Compatibility with PDB format files (§3.6.8.3.2)

DATABASE_PDB_CAVEAT
DATABASE_PDB_MATRIX
DATABASE_PDB_REMARK
DATABASE_PDB_REV
DATABASE_PDB_REV_RECORD
DATABASE_PDB_TVECT

The purpose of entries in the DATABASE category group is to provide pointers that link the mmCIF to all database entries that result from the deposition of the file. For mmCIF, the relevant category is DATABASE_2, which replaces the DATABASE category of the core dictionary.

Note the distinction between the database pointers provided here and those in the STRUCT_REF family of categories. The latter are intended to provide links to external database entries for any aspect of any subset of the structure that the author may wish to record, including previous determinations of the same structure, other structures containing the same ligand or references to the sequence(s) of the macromolecule(s) in sequence databases. In contrast, the links provided in DATABASE_2 refer to the entire contents of the mmCIF and are designed to cover situations in which the entire file is deposited in more than one database (for example, in the PDB and in a database for protein kinases).

3.6.8.3.1. *Related database entries*

Data items in these categories are as follows:

(*a*) DATABASE

- **_database.entry_id**
  → **_entry.id**

_database.code_CAS
_database.code_CSD
_database.code_ICSD
_database.code_MDF
_database.code_NBS
_database.code_PDB
_database.code_PDF
_database.code_depnum_ccdc_archive
_database.code_depnum_ccdc_fiz
_database.code_depnum_ccdc_journal
_database.CSD_history
_database.journal_ASTM
_database.journal_CSD

(*b*) DATABASE_2

- **_database_2.database_id**
- **_database_2.database_code**

The bullet (•) *indicates a category key. Where multiple items within a category are marked with a bullet, they must be taken together to form a compound key. The arrow* (→) *is a reference to a parent data item. Items in italics have aliases in the core CIF dictionary formed by changing the full stop* (**.**) *to an underscore* (_).

The DATABASE category is retained in the mmCIF dictionary, but only for consistency with the core dictionary.

The role of the data items in the DATABASE_2 category is to store identifiers assigned by one or more databases to the structure described in the mmCIF. In the data model used in the core CIF dictionary, each database has an individual data item. The data model in mmCIF is more general. It comprises the data items **_database_2.database_id**, which identifies the database, and **_database_2.database_code**, which is the code assigned by the database to the entry. Thus a new database can be referred to without needing to add an additional data item to the dictionary. If a structure has been deposited in more than one database, the values of **_database_2.database_id** and **_database_2.database_code** can be looped.

The institutions and individual databases recognized in the DATABASE_2 category in the current version of the mmCIF dictionary are CAS (*Chemical Abstracts* Service), CSD (Cambridge Structural Database), ICSD (Inorganic Crystal Structure Database), MDF (Metals Data File), NDB (Nucleic Acid Database), NBS (the Crystal Data database of the National Institute of Standards and Technology, formerly the National Bureau of Standards), PDB (Protein Data Bank), PDF (Powder Diffraction File), RCSB (Research Collaboratory for Structural Bioinformatics) and EBI (European Bioinformatics Institute). It is intended that new databases will be added to this list on an ongoing basis; the purpose of specifying a list of possible databases in the dictionary is to ensure that each database is referenced consistently.

3.6.8.3.2. *Compatibility with PDB format files*

Data items in these categories are as follows:

(*a*) DATABASE_PDB_REV

- **_database_PDB_rev.num**

_database_PDB_rev.author_name
_database_PDB_rev.date
_database_PDB_rev.date_original
_database_PDB_rev.mod_type
_database_PDB_rev.replaced_by
_database_PDB_rev.replaces
_database_PDB_rev.status

(*b*) DATABASE_PDB_REV_RECORD

- **_database_PDB_rev_record.rev_num**
  → **_database_PDB_rev.num**
- **_database_PDB_rev_record.type**

_database_PDB_rev_record.details

(*c*) DATABASE_PDB_MATRIX

- **_database_PDB_matrix.entry_id**
  → **_entry.id**

_database_PDB_matrix.origx[1][1]
_database_PDB_matrix.origx[1][2]
_database_PDB_matrix.origx[1][3]
_database_PDB_matrix.origx[2][1]
_database_PDB_matrix.origx[2][2]
_database_PDB_matrix.origx[2][3]
_database_PDB_matrix.origx[3][1]
_database_PDB_matrix.origx[3][2]
_database_PDB_matrix.origx[3][3]
_database_PDB_matrix.origx_vector[1]
_database_PDB_matrix.origx_vector[2]
_database_PDB_matrix.origx_vector[3]
_database_PDB_matrix.scale[1][1]
_database_PDB_matrix.scale[1][2]
_database_PDB_matrix.scale[1][3]
_database_PDB_matrix.scale[2][1]
_database_PDB_matrix.scale[2][2]
_database_PDB_matrix.scale[2][3]
_database_PDB_matrix.scale[3][1]
_database_PDB_matrix.scale[3][2]
_database_PDB_matrix.scale[3][3]
_database_PDB_matrix.scale_vector[1]
_database_PDB_matrix.scale_vector[2]
_database_PDB_matrix.scale_vector[3]

(*d*) DATABASE_PDB_TVECT

- **_database_PDB_tvect.id**

_database_PDB_tvect.details
_database_PDB_tvect.vector[1]
_database_PDB_tvect.vector[2]
_database_PDB_tvect.vector[3]

(*e*) DATABASE_PDB_CAVEAT

- **_database_PDB_caveat.id**

_database_PDB_caveat.text

(*f*) DATABASE_PDB_REMARK

- **_database_PDB_remark.id**

_database_PDB_remark.text

The bullet (•) *indicates a category key. Where multiple items within a category are marked with a bullet, they must be taken together to form a compound key. The arrow* (→) *is a reference to a parent data item.*

A major goal of the design of the mmCIF data model was that a file could be transformed from Protein Data Bank (PDB) format to mmCIF format and back again without loss of information. This required the creation of mmCIF data items whose sole purpose is to capture PDB-specific records that do not map onto mmCIF data items. These records would never be created for a *de novo* mmCIF. This family of categories also belongs to the PDB category group (see Section 3.6.9.3).

The items in the categories DATABASE_PDB_MATRIX and DATABASE_PDB_TVECT are derived from the elements of transformation matrices and vectors used by the Protein Data Bank. The items in the categories DATABASE_PDB_REV and DATABASE_PDB_REV_RECORD record details about the revision history of the data block as archived by the Protein Data Bank.

The items in the DATABASE_PDB_CAVEAT category record comments about the data block flagged as 'CAVEATS' by the Protein Data Bank at the time the original PDB archive file was created. A PDB CAVEAT record indicates that the entry contains severe errors. In PDB format, extended comments were stored as a sequence of fixed-length (80-character) format records, columns 9 and 10 being reserved for continuation sequence numbering. The mmCIF representation retains each record as a separate data value and does not attempt to merge continuation records to provide more readable running text. Hence the PDB CAVEAT entry

```
CAVEAT     1ABC    THE CRYSTAL TRANSFORMATION IS WRONG
CAVEAT   2 1ABC    BUT IS UNCORRECTABLE AT THIS TIME
```

would be represented in mmCIF as

```
loop_
  _database_PDB_caveat.id
     _database_PDB_caveat.text
    1
; THE CRYSTAL TRANSFORMATION IS WRONG
;
    2
; BUT IS UNCORRECTABLE AT THIS TIME
;
```

The PDB format used 'REMARK' records to store information relating to several aspects of the structure in free or loosely structured text. In some cases, the conventions used for individual types of REMARK record allow structured data to be extracted automatically and translated to specific mmCIF data items. Where this is not possible, the DATABASE_PDB_REMARK category may be used to retain the information that appeared in these parts of PDB format files. Unlike the CAVEAT records, it is possible to collect together several REMARK records sharing a common numbering into a single free-text field. For example, PDB practice has been to repeat the contents of CAVEAT records (see above) as records of type 'REMARK 5'. While each separate CAVEAT record is converted to a separate mmCIF data value, the complete text of a REMARK 5 record may be gathered into a single mmCIF data value. Hence the CAVEAT example above would also appear in a PDB file as part of a 'REMARK 5' as

```
REMARK   5 THE CRYSTAL TRANSFORMATION IS WRONG
REMARK   5 BUT IS UNCORRECTABLE AT THIS TIME
```

and would appear in an mmCIF as

```
loop_
  _database_PDB_remark.id
     _database_PDB_remark.text
    5
; THE CRYSTAL TRANSFORMATION IS WRONG
  BUT IS UNCORRECTABLE AT THIS TIME
;
```

Note that by convention the value of `_database_PDB_remark.id` matches the class of the REMARK record in the PDB file.

3.6.8.4. Article publication

Categories used during the publication of an article are as follows:

IUCR group

Journal housekeeping and reference entries (§3.6.8.4.1)

JOURNAL

JOURNAL_INDEX

Contents of a publication (§3.6.8.4.2)

PUBL

PUBL_AUTHOR

PUBL_BODY

PUBL_MANUSCRIPT_INCL

These categories cover both the metadata for the article (information about the article) and the text of the article itself.

3.6.8.4.1. *Journal housekeeping and citation entries*

Data items in these categories are as follows:

(*a*) JOURNAL

- **_journal.entry_id**
  → **_entry.id**
  _journal.coden_ASTM
  _journal.coden_Cambridge
  _journal.coeditor_address
  _journal.coeditor_code
  _journal.coeditor_email
  _journal.coeditor_fax
  _journal.coeditor_name
  _journal.coeditor_notes
  _journal.coeditor_phone
  _journal.data_validation_number
  _journal.date_accepted
  _journal.date_from_coeditor
  _journal.date_to_coeditor
  _journal.date_printers_final
  _journal.date_printers_first
  _journal.date_proofs_in
  _journal.date_proofs_out
  _journal.date_recd_copyright
  _journal.date_recd_electronic
  _journal.date_recd_hard_copy
  _journal.issue
  _journal.language
  _journal.name_full
  _journal.page_first
  _journal.page_last
  _journal.paper_category
  _journal.suppl_publ_number
  _journal.suppl_publ_pages
  _journal.techeditor_address
  _journal.techeditor_code
  _journal.techeditor_email
  _journal.techeditor_fax
  _journal.techeditor_name
  _journal.techeditor_notes
  _journal.techeditor_phone
  _journal.volume
  _journal.year

(*b*) JOURNAL_INDEX

_journal_index.subterm
_journal_index.term
_journal_index.type

The bullet (•) indicates a category key. The arrow (→) is a reference to a parent data item. Items in italics have aliases in the core CIF dictionary formed by changing the full stop (.) to an underscore (_).

In mmCIF, the families of categories used to contain the text of an article for publication and to record information about the handling and processing of the article by a publisher are assigned to the IUCR category group. The name arose from the fact that CIF is sponsored by the International Union of Crystallography and several of the journals of the IUCr can handle articles submitted for publication in CIF format. However, these data items may be