collection', 'solution', 'crystallization', 'synthesis', could be given to indicate the authors more specific contribution to the particular structure.

IUCr response:

At present, we see the pressure for such additional information as coming from publication metrics, so think this would be better suited to an item _publ_author_contribution to be looped against author names in the 'text_global' block. This would contain set terms using a controlled vocabulary, possibly based on emerging publishing industry standards such as the CrediT taxonomy (https://casrai.org/credit/). However, we would need further consultations with our Editors and other interested parties before putting forward a definite proposal.

CCDC response:

Adding the information into the 'text_global' block could lead to confusion if multiple structures are contained within one CIF. The role a scientist played in the acquisition of each structure may differ e.g. a scientist may have only collected and processed the data for one of the structures. This is something CCDC have seen when depositors are providing multiple CIFs in the same deposition – that the crystallographer (the name is provided by the depositor during the deposition process) is not always the same for each structure.

2 Draft Proposal: addition of ORCiD identifiers to AUDIT_AUTHOR and related categories

If Recommendation [1] is approved, the value of the information recorded in the AUDIT_AUTHOR and AUDIT_CONTACT_AUTHOR categories is enhanced. We therefore propose that these categories should also have provision for capturing ORCiD identifiers associated with the scientist(s) who are responsible for the data collection and analysis.

```
data_audit_author_id_orcid
    _name                       '_audit_author_id_orcid'
    _category                     audit_author
    _type                         char
    _example                      0000-0003-0391-0002
    _definition
;               Identifier in the ORCID Registry of an author of this
                data block. ORCID is an open, non-profit,
                community-driven service to provide a registry of unique
                researcher identifiers (http://orcid.org).
;


data_audit_contact_author_id_orcid
    _name                       '_audit_contact_author_id_orcid'
    _category                     audit_contact_author
    _type                         char
    _example                      0000-0003-0391-0002
    _definition
;               Identifier in the ORCID Registry of the author of the data
                block to whom correspondence should be addressed. ORCID is
                an open, non-profit, community-driven service to provide a
                registry of unique researcher identifiers
                (http://orcid.org).
;


data_audit_author_id_iucr
    _name                       '_audit_author_id_iucr'
    _category                     audit_author
    _type                         char
    _example                      IUCr2895
    _definition
;               Identifier in the IUCr contact database of an author of this
                data block. This identifier may be available from the World
                Directory of Crystallographers (http://wdc.iucr.org).
;


data_audit_contact_author_id_iucr
    _name                       '_audit_contact_author_id_iucr'
    _category                     audit_contact_author
    _type                         char
    _example                      0000-0003-0391-0002
    _definition
;               Identifier in the IUCr contact database of the author of
                the data block to whom correspondence should be addressed.
                This identifier may be available from the World
                Directory of Crystallographers (http://wdc.iucr.org).

;

data_audit_author_email
```

```
    _name                       '_audit_author_email'
    _category                     audit_author
    _type                         char
    _list                         both
    _list_reference             '_audit_author_name'
    loop_ _example                name@host.domain.country
                                  bm@iucr.org
    _definition
;               The e-mail address of the author. If there is more
                than one author, this will be looped with
                _audit_author_name. The format of e-mail addresses
                is given in Section 3.4, Address Specification, of
                Internet Message Format, RFC 2822, P. Resnick
                (Editor), Network Standards Group, April 2001.
;
```

CCDC comment:

There is some concern as to how to ensure the ORCiD included in the CIF has been verified and a) is a real ORCiD and b) relates to the scientist in question.

Perhaps the verification could be indicated by some form of checksum.

IUCr response:

We acknowledge the concern but think that approaches to verification lie outside the scope of this proposal. Note our addition of the *_id_iucr items for the sake of symmetry with the _publ_contact_author_* and _publ_author_* equivalents. We have also added _audit_author_email, since this is already being used by Olex2.

3 Draft Proposal: Addition of AUDIT_SUPPORT category to core dictionary

In recent years there has been an increasing requirement to record details of funding bodies supporting research projects. We propose that the details of supporting bodies be recorded in CIF data files through the use of a new category AUDIT_SUPPORT. [Note: the name of the AUDIT_SUPPORT category derives from the PDBX_AUDIT_SUPPORT category used by the Protein Data Bank;

however, the proposed new items in the AUDIT_SUPPORT category do not share exactly the same definitions as the corresponding PDBX_AUDIT_SUPPORT items, *e.g.* `_pdbx_audit_support.funding_organization` is an enumerated list of specific funders, whereas `_audit_support.funding_organization_name` is a free-text field, complemented by `_audit_support.funding_organization_doi`, which is defined to identify uniquely the funder against an external standard registry (specifically, the funding information currently managed by CrossRef).]

The proposed AUDIT_SUPPORT category is presented below in DDL1 format
(note that the data item `_audit_support.id` provides the unique category key, to ensure ready translation to DDL2 and DDLm).

```
####################
## AUDIT_SUPPORT ##
####################

data_audit_support_[]
    _name                       '_audit_support_[]'
    _category                     category_overview
    _type                         null
    loop_ _example
          _example_detail

# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
- -
;
    loop_
    _audit_support.id
    _audit_support.funding_organization_name
    _audit_support.funding_organization_doi
    _audit_support.award_type
    _audit_support.award_number
    _audit_support.award_recipient

    1    'Engineering and Physical Sciences Research Council'
    'https://doi.org/10.13039/501100000266'
    studentship 'EP-M506515-1' 'E. T. Broadhurst'
```

```
    2   'Swedish Funding Council'
    ?
    grant '2017-05333' 'M. Lightowler'

    3   'Wellcome Trust'
    'https://doi.org/10.13039/100004440'
    grant 'WT087658' 'University of Edinburgh EM facility'

    4   'Scottish Universities Life Sciences Alliance (SULSA)'
    ?
    other ? 'University of Edinburgh EM facility'
    5   'Harvard Medical School'
    'https://doi.org/10.13039/100006691'
    ?     ?      ?
;
;
    Example prepared from funding data published in
    https://doi.org/10.1107/S2052252519016105
;

# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
- -
    _definition
;   Data items in the AUDIT_SUPPORT category record details about the
    funding support for the data collected and analysed in the data block.
;


data_audit_support.id
    _name                       '_audit_support.id'
    _category                     audit_support
    _type                         char
    _list                         both
    _list_mandatory               yes
    _example  '1'
    _definition
;           An arbitrary unique identifier for each source of support for
            the data collected and analysed in the data block.
;

data_audit_support.funding_organization_name
    _name                       '_audit_support.funding_organization_name'
    _category                     audit_support
    _type                         char
    _list                         both
    _list_reference             '_audit_support.id'
    _example  'National Center for Complementary and Alternative Medicine'
    _definition
;             The name of the organization providing funding support for
              the data collected and analysed in the data block. The
              recommended source for such names is the Open Funder
              Registry (https://github.com/CrossRef/open-funder-registry)
;

data_audit_support.funding_organization_doi
    _name                       '_audit_support.funding_organization_doi'
    _category                     audit_support
    _type                         char
    _list                         both
    _list_reference             '_audit_support.id'
    _example                    'https://doi.org/10.13039/100000064'
```

```
    _definition
;               The Digital Object Identifier (DOI) associated with the
                Organization providing funding support for
                the data collected and analysed in the data block. In
                accordance with CrossRef guidelines, the full URI of
                the resolved page describing the funding organization
                should be given (i.e. including the https://doi.org/
                component).
;

data_audit_support.award_type
    _name                       '_audit_support.award_type'
    _category                    audit_support
    _type                        char
    _list                        both
    _list_reference             '_audit_support.id'
    _example  'grant'
    loop_ _enumeration
          _enumeration_detail
             award       'award'
             bursary     'bursary'
             contract    'contract'
             gift        'gift'
             grant       'grant'
             other       'other type of award'
             scholarship 'scholarship'
             studentship 'studentship'
    _definition
;               Type or kind of award.
;

data_audit_support.award_number
    _name                       '_audit_support.award_number'
    _category                    audit_support
    _type                        char
    _list                        both
    _list_reference             '_audit_support.id'
    _example                    'FA9550-14-1-0409'
    _definition
;               The award number associated with this source of support.
;

data_audit_support.award_recipient
    _name                       '_audit_support.award_recipient'
    _category                    audit_support
    _type                        char
    _list                        both
    _list_reference             '_audit_support.id'
    _example                    'Cardiff University'
    _definition
;               The recipient of the support. May be an
                individual or institution.
;
```

IUCr comment:

We have modified the proposed definitions to reflect current practice, and to distinguish practice from that employed in the pdbx definitions, while still acknowledging that they borrow from what the PDB has done. We note that we might expect to find this in the

‘text_global’ data block rather than in individual structure data blocks (given that awards are typically made to the umbrella research project rather than to individual data sets), but we do not think that it is necessary to constrain usage in that way within the definitions.

Here, and in the following section where reference is made to a putative _diffrn_source_facility_canonical_identifier, we suggest that the ‘Research Organization Registry’ (https://ror.org) might be a suitable candidate source for canonical identifiers of a wide variety of organizations. At present, we are not sure that ROR is sufficiently stable or developed for us to endorse it.

CCDC comment:

The CCDC are aware that Ringgold (https://www.ringgold.com/) also offer identifiers for organizations (https://www.ringgold.com/identify/).

4 Draft Proposal: Capturing information about experimental facilities

There is growing interest from large facilities and other institutions to trace research results from data sets collected at specific locations. We propose that the relevant information be stored in a CIF through a new set of data items. Some of these are based on items in the PDB/Biosync extension CIF dictionary.

_diffrn_source_facility_name
(_diffrn_source_facility_canonical_identifier) - does not yet exist
_diffrn_source_beamline_name

Draft definitions follow:

```
data_diffrn_source_facility_name
    _name                       '_diffrn_source_facility_name'
    _category                     diffrn_source
    _type                         char
    _example                    'Diamond Light Source'
    _definition
;               The name of the synchrotron or other large-scale
                experimental facility at which the experiment was
                conducted. Names should conform to the spelling and
                format used in the 'Light Sources of the World' listing
                of lightsources.org
                (https://lightsources.org/lightsources-of-the-world/)
;

data_diffrn_source_beamline_name
    _name                       '_diffrn_source_beamline_name'
    _category                     diffrn_source
    _type                         char
    _example                    'I19'
    _definition
;               The name of the beamline at the synchrotron or other
                large-scale experimental facility at which the experiment
                was conducted.
;
```

Journal notes and/or the updated chapter of *International Tables Volume G* should also provide stronger guidance on the use of recording specific detectors (this may be particularly important for experiments outside the area of X-ray diffraction that involve detectors built for particular purposes). Some database depositors routinely use the CIF data item `_diffrn_measurement_device_type` for this purpose. An analysis, using data in Cambridge Structural Database, was undertaken of CIFs from a range of depositors and to assess whether it would be appropriate to encourage authors to use this item to record the name of the detector used, *e.g.* `'Kookaburra'` (the ultra-small-angle scattering instrument at the OPAL research reactor of ANSTO). **(See Additional Information for analysis)**

IUCr comment:

We suggest removing item 5 ('Topic for discussion: specifying the type of research study') from the current draft proposal, because it is not yet sufficiently well developed as a proposal in its own right. However, we recommend introducing it as a topic of discussion on the coreDMG mailing list to canvass other opinions on a suitable list of terms. Such a discussion could take place in parallel with the presentation of the main proposal.

Additional Information

`Additional information for proposal 4 (Capturing information about experimental facilities)`

This information is from a CCDC study where CIF information was used to categorize synchrotron studies (v540, October 2018). These studies were recognized by looking for a variety of key words in certain CIF fields – including the names of synchrotron facilities and their abbreviations. This list did not include any named detectors.

From 10,100 structures identified as synchrotron studies using CIF fields (as of October 2018), the table below show the number of structures which had synchrotron or facility identifying information in each CIF field. This shows that ~ 8% of identified synchrotron structures had synchrotron identifying information in the measurement_device field. 86% of this information contained facility identifying information – usually of the form of the facility and beamline the data was collected at.

	CIF Attribute	**Synchrotron Identification**	**Facility Identification**	**%**
Radiation type	_diffrn_radiation_type	8929	75	0.84%
Source	_diffrn_(source/radiation_source)	7960	5209	65.44%
Measurement	_diffrn_measurement_device/(_type)	867	749	86.39%
Collection software	_computing_data_collection	850	754	88.71%
Source type	_diffrn_source_type	705	598	84.82%
Monochromator	_diffrn_radiation_monochromator	615	0	0.00%
Probe	_diffrn_radiation_probe	12	0	0.00%

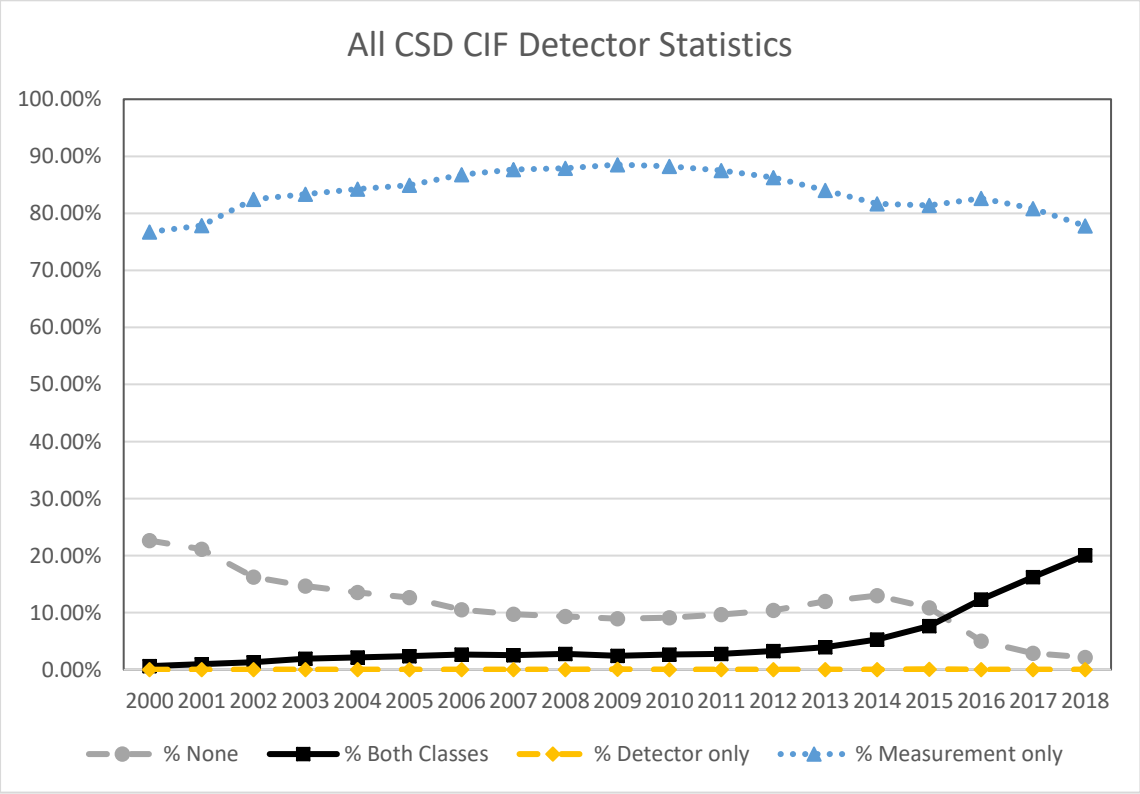
Structures identified from neutron sources, however, did contain named detectors/beamlines in these fields, such as Echidna, Koala (ANSTO), Pearl (ISIS), SENJU (J-Parc), Topaz (ORNL).

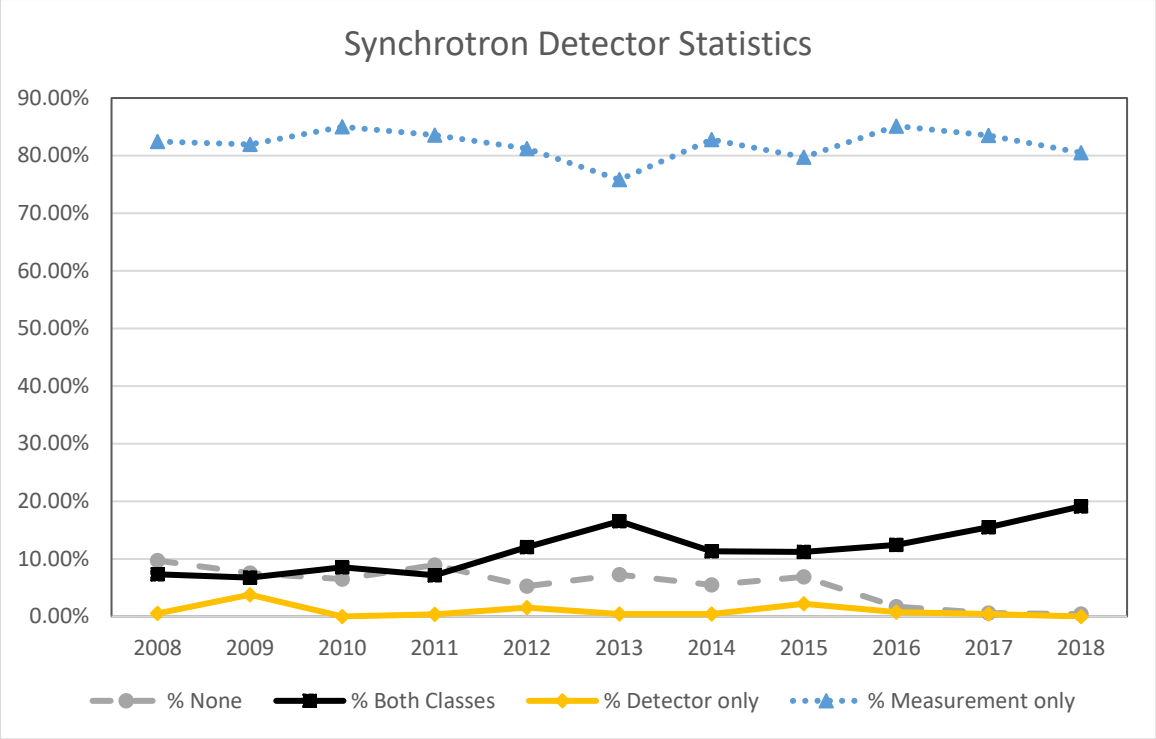
The usage of fields which commonly contain detector information, which has been separated into two classes – detector fields (_diffrn_detector/_diffrn_detector_type) and measurement fields (_diffrn_measurement_device/_diffrn_measurement_device_type) – has been studied and the number of CIFs have been counted where there is:

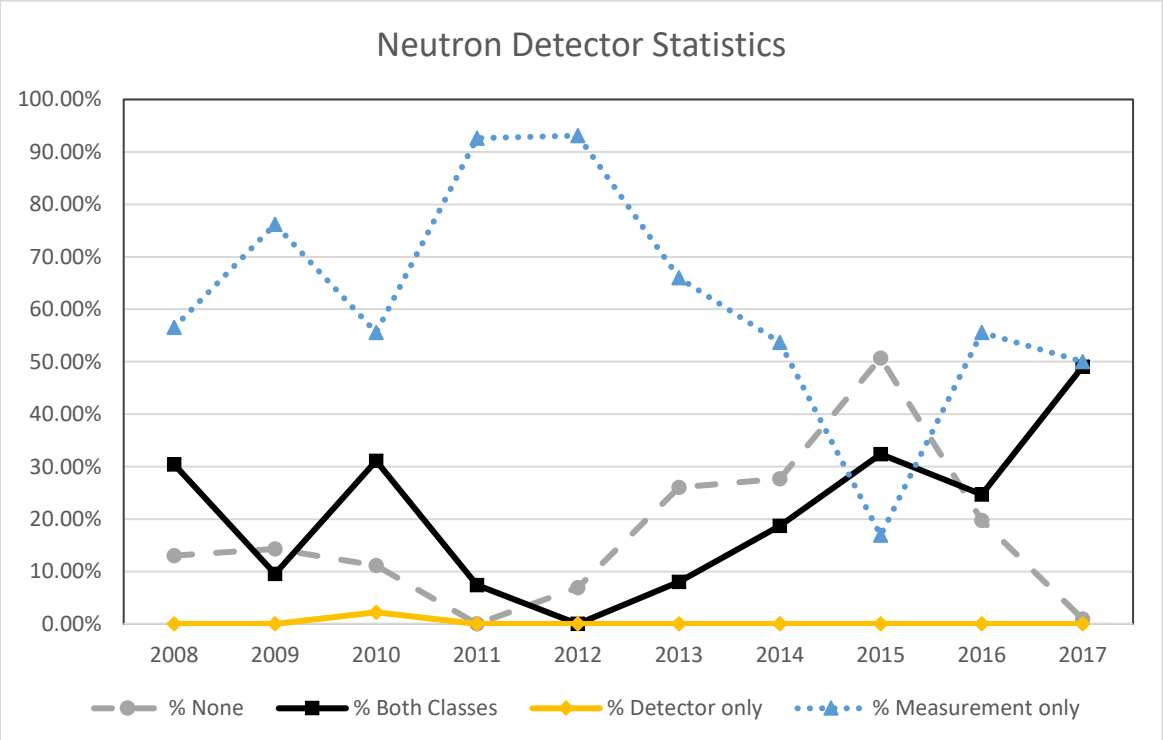
- information in at least one of the detector **and** at least one of the measurement fields [**% Both Classes**],
- information only in the detector fields (with no information in the measurement fields) [**% Detector only**]
- information only in the measurement fields [**% Measurement only**]
- no information in either one of the measurement fields (_diffrn_measurement_device/_diffrn_measurement_device_type) and either one of the detector fields (_diffrn_detector/_diffrn_detector_type) [**% None**],

The graphs show the % of CIFs by year for each of the four categories for all CIFs in the CSD, Synchrotron CIFs and Neutron CIFs.

In all of the graphs, although it is most likely that there will only to be information in one of the measurements fields, the proportion of CIFs that have information in at least one of the measurement and at least one of the detector fields is increasing. Measurement fields should contain goniometer information, not the detector type, according to the IUCr CIF dictionary v2.4.5 (_diffrn_measurement_device/_type: "`The general class of goniometer or device used to support and orient the specimen`"/"`The make, model or name of the measurement device (goniometer) used`"), however, these fields often contain the specific detector instead. The exception is in those CIFs where both classes have information; in this case the information is included in line with the CIF dictionary definitions.







Recommendation: As the % of CIFs where both classes contain information is increasing, it could be useful to also recommend the usage of the detector and measurement fields to include specific information about the experiments, including the name of the detectors.