International Tables for Crystallography (2006). Vol. G, Chapter 3.6, pp. 144–198.

3.6. Classification and use of macromolecular data

BY P. M. D. FITZGERALD, J. D. WESTBROOK, P. E. BOURNE, B. MCMAHON, K. D. WATENPAUGH AND H. M. BERMAN
WITH APPENDIX 3.6.2 BY J. D. WESTBROOK, K. HENRICK, E. L. ULRICH AND H. M. BERMAN

3.6.1. Introduction

As described in Chapter 1.1, the macromolecular crystallographic information file (mmCIF) dictionary (Fitzgerald *et al.*, 1996; Bourne *et al.*, 1997) was initially commissioned as an extension to the core CIF dictionary (Hall *et al.*, 1991), with the intention of adding data names suitable for a full description of a macromolecular crystallographic experiment and its results. However, the need to specify relationships between the data items describing different components of a complex macromolecular structure led to the development of a richer dictionary definition language (DDL2). The data names were then defined according to the DDL2 formalism. For consistency, the existing core dictionary data items were also recast in the DDL2 formalism. Since no other DDL2 applications were envisaged at that time, the core items were then embedded in the mmCIF dictionary as a subset of the complete dictionary. The current release of the mmCIF dictionary described in this chapter includes all the data items in version 2.3.1 of the core dictionary. The mmCIF dictionary is not routinely updated to match additions to the core dictionary, but it is expected that when new versions of the mmCIF dictionary are released to meet the requirements of the macromolecular community, the most recent version of the core dictionary will be incorporated in the new mmCIF dictionary as part of the revision.

The resulting stand-alone dictionary is very large and is described in detail in this chapter. The philosophy behind the design of the dictionary is discussed in Section 3.6.2 and an example of its use is given and discussed in Section 3.2.3. The contents of the dictionary are then described in the remainder of the chapter, starting at Section 3.6.4. The discussion follows the sequence of Table 3.1.10.1: experimental measurements, analysis, structure, publication and file metadata are considered in turn. The discussion of individual categories may be found by using the overview of the dictionary structure given in Appendix 3.6.1.

The data names in the mmCIF dictionary derived from the core CIF dictionary differ from their DDL1 counterparts in that a full stop (.) is used to designate explicitly the category to which the data name belongs, *e.g.* `_cell.length_a` is used in place of `_cell_length_a`. Sometimes the mmCIF counterpart of a core data name may have a different form, for example to enforce the rule in DDL2 that the category name is the initial part of any data name within that category. This convention is generally observed in DDL1, but is not mandatory. Formally, the corresponding DDL1 core data name is obtained from the `_item_aliases.alias_name` attribute of the definition. The provision of a formal alias for all data names derived from the core dictionary allows a DDL2-compliant parser to read and interpret a data file constructed according to the DDL1 dictionary described in Chapter 3.2. Achieving this compatibility with CIFs built using DDL1 dictionaries was a very important goal in the design of DDL2 and the mmCIF dictionary.

In this chapter, categories and individual data names that correspond to matching entries in the core dictionary are not discussed in detail unless they are used in a different way in mmCIF. Chapter 3.2 should therefore be read first for a description of the categories common to both the core and mmCIF dictionaries. This chapter concentrates on the categories specific to mmCIF. Formal differences between mmCIF categories and core CIF categories are also summarized.

3.6.2. Considerations underlying the design of the dictionary

From the outset, mmCIF was envisaged as a providing a more detailed description of macromolecular structures than the existing Protein Data Bank (PDB) format (Chapter 1.1). A number of considerations guided the development of version 1 of the mmCIF dictionary. These included:

(i) Every field of every PDB record type should be represented by an mmCIF data item if the PDB field is important for describing the structure, the experiment that was conducted in determining the structure or the revision history of the entry. It is important to note that it is straightforward to convert an mmCIF data file to a PDB file without loss of information, since all the information is parsable. It is not possible, however, to automate completely the conversion of a PDB file to an mmCIF, since many mmCIF data items are either not present in the PDB file or are present in PDB REMARK records that in some cases cannot be parsed. The contents of PDB REMARK records are maintained as separate data items within mmCIF so as to preserve all the information, even if the information is not parsable.

(ii) Data items should be defined so that all the information given in the materials and methods section of an article describing the structure can be referenced. This includes major features of the crystal, the diffraction experiment, the phasing calculations and the refinement.

(iii) Data items should be provided for describing the biologically active molecule and any important structural subcomponents.

(iv) It should be possible to represent atom positions using either orthogonal ångström or fractional coordinates.

(v) Data items should be provided for describing the initial experimental reflection data, including all the data sets used in the phasing of the structure, and the final processed data.

(vi) Crystallographic and noncrystallographic symmetry should be described.

Affiliations: PAULA M. D. FITZGERALD, Merck Research Laboratories, Rahway, New Jersey, USA; JOHN D. WESTBROOK, Protein Data Bank, Research Collaboratory for Structural Bioinformatics, Rutgers, The State University of New Jersey, Department of Chemistry and Chemical Biology, 610 Taylor Road, Piscataway, New Jersey, USA; PHILLIP E. BOURNE, Research Collaboratory for Structural Bioinformatics, San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0537, USA; BRIAN MCMAHON, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, England; KEITH D. WATENPAUGH, retired; formerly Structural, Analytical and Medicinal Chemistry, Pharmacia Corporation, Kalamazoo, Michigan, USA; HELEN M. BERMAN, Protein Data Bank, Research Collaboratory for Structural Bioinformatics, Rutgers, The State University of New Jersey, Department of Chemistry and Chemical Biology, 610 Taylor Road, Piscataway, New Jersey, USA; KIM HENRICK, EMBL Outstation, The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, England; ELDON L. ULRICH, Department of Biochemistry, University of Wisconsin Madison, 433 Babcock Drive, Madison, WI 53706-1544, USA.

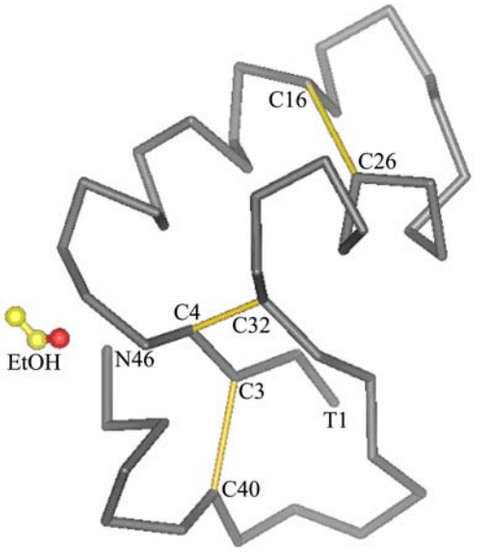

Fig. 3.6.3.1. A representation of crambin (PDB 3CNR) with a co-crystallized ethanol molecule.

(vii) Data items should be present for describing the characteristics and geometry of canonical and non-canonical amino acids, nucleotides, sugars and ligand groups.

(viii) Data items should be provided that permit a detailed description of the chemistry of the component parts of the macromolecule to be given.

(ix) Data items should be present that provide specific pointers from elements of the structure (*e.g.* the sequence, bound inhibitors) to appropriate entries in publicly available databases.

(x) Data items should be present that provide meaningful three-dimensional views of the structure so as to highlight functional and structural aspects of the macromolecule.

(xi) Data items specific to an NMR experiment or modelling study would not in general be included in version 1. However, data items that summarize the features of an ensemble of structures and permit a description of each member of the ensemble to be given should be available.

(xii) A comprehensive set of data items for providing a higher-order structure description (for example, to cover supersecondary structure and functional classification) was considered to be beyond the scope of version 1.

Based on the above, the first version of the mmCIF dictionary with approximately 1700 data items (including those data items taken from the core CIF dictionary) was developed and officially approved in October 1997. Subsequent revisions have increased the number of data items to over 2000. It is not expected that all the data items will be present in every mmCIF data file. Instead, the goal was to provide a wide range of data items from which users can select those that best suit the structure they wish to describe.

3.6.3. Overview of the mmCIF data model

The solution and refinement of a macromolecular structure is complex and often difficult, as there are a large number of atoms in a typical macromolecule, the molecular conformation can be complex and it can be difficult to model included solvent molecules. However, even when a satisfactory structural model has been derived, describing the structure can be a considerable challenge. Using diagrams can help, but two-dimensional projections are often inadequate for illustrating important features and a complete understanding of the three-dimensional structure of a macromolecule can often only be reached by using interactive molecular graphics software.

Example 3.6.3.1. *Specification of the three distinct components of the crambin structure.*

```
loop_
  _struct_asym.id
  _struct_asym.entity_id
  _struct_asym.details
   chain_a A       'single polypeptide chain'
   ethanol ethanol 'cocrystallized ethanol molecule'
   water   HOH     .
```

The mmCIF dictionary provides several ways for describing the structure. The PUBL categories can be used to record text describing the structure. The complete list of atomic coordinates may be used as input for visualization programs that allow a range of wire-frame, stick, space-filling, ribbon or cartoon representations to be generated based upon inbuilt heuristics and user interaction. However, most importantly, the mmCIF approach also offers a large collection of categories which are designed to provide descriptions of the structure at different levels of detail, and the relationships between data items in different categories permit the function of an individual atom site at any particular level of detail to be traced.

Before beginning the detailed description of the full mmCIF dictionary, it is helpful to demonstrate how it is used to describe the structure of a biological macromolecule. Fig. 3.6.3.1 shows the small protein crambin, which is a single polypeptide chain of 48 residues. The molecule co-crystallizes with a molecule of ethanol, although this is not thought to have any biological effect. Almost a quarter of the residues have side chains that adopt alternative conformations, and there is sequence heterogeneity at positions 22 (Pro/Ser) and 25 (Leu/Ile). Three disulfide links stabilize the structure.

The highest level of the description of the structure uses data items from the STRUCT category group. The crystallographic asymmetric unit contains one protein molecule, one co-crystallization ethanol molecule and a water solvent molecule. These are described with data items from the STRUCT_ASYM category (Example 3.6.3.1).

Each entry in this list assigns a label to a discrete component of the asymmetric unit and associates it with an entry in the entity list that defines each distinct chemical species in the crystal (Example 3.6.3.2).

The biological functions of the components of the crystal structure are described using data items in the STRUCT_BIOL and related categories. For crambin, the biological function is still unknown (see Example 3.6.3.3). This example also shows how the biological unit is generated from specific discrete objects in the asymmetric unit. In this case the relationship is trivial, but it will often be much more complex.

The secondary structure of the protein is described using data items in the STRUCT_CONF category (and in the STRUCT_SHEET category where relevant). The beginning and end labels for each

Example 3.6.3.2. *Specification of the distinct chemical entities in the crambin structure.*

```
loop_
  _entity.id
  _entity.type
  _entity.formula_weight
  _entity.src_method
    A        polymer      4716     natural
    ethanol  non-polymer    52     synthetic
    HOH      water          18     .
```

Example 3.6.3.3. *Identification of the biological function of the components of the crambin structure.*

```
_struct_biol.id                 crambin_1
_struct_biol.details
; The function of this protein is unknown and
  therefore the biological unit is assumed to be
  the single polypeptide chain without
  co-crystallization factors i.e. ethanol.
;
_struct_biol_gen.biol_id        crambin_1
_struct_biol_gen.asym_id        chain_a
_struct_biol_gen.symmetry       1_555
```

α-helix, β-strand or turn in Example 3.6.3.4 refer to the chemical components of the structural unit labelled chain_a at the given locations in the sequence (*e.g.* helix H1 runs from the isoleucine at position number 7 to the proline at position number 19 in the amino-acid sequence).

Interactions between different parts of the structure are described using data items in the STRUCT_CONN and related categories. In Example 3.6.3.5, some of the disulfide bridges and intramolecular hydrogen bonds are reported. As with the secondary structural elements, the partners in the links are identified by complex labels that include the chemical component involved, the object within the asymmetric unit that is under consideration, the position in the amino-acid (or nucleotide) sequence and the individual atom.

The objects identified at the highest level of the description of the structure are arbitrary. To discover their chemical identity, one needs to consult the ENTITY category group. As indicated above, each separate chemical species in the crystal should be specified in the entity table. Chemical entities are classified as polymer, non-polymer or water. Non-polymeric molecules, such as the co-crystallized ethanol in this example, are described as distinct chemical components using data items in the CHEM_COMP family of categories. Polymeric molecules are described using data items in the ENTITY_POLY family of categories.

In Example 3.6.3.6, the natural source for crambin is described, the overall features of the polypeptide chain are listed and the component parts (in effect the amino-acid sequence) are tabulated. Note that sequence heterogeneity is described by allowing a sequence number to be correlated with more than one monomer identifier (in the example, sequence number 22 is assigned both to proline and serine, while 25 is assigned to both leucine and isoleucine). Sequence heterogeneity can be defined by assigning suitable labels in the ATOM_SITE list.

The individual amino acids in the protein sequence of Example 3.6.3.6 are labelled by the data item `_entity_poly_seq.mon_id`; this refers to the separate chemical components listed in the CHEM_COMP family of categories (Example 3.6.3.7). As mentioned above, entries in these categories may be individual monomeric species within the crystal structure, or they may be amino acids or nucleotide bases that form the macromolecular polymer. In most cases, the entries recorded in these categories will be summaries of chemical information for standard amino acids and nucleotides, or references to external libraries of standard data for these. However, the categories contain enough data items to describe modified residues or co-crystallization factors in full if necessary.

At the most detailed level, the individual atom sites are described with data items in the ATOM category group, as shown for crambin in Example 3.6.3.8. A few points about this

Example 3.6.3.4. *Description of the secondary structure of crambin.*

```
loop_
  _struct_conf.id
  _struct_conf.conf_type_id
  _struct_conf.beg_label_comp_id
  _struct_conf.beg_label_asym_id
  _struct_conf.beg_label_seq_id
  _struct_conf.end_label_comp_id
  _struct_conf.end_label_asym_id
  _struct_conf.end_label_seq_id
  _struct_conf.details
    H1 HELX_RH_AL_P ILE chain_a  7 PRO chain_a 19
                                 'HELX-RH3T 17-19'
    H2 HELX_RH_AL_P GLU chain_a 23 THR chain_a 30
                                'Alpha-N start'
    S1 STRN_P       CYS chain_a 32 ILE chain_a 35 .
    S2 STRN_P       THR chain_a  1 CYS chain_a  4 .
    S3 STRN_P       ASN chain_a 46 ASN chain_a 46 .
    S4 STRN_P       THR chain_a 39 PRO chain_a 41 .
    T1 TURN-TY1_P   ARG chain_a 17 GLY chain_a 20 .
    T2 TURN-TY1_P   PRO chain_a 41 TYR chain_a 44 .
```

Example 3.6.3.5. *Interactions between parts of the crambin structure.*

```
loop_
  _struct_conn.id
  _struct_conn.conn_type_id
  _struct_conn.ptnr1_label_comp_id
  _struct_conn.ptnr1_label_asym_id
  _struct_conn.ptnr1_label_seq_id
  _struct_conn.ptnr1_label_atom_id
  _struct_conn.ptnr1_role
  _struct_conn.ptnr1_symmetry
  _struct_conn.ptnr2_label_comp_id
  _struct_conn.ptnr2_label_asym_id
  _struct_conn.ptnr2_label_seq_id
  _struct_conn.ptnr2_label_atom_id
  _struct_conn.ptnr2_role
  _struct_conn.ptnr2_symmetry
  _struct_conn.details
    SS1 disulf CYS chain_a  3 S  .        1_555
               CYS chain_a 40 S  .        1_555 .
    SS2 disulf CYS chain_a  4 S  .        1_555
               CYS chain_a 32 S  .        1_555 .
    HB1 hydrog SER chain_a  6 OG positive 1_555
               LEU chain_a  8 O  negative 1_556 .
    HB2 hydrog ARG chain_a 17 N  positive 1_555
               ASP chain_a 43 O  negative 1_554 .
```

Example 3.6.3.6. *Description of the crambin polypeptide.*

```
_entity_name_com.entity_id      A
_entity_name_com.name           crambin
_entity_src_nat.entity_id       A
_entity_src_nat.common_name     'Abyssinian Cabbage'
_entity_src_nat.genus           Crambe
_entity_src_nat.species         abyssinica
_entity_src_nat.details         ?
_entity_poly.entity_id          A
_entity_poly.type               polypeptide(L)
_entity_poly.nstd_chirality     no
_entity_poly.nstd_linkage       no
_entity_poly.nstd_monomer       no
_entity_poly.type_details
 'Sequence heterogeneity at residues 22 and 25'
loop_
  _entity_poly_seq.entity_id
  _entity_poly_seq.num
  _entity_poly_seq.mon_id
    A    1   THR     A    2   THR
# - - abbreviated - - -
    A   22   PRO     A   22   SER
    A   23   GLU     A   24   ALA
    A   25   LEU     A   25   ILE
# - - abbreviated - - -
    A   47   ALA     A   48   ASN
```

Example 3.6.3.7. *Separate chemical components forming the crambin polypeptide.*

```
loop_
_chem_comp.id
_chem_comp.mon_nstd_flag
_chem_comp.formula
_chem_comp.name
  ethanol .   'C2 H6 O1'         "ethanol"
  ALA     yes 'C3 H7 N1 O2'      "alanine"
  ARG     yes 'C6 H14 N4 O2'     "arginine"
  ASN     yes 'C4 H8 N2 O3'      "asparagine"
  ASP     yes 'C4 H7 N1 O4'      "aspartic acid"
  CYS     yes 'C3 H7 N1 O2 S1'   "cysteine"
  GLU     yes 'C5 H9 N1 O4'      "glutamic acid"
  GLY     yes 'C2 H5 N1 O2'      "glycine"
  ILE     yes 'C6 H13 N1 O2'     "isoleucine"
  LEU     yes 'C6 H13 N1 O2'     "leucine"
  PHE     yes 'C9 H11 N1 O2'     "phenylalanine"
  PRO     yes 'C5 H9 N1 O2'      "proline"
  SER     yes 'C3 H7 N1 O3'      "serine"
  THR     yes 'C4 H9 N1 O3'      "threonine"
  TYR     yes 'C9 H11 N1 O3'     "tyrosine"
  VAL     yes 'C5 H11 N1 O2'     "valine"
```

example should be noted. The composite labelling of each site includes a pointer to the description of the parent molecule as a specific object in the asymmetric unit (`_atom_site.label_asym_id`) and to the relevant monomeric building block of which the atom is a member (`_atom_site.label_comp_id`). The label component `_atom_site.label_alt_id` indicates alternative conformations in which an atom site may be found. For example, the atom sites numbered 3 and 4 are alternative locations for the α-carbon of the terminal residue. It may be deduced from the occupancies that the alternative conformations A and B are modelled with 80% and 20% occupancy, respectively, but this can be stated explicitly using the ATOM_SITES_ALT category. The sequence heterogeneity at residue 22 is shown by the presence of pointers to proline and serine, and the occupancy factors show that proline and serine are present in the ratio 60 to 40. There is also an alternative conformation within the serine at residue 22, split equally across two sites.

3.6.4. Content of the macromolecular CIF dictionary

Because it is derived from the core CIF dictionary, the mmCIF dictionary shares the same general structure as outlined in Chapter 3.2. However, DDL2 permits the formal assignment of categories to *category groups*. Table 3.6.4.1 lists the major category groups in the mmCIF dictionary (a full list is given in Appendix 3.6.1 and at the beginning of Chapter 4.5).

Small capitals are used for the names of category groups and individual categories in this volume, but the identifiers in the dictionary are actually lower-case strings.

The ordering of category groups in the remainder of this chapter follows the thematic scheme of Table 3.1.10.1. The discussion proceeds under the headings *Experimental measurements* (Section 3.6.5), *Analysis* (Section 3.6.6), *Atomicity, chemistry and structure* (Section 3.6.7), *Publication* (Section 3.6.8) and *File metadata* (Section 3.6.9).

Certain conventions of style and layout have been followed to summarize the large amount of information in the mmCIF dictionary and to help the reader navigate their way through this chapter. Appendix 3.6.1 is an overview of the mmCIF dictionary structure by category and lists all the categories with the number of the section in which they are discussed. This acts as an index between the alphabetical ordering within the dictionary and the thematic ordering of this chapter. Each thematic section lists the

Example 3.6.3.8. *Partial listing of the atomic coordinates of crambin.*

```
loop_
_atom_site.label_seq_id
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.label_alt_id
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.footnote_id
_atom_site.label_entity_id
_atom_site.id
1  N  N    THR chain_a A  16.864  14.059   3.442
   0.80  6.22  .   A    1
1  N  N    THR chain_a B  17.633  14.126   4.146
   0.20  8.40  .   A    2
1  C  CA   THR chain_a A  16.868  12.814   4.233
   0.80  4.45  .   A    3
1  C  CA   THR chain_a B  17.282  12.671   4.355
   0.20  7.82  .   A    4
1  C  C    THR chain_a .  15.583  12.775   4.990
   1.00  4.39  .   A    5
1  O  O    THR chain_a .  15.112  13.824   5.431
   1.00  7.04  .   A    6
1  C  CB   THR chain_a A  18.060  12.807   5.200
   0.80  5.42  .   A    7
1  C  CB   THR chain_a B  18.202  11.709   5.108
   0.20 11.07  .   A    8
1  O  OG1  THR chain_a A  19.233  12.892   4.380
   0.80  7.87  .   A    9
1  O  OG1  THR chain_a B  17.662  10.381   4.831
   0.20 14.39  .   A   10
1  C  CG2  THR chain_a A  18.117  11.578   6.092
   0.80  6.88  .   A   11
1  C  CG2  THR chain_a B  17.973  11.955   6.599
   0.20 19.74  .   A   12
# - - abbreviated - - -
22 N  N    PRO chain_a .   4.909  12.659  -3.127
   0.60  3.03  .   A  352
22 C  CA   PRO chain_a .   6.035  13.459  -2.622
   0.60  3.04  .   A  353
22 C  C    PRO chain_a .   6.362  13.139  -1.174
   0.60  3.08  .   A  354
22 O  O    PRO chain_a .   5.473  12.959  -0.323
   0.60  3.67  .   A  355
22 C  CB   PRO chain_a .   5.528  14.895  -2.825
   0.60  4.19  .   A  356
22 C  CG   PRO chain_a .   4.614  14.846  -4.059
   0.60  3.91  .   A  357
22 C  CD   PRO chain_a .   3.904  13.493  -3.885
   0.60  3.25  .   A  358
22 N  N    SER chain_a .   4.909  12.659  -3.127
   0.40  3.03  .   A  366
22 C  CA   SER chain_a .   6.035  13.459  -2.622
   0.40  3.04  .   A  367
22 C  C    SER chain_a .   6.362  13.139  -1.174
   0.40  3.08  .   A  368
22 O  O    SER chain_a .   5.473  12.959  -0.323
   0.40  3.67  .   A  369
22 C  CB   SER chain_a .   5.644  14.934  -2.679
   0.40  3.96  .   A  370
22 O  OG   SER chain_a C   4.712  15.250  -1.677
   0.20  3.53  .   A  371
22 O  OG   SER chain_a D   6.688  15.800  -2.315
   0.20  7.09  .   A  372
```

categories discussed in that section. Within each subsection, the data names within the relevant categories are listed. Category keys, pointers to parent data items and aliases to data items in the core CIF dictionary are indicated. For each category, the data item (or set of data items that must be considered together) that forms the category key is marked by a bullet (•) and listed first; the other data names follow in alphabetical order.

For measured or derived numerical quantities that should be specified with a standard uncertainty (in older terminology, an estimated standard deviation), the core dictionary uses the DDL1

Table 3.6.4.1. *Major category groups defined in the mmCIF dictionary*

The groups are listed in the order in which they are described in this chapter. There is also an INCLUSIVE category group, which serves as a formal higher-order container group to which all other category groups belong.

Section	Category group	Subject covered
(*a*) Experimental measurements		
3.6.5.1	CELL	Unit cell
3.6.5.2	DIFFRN	Diffraction experiment
3.6.5.3	EXPTL	Experimental conditions
(*b*) Analysis		
3.6.6.1	PHASING	Phasing techniques
3.6.6.2	REFINE	Refinement procedures
3.6.6.3	REFLN	Reflection measurements
(*c*) Atomicity, chemistry and structure		
3.6.7.1	ATOM	Atom sites
3.6.7.2	CHEMICAL	Chemical properties and nomenclature
3.6.7.3	ENTITY	Chemical entities
3.6.7.4	GEOM	Geometry of atom sites
3.6.7.5	STRUCT	Crystallographic structure
3.6.7.6	SYMMETRY	Symmetry information
3.6.7.7	VALENCE	Bond-valence information
(*d*) Publication		
3.6.8.1	CITATION	Bibliographic references
3.6.8.2	COMPUTING	Computational details of the experiment
3.6.8.3	DATABASE	Database information
3.6.8.4	IUCR	Journal housekeeping and the contents of a published article
(*e*) File metadata		
3.6.9.1	AUDIT	Dictionary maintenance and identification
3.6.9.2	ENTRY	Links between data blocks
3.6.9.3	COMPLIANCE	Compliance with previous dictionaries

attribute **_type_conditions_esd** and allows the standard uncertainty of the value to be placed in parentheses after the numerical value, as in

```
_cell_length_a      58.39(5)
```

This is also permitted in mmCIF, but it is preferable to use a separate data item to record the standard uncertainty, as in

```
_cell.length_a        58.39
_cell.length_a_esd     0.05
```

There are many of these kinds of data names in the mmCIF dictionary. The name of each is derived by adding **_esd** to the data name for the value. They are indicated by a + symbol in the category summaries in this chapter.

3.6.5. Experimental measurements

The CELL, DIFFRN and EXPTL category groups are used to describe the crystallographic experiment. The data items used for this purpose in mmCIF are for the most part identical to those in the core CIF dictionary. A complete discussion of the data names in each category may be found in Section 3.2.2.

mmCIF also contains the new categories EXPTL_CRYSTAL_GROW and EXPTL_CRYSTAL_GROW_COMP (Section 3.6.5.3.2), which are used to provide a more structured description of crystallization than is available in the core CIF dictionary.

3.6.5.1. Crystal cell parameters and measurement conditions

The categories describing the crystal unit cell and its determination are as follows:

CELL group
- CELL
- CELL_MEASUREMENT
- CELL_MEASUREMENT_REFLN

The mmCIF dictionary differs from the core CIF dictionary in assigning separate categories to data names that define the crystal unit-cell parameters and to data names relating to the experimental determination of the unit cell. Details of the unit-cell parameters are given in the CELL category and data items in the distinct CELL_MEASUREMENT category are used to describe how the unit-cell parameters were measured. The category CELL_MEASUREMENT_REFLN, which is used to list the reflections used in the unit-cell determination, is common to the core and mmCIF dictionaries.

The data items in these categories are as follows:

(*a*) CELL

```
• _cell.entry_id
        → _entry.id
+ _cell.angle_alpha
+ _cell.angle_beta
+ _cell.angle_gamma
  _cell.details (~ _cell_special_details)
  _cell.formula_units_Z
+ _cell.length_a
+ _cell.length_b
+ _cell.length_c
+ _cell.reciprocal_angle_alpha
+ _cell.reciprocal_angle_beta
+ _cell.reciprocal_angle_gamma
+ _cell.reciprocal_length_a
+ _cell.reciprocal_length_b
+ _cell.reciprocal_length_c
+ _cell.volume
  _cell.Z_PDB
```

(*b*) CELL_MEASUREMENT

```
• _cell_measurement.entry_id
        → _entry.id
+ _cell_measurement.pressure
  _cell_measurement.radiation
  _cell_measurement.reflns_used
+ _cell_measurement.temp
        (~ _cell_measurement_temperature)
  _cell_measurement.theta_max
  _cell_measurement.theta_min
  _cell_measurement.wavelength
```

(*c*) CELL_MEASUREMENT_REFLN

```
• _cell_measurement_refln.index_h
• _cell_measurement_refln.index_k
• _cell_measurement_refln.index_l
  _cell_measurement_refln.theta
```

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. Items in italics have aliases in the core CIF dictionary formed by changing the full stop (.) to an underscore (_) except where indicated by the ~ symbol. Data items marked with a plus (+) have companion data names for the standard uncertainty in the reported value, formed by appending the string **_esd** *to the data name listed.*

The summary above includes the formal category keys that have been introduced in mmCIF because the corresponding core categories do not expect looped data, and therefore do not require the specification of a unique identifier. In the relational model of DDL2, all categories are considered to be tables and therefore each category must have a unique identifier. Where core CIF categories have one or more data names that fulfil the role of table-row identifiers, these have generally been carried over as category keys in the mmCIF dictionary (for example, the data items that correspond to the *h*, *k*, and *l* Miller indices of a reflection in the CELL_MEASUREMENT_REFLN category).

Example 3.6.5.1. *Cell constants and their measurement for an HIV-1 protease crystal (PDB 5HVP) described with data items in the* CELL *and* CELL_MEASUREMENT *categories (Fitzgerald et al., 1990).*

```
_cell.entry_id                          '5HVP'
_cell.length_a                          58.39
_cell.length_a_esd                       0.05
_cell.length_b                          86.70
_cell.length_b_esd                       0.12
_cell.length_c                          46.27
_cell.length_c_esd                       0.06
_cell.angle_alpha                       90.00
_cell.angle_beta                        90.00
_cell.angle_gamma                       90.00
_cell.volume                           234237
_cell.details
; The cell parameters were refined every twenty
  frames during data integration. The cell lengths
  given are the mean of 55 such refinements; the
  esds given are the root-mean-square deviations
  of these 55 observations from that mean.
;
_cell_measurement.entry_id              '5HVP'
_cell_measurement.temp                    293
_cell_measurement.temp_esd                  3
_cell_measurement.theta_min                11
_cell_measurement.theta_max                31
_cell_measurement.wavelength             1.54
```

Example 3.6.5.1 shows how data items from these categories are used in practice and shows the use of separate data items to record standard uncertainties of measurable quantities.

3.6.5.2. Data collection

The categories describing data collection are as follows:

DIFFRN group
- DIFFRN
- DIFFRN_ATTENUATOR
- DIFFRN_DETECTOR
- DIFFRN_MEASUREMENT
- DIFFRN_ORIENT_MATRIX
- DIFFRN_ORIENT_REFLN
- DIFFRN_RADIATION
- DIFFRN_RADIATION_WAVELENGTH
- DIFFRN_REFLN
- DIFFRN_REFLNS
- DIFFRN_REFLNS_CLASS
- DIFFRN_SCALE
- DIFFRN_SOURCE
- DIFFRN_STANDARD_REFLN
- DIFFRN_STANDARDS

The categories in the DIFFRN category group describe the diffraction experiment. Data items in the DIFFRN category itself can be used to give overall information about the experiment, such as the temperature and pressure. Examples of the other categories are DIFFRN_DETECTOR, which is used for describing the detector used for data collection, and DIFFRN_SOURCE, which is used to give details of the source of the radiation used in the experiment. Data items in the DIFFRN_REFLN category can be used to give information about the raw data and data items in the DIFFRN_REFLNS category can be used to give information about all the reflection data collectively.

The data items in the categories in the DIFFRN group are as follows:

(*a*) DIFFRN

• **_diffrn.id**
  _diffrn.ambient_environment
\+ **_diffrn.ambient_pressure**
  _diffrn.ambient_pressure_gt
  _diffrn.ambient_pressure_lt
\+ **_diffrn.ambient_temp** (∼ **_diffrn_ambient_temperature**)
  _diffrn.ambient_temp_details
  _diffrn.ambient_temp_gt
  _diffrn.ambient_temp_lt
  _diffrn.crystal_id (∼ **_diffrn_refln_crystal_id**)
  _diffrn.crystal_support
  _diffrn.crystal_treatment
  _diffrn.details (∼ **_diffrn_special_details**)

(*b*) DIFFRN_ATTENUATOR

• **_diffrn_attenuator.code**
  _diffrn_attenuator.material
  _diffrn_attenuator.scale

(*c*) DIFFRN_DETECTOR

• **_diffrn_detector.diffrn_id**
  → **_diffrn.id**
  _diffrn_detector.area_resol_mean
  _diffrn_detector.details
  _diffrn_detector.detector (∼ **_diffrn_detector**)
  _diffrn_detector.dtime
  _diffrn_detector.type

(*d*) DIFFRN_MEASUREMENT

• **_diffrn_measurement.diffrn_id**
  → **_diffrn.id**
  _diffrn_measurement.details
  _diffrn_measurement.device
  _diffrn_measurement.device_details
  _diffrn_measurement.device_type
  _diffrn_measurement.method
  _diffrn_measurement.specimen_support

(*e*) DIFFRN_ORIENT_MATRIX

• **_diffrn_orient_matrix.diffrn_id**
  → **_diffrn.id**
  _diffrn_orient_matrix.type
  _diffrn_orient_matrix.UB[1][1]
  (∼ **_diffrn_orient_matrix_UB_11**)
  _diffrn_orient_matrix.UB[1][2]
  (∼ **_diffrn_orient_matrix_UB_12**)
  _diffrn_orient_matrix.UB[1][3]
  (∼ **_diffrn_orient_matrix_UB_13**)
  _diffrn_orient_matrix.UB[2][1]
  (∼ **_diffrn_orient_matrix_UB_21**)
  _diffrn_orient_matrix.UB[2][2]
  (∼ **_diffrn_orient_matrix_UB_22**)
  _diffrn_orient_matrix.UB[2][3]
  (∼ **_diffrn_orient_matrix_UB_23**)
  _diffrn_orient_matrix.UB[3][1]
  (∼ **_diffrn_orient_matrix_UB_31**)
  _diffrn_orient_matrix.UB[3][2]
  (∼ **_diffrn_orient_matrix_UB_32**)
  _diffrn_orient_matrix.UB[3][3]
  (∼ **_diffrn_orient_matrix_UB_33**)

(*f*) DIFFRN_ORIENT_REFLN

• **_diffrn_orient_refln.diffrn_id**
  → **_diffrn.id**
• **_diffrn_orient_refln.index_h**
• **_diffrn_orient_refln.index_k**
• **_diffrn_orient_refln.index_l**
  _diffrn_orient_refln.angle_chi
  _diffrn_orient_refln.angle_kappa
  _diffrn_orient_refln.angle_omega
  _diffrn_orient_refln.angle_phi
  _diffrn_orient_refln.angle_psi
  _diffrn_orient_refln.angle_theta

(*g*) DIFFRN_RADIATION

• **_diffrn_radiation.diffrn_id**
  → **_diffrn.id**
  _diffrn_radiation.collimation
  _diffrn_radiation.filter_edge
  _diffrn_radiation.inhomogeneity
  _diffrn_radiation.monochromator
  _diffrn_radiation.polarisn_norm
  _diffrn_radiation.polarisn_ratio
  _diffrn_radiation.probe
  _diffrn_radiation.type

_diffrn_radiation.wavelength_id
→ _diffrn_radiation_wavelength.id
_diffrn_radiation.xray_symbol

(*h*) DIFFRN_RADIATION_WAVELENGTH
• _diffrn_radiation_wavelength.id
• _diffrn_radiation_wavelength.wavelength
(∼ _diffrn_radiation_wavelength.wavelength)
• _diffrn_radiation_wavelength.wt

(*i*) DIFFRN_REFLN
• _diffrn_refln.diffrn_id
→ _diffrn.id
• _diffrn_refln.id
_diffrn_refln.angle_chi
_diffrn_refln.angle_kappa
_diffrn_refln.angle_omega
_diffrn_refln.angle_phi
_diffrn_refln.angle_psi
_diffrn_refln.angle_theta
_diffrn_refln.attenuator_code
→ _diffrn_attenuator.code
_diffrn_refln.class_code
_diffrn_refln.counts_bg_1
_diffrn_refln.counts_bg_2
_diffrn_refln.counts_net
_diffrn_refln.counts_peak
_diffrn_refln.counts_total
_diffrn_refln.detect_slit_horiz
_diffrn_refln.detect_slit_vert
_diffrn_refln.elapsed_time
_diffrn_refln.index_h
_diffrn_refln.index_k
_diffrn_refln.index_l
_diffrn_refln.intensity_net
_diffrn_refln.intensity_sigma
_diffrn_refln.intensity_u
_diffrn_refln.scale_group_code
→ _diffrn_scale_group.code
_diffrn_refln.scan_mode
_diffrn_refln.scan_mode_backgd
_diffrn_refln.scan_rate
_diffrn_refln.scan_time_backgd
_diffrn_refln.scan_width
_diffrn_refln.sint_over_lambda
(∼ _diffrn_refln.sint/lambda)
_diffrn_refln.standard_code
→ _diffrn_standard_refln.code
_diffrn_refln.wavelength
_diffrn_refln.wavelength_id
→ _diffrn_radiation.wavelength_id

(*j*) DIFFRN_REFLNS
• _diffrn_reflns.diffrn_id
→ _diffrn.id
_diffrn_reflns.av_R_equivalents
_diffrn_reflns.av_sigmaI_over_netI
_diffrn_reflns.av_unetI/netI
_diffrn_reflns.limit_h_max
_diffrn_reflns.limit_h_min
_diffrn_reflns.limit_k_max
_diffrn_reflns.limit_k_min
_diffrn_reflns.limit_l_max
_diffrn_reflns.limit_l_min
_diffrn_reflns.number
_diffrn_reflns.reduction_process
_diffrn_reflns.theta_max
_diffrn_reflns.theta_min
_diffrn_reflns.transf_matrix[1][1]
(∼ _diffrn_reflns_transf_matrix_11)
_diffrn_reflns.transf_matrix[1][2]
(∼ _diffrn_reflns_transf_matrix_12)
_diffrn_reflns.transf_matrix[1][3]
(∼ _diffrn_reflns_transf_matrix_13)
_diffrn_reflns.transf_matrix[2][1]
(∼ _diffrn_reflns_transf_matrix_21)
_diffrn_reflns.transf_matrix[2][2]
(∼ _diffrn_reflns_transf_matrix_22)
_diffrn_reflns.transf_matrix[2][3]
(∼ _diffrn_reflns_transf_matrix_23)
_diffrn_reflns.transf_matrix[3][1]
(∼ _diffrn_reflns_transf_matrix_31)
_diffrn_reflns.transf_matrix[3][2]
(∼ _diffrn_reflns_transf_matrix_32)
_diffrn_reflns.transf_matrix[3][3]
(∼ _diffrn_reflns_transf_matrix_33)

(*k*) DIFFRN_REFLNS_CLASS
• _diffrn_reflns_class.code
_diffrn_reflns_class.av_R_eq
_diffrn_reflns_class.av_sgI/I
_diffrn_reflns_class.av_uI/I
_diffrn_reflns_class.d_res_high
_diffrn_reflns_class.d_res_low
_diffrn_reflns_class.description
_diffrn_reflns_class.number

(*l*) DIFFRN_SCALE_GROUP
• _diffrn_scale_group.code
_diffrn_scale_group.I_net

(*m*) DIFFRN_SOURCE
• _diffrn_source.diffrn_id
→ _diffrn.id
_diffrn_source.current
_diffrn_source.details
_diffrn_source.power
_diffrn_source.size
_diffrn_source.source (∼ _diffrn_source)
_diffrn_source.take-off_angle
_diffrn_source.target
_diffrn_source.type
_diffrn_source.voltage

(*n*) DIFFRN_STANDARD_REFLN
• _diffrn_standard_refln.code
• _diffrn_standard_refln.diffrn_id
→ _diffrn.id
_diffrn_standard_refln.index_h
_diffrn_standard_refln.index_k
_diffrn_standard_refln.index_l

(*o*) DIFFRN_STANDARDS
• _diffrn_standards.diffrn_id
→ _diffrn.id
_diffrn_standards.decay_%
_diffrn_standards.interval_count
_diffrn_standards.interval_time
_diffrn_standards.number
_diffrn_standards.scale_sigma
_diffrn_standards.scale_u

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* (_) *except where indicated by the* ∼ *symbol. Data items marked with a plus* (+) *have companion data names for the standard uncertainty in the reported value, formed by appending the string* _esd *to the data name listed.*

To a very great extent, data items in the DIFFRN category group are used in the same way in the mmCIF and core CIF dictionaries, and Section 3.2.2.2 can be consulted for details. Example 3.6.5.2 shows how these categories are used to describe the data collection for a macromolecule.

There is, however, one important difference. An mmCIF may describe several separate diffraction experiments that were conducted with a common purpose; each such experiment would be given a unique value of _diffrn.id, the key for the DIFFRN category. Descriptions of features of that experiment in related categories would be given a matching identifier with the same value (*e.g.* _diffrn_detector.diffrn_id). The use of the suffix *.diffrn_id for the key data names in each related category emphasizes the connection to the parent experiment.

As a consequence, there are differences between the mmCIF and core CIF dictionaries in the definition of the category keys for

Example 3.6.5.2. *Data collection for an HIV-1 protease crystal (PDB 5HVP) described with data items in the* DIFFRN *and related categories.*

```
_diffrn.id                              'set1'
_diffrn.crystal_id                       1
_diffrn.ambient_temp                     293(3)
_diffrn.ambient_environment
; Mother liquor from the reservoir of the vapor
  diffusion experiment, mounted in room air
;
_diffrn.crystal_support
;  0.7 mm glass capillary, sealed with dental wax
;
_diffrn.crystal_treatment
; Equilibrated in rotating anode radiation enclosure
  for 18 hours prior to beginning of data collection.
;
_diffrn_detector.diffrn_id               'set1'
_diffrn_detector.detector                'multiwire'
_diffrn_detector.type                    'Siemens'

_diffrn_measurement.diffrn_id            'd1'
_diffrn_measurement.device          '3-circle camera'
_diffrn_measurement.device_type      'Supper model x'
_diffrn_measurement.device_details       'none'
_diffrn_measurement.method               'omega scan'
_diffrn_measurement.details
; 440 frames, 0.20 degrees, 150 sec, detector
  distance 12 cm, detector angle 22.5 degrees
;
_diffrn_radiation.diffrn_id               'set1'
_diffrn_radiation.collimation
                              '0.3 mm double pinhole'
_diffrn_radiation.monochromator           'graphite'
_diffrn_radiation.type                    'Cu Kalpha'
_diffrn_radiation.wavelength_id            1
_diffrn_radiation_wavelength.id            1
_diffrn_radiation_wavelength.wavelength   1.54
_diffrn_radiation_wavelength.wt            1.0
_diffrn_source.diffrn_id                'set1'
_diffrn_source.source                   'rotating anode'
_diffrn_source.type                     'Rigaku RU-200'
_diffrn_source.power                     50
_diffrn_source.current                   180
_diffrn_source.target       '8mm x 0.4 mm broad-focus'
```

the DIFFRN categories. These differences were introduced in order to accommodate data from more than one experiment in the same table. For example, in the core CIF dictionary, the Miller indices `_diffrn_refln_index_h`, `*_k` and `*_l` play the role of the category key for the DIFFRN_REFLN category. In the mmCIF dictionary, the category key is formed by the data items `_diffrn_refln.id` and `_diffrn_refln.diffrn_id`.

3.6.5.3. Growth, description and analysis of the crystal

The categories describing the crystal properties and growth are as follows:

EXPTL group

Crystal properties (§3.6.5.3.1)

EXPTL

EXPTL_CRYSTAL

EXPTL_CRYSTAL_FACE

Crystal growth (§3.6.5.3.2)

EXPTL_CRYSTAL_GROW

EXPTL_CRYSTAL_GROW_COMP

Categories in the EXPTL category group are used to describe experimental measurements on the crystal (*e.g.* of its shape, size and density) and the growth of the crystal. Data items in the EXPTL category are used to describe the gross properties of the crystal or crystals used in the experiment. Data items in the EXPTL_CRYSTAL category are used to describe the crystal properties in detail and allow for cases where multiple crystals are used. The data items in the EXPTL_CRYSTAL_FACE category are used to describe the crystal faces.

Data items for describing crystal growth are given in two categories that are not found in the current version of the core CIF dictionary. Data items in the EXPTL_CRYSTAL_GROW category are used to describe the conditions and methods used to grow the crystals, and data items in the EXPTL_CRYSTAL_GROW_COMP category can be used to list the components of the solutions in which the crystals were grown.

3.6.5.3.1. *Crystal properties*

The data items in these categories are as follows:

(*a*) EXPTL

- `_exptl.entry_id`
  → `_entry.id`

  `_exptl.absorpt_coefficient_mu`
  `_exptl.absorpt_correction_T_max`
  `_exptl.absorpt_correction_T_min`
  `_exptl.absorpt_correction_type`
  `_exptl.absorpt_process_details`
  `_exptl.crystals_number`
  `_exptl.details` (∼ `_exptl_special_details`)
  `_exptl.method`
  `_exptl.method_details`

(*b*) EXPTL_CRYSTAL

- `_exptl_crystal.id`

  `_exptl_crystal.colour`
  `_exptl_crystal.colour_lustre`
  `_exptl_crystal.colour_modifier`
  `_exptl_crystal.colour_primary`
  `_exptl_crystal.density_diffrn`
  `_exptl_crystal.density_Matthews`
+ `_exptl_crystal.density_meas`
  `_exptl_crystal.density_meas_gt`
  `_exptl_crystal.density_meas_lt`
+ `_exptl_crystal.density_meas_temp`
  `_exptl_crystal.density_meas_temp_gt`
  `_exptl_crystal.density_meas_temp_lt`
  `_exptl_crystal.density_method`
  `_exptl_crystal.density_percent_sol`
  `_exptl_crystal.description`
  `_exptl_crystal.F_000`
  `_exptl_crystal.preparation`
  `_exptl_crystal.size_max`
  `_exptl_crystal.size_mid`
  `_exptl_crystal.size_min`
  `_exptl_crystal.size_rad`

(*c*) EXPTL_CRYSTAL_FACE

- `_exptl_crystal_face.crystal_id`
  → `_exptl_crystal.id`
- `_exptl_crystal_face.index_h`
- `_exptl_crystal_face.index_k`
- `_exptl_crystal_face.index_l`

  `_exptl_crystal_face.diffr_chi`
  `_exptl_crystal_face.diffr_kappa`
  `_exptl_crystal_face.diffr_phi`
  `_exptl_crystal_face.diffr_psi`
  `_exptl_crystal_face.perp_dist`

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* (_) *except where indicated by the* ∼ *symbol. Data items marked with a plus* (+) *have companion data names for the standard uncertainty in the reported value, formed by appending the string* `_esd` *to the data name listed.*

Data items in these categories are used in the same way in the mmCIF and core CIF dictionaries, and Section 3.2.2.3 can be consulted for details (see Example 3.6.5.3). Identifiers have been introduced to the categories to provide the formal category keys required by the DDL2 data model.

Example 3.6.5.3. *The crystal used in the determination of an HIV-1 protease structure (PDB 5HVP) described using data items in the* EXPTL *and* EXPTL_CRYSTAL *categories.*

```
_exptl.entry_id                    '5HVP'
_exptl.crystals_number             1
_exptl.method     'single-crystal x-ray diffraction'
_exptl.method_details
; graphite monochromatized Cu K(alpha) fixed tube
  and Siemens multiwire detector used
;
_exptl_crystal.id                  1
_exptl_crystal.colour              'colorless'
_exptl_crystal.density_percent_sol0.57
_exptl_crystal.description        'rectangular plate'
_exptl_crystal.size_max            0.30
_exptl_crystal.size_mid            0.20
_exptl_crystal.size_min            0.05
```

3.6.5.3.2. *Crystal growth*

The data items in these categories are as follows:

(*a*) EXPTL_CRYSTAL_GROW

```
• _exptl_crystal_grow.crystal_id
        → _exptl_crystal.id
  _exptl_crystal_grow.apparatus
  _exptl_crystal_grow.atmosphere
  _exptl_crystal_grow.details
  _exptl_crystal_grow.method
  _exptl_crystal_grow.method_ref
  _exptl_crystal_grow.pH
+ _exptl_crystal_grow.pressure
  _exptl_crystal_grow.seeding
  _exptl_crystal_grow.seeding_ref
+ _exptl_crystal_grow.temp
  _exptl_crystal_grow.temp_details
  _exptl_crystal_grow.time
```

(*b*) EXPTL_CRYSTAL_GROW_COMP

```
• _exptl_crystal_grow_comp.crystal_id
        → _exptl_crystal.id
• _exptl_crystal_grow_comp.id
  _exptl_crystal_grow_comp.conc
  _exptl_crystal_grow_comp.details
  _exptl_crystal_grow_comp.name
  _exptl_crystal_grow_comp.sol_id
  _exptl_crystal_grow_comp.volume
```

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. Data items marked with a plus* (+) *have companion data names for the standard uncertainty in the reported value, formed by appending the string* `_esd` *to the data name listed.*

Crystallization strategies and protocols are very varied and may not lend themselves to a formal tabulation. Common or well defined techniques may be indicated using the data item `_exptl_crystal_grow.method`, and a literature reference, where appropriate, may be given using `_exptl_crystal_grow.method_ref`. Frequently, however, a detailed description of methodology is required; this can be given in `_exptl_crystal_grow.details`. Example 3.6.5.4 shows how information about strategies that were attempted and proved unsuccessful can be recorded. In circumstances such as this, the data item `_exptl_crystal_grow.pH` would record the final pH.

Where the crystallization protocol is well defined, it is useful to list the individual components of the solution in the category EXPTL_CRYSTAL_GROW_COMP. Example 3.6.5.4 labels the solutions used as 1 and 2, in accordance with the convention that solution 1 contains the molecule to be crystallized and solution 2 (and if necessary additional solutions) contains the precipitant. However, it is permissible and may be preferable to use more explicit labels such as 'well solution' in the `_exptl_crystal_grow_comp.sol_id` field.

Example 3.6.5.4. *The growth of HIV-1 protease crystals (PDB 5HVP) described with data items in the* EXPTL_CRYST_GROW *and* EXPTL_CRYSTAL_GROW_COMP *categories.*

```
_exptl_crystal_grow.crystal_id     1
_exptl_crystal_grow.method         'hanging drop'
_exptl_crystal_grow.apparatus      'Linbro plates'
_exptl_crystal_grow.atmosphere     'room air'
_exptl_crystal_grow.pH             4.7
_exptl_crystal_grow.temp           18(3)
_exptl_crystal_grow.time    'approximately 2 days'
_exptl_crystal_grow.details
; The dependence on pH for successful crystal growth
  is very sharp. At pH 7.4 only showers of tiny
  crystals grew, at pH 7.5 well formed single
  crystals grew, at pH 7.6 no crystallization
  occurred at all.
;
loop_
  _exptl_crystal_grow_comp.crystal_id
  _exptl_crystal_grow_comp.id
  _exptl_crystal_grow_comp.sol_id
  _exptl_crystal_grow_comp.name
  _exptl_crystal_grow_comp.volume
  _exptl_crystal_grow_comp.conc
  _exptl_crystal_grow_comp.details
  1  1  1  'HIV-1 protease'  '0.002 ml'  '6 mg/ml'
; The protein solution was in a buffer containing
  25 mM NaCl, 100 mM NaMES/MES buffer, pH 7.5,
  3 mM NaAzide
;
  1  2  2  'NaCl'  '0.200 ml'  '4 M'
  'in 3 mM NaAzide'
  1  3  2  'Acetic Acid'  '0.047 ml'  '100 mM'
  'in 3 mM NaAzide'
  1  4  2  'Na Acetate'  '0.053 ml'  '100 mM'
; in 3 mM NaAzide. Buffer components were mixed
  to produce a pH of 4.7 according to a ratio
  calculated from the pKa. The actual pH of
  solution 2 was not measured.
;
  1  5  2  'water'  '0.700 ml'  'neat'
 'in 3 mM NaAzide'
```

3.6.6. Analysis

The mmCIF dictionary contributes several new categories and data items to the REFINE and REFLN category groups. These reflect common practices in macromolecular crystallography in refinement and in the handling of experimental observations.

A new category group, the PHASING group, has been introduced to provide a structured description of phasing strategies, as macromolecular crystallography differs strongly from small-molecule crystallography in how phases are determined. The data model for phasing in the current version of the mmCIF dictionary cannot describe all approaches to phasing yet. Additions and revisions to the data items in the PHASING group of categories are anticipated in future versions of the dictionary.

3.6.6.1. Phasing

The categories describing phasing are as follows:

PHASING group

Overall description of phasing (§3.6.6.1.1)

PHASING

Phasing via molecular averaging (§3.6.6.1.2)

PHASING_AVERAGING

Phasing via isomorphous replacement (§3.6.6.1.3)

PHASING_ISOMORPHOUS

Phasing via multiple-wavelength anomalous dispersion (§3.6.6.1.4)

PHASING_MAD

PHASING_MAD_CLUST

PHASING_MAD_EXPT
PHASING_MAD_RATIO
PHASING_MAD_SET

Phasing via multiple isomorphous replacement (§3.6.6.1.5)

PHASING_MIR
PHASING_MIR_DER
PHASING_MIR_DER_REFLN
PHASING_MIR_DER_SHELL
PHASING_MIR_DER_SITE
PHASING_MIR_DER_SHELL

Phasing data sets (§3.6.6.1.6)

PHASING_SET
PHASING_SET_REFLN

The data items in the PHASING category group can be used to record details about the phasing of the structure and cover the various methods used in the phasing process. Many data items are provided for multiple isomorphous replacement (MIR) and multiple-wavelength anomalous dispersion (MAD). More limited sets of data items are provided for phasing using molecular averaging and phasing *via* using a structure that is isomorphous to the present structure. The current version of the mmCIF dictionary does not provide specific data items for recording the details of phasing *via* molecular replacement.

3.6.6.1.1. *Overall description of phasing*

The single data item in this category is as follows:

PHASING
• **_phasing.method**

The bullet (•) *indicates a category key.*

Phasing of macromolecular structures often involves the application of more than one of the methods described in the PHASING section of the mmCIF dictionary, such as when phases generated from a multiple isomorphous replacement experiment are improved by molecular averaging. The PHASING category is used to list the methods that were used.

At present, the category contains a single data item, the purpose of which is to specify the method employed in the structure determination. It may have one or more of the values listed in the dictionary (Example 3.6.6.1).

3.6.6.1.2. *Phasing via molecular averaging*

The data items in this category are as follows:

PHASING_AVERAGING
• **_phasing_averaging.entry_id**
  → _entry.id
  _phasing_averaging.details
  _phasing_averaging.method

The bullet (•) *indicates a category key. The arrow* (→) *is a reference to a parent data item.*

When more than one copy of a molecule is present in the asymmetric unit, phases can be improved by averaging an electron-density map over the multiple images of the molecule. In some special cases with very high noncrystallographic symmetry, *de novo* phases have been derived by iterative application of molecular averaging, but more often averaging is used to improve phases determined by another method.

There are many protocols used for phasing with averaging and they are very varied. It was not thought to be appropriate to specify data items for any one approach in the current version of the mmCIF dictionary. The data items that are provided allow a text-based description of the protocol to be given; a formalism for recording a fully parsable description of molecular averaging needs to be developed for future revisions of the dictionary.

Example 3.6.6.1. *The methods used to generate the phases for a hypothetical structure described with the data item in the* PHASING *category.*

```
loop_
_phasing.method
   'mir'
   'averaging'
```

Example 3.6.6.2. *Phase improvement with molecular averaging for a hypothetical structure described with data items in the* PHASING_AVERAGING *category.*

```
_phasing_averaging.entry_id     'EXAMHYPO'
_phasing_averaging.method
; Iterative threefold averaging alternating with
  phase extensions by 0.5 reciprocal lattice units
  per cycle.
;

_phasing_averaging.details
; The position of the threefold axis was redetermined
  every five cycles.
;
```

Data items in the PHASING_AVERAGING category allow free-text descriptions to be given of the method used for structure determination or phase improvement using averaging over multiple observations of the molecule in the asymmetric unit and of any specific details of the application of the method to the current structure determination (Example 3.6.6.2). Note that the reference to the method is to be used to describe the method itself, and not as a reference to a software package; references to software packages would be made using data items in the SOFTWARE category.

3.6.6.1.3. *Phasing via isomorphous replacement*

The data items in this category are as follows:

PHASING_ISOMORPHOUS
• **_phasing_isomorphous.entry_id**
  → _entry.id
  _phasing_isomorphous.details
  _phasing_isomorphous.method
  _phasing_isomorphous.parent

The bullet (•) *indicates a category key. The arrow* (→) *is a reference to a parent data item.*

Phases for many macromolecular structures are obtained from a previous determination of the same structure in the same crystal lattice. Examples of this are the determination of the structure of a point mutant or the determination of a structure in which a ligand is bound to an active site that was empty in the previous structure determination. In these cases, the new structure is essentially isomorphous with the parent structure, hence this method of phasing is termed 'isomorphous phasing' in the mmCIF dictionary. It is not to be confused with multiple isomorphous phasing (MIR), a phasing technique that involves the use of heavy-atom derivatives. MIR phasing is discussed in Section 3.6.6.1.5.

Not much information is needed to characterize isomorphous phasing. The 'parent' structure (the structure used to generate the initial phases for the present structure) is described in a free-text field and a second free-text field can be used to give details of the application of the method to the determination of the present structure (for instance, the removal of solvent or a bound ligand). In Example 3.6.6.3, the parent structure is the PDB entry 5HVP and the structure that is the subject of the present data block is identified as 'HVP+CmpdA'. **_phasing_isomorphous.method** allows

Example 3.6.6.3. *Isomorphous replacement phasing of an HIV-1 protease structure described using data items in the* PHASING_ISOMORPHOUS *category.*

```
_phasing_isomorphous.entry_id     'HVP+CmpdA'
_phasing_isomorphous.parent       'PDB entry 5HVP'
_phasing_isomorphous.details
; The inhibitor and all solvent atoms were removed
  from the parent structure before beginning
  refinement. All static disorder present in the
  parent structure was also removed.
;
```

any formal techniques that were used in the application of the method to the present structure determination to be described, for example rigid-body refinement. Note that this data item is not to be used to reference a software package; this would be done using data items in the SOFTWARE category.

3.6.6.1.4. *Phasing via multiple-wavelength anomalous dispersion*

The data items in these categories are as follows:

(*a*) PHASING_MAD

- _phasing_MAD.entry_id
  → _entry.id
  _phasing_MAD.details
  _phasing_MAD.method

(*b*) PHASING_MAD_CLUST

- _phasing_MAD_clust.expt_id
  → _phasing_MAD_clust.expt_id
- _phasing_MAD_clust.id
  _phasing_MAD_clust.number_set

(*c*) PHASING_MAD_EXPT

- _phasing_MAD_expt.id
  _phasing_MAD_expt.delta_delta_phi
  _phasing_MAD_expt.delta_phi
  _phasing_MAD_expt.delta_phi_sigma
  _phasing_MAD_expt.mean_fom
  _phasing_MAD_expt.number_clust
  _phasing_MAD_expt.R_normal_all
  _phasing_MAD_expt.R_normal_anom_scat

(*d*) PHASING_MAD_RATIO

- _phasing_MAD_ratio.expt_id
  → _phasing_MAD_expt.id
- _phasing_MAD_ratio.clust_id
  → _phasing_MAD_clust.id
- _phasing_MAD_ratio.wavelength_1
  → _phasing_MAD_set.wavelength
- _phasing_MAD_ratio.wavelength_2
  → _phasing_MAD_set.wavelength
  _phasing_MAD_ratio.d_res_high
  _phasing_MAD_ratio.d_res_low
  _phasing_MAD_ratio.ratio_one_wl
  _phasing_MAD_ratio.ratio_one_wl_centric
  _phasing_MAD_ratio.ratio_two_wl

(*e*) PHASING_MAD_SET

- _phasing_MAD_set.clust_id
  → _phasing_MAD_clust.id
- _phasing_MAD_set.expt_id
  → _phasing_MAD_expt.id
- _phasing_MAD_set.set_id
  → _phasing_set.id
- _phasing_MAD_set.wavelength
  _phasing_MAD_set.d_res_high
  _phasing_MAD_set.d_res_low
  _phasing_MAD_set.f_double_prime
  _phasing_MAD_set.f_prime
  _phasing_MAD_set.wavelength_details

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.

PHASING_MAD and related categories are used to provide information about phasing using the multiple-wavelength anomalous dispersion (MAD) technique. The data model used for MAD phasing in the current version of the mmCIF dictionary is that of Hendrickson, as exemplified in the structure determination of N-cadherin (Shapiro *et al.*, 1995; Example 3.6.6.4). In current practice, MAD phasing is often treated as a special case of MIR phasing and the PHASING_MIR categories would be more appropriate to describe the results.

Unlike the PHASING_MIR categories, there is no provision in the current mmCIF model of MAD phasing for analysis of the overall phasing statistics and the contribution to the phasing of each data set by bins of resolution, and no provision for giving a list of the phased reflections. This will need to be addressed in future versions of the mmCIF dictionary.

Example 3.6.6.4. *MAD phasing of the structure of N-cadherin (Shapiro et al., 1995) described using data items in the* PHASING_MAD *and related categories.*

```
_phasing_MAD.entry_id       'NCAD'

loop_
_phasing_MAD_expt.id
_phasing_MAD_expt.number_clust
_phasing_MAD_expt.R_normal_all
_phasing_MAD_expt.R_normal_anom_scat
_phasing_MAD_expt.delta_delta_phi
_phasing_MAD_expt.delta_phi_sigma
_phasing_MAD_expt.mean_fom
   1  2  0.063  0.451  58.5  20.3  0.88
   2  1  0.051  0.419  36.8  18.2  0.93

loop_
_phasing_MAD_clust.id
_phasing_MAD_clust.expt_id
_phasing_MAD_clust.number_set
  'four wavelength'  1  4
  'five wavelength'  1  5
  'five wavelength'  2  5

loop_
_phasing_MAD_ratio.expt_id
_phasing_MAD_ratio.clust_id
_phasing_MAD_ratio.wavelength_1
_phasing_MAD_ratio.wavelength_2
_phasing_MAD_ratio.d_res_low
_phasing_MAD_ratio.d_res_high
_phasing_MAD_ratio.ratio_two_wl
_phasing_MAD_ratio.ratio_one_wl
_phasing_MAD_ratio.ratio_one_wl_centric
   1  'four wavelength'  1.4013 1.4013  20.00  4.00
        .  0.084 0.076
   1  'four wavelength'  1.4013 1.3857  20.00  4.00
     0.067      .     .
   1  'four wavelength'  1.4013 1.3852  20.00  4.00
     0.051      .     .
   1  'four wavelength'  1.4013 1.3847  20.00  4.00
     0.044      .     .
   1  'four wavelength'  1.3857 1.3857  20.00  4.00
        .  0.110 0.049
   1  'four wavelength'  1.3857 1.3852  20.00  4.00
     0.049      .     .
# - - - abbreviated - - -

loop_
_phasing_MAD_set.expt_id
_phasing_MAD_set.clust_id
_phasing_MAD_set.set_id
_phasing_MAD_set.wavelength
_phasing_MAD_set.wavelength_details
_phasing_MAD_set.d_res_low
_phasing_MAD_set.d_res_high
_phasing_MAD_set.f_prime
_phasing_MAD_set.f_double_prime
   1 'four wavelength' aa 1.4013 'pre-edge'  20.00
      3.00   -12.48   3.80
   1 'four wavelength' bb 1.3857 'peak'      20.00
      3.00.  -31.22  17.20
   1 'four wavelength' cc 1.3852 'edge'      20.00
      3.00   -13.97  29.17
```

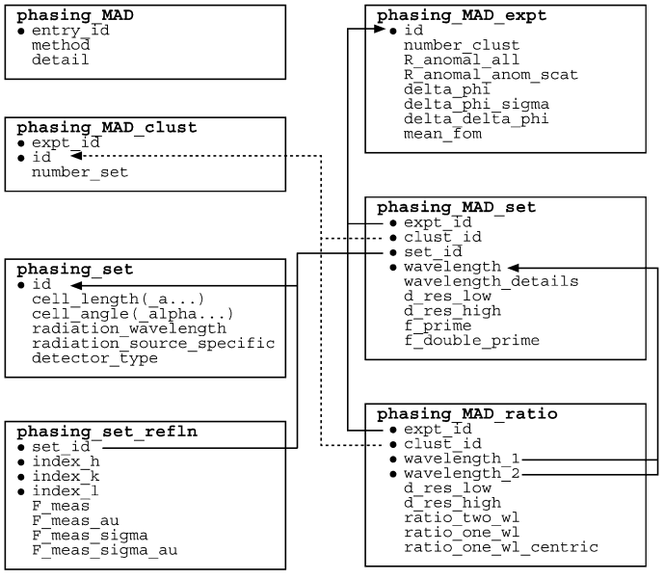

Fig. 3.6.6.1. The family of categories used to describe MAD phasing. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

The relationships between categories describing MAD phasing are shown in Fig. 3.6.6.1.

Data items in the PHASING_MAD category allow a brief overview of the method that was used to be given and allow special aspects of the phasing strategy to be noted; data items in this category are analogous to the data items in the other overview categories describing phasing techniques.

In the data model for MAD phasing used in the present version of the mmCIF dictionary, a collection of data sets measured at different wavelengths can be used to construct more than one set of phases. These phase sets will produce electron-density maps with different local properties. The model of the structure is often constructed using information from a collection of these maps. The collections of multiple phase sets are referred to as 'experiments' and the groups of data sets that contribute to each experiment are referred to as 'clusters'. Data items in PHASING_MAD_EXPT identify each experiment and give the number of contributing clusters. Additional data items record the phase difference between the structure factors due to normal scattering from all atoms and from only the anomalous scatterers, the standard uncertainty of this quantity, the mean figure of merit, and a number of other indicators of the quality of the phasing.

Data items in the PHASING_MAD_CLUST category can be used to label the clusters of data sets and give the number of data sets allocated to each cluster. In Example 3.6.6.4 two experiments are described. The first experiment contains two clusters, one of which contains four data sets and the second of which contains five data sets. The second experiment contains a single cluster of five data sets. Note that the author has chosen informative labels to identify the clusters ('four wavelength', 'five wavelength'). Carefully chosen labels can help someone reading the mmCIF to trace the complex relationships between the categories.

Data items in the PHASING_MAD_RATIO category can be used to record the ratios of phasing statistics (Bijvoet differences) between pairs of data sets in a MAD phasing experiment, within shells of resolution characterized by **_phasing_MAD_ratio.d_res_high** and ***.d_res_low**.

The data sets used in the MAD phasing experiments are described using data items in the PHASING_MAD_SET category. Each data set is characterized by resolution shell and wavelength, and by the f' and f'' components of the anomalous scattering factor at that wavelength. The actual observations in each data set and the experimental conditions under which they were made are recorded using data items in the PHASING_SET and PHASING_SET_REFLN categories.

3.6.6.1.5. *Phasing via multiple isomorphous replacement*

The data items in these categories are as follows:

(*a*) PHASING_MIR

- **_phasing_MIR.entry_id**
  → **_entry.id**

_phasing_MIR.details
_phasing_MIR.d_res_high
_phasing_MIR.d_res_low
_phasing_MIR.FOM
_phasing_MIR.FOM_acentric
_phasing_MIR.FOM_centric
_phasing_MIR.method
_phasing_MIR.reflns
_phasing_MIR.reflns_acentric
_phasing_MIR.reflns_centric
_phasing_MIR.reflns_criterion

(*b*) PHASING_MIR_SHELL

- **_phasing_MIR_shell.d_res_high**
- **_phasing_MIR_shell.d_res_low**

_phasing_MIR_shell.FOM
_phasing_MIR_shell.FOM_acentric
_phasing_MIR_shell.FOM_centric
_phasing_MIR_shell.loc
_phasing_MIR_shell.mean_phase
_phasing_MIR_shell.power
_phasing_MIR_shell.R_cullis
_phasing_MIR_shell.R_kraut
_phasing_MIR_shell.reflns
_phasing_MIR_shell.reflns_acentric
_phasing_MIR_shell.reflns_anomalous
_phasing_MIR_shell.reflns_centric

(*c*) PHASING_MIR_DER

- **_phasing_MIR_der.id**

_phasing_MIR_der.d_res_high
_phasing_MIR_der.d_res_low
_phasing_MIR_der.der_set_id
→ **_phasing_set.id**
_phasing_MIR_der.details
_phasing_MIR_der.native_set_id
→ **_phasing_set.id**
_phasing_MIR_der.number_of_sites
_phasing_MIR_der.power_acentric
_phasing_MIR_der.power_centric
_phasing_MIR_der.R_cullis_acentric
_phasing_MIR_der.R_cullis_anomalous
_phasing_MIR_der.R_cullis_centric
_phasing_MIR_der.reflns_acentric
_phasing_MIR_der.reflns_anomalous
_phasing_MIR_der.reflns_centric
_phasing_MIR_der.reflns_criteria

(*d*) PHASING_MIR_DER_REFLN

- **_phasing_MIR_der_refln.der_id**
  → **_phasing_MIR_der.id**
- **_phasing_MIR_der_refln.index_h**
- **_phasing_MIR_der_refln.index_k**
- **_phasing_MIR_der_refln.index_l**
- **_phasing_MIR_der_refln.set_id**
  → **_phasing_set.id**

_phasing_MIR_der_refln.F_calc
_phasing_MIR_der_refln.F_calc_au
_phasing_MIR_der_refln.F_meas
_phasing_MIR_der_refln.F_meas_au
_phasing_MIR_der_refln.F_meas_sigma
_phasing_MIR_der_refln.F_meas_sigma_au
_phasing_MIR_der_refln.HL_A_iso
_phasing_MIR_der_refln.HL_B_iso
_phasing_MIR_der_refln.HL_C_iso

```
  _phasing_MIR_der_refln.HL_D_iso
  _phasing_MIR_der_refln.phase_calc

(e) PHASING_MIR_DER_SHELL
• _phasing_MIR_der_shell.d_res_high
• _phasing_MIR_der_shell.d_res_low
• _phasing_MIR_der_shell.der_id
        → _phasing_MIR_der.id
  _phasing_MIR_der_shell.fom
  _phasing_MIR_der_shell.ha_ampl
  _phasing_MIR_der_shell.loc
  _phasing_MIR_der_shell.phase
  _phasing_MIR_der_shell.power
  _phasing_MIR_der_shell.R_cullis
  _phasing_MIR_der_shell.R_kraut
  _phasing_MIR_der_shell.reflns

(f) PHASING_MIR_DER_SITE
• _phasing_MIR_der_site.der_id
        → _phasing_MIR_der.id
• _phasing_MIR_der_site.id
  _phasing_MIR_der_site.atom_type_symbol
        → _atom_type.symbol
+ _phasing_MIR_der_site.B_iso
+ _phasing_MIR_der_site.Cartn_x
+ _phasing_MIR_der_site.Cartn_y
+ _phasing_MIR_der_site.Cartn_z
  _phasing_MIR_der_site.details
+ _phasing_MIR_der_site.fract_x
+ _phasing_MIR_der_site.fract_y
+ _phasing_MIR_der_site.fract_z
  _phasing_MIR_der_site.occupancy
  _phasing_MIR_der_site.occupancy_anom
  _phasing_MIR_der_site.occupancy_anom_su
  _phasing_MIR_der_site.occupancy_iso
  _phasing_MIR_der_site.occupancy_iso_su
```

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. Data items marked with a plus (+) have companion data names for the standard uncertainty in the reported value, formed by appending the string **_esd** *to the data name listed.*

Fig. 3.6.6.2. The family of categories used to describe MIR phasing. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

PHASING_MIR and related categories provide information about phasing by methods involving multiple isomorphous replacement (MIR). These same categories may also be used to describe phasing by related techniques, such as single isomorphous replacement (SIR) and single or multiple isomorphous replacement plus anomalous scattering (SIRAS, MIRAS). The relationships between the categories describing MIR phasing are shown in Fig. 3.6.6.2.

As with the other overview categories described in this section, the PHASING_MIR category contains data items that can be used for text-based descriptions of the method used and any special aspects of its application. There are also items for describing the resolution limit of the reflections that were phased, the figures of merit for all reflections and for the acentric reflections phased in the native data set, and the total numbers of reflections and their inclusion threshold in the native data set. Statistics for the phasing can be given by shells of resolution using data items in the PHASING_MIR_SHELL category.

An MIR phasing experiment involves one or more derivatives. The remaining categories in this group are used to describe aspects of each derivative (Example 3.6.6.5). A derivative in this context does not necessarily correspond to a data set; for instance, the same data set could be used to one resolution limit as an isomorphous scatterer and to a different resolution (and with a different sigma cutoff) as an anomalous scatterer. These would be treated as two distinct derivatives, although both derivatives would point to the same data sets *via* **_phasing_MIR_der.der_set_id** and **_phasing_MIR_der.native_set_id** (see Fig. 3.6.6.2).

Data items in the PHASING_MIR_DER category can be used to identify and describe each derivative. The resolution limits for the individual derivatives need not match those of the overall phasing experiment, as the phasing power of each derivative as a function of resolution will vary. Many of the statistical descriptors of phasing given in the PHASING_MIR category are repeated in this category, as derivatives vary in quality and their contribution to the phasing must be assessed individually. These same statistical measures can be given for shells of resolution in the PHASING_MIR_DER_SHELL category.

Data items in the PHASING_MIR_DER_REFLN category can be used to provide details of each reflection used in an MIR phasing experiment. The pointer **_phasing_MIR_der_refln.set_id** links the reflection to a particular set of experimental data and **_phasing_MIR_der_refln.der_id** points to a particular derivative used in the phasing (as mentioned above, derivatives in this context do not equate to data sets). The phase assigned to each reflection and the measured and calculated values of its structure factor can be given. (It is not necessary to include the measured values of the structure factors in this list, since they are accessible in the PHASING_SET_REFLN category, but it may be convenient to present them here). Data items are also provided for the *A*, *B*, *C* and *D* phasing coefficients of Hendrickson & Lattman (1970).

The heavy atoms identified in each derivative can be listed using data items in the PHASING_MIR_DER_SITE category. Most of the data names are clear analogues of similar items in the ATOM_SITE category; an exception is **_phasing_MIR_der_site.occupancy_anom**, which specifies the relative anomalous occupancy of the atom type present at a heavy-atom site in a particular derivative.

Example 3.6.6.5. *Phasing of the structure of bovine plasma retinol-binding protein (Zanotti et al., 1993) described using data items in the* PHASING_MIR *and related categories.*

```
_phasing_MIR.entry_id          '1HBP'
_phasing_MIR.method
;  Standard phase refinement (Blow & Crick, 1959)
;

loop_
_phasing_MIR_shell.d_res_low
_phasing_MIR_shell.d_res_high
_phasing_MIR_shell.reflns
_phasing_MIR_shell.FOM
 15.0  8.3   80  0.69       8.3  6.4  184  0.73
  6.4  5.2  288  0.72       5.2  4.4  406  0.65
  4.4  3.8  554  0.54       3.8  3.4  730  0.53
  3.4  3.0  939  0.50

loop_
_phasing_MIR_der.id
_phasing_MIR_der.number_of_sites
_phasing_MIR_der.details
KAu(CN)2  3
     'major site interpreted in difference Patterson'
K2HgI4    6 'sites found in cross-difference Fourier'
K3IrCl6   2 'sites found in cross-difference Fourier'
All      11 'data for all three derivatives combined'

loop_
_phasing_MIR_der_shell.der_id
_phasing_MIR_der_shell.d_res_low
_phasing_MIR_der_shell.d_res_high
_phasing_MIR_der_shell.ha_ampl
_phasing_MIR_der_shell.loc
   KAu(CN)2  15.0  8.3   54   26
   KAu(CN)2   8.3  6.4   54   20
# - - - abbreviated - - -
   K2HgI4    15.0  8.3  149   87
   K2HgI4     8.3  6.4  121   73
# - - - abbreviated - - -
   K3IrCl6   15.0  8.3   33   27
   K3IrCl6    8.3  6.4   40   23
# - - - abbreviated - - -

loop_
_phasing_MIR_der_site.der_id
_phasing_MIR_der_site.id
_phasing_MIR_der_site.atom_type_symbol
_phasing_MIR_der_site.occupancy
_phasing_MIR_der_site.fract_x
_phasing_MIR_der_site.fract_y
_phasing_MIR_der_site.fract_z
_phasing_MIR_der_site.B_iso
   KAu(CN)2  1  Au  0.40  0.082  0.266  0.615  33.0
   KAu(CN)2  2  Au  0.03  0.607  0.217  0.816  25.9
   K2HgI4    1  Hg  0.63  0.048  0.286  0.636  33.7
   K2HgI4    2  Hg  0.34  0.913  0.768  0.889  36.7
# - - - abbreviated - - -

_phasing_MIR_der_refln.index_h           6
_phasing_MIR_der_refln.index_k           1
_phasing_MIR_der_refln.index_l          25
_phasing_MIR_der_refln.der_id        HGPT1
_phasing_MIR_der_refln.set_id      'NS1-96'
_phasing_MIR_der_refln.F_calc_au    106.66
_phasing_MIR_der_refln.F_meas_au    204.67
_phasing_MIR_der_refln.F_meas_sigma   6.21
_phasing_MIR_der_refln.HL_A_iso      -3.15
_phasing_MIR_der_refln.HL_B_iso      -0.76
_phasing_MIR_der_refln.HL_C_iso       0.65
_phasing_MIR_der_refln.HL_D_iso       0.23
_phasing_MIR_der_refln.phase_calc   194.48
```

3.6.6.1.6. *Phasing data sets*

The data items in these categories are as follows:

(*a*) PHASING_SET

```
• _phasing_set.id
  _phasing_set.cell_angle_alpha
  _phasing_set.cell_angle_beta
  _phasing_set.cell_angle_gamma
  _phasing_set.cell_length_a
  _phasing_set.cell_length_b
  _phasing_set.cell_length_c
  _phasing_set.detector_specific
  _phasing_set.detector_type
  _phasing_set.radiation_source_specific
  _phasing_set.radiation_wavelength
  _phasing_set.temp
```

(*b*) PHASING_SET_REFLN

```
• _phasing_set_refln.index_h
• _phasing_set_refln.index_k
• _phasing_set_refln.index_l
• _phasing_set_refln.set_id
        → _phasing_set.id
  _phasing_set_refln.F_meas
  _phasing_set_refln.F_meas_au
  _phasing_set_refln.F_meas_sigma
  _phasing_set_refln.F_meas_sigma_au
```

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.

Data items in the PHASING_SET family of categories are homologous to items with related names in the CELL and DIFFRN families of categories. The PHASING_SET categories were added to the mmCIF data model so that intensity and phase information for the data sets used in phasing could be stored in the same data block as the information for the refined structure. It is not necessary to store all the experimental information for each data set (*e.g.* the raw data sets or crystal growth conditions); it is assumed that the full experimental description of each phasing set would be recorded in a separate data block (see Example 3.6.6.6).

Data items in the PHASING_SET category identify each set of diffraction data used in a phasing experiment and can be used to summarize relevant experimental conditions. Because a given data set may be used in a number of different ways (for example, as an isomorphous derivative and as a component of a multiple-wavelength calculation), it is appropriate to store the reflections in a category distinct from either the PHASING_MAD or PHASING_MIR family of categories, but accessible to both these families (and any similar categories that might be introduced later to describe new phasing methods). Figs. 3.6.6.1 and 3.6.6.2 show how reference is made to the relevant sets from within the PHASING_MAD and PHASING_MIR categories.

Each phasing set is given a unique value of **_phasing_set.id**. The other PHASING_SET data items record the cell dimensions and

Example 3.6.6.6. *The phasing sets used in the structure determination of bovine plasma retinol-binding protein (Zanotti et al., 1993) described with data items in the* PHASING_SET *and* PHASING_SET_REFLN *categories.*

```
_phasing_set.id                    'NS1-96'
_phasing_set.cell_angle_alpha      90.0
_phasing_set.cell_angle_beta       90.0
_phasing_set.cell_angle_gamma      90.0
_phasing_set.cell_length_a         38.63
_phasing_set.cell_length_b         38.63
_phasing_set.cell_length_c         82.88
_phasing_set.radiation_wavelength  1.5145
_phasing_set.detector_type         'image plate'
_phasing_set.detector_specific     'RXII'

_loop
_phasing_set_refln.set_id
_phasing_set_refln.index_h
_phasing_set_refln.index_k
_phasing_set_refln.index_l
_phasing_set_refln.F_meas_au
_phasing_set_refln.F_meas_sigma_au
  'NS1-96'  15  15  32  181.79  3.72
  'NS1-96'  15  15  33   34.23  1.62
# - - - abbreviated - - -
```

angles associated with each phasing set, the wavelength of the radiation used in the experiment, the source of the radiation, the detector type, and the ambient temperature.

Data items in the PHASING_SET_REFLN category are used to record the values of the measured structure factors and their uncertainties. Several distinct data sets may be present in this list, with reflections in each set identified by the appropriate value of **_phasing_set_refln.set_id**.

3.6.6.2. Refinement

The categories describing refinement are as follows:

REFINE group

Overall description of the refinement (§3.6.6.2.1)
- REFINE
- REFINE_FUNCT_MINIMIZED

Analysis of the refined structure (§3.6.6.2.2)
- REFINE_ANALYZE

Restraints and refinement by shells of resolution (§3.6.6.2.3)
- REFINE_LS_RESTR
- REFINE_LS_RESTR_NCS
- REFINE_LS_RESTR_TYPE
- REFINE_LS_SHELL
- REFINE_LS_CLASS

Equivalent atoms in the refinement (§3.6.6.2.4)
- REFINE_B_ISO
- REFINE_OCCUPANCY

History of the refinement (§3.6.6.2.5)
- REFINE_HIST

The macromolecular CIF dictionary contains many more data items for describing the refinement process than the core CIF dictionary does. In addition to new items in the REFINE category itself, additional categories have been introduced to describe in great detail the function minimized and the restraints applied, and the history of the refinement process, which often has many cycles. The REFINE_ANALYZE category can be used to give details of many of the quantities that may be used to assess the quality of the refinement. The REFINE_LS_SHELL category allows results to be reported by shells of resolution, and in effect replaces the more general core CIF category REFINE_LS_CLASS.

3.6.6.2.1. *Overall description of the refinement*

The data items in these categories are as follows:

(*a*) REFINE

- **_refine.entry_id**
  → **_entry.id**
  _refine.aniso_B[1][1]
  _refine.aniso_B[1][2]
  _refine.aniso_B[1][3]
  _refine.aniso_B[2][2]
  _refine.aniso_B[2][3]
  _refine.aniso_B[3][3]
  _refine.B_iso_max
  _refine.B_iso_mean
  _refine.B_iso_min
  _refine.correlation_coeff_Fo_to_Fc
  _refine.correlation_coeff_Fo_to_Fc_free
  _refine.details (∼ ***_refine_special_details***)
+ ***_refine.diff_density_max***
+ ***_refine.diff_density_min***
+ ***_refine.diff_density_rms***
  _refine.ls_abs_structure_details
+ ***_refine.ls_abs_structure_Flack***
+ ***_refine.ls_abs_structure_Rogers***
  _refine.ls_d_res_high
  _refine.ls_d_res_low
+ ***_refine.ls_extinction_coef***
  _refine.ls_extinction_expression
  _refine.ls_extinction_method
+ ***_refine.ls_goodness_of_fit_all***
  _refine.ls_goodness_of_fit_gt
+ ***_refine.ls_goodness_of_fit_obs***
  _refine.ls_goodness_of_fit_ref
  _refine.ls_hydrogen_treatment
  _refine.ls_matrix_type
  _refine.ls_number_constraints
  _refine.ls_number_parameters
  _refine.ls_number_reflns_all
  _refine.ls_number_reflns_obs
  (∼ ***_refine_ls_number_reflns***)
  _refine.ls_number_reflns_R_free
  _refine.ls_number_reflns_R_work
  _refine.ls_number_restraints
  _refine.ls_percent_reflns_obs
  _refine.ls_percent_reflns_R_free
  _refine.ls_R_factor_all
  _refine.ls_R_factor_gt
  _refine.ls_R_factor_obs
  _refine.ls_R_factor_R_free
  _refine.ls_R_factor_R_free_error
  _refine.ls_R_factor_R_free_error_details
  _refine.ls_R_factor_R_work
  _refine.ls_R_Fsqd_factor_obs
  (∼ ***_refine_ls_R_Fsqd_factor***)
  _refine.ls_R_I_factor_obs (∼ ***_refine_ls_R_I_factor***)
  _refine.ls_redundancy_reflns_all
  _refine.ls_redundancy_reflns_obs
  _refine.ls_restrained_S_all
  _refine.ls_restrained_S_obs
  _refine.ls_shift_over_esd_max
  (∼ ***_refine_ls_shift/esd_max***)
  _refine.ls_shift_over_esd_mean
  (∼ ***_refine_ls_shift/esd_mean***)
  _refine.ls_shift_over_su_max
  (∼ ***_refine_ls_shift/su_max***)
  _refine.ls_shift_over_su_max_lt
  (∼ ***_refine_ls_shift/su_max_lt***)
  _refine.ls_shift_over_su_mean
  (∼ ***_refine_ls_shift/su_mean***)
  _refine.ls_shift_over_su_mean_lt
  (∼ ***_refine_ls_shift/su_mean_lt***)
  _refine.ls_structure_factor_coef
  _refine.ls_weighting_details
  _refine.ls_weighting_scheme
  _refine.ls_wR_factor_all
  _refine.ls_wR_factor_obs
  _refine.ls_wR_factor_R_free
  _refine.ls_wR_factor_R_work
  _refine.occupancy_max
  _refine.occupancy_min
  _refine.overall_FOM_free_R_set
  _refine.overall_FOM_work_R_set
  _refine.overall_SU_B
  _refine.overall_SU_ML
  _refine.overall_SU_R_Cruickshank_DPI
  _refine.overall_SU_R_free
  _refine.solvent_model_details
  _refine.solvent_model_param_bsol
  _refine.solvent_model_param_ksol

(*b*) REFINE_FUNCT_MINIMIZED

- **_refine_funct_minimized.type**
  _refine_funct_minimized.number_terms
  _refine_funct_minimized.residual
  _refine_funct_minimized.weight

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 (_) except where indicated by the ∼ symbol. Data items marked with a plus (+) have companion data names for the standard uncertainty in the reported value, formed by appending the string **_esd** *to the data name listed.*

There is already an extensive set of data names in the REFINE category of the core dictionary, and Section 3.2.3.1 should be read with the present section. The only data items discussed in this section are entries in the mmCIF dictionary that do not have a counterpart in the core CIF dictionary. Analogues of a number of *R* factors in the core CIF dictionary have been added to the mmCIF dictionary to express these same *R* factors indepen-

dently for the free and working sets of reflections. The remaining new data items have more specialized roles, which are discussed below.

The data item **_refine.entry_id** has been added to the REFINE category to provide the formal category key required by the DDL2 data model.

Many macromolecular structure refinements now use the statistical cross-validation technique of monitoring a 'free' *R* factor (Brünger, 1997). $R_{\rm free}$ is calculated the same way as the conventional least-squares *R* factor, but using a small subset of reflections that are not used in the refinement of the structural model. Thus $R_{\rm free}$ tests how well the model predicts experimental observations that are not themselves used to fit the model.

The mmCIF dictionary provides data names for $R_{\rm free}$ and for the complementary $R_{\rm work}$ values for the 'working' set of reflections, which are the reflections that are used in the refinement. Separate data items are provided for unweighted and weighted versions of each *R* factor. A fixed percentage of the total number of reflections is usually assigned to the free group, and this percentage can be specified. Further details about the method used for selecting the free reflections can be given using **_reflns.R_free_details**. The estimated error in the $R_{\rm free}$ value may also be given, along with the method used for determining its value.

The purposes of having a set of reflections that are not used in the refinement are to monitor the progress of the refinement and to ensure that the *R* factor is not being artificially reduced by the introduction of too many parameters. However, as the refinement converges, the working and free *R* factors both approach stable values. It is common practice, particularly in structures at high resolution, to stop monitoring $R_{\rm free}$ at this point and to include all the reflections in the final rounds of refinement. It is thus worth noting a distinction between **_refine.ls_R_factor_obs** and **_refine.ls_R_factor_R_work**: **_refine.ls_R_factor_obs** relates to a refinement in which all reflections more intense than a specified threshold were used, while **_refine.ls_R_factor_R_work** relates to a refinement in which a subset of the observed reflections were excluded from the refinement and were used to calculate the free *R* factor. The dictionary allows the use of both values if a free *R* factor were calculated for most of the refinement, but all of the observed reflections were used in the final rounds of refinement; the protocol for this may be explained in **_refine.details**. When a full history of the refinement is provided using data items in the REFINE_HIST category, it is preferable to specify a change in protocol using data items in this category.

Other data items help to provide an assessment of the quality of the refinement. The scale-independent correlation coefficient between the observed and calculated structure factors may be recorded for the reflections included in the refinement using the data item **_refine.correlation_coeff_Fo_to_Fc**. There is a similar data item for the reflections that were not included in the refinement.

Overall standard uncertainties for positional and displacement parameters can be recorded according to a number of conventions. A maximum-likelihood residual for the positional parameters can be given using **_refine.overall_SU_ML** and the corresponding value for the displacement parameters can be given using **_refine.overall_SU_B**. Diffraction-component precision indexes for the displacement parameters based on the crystallographic *R* factor (the Cruickshank DPI; Cruickshank, 1999) can be given using **_refine.overall_SU_R_Cruickshank_DPI**. The corresponding value for $R_{\rm free}$ can be given using **_refine.overall_SU_R_free**.

Example 3.6.6.7. *Results of the overall refinement of an HIV-1 protease structure (PDB 5HVP) described using data items in the* REFINE *and* REFINE_FUNCT_MINIMIZED *categories.*

```
_refine.entry_id                          '5HVP'
_refine.ls_number_reflns_obs              12901
_refine.ls_number_restraints              6609
_refine.ls_number_parameters              7032
_refine.ls_R_factor_obs                   0.176
_refine.ls_weighting_scheme               calc
_refine.ls_weighting_details
;  Sigdel model of Konnert-Hendrickson:
   Sigdel: Afsig +  Bfsig*(sin(theta)/lambda-1/6)
   Afsig = 22.0, Bfsig = -150.0 at the beginning
      of refinement.
   Afsig = 15.5, Bfsig =  -50.0 at the end of
      refinement.
;
loop_
  _refine_funct_minimized.type
  _refine_funct_minimized.number_terms
  _refine_funct_minimized.residual
    'sum(W*Delta(Amplitude)^2^'       3009   1621.3
    'sum(W*Delta(Plane+Rigid)^2^'       85     56.68
    'sum(W*Delta(Distance)^2^'        1219    163.59
    'sum(W*Delta(U-tempfactors)^2^'   1192     69.338
```

The quality of a data set used for the refinement of a macromolecular structure is often given not only in terms of the scaling residuals, but also in terms of the data redundancy (the ratio of the number of reflections measured to the number of crystallographically unique reflections). Data items are provided to express the redundancy of all reflections, as well as those that have been marked as 'observed' (*i.e.* exceeding the threshold for inclusion in the refinement). The percentage of the total number of reflections that are considered observed is another metric of the quality of the data set, and a data item is provided for this (**_refine.ls_percent_reflns_obs**).

The limited resolution of many macromolecular data sets makes it inappropriate to refine anisotropic displacement factors for each atom. For these low- to medium-resolution studies, an overall anisotropic displacement model may be refined. The data items **_refine.aniso_B*** are provided for recording the unique elements of the matrix that describes the refined anisotropy.

The two-parameter method for modelling the contribution of the bulk solvent to the scattering proposed by Tronrud is used in several refinement programs. The data items **_refine.solvent_model_*** can be used to record the scale and displacement factors of this model, and any special aspects of its application to the refinement.

The average phasing figure of merit can be given for the working and free reflections. Unusually high or low values of displacement factors or occupancies can be a sign of problems with the refinement, so data items are provided to record the high, low and mean values of each. Further indicators of the quality of the refinement are found in the REFINE_ANALYZE category (Section 3.6.6.2.2).

The data items in the REFINE_FUNCT_MINIMIZED category allow a brief description of the function minimized during refinement to be given (Example 3.6.6.7). It is not possible to reconstruct the functioned minimized during the refinement by automatic parsing of the values of these data items, but the details given in them may still be helpful to someone reading the mmCIF.

3.6.6.2.2. *Analysis of the refined structure*

The data items in this category are as follows:

REFINE_ANALYZE

- **_refine_analyze.entry_id**
  → **_entry.id**

```
_refine_analyze.Luzzati_coordinate_error_free
_refine_analyze.Luzzati_coordinate_error_obs
_refine_analyze.Luzzati_d_res_low_free
_refine_analyze.Luzzati_d_res_low_obs
_refine_analyze.Luzzati_sigma_a_free
_refine_analyze.Luzzati_sigma_a_free_details
_refine_analyze.Luzzati_sigma_a_obs
_refine_analyze.Luzzati_sigma_a_obs_details
_refine_analyze.number_disordered_residues
_refine_analyze.occupancy_sum_hydrogen
_refine_analyze.occupancy_sum_non_hydrogen
_refine_analyze.RG_d_res_high
_refine_analyze.RG_d_res_low
_refine_analyze.RG_free
_refine_analyze.RG_free_work_ratio
_refine_analyze.RG_work
```

The bullet (•) indicates a category key. The arrow (→) is a reference to a parent data item.

In small-molecule crystallography, there is general agreement on the metrics that should be used to assess the quality of a structure determination, and data items in the REFINE category of the core CIF dictionary can be used to record them. For macromolecular structure determinations, no such agreement has been achieved yet and new metrics are frequently suggested as the field evolves. The REFINE_ANALYZE category can be used to record the metrics that were in common use at the time that the mmCIF dictionary was constructed; it is anticipated that new metrics will be added in future versions of the dictionary, and that some of the current metrics may fall into disuse.

Luzzati (1952) devised a method for estimating the average positional shift that would be needed in an idealized refinement to reach an R factor of zero by using a plot of R factors against resolution. For some time, macromolecular crystallographers have used a modification of this approach to assess the average positional error. Recent practice has used Luzzati plots based on the free R values to yield a cross-validated error estimate. Data items are provided for recording these coordinate-error estimates and the range of resolution included in the plot (Example 3.6.6.8). Related data names allow the specification of the value of σ_a used in constructing the Luzzati plot.

A general feature of introducing more parameters in the model of the structure is a reduction in the R factor, but the statistical significance of this is often obscured by the simultaneous reduction in the ratio of observations to parameters. Attempts to extend Hamilton's (1965) test to macromolecular structures are usually confounded by the use of restraints. Tickle *et al.* (1998) proposed the use of a Hamilton generalized R factor analyzed separately for reflections in the working set (those used in the refinement) and for reflections in the free set (those set aside for cross validation), and these metrics are often reported in the literature. Data items are provided for recording the Hamilton generalized R factor for the working and free set of reflections, and for the ratio of the two.

Other indicators of a successful refinement involve the relative order of the model. Data items are provided for recording the sum of the occupancies of the hydrogen and non-hydrogen atoms in the model. The number of disordered residues may also be recorded.

3.6.6.2.3. *Restraints and refinement by shells of resolution*

The data items in these categories are as follows:

(*a*) REFINE_LS_RESTR

```
• _refine_ls_restr.type
  _refine_ls_restr.criterion
  _refine_ls_restr.dev_ideal
  _refine_ls_restr.dev_ideal_target
  _refine_ls_restr.number
  _refine_ls_restr.rejects
  _refine_ls_restr.weight
```

Example 3.6.6.8. *Aspects of the refinement of an HIV-1 protease structure (PDB 5HVP) described with data items in the REFINE_ANALYZE category.*

```
loop_
_refine_analyze.entry_id                          '5HVP'
_refine_analyze.Luzzati_coordinate_error_obs       0.32
_refine_analyze.Luzzati_d_res_low_obs              5.0
```

(*b*) REFINE_LS_RESTR_NCS

```
• _refine_ls_restr_ncs.dom_id
        → _struct_ncs_dom.id
  _refine_ls_restr_ncs.ncs_model_details
  _refine_ls_restr_ncs.rms_dev_B_iso
  _refine_ls_restr_ncs.rms_dev_position
  _refine_ls_restr_ncs.weight_B_iso
  _refine_ls_restr_ncs.weight_position
```

(*c*) REFINE_LS_RESTR_TYPE

```
• _refine_ls_restr_type.type
        → _refine_ls_restr.type
  _refine_ls_restr_type.distance_cutoff_high
  _refine_ls_restr_type.distance_cutoff_low
```

(*d*) REFINE_LS_SHELL

```
• _refine_ls_shell.d_res_high
• _refine_ls_shell.d_res_low
  _refine_ls_shell.number_reflns_all
  _refine_ls_shell.number_reflns_obs
  _refine_ls_shell.number_reflns_R_free
  _refine_ls_shell.number_reflns_R_work
  _refine_ls_shell.percent_reflns_obs
  _refine_ls_shell.percent_reflns_R_free
  _refine_ls_shell.R_factor_all
  _refine_ls_shell.R_factor_obs
  _refine_ls_shell.R_factor_R_free
  _refine_ls_shell.R_factor_R_free_error
  _refine_ls_shell.R_factor_R_work
  _refine_ls_shell.redundancy_reflns_all
  _refine_ls_shell.redundancy_reflns_obs
  _refine_ls_shell.wR_factor_all
  _refine_ls_shell.wR_factor_obs
  _refine_ls_shell.wR_factor_R_free
  _refine_ls_shell.wR_factor_R_work
```

(*e*) REFINE_LS_CLASS

```
• _refine_ls_class.code
  _refine_ls_class.d_res_high
  _refine_ls_class.d_res_low
  _refine_ls_class.R_factor_all
  _refine_ls_class.R_factor_gt
  _refine_ls_class.R_Fsqd_factor
  _refine_ls_class.R_I_factor
  _refine_ls_class.wR_factor_all
```

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.

These categories were introduced in the mmCIF dictionary to allow a detailed description of several aspects of structure refinement to be given. Data items in the REFINE_LS_RESTR category allow geometric restraints to be specified and the deviations of restrained parameters from ideal values in the final model to be given. The type of the geometric restraints can be described in more detail using data items in the REFINE_LS_RESTR_TYPE category. Data items in the REFINE_LS_RESTR_NCS category can be used to give information about any restraints on noncrystallographic symmetry used in the refinement and the category REFINE_LS_SHELL contains data items that allow the results of refinement to be given by shells of resolution.

Data items in the REFINE_LS_RESTR category can be used to record details about the restraints applied to various classes of parameters during least-squares refinement (Example 3.6.6.9). It is clearly useful to tabulate the various classes of restraint, their deviation from ideal target values and the criteria used to reject

Example 3.6.6.9. *Results of the refinement of an HIV-1 protease structure (PDB 5HVP) described with data items in the* REFINE_LS_RESTR *and* REFINE_LS_SHELL *categories.*

```
loop_
_refine_ls_restr.type
_refine_ls_restr.dev_ideal_target
_refine_ls_restr.dev_ideal
_refine_ls_restr.number
_refine_ls_restr.criterion
_refine_ls_restr.rejects
'p_bond_d'          0.020  0.018  1654 '>2 sigma' 22
'p_angle_d'         0.030  0.038  2246 '>2 sigma' 139
'p_planar_d'        0.040  0.043  498  '>2 sigma' 21
'p_planar'          0.020  0.015  270  '>2 sigma' 1
'p_chiral'          0.150  0.177  278  '>2 sigma' 2
'p_singtor_nbd'     0.500  0.216  582  '>2 sigma' 0
'p_multtor_nbd'     0.500  0.207  419  '>2 sigma' 0
'p_xyhbond_nbd'     0.500  0.245  149  '>2 sigma' 0
'p_planar_tor'      3.0    2.6    203  '>2 sigma' 9
'p_staggered_tor'   15.0   17.4   298  '>2 sigma' 31
'p_orthonormal_tor' 20.0   18.1   12   '>2 sigma' 1

loop_
_refine_ls_shell.d_res_low
_refine_ls_shell.d_res_high
_refine_ls_shell.number_reflns_obs
_refine_ls_shell.R_factor_obs
   8.00   4.51   1226   0.196
   4.51   3.48   1679   0.146
   3.48   2.94   2014   0.160
   2.94   2.59   2147   0.182
   2.59   2.34   2127   0.193
   2.34   2.15   2061   0.203
   2.15   2.00   1647   0.188
```

parameters that lie too far from a target, as these data are often published as part of a description of the refinement and are often deposited with the coordinates in an archive. However, the types of restraints applied depend strongly on the software package used, and as new refinement packages regularly become available, it was clearly not advisable to provide program-specific data items in the mmCIF dictionary. The approach taken in the mmCIF dictionary has been to allow the value of **_refine_ls_restr.type** to be a free-text field, so that arbitrary labels can be given to restraints that are particular to a software package, but to recommend the use of specific labels for restraints applied by particular programs. The dictionary provides examples for labels specific to the programs *PROTIN/PROLSQ* (Hendrickson & Konnert, 1979) and *RESTRAIN* (Driessen *et al.*, 1989). These program-specific representations have particular prefixes; thus the value p_bond_d is a bond-distance restraint as applied by *PROTIN/PROLSQ*. Values for **_refine_ls_restr.type** appropriate for other refinement programs may be suggested in future versions of the mmCIF dictionary.

Data items in the REFINE_LS_RESTR_TYPE category can be used to specify the ranges within which quantities are allowed to vary for each type of restraint. The special value indicated by a full stop (.) represents a restraint unbounded on the high or low side.

Data items in the REFINE_LS_RESTR_NCS category can be used to record details about the restraints applied to atom positions in domains related by noncrystallographic symmetry during least-squares refinement, and also to record the deviation of the restrained atomic parameters at the end of the refinement. The domains related by noncrystallographic symmetry are defined in the STRUCT_NCS_DOM and related categories (see Section 3.6.7.5.5). The quantities that can be recorded for each restrained domain are the root-mean-square deviations of the displacement and positional parameters, and the weighting coefficients used in the noncrystallographic restraint of each type of parameter. Any special aspects of the way the restraints were applied may be described using **_refine_ls_restr_ncs.ncs_model_details**.

Data items in the REFINE_LS_SHELL category are used to summarize details of the results of the least-squares refinement by shells of resolution (Example 3.6.6.9). The resolution range, in ångströms, forms the category key; for each shell the quantities reported, such as the number of reflections above the threshold for counting as significantly intense, are all defined in the same way as the corresponding data items used to describe the results of the overall refinement in the REFINE category.

The core dictionary category REFINE_LS_CLASS was introduced after the release of the first version of the mmCIF dictionary. It provides a more general way of describing the treatment of particular subsets of the observations, but it is not expected to be used in macromolecular structural studies, where partition by shells of resolution is traditional.

3.6.6.2.4. *Equivalent atoms in the refinement*

The data items in these categories are as follows:

(*a*) REFINE_B_ISO

- **_refine_B_iso.class**
  _refine_B_iso.details
  _refine_B_iso.treatment
  _refine_B_iso.value

(*b*) REFINE_OCCUPANCY

- **_refine_occupancy.class**
  _refine_occupancy.details
  _refine_occupancy.treatment
  _refine_occupancy.value

The bullet (•) *indicates a category key.*

In macromolecular structure refinement, displacement factors or occupancies are often treated as equivalent for groups of atoms. An example would be the case where most of the atoms in the structure are refined with isotropic displacement factors, but a bound metal atom is allowed to refine anisotropically. Another example would be where the occupancies for all of the atoms in the protein part of a macromolecular complex are fixed at 1.0, but the occupancies of atoms in a bound inhibitor are refined. The REFINE_B_ISO and REFINE_OCCUPANCY categories can be used to record this information (Example 3.6.6.10).

Example 3.6.6.10. *The handling of displacement factors and occupancies during the refinement of an HIV-1 protease structure (PDB 5HVP) described with data items in the* REFINE_B_ISO *and* REFINE_OCCUPANCY *categories.*

```
loop_
_refine_B_iso.class
_refine_B_iso.treatment
   'protein'    isotropic
   'solvent'    isotropic
   'inhibitor'  isotropic

loop_
_refine_occupancy.class
_refine_occupancy.treatment
_refine_occupancy.value
_refine_occupancy.details
   'protein'                   fix  1.00  .
   'solvent'                   fix  1.00  .
   'inhibitor orientation 1'   fix  0.65  .
   'inhibitor orientation 2'   fix  0.35
; The inhibitor binds to the enzyme in two
  alternative conformations. The occupancy of
  each conformation was adjusted so as to result
  in approximately equal mean thermal factors
  for the atoms in each conformation.
;
```

Example 3.6.6.11. *An example of one cycle of refinement described with data items in the* REFINE_HIST *category.*

```
_refine_hist.cycle_id                C134
_refine_hist.d_res_high              1.85
_refine_hist.d_res_low               20.0
_refine_hist.number_atoms_solvent     217
_refine_hist.number_atoms_total       808
_refine_hist.number_reflns_all       6174
_refine_hist.number_reflns_obs       4886
_refine_hist.number_reflns_R_free     476
_refine_hist.number_reflns_R_work    4410
_refine_hist.R_factor_all            .265
_refine_hist.R_factor_obs            .195
_refine_hist.R_factor_R_free         .274
_refine_hist.R_factor_R_work         .160
_refine_hist.details
; Add majority of solvent molecules. B factors
  refined by group. Continued to remove
  misplaced water molecules.
;
```

Data items in the REFINE_B_ISO category can be used to record details of the treatment of isotropic *B* (displacement) factors during refinement. There is no formal link between the classes identified by `_refine_B_iso.class` and individual atom sites, although relationships may be inferred if the class names are carefully chosen. The category allows the treatment of the atoms in each class (isotropic, anisotropic or fixed) and the value assigned for fixed isotropic *B* factors to be recorded. Any special details can be given in a free-text field.

Data items in the REFINE_OCCUPANCY category can be used to record details of the treatment of occupancies of groups of atom sites during refinement. As with the treatment of displacement factors in the REFINE_B_ISO category, the classes itemized by `_refine_occupancy.class` are not formally linked to the individual atom sites, but the relationships may be deduced if the class names are chosen carefully.

3.6.6.2.5. *History of the refinement*

The data items in this category are as follows:

REFINE_HIST
- • `_refine_hist.cycle_id`
  - `_refine_hist.details`
  - `_refine_hist.d_res_high`
  - `_refine_hist.d_res_low`
  - `_refine_hist.number_atoms_solvent`
  - `_refine_hist.number_atoms_total`
  - `_refine_hist.number_reflns_all`
  - `_refine_hist.number_reflns_obs`
  - `_refine_hist.number_reflns_R_free`
  - `_refine_hist.number_reflns_R_work`
  - `_refine_hist.R_factor_all`
  - `_refine_hist.R_factor_obs`
  - `_refine_hist.R_factor_R_free`
  - `_refine_hist.R_factor_R_work`

The bullet (•) *indicates a category key.*

Data items in the REFINE_HIST category can be used to record details about the various steps in the refinement of the structure. They do not provide as thorough a description of the refinement as can be given in other categories for the final model, but instead allow a summary of the progress of the refinement to be given and supported by a small set of representative statistics.

The category is sufficiently compact that a large number of cycles could be summarized, but it is not expected that every cycle of refinement would be routinely reported. Example 3.6.6.11 shows an entry for a single cycle of refinement. It is likely that an author would present a representative sequence of entries in a looped list.

3.6.6.3. Reflection measurements

The categories describing the reflections used in the refinement are as follows:

REFLN group
- *Individual reflections* (§3.6.6.3.1)
  - REFLN
  - REFLN_SYS_ABS
- *Groups of reflections* (§3.6.6.3.2)
  - REFLNS
  - REFLNS_SCALE
  - REFLNS_SHELL
  - REFLNS_CLASS

Data items in the REFLN category can be used to give information about the individual reflections that were used to derive the final model. The related category REFLN_SYS_ABS allows the reflections that should be systematically absent for the space group in which the structure was refined to be tabulated. Data items in the REFLNS category can be used to record information that applies to all of the reflections. Scale factors can be listed in the REFLNS_SCALE category, while the data items in REFLNS_SHELL can be used to record information about the reflection set by shells of resolution. The core CIF dictionary category REFLNS_CLASS, which can be used for a general classification of reflection groups according to criteria other than resolution shell, is not expected to be used in mmCIF applications.

3.6.6.3.1. *Individual reflections*

The data items in these categories are as follows:

(*a*) REFLN
- • ***_refln.index_h***
- • ***_refln.index_k***
- • ***_refln.index_l***
  - ***_refln.A_calc***
  - `_refln.A_calc_au`
  - ***_refln.A_meas***
  - `_refln.A_meas_au`
  - ***_refln.B_calc***
  - `_refln.B_calc_au`
  - ***_refln.B_meas***
  - `_refln.B_meas_au`
  - ***_refln.class_code***
  - `_refln.crystal_id`
    → `_exptl_crystal.id`
  - ***_refln.d_spacing***
  - ***_refln.F_calc***
  - `_refln.F_calc_au`
  - ***_refln.F_meas***
  - `_refln.F_meas_au`
  - ***_refln.F_meas_sigma*** (∼ ***_refln_F_sigma***)
  - `_refln.F_meas_sigma_au`
  - ***_refln.F_squared_calc***
  - ***_refln.F_squared_meas***
  - ***_refln.F_squared_sigma***
  - `_refln.fom`
  - ***_refln.include_status***
  - ***_refln.intensity_calc***
  - ***_refln.intensity_meas***
  - ***_refln.intensity_sigma***
  - ***_refln.mean_path_length_tbar***
  - ***_refln.phase_calc***
  - ***_refln.phase_meas***
  - ***_refln.refinement_status***
  - `_refln.scale_group_code`
    → `_reflns_scale.group_code`
  - ***_refln.sint_over_lambda*** (∼ ***_refln_sint/lambda***)
  - ***_refln.status*** (∼ ***_refln_observed_status***)
  - ***_refln.symmetry_epsilon***
  - ***_refln.symmetry_multiplicity***
  - ***_refln.wavelength***

Example 3.6.6.12. *Part of the reflection list for an HIV-1 protease structure (PDB 5HVP) described with data items in the* REFLN *category.*

```
loop_
_refln.index_h
_refln.index_k
_refln.index_l
_refln.F_squared_calc
_refln.F_squared_meas
_refln.F_squared_sigma
_refln.status
   2   0   0       85.57       58.90      1.45 o
   3   0   0    15718.18    15631.06     30.40 o
   4   0   0    55613.11    49840.09     61.86 o
   5   0   0      246.85      241.86     10.02 o
   6   0   0       82.16       69.97      1.93 o
   7   0   0     1133.62      947.79     11.78 o
   8   0   0     2558.04     2453.33     20.44 o
   9   0   0      283.88      393.66      7.79 o
  10   0   0      283.70      171.98      4.26 o
```

```
  _refln.wavelength_id
        →  _diffrn_radiation.wavelength_id
```

(*b*) REFLN_SYS_ABS

```
• _refln_sys_abs.index_h
• _refln_sys_abs.index_k
• _refln_sys_abs.index_l
  _refln_sys_abs.I
  _refln_sys_abs.I_over_sigmaI
  _refln_sys_abs.sigmaI
```

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* (_) *except where indicated by the* ∼ *symbol.*

Data items in the REFLN category are used in the same way in the mmCIF and core CIF dictionaries, and Section 3.2.3.2.1 can be consulted for details. However, in macromolecular crystallography it is not usual for reflection intensities to be given in units of electrons (the units specified by the core CIF dictionary). Thus it was necessary to introduce in the mmCIF dictionary data items for the magnitudes of structure factors and their *A* and *B* components in arbitrary units (Example 3.6.6.12). A figure of merit (`_refln.fom`) can also be included for reflections that were phased using experimental methods.

The REFLN_SYS_ABS category allows the intensities of the reflections that should be systematically absent to be tabulated. The ratio of the intensity to its standard uncertainty, given in the data item `_refln_sys_abs.I_over_sigmaI`, can be used to assess whether the reflection is indeed absent. The decision as to whether it is absent is left to the user of the mmCIF and is not recorded in the mmCIF.

3.6.6.3.2. *Groups of reflections*

The data items in these categories are as follows:

(*a*) REFLNS

```
• _reflns.entry_id
        → _entry.id
  _reflns.B_iso_Wilson_estimate
  _reflns.data_reduction_details
  _reflns.data_reduction_method
  _reflns.d_resolution_high
  _reflns.d_resolution_low
  _reflns.details (∼ _reflns_special_details)
  _reflns.Friedel_coverage
  _reflns.limit_h_max
  _reflns.limit_h_min
  _reflns.limit_k_max
  _reflns.limit_k_min
  _reflns.limit_l_max
  _reflns.limit_l_min
  _reflns.number_all (∼ _reflns_number_total)
  _reflns.number_gt
  _reflns.number_obs (∼ _reflns_number_observed)
  _reflns.observed_criterion
  _reflns.observed_criterion_F_max
  _reflns.observed_criterion_F_min
  _reflns.observed_criterion_I_max
  _reflns.observed_criterion_I_min
  _reflns.observed_criterion_sigma_F
  _reflns.observed_criterion_sigma_I
  _reflns.percent_possible_obs
  _reflns.R_free_details
  _reflns.Rmerge_F_all
  _reflns.Rmerge_F_obs
  _reflns.threshold_expression
```

(*b*) REFLNS_SCALE

```
• _reflns_scale.group_code
  _reflns_scale.meas_F
  _reflns_scale.meas_F_squared
  _reflns_scale.meas_intensity
```

(*c*) REFLNS_SHELL

```
• _reflns_shell.d_res_high
• _reflns_shell.d_res_low
  _reflns_shell.meanI_over_sigI_all
  _reflns_shell.meanI_over_sigI_gt
  _reflns_shell.meanI_over_sigI_obs
  _reflns_shell.meanI_over_uI_all
  _reflns_shell.meanI_over_uI_gt
  _reflns_shell.number_measured_all
  _reflns_shell.number_measured_gt
  _reflns_shell.number_measured_obs
  _reflns_shell.number_possible
  _reflns_shell.number_unique_all
  _reflns_shell.number_unique_gt
  _reflns_shell.number_unique_obs
  _reflns_shell.percent_possible_all
  _reflns_shell.percent_possible_gt
  _reflns_shell.percent_possible_obs
  _reflns_shell.Rmerge_F_all
  _reflns_shell.Rmerge_F_gt
  _reflns_shell.Rmerge_F_obs
  _reflns_shell.Rmerge_I_all
  _reflns_shell.Rmerge_I_gt
  _reflns_shell.Rmerge_I_obs
```

(*d*) REFLNS_CLASS

```
• _reflns_class.code
  _reflns_class.d_res_high
  _reflns_class.d_res_low
  _reflns_class.description
  _reflns_class.number_gt
  _reflns_class.number_total
  _reflns_class.R_factor_all
  _reflns_class.R_factor_gt
  _reflns_class.R_Fsqd_factor
  _reflns_class.R_I_factor
  _reflns_class.wR_factor_all
```

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* (_) *except where indicated by the* ∼ *symbol.*

Data items in the REFLNS category of the core CIF dictionary can be used to summarize the properties or attributes of the complete set of reflections used in refinement (Section 3.2.3.2.2). The mmCIF dictionary adds a number of data items to this category, including the formal category key required by the DDL2 data model. There are also data items for describing the data-reduction method and recording any relevant details about data reduction, and for giving an estimate of the overall Wilson *B* factor for the data set.

A number of the new data items relate to the issue of how reflections are flagged as being observed and are thus used in the refinement. In the core CIF dictionary, the criteria used to consider

Example 3.6.6.13. *The data set used in the refinement of an HIV-1 protease structure (PDB 5HVP) described using data items in the* REFLNS *and* REFLNS_SHELL *categories.*

```
_reflns.entry_id                         '5HVP'
_reflns.data_reduction_method
; Xengen program scalei. Anomalous pairs were merged.
  Scaling proceeded in several passes, beginning with
  1-parameter fit and ending with 3-parameter fit.
;
_reflns.data_reduction_details
; Merging and scaling based on only those reflections
  with I > sigma(I).
;

_reflns.d_resolution_high                2.00
_reflns.d_resolution_low                 8.00

_reflns.limit_h_max                      22
_reflns.limit_h_min                      0
_reflns.limit_k_max                      46
_reflns.limit_k_min                      0
_reflns.limit_l_max                      57
_reflns.limit_l_min                      0

_reflns.number_obs                       7228
_reflns.observed_criterion_sigma_I       1.0
_reflns.details                          none

loop_
_reflns_shell.d_res_high
_reflns_shell.d_res_low
_reflns_shell.meanI_over_sigI_obs
_reflns_shell.number_measured_obs
_reflns_shell.number_unique_obs
_reflns_shell.percent_possible_obs
_reflns_shell.Rmerge_F_obs
   31.38  3.82  69.8  9024  2540  96.8   1.98
    3.82  3.03  26.1  7413  2364  95.1   3.85
    3.03  2.65  10.5  5640  2123  86.2   6.37
    2.65  2.41   6.4  4322  1882  76.8   8.01
    2.41  2.23   4.3  3247  1714  70.4   9.86
    2.23  2.10   3.1  1140   812  33.3  13.99
```

a reflection as being observed are given using the data item `_reflns.observed_criterion`. This is a free-text field so is not automatically parsable. Therefore it is supplemented in the mmCIF dictionary by data items that can be used to stipulate the criterion in terms of the values of F, I or the uncertainties in these quantities (Example 3.6.6.13). The percentage of the total number of reflections that meet the criterion can be recorded.

Data items are also provided for describing the selection of the reflections used to calculate the free R factor, and for giving the $R_{\rm merge}$ values for all reflections and for the subset of 'observed' reflections. Data items in the REFLNS_SCALE and REFLNS_SHELL categories are used in the same way in the mmCIF and core CIF dictionaries, and Section 3.2.3.2.2 can be consulted for details.

As with the related categories DIFFRN_REFLNS_CLASS and REFINE_LS_CLASS, the core dictionary category REFLNS_CLASS was introduced after the release of the first version of the mmCIF dictionary. It provides a more general way of describing the treatment of particular subsets of the observations, but it is not expected to be used in macromolecular structural studies, where partition by shells of resolution is traditional.

3.6.7. Atomicity, chemistry and structure

The basic concepts of the mmCIF model for describing a macromolecular structure were outlined in Section 3.6.3. The present section describes the components of the model in more detail. The category groups used to describe the molecular chemistry and structure are: the ATOM group describing atom positions (Section 3.6.7.1); the CHEMICAL, CHEM_COMP and CHEM_LINK groups describing molecular chemistry (Section 3.6.7.2); the ENTITY group describing distinct chemical species (Section 3.6.7.3); the GEOM group describing molecular or packing geometry (Section 3.6.7.4); the STRUCT group describing the large-scale features of molecular structure (Section 3.6.7.5); and the SYMMETRY group describing the symmetry and space group (Section 3.6.7.6).

The CHEMICAL category group itself is not generally used in an mmCIF. The purpose of this category group in the core CIF dictionary is to specify the chemical identity and connectivity of the relatively simple molecular or ionic species in a small-molecule or inorganic crystal. In principle, a macromolecular structure determined to atomic resolution could be represented as a coherent chemical entity with a complete connectivity graph. However, in practice, biological macromolecules are built from units from a library of models of standard amino acids, nucleotides and sugars. Data items in the CHEM_COMP and CHEM_LINK category groups of the mmCIF dictionary describe the internal connectivity and standard bonding processes between these units.

Molecular or packing geometry is also rarely tabulated for large macromolecular complexes, so the GEOM category group is rarely used in an mmCIF.

3.6.7.1. Atom sites

The categories describing atom sites are as follows:

ATOM group
Individual atom sites (§3.6.7.1.1)
  ATOM_SITE
  ATOM_SITE_ANISOTROP
Collections of atom sites (§3.6.7.1.2)
  ATOM_SITES
  ATOM_SITES_FOOTNOTE
Atom types (§3.6.7.1.3)
  ATOM_TYPE
Alternative conformations (§3.6.7.1.4)
  ATOM_SITES_ALT
  ATOM_SITES_ALT_ENS
  ATOM_SITES_ALT_GEN

The ATOM category group represents a compromise between the representation of a small-molecule structure as an annotated list of atomic coordinates and the need in macromolecular crystallography to present a more structured view organized around residues, chains, sheets, turns, helices *etc.* The locations of individual atoms and other information about the atom sites are given using data items in this category group. The categories within the group may be classified as shown in the summary above.

The ATOM_SITE, ATOM_SITES and ATOM_TYPE categories have many data items that are aliases of equivalent data items in the same categories in the core CIF dictionary, but the conventions for the labelling of the atom sites are different.

The ATOM_SITE_ANISOTROP and ATOM_SITES_FOOTNOTE categories are new to the mmCIF dictionary, as are the categories related to alternative conformations: ATOM_SITES_ALT, ATOM_SITES_ALT_ENS and ATOM_SITES_ALT_GEN.

3.6.7.1.1. *Individual atom sites*

The data items in these categories are as follows:

(*a*) ATOM_SITE

• **_atom_site.id** (∼ ***_atom_site_label***)
  _atom_site.adp_type
\+ **_atom_site.aniso_B[1][1]**
  ⇌ ***_atom_site_anisotrop.B[1][1]***

+ _atom_site.aniso_B[1][2]
  ⇌ _atom_site_anisotrop.B[1][2]
+ _atom_site.aniso_B[1][3]
  ⇌ _atom_site_anisotrop.B[1][3]
+ _atom_site.aniso_B[2][2]
  ⇌ _atom_site_anisotrop.B[2][2]
+ _atom_site.aniso_B[2][3]
  ⇌ _atom_site_anisotrop.B[2][3]
+ _atom_site.aniso_B[3][3]
  ⇌ _atom_site_anisotrop.B[3][3]
  _atom_site.aniso_ratio
  ⇌ _atom_site_anisotrop.ratio
+ _atom_site.aniso_U[1][1]
  ⇌ _atom_site_anisotrop.U[1][1]
+ _atom_site.aniso_U[1][2]
  ⇌ _atom_site_anisotrop.U[1][2]
+ _atom_site.aniso_U[1][3]
  ⇌ _atom_site_anisotrop.U[1][3]
+ _atom_site.aniso_U[2][2]
  ⇌ _atom_site_anisotrop.U[2][2]
+ _atom_site.aniso_U[2][3]
  ⇌ _atom_site_anisotrop.U[2][3]
+ _atom_site.aniso_U[3][3]
  ⇌ _atom_site_anisotrop.U[3][3]
  _atom_site.attached_hydrogens
  _atom_site.auth_asym_id
  _atom_site.auth_atom_id
  _atom_site.auth_comp_id
  _atom_site.auth_seq_id
+ _atom_site.B_equiv_geom_mean
+ _atom_site.B_iso_or_equiv
  _atom_site.calc_attached_atom
  _atom_site.calc_flag
+ _atom_site.Cartn_x
+ _atom_site.Cartn_y
+ _atom_site.Cartn_z
  _atom_site.chemical_conn_number
  → _chemical_conn_atom.number
  _atom_site.constraints
  _atom_site.details (∼ _atom_site_description)
  _atom_site.disorder_assembly
  _atom_site.disorder_group
  _atom_site.footnote_id
+ _atom_site.fract_x
+ _atom_site.fract_y
+ _atom_site.fract_z
  _atom_site.group_PDB
  _atom_site.label_alt_id
  → _atom_sites_alt.id
  _atom_site.label_asym_id
  → _struct_asym.id
  _atom_site.label_atom_id
  → _chem_comp_atom.atom_id
  _atom_site.label_comp_id
  → _chem_comp.id
  _atom_site.label_entity_id
  → _entity.id
  _atom_site.label_seq_id
  → _entity_poly_seq.num
+ _atom_site.occupancy
  _atom_site.refinement_flags
  _atom_site.refinement_flags_adp
  _atom_site.refinement_flags_occupancy
  _atom_site.refinement_flags_posn
  _atom_site.restraints
  _atom_site.symmetry_multiplicity
  _atom_site.thermal_displace_type
  _atom_site.type_symbol
  → _atom_type.symbol
+ _atom_site.U_equiv_geom_mean
+ _atom_site.U_iso_or_equiv
  _atom_site.Wyckoff_symbol

(*b*) ATOM_SITE_ANISOTROP

• _atom_site_anisotrop.id
+ _atom_site_anisotrop.B[1][1] (∼ _atom_site_aniso_B_11)
+ _atom_site_anisotrop.B[1][2] (∼ _atom_site_aniso_B_12)
+ _atom_site_anisotrop.B[1][3] (∼ _atom_site_aniso_B_13)
+ _atom_site_anisotrop.B[2][2] (∼ _atom_site_aniso_B_22)
+ _atom_site_anisotrop.B[2][3] (∼ _atom_site_aniso_B_23)
+ _atom_site_anisotrop.B[3][3] (∼ _atom_site_aniso_B_33)
  _atom_site_anisotrop.ratio (∼ _atom_site_aniso_ratio)
  → _atom_site.id
  _atom_site_anisotrop.type_symbol
  (∼ _atom_site_aniso_type_symbol)
  → _atom_type.symbol
+ _atom_site_anisotrop.U[1][1] (∼ _atom_site_aniso_U_11)
+ _atom_site_anisotrop.U[1][2] (∼ _atom_site_aniso_U_12)
+ _atom_site_anisotrop.U[1][3] (∼ _atom_site_aniso_U_13)
+ _atom_site_anisotrop.U[2][2] (∼ _atom_site_aniso_U_22)
+ _atom_site_anisotrop.U[2][3] (∼ _atom_site_aniso_U_23)
+ _atom_site_anisotrop.U[3][3] (∼ _atom_site_aniso_U_33)

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 (_) except where indicated by the ∼ symbol. Data items marked with a plus (+) have companion data names for the standard uncertainty in the reported value, formed by appending the string _esd to the data name listed. The double arrow (⇌) indicates alternative names in a distinct category.

The refined coordinates of the atoms in the crystallographic asymmetric unit are stored in the ATOM_SITE category. Atom positions and their associated uncertainties may be given using either Cartesian or fractional coordinates, and anisotropic displacement factors and occupancies may be given for each position.

The relationships between categories describing atom sites are shown in Fig. 3.6.7.1.

Several of the mmCIF data names arise from the need to associate atom sites with residues and chains. As in the core CIF dictionary, the identifier for the atom site is the data item **_atom_site_label**. To accommodate standard practice in macromolecular crystallography, the mmCIF atom identifier is the aggregate of **_atom_site.label_alt_id**, ***.label_asym_id**, ***.label_atom_id**, ***.label_comp_id** and ***.label_seq_id**. For the two types of files to be compatible, the data item **_atom_site.id**, which is independent of the different modes of identifying atoms (discussed below), was introduced. The mmCIF identifier **_atom_site.id** is aliased to the core CIF identifier **_atom_site_label**.

Since the identifier does not need to be a number, it is quite possible (although it is not recommended) to use a complex label with an internal structure corresponding to the label components that the mmCIF dictionary provides as separate data items. This scheme is described in Section 3.2.4.1.1. However, normal practice in mmCIFs should be to label sites with the functional components available and to assign a simple numeric sequence to the values of **_atom_site.id** (see Example 3.6.7.1).

In addition to labelling information, each entry in the ATOM_SITE list must contain a value for the data item **_atom_site.type_symbol**, which is a pointer to the table of element symbols in the ATOM_TYPE category. All other data items in the ATOM_SITE category are optional, but it is normal practice to

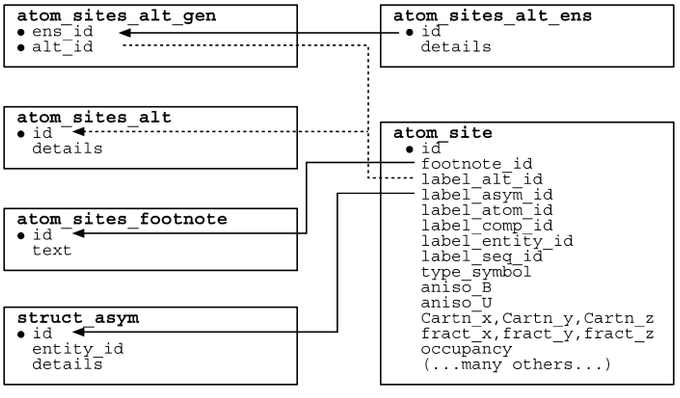

Fig. 3.6.7.1. The family of categories used to describe atom sites. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

Example 3.6.7.1. *Part of the coordinate list for an HIV-1 protease structure (PDB 5HVP) described with data items in the* ATOM_SITE *category. Atoms are given for both polymer and non-polymer regions of the structure, and atoms in the side chain of residue 12 adopt alternative conformations.*

```
loop_
_atom_site.group_PDB
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.label_seq_id
_atom_site.label_alt_id
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.footnote_id
_atom_site.auth_seq_id
_atom_site.id
 ATOM N N   THR  A  12  .  26.095  32.930  14.590
      1.00  18.97  4  12 8
 ATOM C CA  THR  A  12  .  25.734  32.995  16.032
      1.00  19.80  4  12 9
 ATOM C C   THR  A  12  .  24.695  34.106  16.113
      1.00  20.92  4  12 10
 ATOM O O   THR  A  12  .  24.869  35.118  15.421
      1.00  21.84  4  12 11
 ATOM C CB  THR  A  12  .  26.911  33.346  17.018
      1.00  20.51  4  12 12
 ATOM O OG1 THR  A  12  3  27.946  33.921  16.183
      0.50  20.29  4  12 13
 ATOM O OG1 THR  A  12  4  27.769  32.142  17.103
      0.50  20.59  4  12 14
 ATOM C CG2 THR  A  12  3  27.418  32.181  17.878
      0.50  20.47  4  12 15
 ATOM C CG2 THR  A  12  4  26.489  33.778  18.426
      0.50  20.00  4  12 16
# - - - abbreviated - - -
 HETATM C C1 APS  C  .  1  4.171  29.012   7.116
      0.58  17.27  1 300 101
 HETATM C C2 APS  C  .  1  4.949  27.758   6.793
      0.58  16.95  1 300 102
 HETATM O O3 APS  C  .  1  4.800  26.678   7.393
      0.58  16.85  1 300 103
 HETATM N N4 APS  C  .  1  5.930  27.841   5.869
      0.58  16.43  1 300 104
# - - - abbreviated - - -
```

give either the Cartesian or fractional coordinates. Most macromolecular structures use Cartesian coordinates. Isotropic displacement factors are normally placed directly in the ATOM_SITE category, using **_atom_site.B_iso_or_equiv**. Anisotropic displacement factors may be placed directly in the ATOM_SITE category *or* in the ATOM_SITE_ANISOTROP category. *U*'s may be used instead of *B*'s. It is not acceptable to use both *U*'s and *B*'s, nor is it acceptable to have anisotropic displacement factors in both the ATOM_SITE category and the ATOM_SITE_ANISOTROP category.

Each atom within each chemical component is uniquely identified using the data item **_atom_site.label_atom_id**, which is a reference to the data item **_chem_comp_atom.atom_id** in the CHEM_COMP_ATOM category.

The specific object in the asymmetric unit to which the atom belongs is indicated using the data item **_atom_site.label_asym_id**, which is a reference to the data item **_struct_asym.id** in the STRUCT_ASYM category. For macromolecules, it is useful to think of this identifier as a chain ID.

The chemical component to which the atom belongs is indicated using the data item **_atom_site.label_comp_id**, which is a reference to the data item **_chem_comp.id** in the CHEM_COMP category. The chemical component that is referenced in this way may be either a non-polymer or a monomer in a polymer; if it is a monomer in a polymer, it is useful to think of this identifier as the residue name.

The correspondence between the sequence of an entity in a polymer and the sequence information in the coordinate list (and in the STRUCT categories) is established using the data item **_atom_site.label_seq_id**, which is a reference to the data item **_entity_poly_seq.num** in the ENTITY_POLY_SEQ category. This identifier has no meaning for entities that are not part of a polymer; in a polymer it is useful to think of this identifier as the residue number. Note that this is strictly a number. If the combination of a number with an insertion code is needed, **_atom_site.auth_seq_id** should be used (see below).

An alternative set of identifiers can be used for the ***_asym_id**, ***_atom_id**, ***_comp_id** and ***_seq_id** identifiers, but not for ***_alt_id**. The **_atom_site.label_*** data names are standard; there are rules for these identifiers such as the requirement that residue numbers are sequential integers. Different databases may also have their own rules. However, the author of an mmCIF may wish to use a nonstandard labelling scheme, *e.g.* to reflect the residue numbering scheme of a structure to which the present structure is homologous, apart from insertions and gaps. Another situation in which a nonstandard labelling scheme might be used is to follow a local convention for atom names in a non-polymer, such as a haem, that conflicts with the scheme required by a database in which the structure is to be deposited. In these situations, alternative identifiers can be given using the data names (**_atom_site.auth_***).

In regions of the structure with alternative conformations, the specific conformation to which an atom belongs can be indicated using the data item **_atom_site.label_alt_id**, which is a reference to the data item **_atom_sites_alt.id** in the ATOM_SITES_ALT category.

The chemically distinct part of the structure (*e.g.* polymer chain, ligand, solvent) to which an atom belongs can be indicated using the data item **_atom_site.label_entity_id**, which is a reference to the data item **_entity.id** in the ENTITY category.

Most of the information that needs to be associated with an atom site is conveyed by the values of specific data names in mmCIF. However, for historical reasons, a pointer to additional free-text information about an atom site or about a group of atom sites can be given using the data item **_atom_site.footnote_id**, which is a reference to the data item **_atom_sites_footnote.id** in the ATOM_SITES_FOOTNOTE category.

The data item **_atom_site.group_PDB** is a place holder for the tags used by the PDB to identify types of coordinate records. It allows interconversion between mmCIFs and PDB format files. The only permitted values are ATOM and HETATM.

As in the core CIF dictionary, anisotropic displacement parameters in an mmCIF can be given in the same list as the atom positions and occupancies, or can be given in a separate list. However, DDL2 does not permit the same data names to be used for both constructs. Therefore, in mmCIF, anisotropic displacement parameters presented in a separate list are handled in a separate category with its own key, **_atom_site_anisotrop.id**, which must match a corresponding label in the atom-site list, **_atom_site.id**.

The individual elements of the anisotropic displacement matrix are labelled slightly differently in the mmCIF dictionary than in the core CIF dictionary in order to emphasize their matrix character. However, the definitions of the corresponding data items are identical in the two dictionaries.

3.6.7.1.2. *Collections of atom sites*

The data items in these categories are as follows:

(*a*) ATOM_SITES

- **_atom_sites.entry_id**
  → **_entry.id**
  _atom_sites.Cartn_transf_matrix[1][1]
  (∼ **_atom_sites_Cartn_tran_matrix_11**)
  _atom_sites.Cartn_transf_matrix[1][2]
  (∼ **_atom_sites_Cartn_tran_matrix_12**)
  _atom_sites.Cartn_transf_matrix[1][3]
  (∼ **_atom_sites_Cartn_tran_matrix_13**)
  _atom_sites.Cartn_transf_matrix[2][1]
  (∼ **_atom_sites_Cartn_tran_matrix_21**)
  _atom_sites.Cartn_transf_matrix[2][2]
  (∼ **_atom_sites_Cartn_tran_matrix_22**)
  _atom_sites.Cartn_transf_matrix[2][3]
  (∼ **_atom_sites_Cartn_tran_matrix_23**)
  _atom_sites.Cartn_transf_matrix[3][1]
  (∼ **_atom_sites_Cartn_tran_matrix_31**)
  _atom_sites.Cartn_transf_matrix[3][2]
  (∼ **_atom_sites_Cartn_tran_matrix_32**)
  _atom_sites.Cartn_transf_matrix[3][3]
  (∼ **_atom_sites_Cartn_tran_matrix_33**)
  _atom_sites.Cartn_transf_vector[1]
  (∼ **_atom_sites_Cartn_tran_vector_1**)
  _atom_sites.Cartn_transf_vector[2]
  (∼ **_atom_sites_Cartn_tran_vector_2**)
  _atom_sites.Cartn_transf_vector[3]
  (∼ **_atom_sites_Cartn_tran_vector_3**)
  _atom_sites.Cartn_transform_axes
  _atom_sites.fract_transf_matrix[1][1]
  (∼ **_atom_sites_fract_tran_matrix_11**)
  _atom_sites.fract_transf_matrix[1][2]
  (∼ **_atom_sites_fract_tran_matrix_12**)
  _atom_sites.fract_transf_matrix[1][3]
  (∼ **_atom_sites_fract_tran_matrix_13**)
  _atom_sites.fract_transf_matrix[2][1]
  (∼ **_atom_sites_fract_tran_matrix_21**)
  _atom_sites.fract_transf_matrix[2][2]
  (∼ **_atom_sites_fract_tran_matrix_22**)
  _atom_sites.fract_transf_matrix[2][3]
  (∼ **_atom_sites_fract_tran_matrix_23**)
  _atom_sites.fract_transf_matrix[3][1]
  (∼ **_atom_sites_fract_tran_matrix_31**)
  _atom_sites.fract_transf_matrix[3][2]
  (∼ **_atom_sites_fract_tran_matrix_32**)
  _atom_sites.fract_transf_matrix[3][3]
  (∼ **_atom_sites_fract_tran_matrix_33**)
  _atom_sites.fract_transf_vector[1]
  (∼ **_atom_sites_fract_tran_vector_1**)
  _atom_sites.fract_transf_vector[2]
  (∼ **_atom_sites_fract_tran_vector_2**)
  _atom_sites.fract_transf_vector[3]
  (∼ **_atom_sites_fract_tran_vector_3**)
  _atom_sites.solution_hydrogens
  _atom_sites.solution_primary
  _atom_sites.solution_secondary
  _atom_sites.special_details

(*b*) ATOM_SITES_FOOTNOTE

- **_atom_sites_footnote.id**
  _atom_sites_footnote.text

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* (_) *except where indicated by the* ∼ *symbol.*

The ATOM_SITES category of the core dictionary, which is used to record information that applies collectively to all the atom sites in the model of the structure, is incorporated without change into the mmCIF dictionary, and Section 3.2.4.1.2 can be consulted for details.

In practice, the data names in the PHASING categories are preferred to the aliases to the core CIF data items **_atom_sites.solution_primary**, ***_secondary** and ***_hydrogens**. The data items in the mmCIF PHASING categories are designed to allow a much more detailed description of how a macromolecular structure was solved.

Example 3.6.7.2. *Footnotes for particular groups of atom sites in an HIV-1 protease structure (PDB 5HVP) using data items in the ATOM_SITES_FOOTNOTE category.*

```
loop_
  _atom_sites_footnote.id
  _atom_sites_footnote.text
     3
; The positions of these water molecules correlate
  with the alternative orientations of the
  inhibitor. Water molecules with alternative ID
  "1" and occupancy 0.58 correlate with
  inhibitor orientation "1". Water molecules with
  alternative ID "2" and occupancy 0.42 correlate
  with inhibitor orientation "2".
;
     4
; Side chains of these residues adopt alternative
  orientations that do not correlate with the
  alternative orientation of the inhibitor.
;
```

The data item **_atom_sites.entry_id** has been added to the ATOM_SITES category to provide the formal category key required by the DDL2 data model.

The ATOM_SITES_FOOTNOTE category can be used to note something about a group of sites in the ATOM_SITE coordinate list, each of which is flagged with the same value of **_atom_site.footnote_id**. For example, an author may wish to note atoms for which the electron density is very weak, or atoms for which static disorder has been modelled. Example 3.6.7.2 shows how an author has used these data items to describe alternative orientations in part of a structure. However, the very large number of data names describing specific structural characteristics in the mmCIF dictionary mean that these rather general data names are rarely needed.

3.6.7.1.3. *Atom types*

The data items in this category are as follows:

ATOM_TYPE

- **_atom_type.symbol**
  _atom_type.analytical_mass_percent
  (∼ **_atom_type_analytical_mass_%**)
  _atom_type.description
  _atom_type.number_in_cell
  _atom_type.oxidation_number
  _atom_type.radius_bond
  _atom_type.radius_contact
  _atom_type.scat_Cromer_Mann_a1
  _atom_type.scat_Cromer_Mann_a2
  _atom_type.scat_Cromer_Mann_a3
  _atom_type.scat_Cromer_Mann_a4
  _atom_type.scat_Cromer_Mann_b1
  _atom_type.scat_Cromer_Mann_b2
  _atom_type.scat_Cromer_Mann_b3
  _atom_type.scat_Cromer_Mann_b4
  _atom_type.scat_Cromer_Mann_c
  _atom_type.scat_dispersion_imag
  _atom_type.scat_dispersion_real
  _atom_type.scat_dispersion_source
  _atom_type.scat_length_neutron
  _atom_type.scat_source
  _atom_type.scat_versus_stol_list

The bullet (•) *indicates a category key. Items in italics have aliases in the core CIF dictionary formed by changing the full stop* (**.**) *to an underscore* (_) *except where indicated by the* ∼ *symbol.*

The ATOM_TYPE category, which provides information about the atomic species associated with each atom site in the model of the structure, is used in the same way in the mmCIF dictionary as in the core CIF dictionary. See Section 3.2.4.1.3 for details.

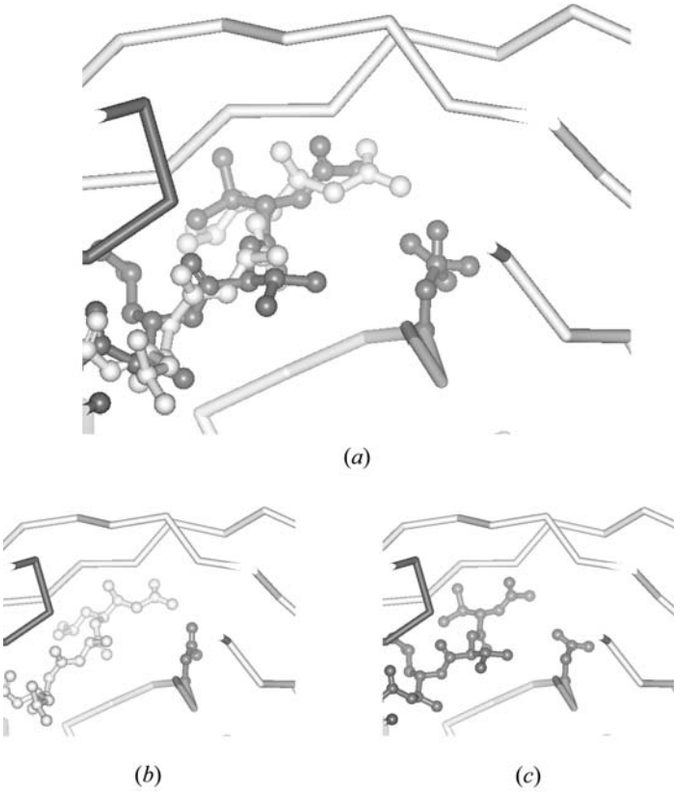

Fig. 3.6.7.2. Alternative conformations in an HIV-1 protease structure (PDB 5HVP) to be described with data items in the ATOM_SITES_ALT, ATOM_SITES_ALT_ENS and ATOM_SITES_ALT_GEN categories. (*a*) Complete structure, (*b*) ensemble 1, (*c*) ensemble 2.

3.6.7.1.4. *Alternative conformations*

The data items in these categories are as follows:

(*a*) ATOM_SITES_ALT

- **_atom_sites_alt.id**
  _atom_sites_alt.details

(*b*) ATOM_SITES_ALT_ENS

- **_atom_sites_alt_ens.id**
  _atom_sites_alt_ens.details

(*c*) ATOM_SITES_ALT_GEN

- **_atom_sites_alt_gen.alt_id**
  → **_atom_sites_alt.id**
- **_atom_sites_alt_gen.ens_id**
  → **_atom_sites_alt_gen.ens_id**

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.

Biological macromolecules are often very flexible, and as the resolution of a structure determination increases, it becomes increasingly possible to model reliably the alternative conformations that the structure adopts. Typically, partial occupancies are assigned to atom sites within the alternative conformations to indicate the relative frequency of occurrence of each conformation. It can, however, be difficult to deduce the possible different conformations of the whole structure from inspection of the atom-site occupancies alone. For instance, a segment of protein main chain might adopt one of three slightly different conformations, and within each conformation a particular side chain might adopt one of two possible conformations, one of which sterically distorts an adjacent residue sequence, while the other does not. The data model in the mmCIF dictionary allows these kinds of correlations in positions to be described.

The relationships between the categories used to describe alternative conformations are shown in Fig. 3.6.7.1.

In the core CIF dictionary, alternative conformations are indicated by using the **_atom_site.disorder_assembly** and ***.disorder_group** data items. Aliases to these data items are present in the mmCIF dictionary, but it is not intended that they should be used to describe disorder in a macromolecular structure.

The model for describing alternative conformations in mmCIF uses the ATOM_SITES_ALT family of categories. Ensembles of correlated alternative conformations can be identified using the category ATOM_SITES_ALT_ENS. Each ensemble is generated from one or more of the alternative conformations given in the list of alternative sites in the ATOM_SITES_ALT category. Data items in the

Example 3.6.7.3. *Alternative conformations in an HIV-1 protease structure (PDB 5HVP) described with data items in the ATOM_SITES_ALT, ATOM_SITES_ALT_ENS and ATOM_SITES_ALT_GEN categories.*

```
loop_
_atom_sites_alt.id
_atom_sites_alt.details
  .
; Atom sites with the alternative ID set to null are
  not modelled in alternative conformations
;
  1
; Atom sites with the alternative ID set to 1 have
  been modelled in alternative conformations with
  respect to atom sites marked with alternative
  ID 2. The conformations of amino-acid side chains
  with alternative ID set to 1 correlate with the
  conformation of the inhibitor marked with
  alternative ID 1. Atoms in these side chains have
  been given an occupancy of 0.58 to match the
  occupancy assigned to the inhibitor.
;
  2
; Atom sites with the alternative ID set to 2 have
  been modelled in alternative conformations with
  respect to atom sites marked with alternative
  ID 1. The conformations of amino-acid side chains
  with alternative ID set to 2 correlate with the
  conformation of the inhibitor marked with
  alternative ID 2. Atoms in these side chains have
  been given an occupancy of 0.42 to match the
  occupancy assigned to the inhibitor.
;

loop_
_atom_sites_alt_ens.id
_atom_sites_alt_ens.details
  'Ensemble 1'
; The inhibitor binds to the enzyme in two, roughly
  twofold symmetric, alternative conformations.

  This conformational ensemble includes the more-
  populated conformation of the inhibitor (ID=1) and
  the amino-acid side chains that correlate with this
  inhibitor conformation.
;
  'Ensemble 2'
; The inhibitor binds to the enzyme in two, roughly
  twofold symmetric, alternative conformations.

  This conformational ensemble includes the less-
  populated conformation of the inhibitor (ID=2) and
  the amino-acid side chains that correlate with this
  inhibitor conformation.
;

loop_
_atom_sites_alt_gen.ens_id
_atom_sites_alt_gen.alt_id
   'Ensemble 1'  .
   'Ensemble 1'  1
   'Ensemble 2'  .
   'Ensemble 2'  2
```

ATOM_SITES_ALT_GEN category explicitly tie together the alternative conformations that contribute to each ensemble. Finally, the atoms in each alternative conformation are identified in the ATOM_SITE category by the data item **_atom_site.label_alt_id**.

The current version of the mmCIF dictionary cannot be used to describe an NMR structure determination completely. However, an mmCIF can be used to store the multiple models usually used to describe a structure determined by NMR using the data items in these categories.

Example 3.6.7.3 is a simplified version of the example given in the mmCIF dictionary (see Fig. 3.6.7.2).

3.6.7.2. Molecular chemistry

The categories describing molecular chemistry are as follows:

Molecular chemistry in the core CIF dictionary (§3.6.7.2.1)

CHEMICAL group
- CHEMICAL
- CHEMICAL_CONN_ATOM
- CHEMICAL_CONN_BOND
- CHEMICAL_FORMULA

Chemical components (§3.6.7.2.2)

CHEM_COMP group
- CHEM_COMP
- CHEM_COMP_ANGLE
- CHEM_COMP_ATOM
- CHEM_COMP_BOND
- CHEM_COMP_CHIR
- CHEM_COMP_CHIR_ATOM
- CHEM_COMP_PLANE
- CHEM_COMP_PLANE_ATOM
- CHEM_COMP_TOR
- CHEM_COMP_TOR_VALUE

Chemical links (§3.6.7.2.3)

CHEM_LINK group
- CHEM_COMP_LINK
- CHEM_LINK
- CHEM_LINK_ANGLE
- CHEM_LINK_BOND
- CHEM_LINK_CHIR
- CHEM_LINK_CHIR_ATOM
- CHEM_LINK_PLANE
- CHEM_LINK_PLANE_ATOM
- CHEM_LINK_TOR
- CHEM_LINK_TOR_VALUE
- ENTITY_LINK

The detailed chemistry of the components of a macromolecular structure can be described using data items in the CHEM_COMP and CHEM_LINK category groups. These mmCIF categories are used in preference to those in the CHEMICAL category group in the core CIF dictionary, as macromolecules are in most cases linked assemblies of a limited number of monomers and so they are most efficiently described by defining the monomers and the links between them, rather than by a formal definition of every bond and angle.

All the categories relevant to molecular chemistry are listed in the summary above; note in particular the presence of the category ENTITY_LINK within the formal CHEM_LINK category group.

3.6.7.2.1. *Molecular chemistry in the core CIF dictionary*

The data items in these categories are as follows:

(*a*) CHEMICAL

• **_chemical.entry_id**
  → **_entry.id**
  _chemical.absolute_configuration
  _chemical.compound_source
  _chemical.melting_point
  _chemical.melting_point_gt
  _chemical.melting_point_lt
  _chemical.name_common
  _chemical.name_mineral
  _chemical.name_structure_type
  _chemical.name_systematic
  _chemical.optical_rotation
  _chemical.properties_biological
  _chemical.properties_physical
+ ***_chemical.temperature_decomposition***
  _chemical.temperature_decomposition_gt
  _chemical.temperature_decomposition_lt
+ ***_chemical.temperature_sublimation***
  _chemical.temperature_sublimation_gt
  _chemical.temperature_sublimation_lt

(*b*) CHEMICAL_CONN_ATOM

• ***_chemical_conn_atom.number***
  _chemical_conn_atom.charge
  _chemical_conn_atom.display_x
  _chemical_conn_atom.display_y
  _chemical_conn_atom.NCA
  _chemical_conn_atom.NH
  _chemical_conn_atom.type_symbol

(*c*) CHEMICAL_CONN_BOND

• ***_chemical_conn_bond.atom_1***
• ***_chemical_conn_bond.atom_2***
  _chemical_conn_bond.type

(*d*) CHEMICAL_FORMULA

• **_chemical_formula.entry_id**
  → **_entry.id**
  _chemical_formula.analytical
  _chemical_formula.iupac
  _chemical_formula.moiety
  _chemical_formula.structural
  _chemical_formula.sum
  _chemical_formula.weight
  _chemical_formula.weight_meas

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 (_). Data items marked with a plus (+) have companion data names for the standard uncertainty in the reported value, formed by appending the string ***_esd*** *to the data name listed.*

Descriptions of molecular chemistry in an mmCIF are normally made using data items in the CHEM_COMP and CHEM_LINK category groups. The CHEMICAL category group is retained in the mmCIF dictionary solely for consistency with the core CIF dictionary and Section 3.2.4.2 may be consulted for details.

Two of the categories in this group, CHEMICAL_CONN_ATOM and CHEMICAL_CONN_BOND, have existing category keys in the core dictionary. The formal keys **_chemical.entry_id** and **_chemical_formula.entry_id** have been added to CHEMICAL and CHEMICAL_FORMULA, respectively, to provide the category keys required by the DDL2 data model.

It is emphasized that these items will not appear in the description of a macromolecular structure, but they are retained to allow the representation of small-molecule or inorganic structures in the DDL2 formalism of mmCIF.

3.6.7.2.2. *Chemical components*

Data items in these categories are as follows:

(*a*) CHEM_COMP

• **_chem_comp.id**
  _chem_comp.formula
  _chem_comp.formula_weight
  _chem_comp.model_details
  _chem_comp.model_erf

```
  _chem_comp.model_source
  _chem_comp.mon_nstd_class
  _chem_comp.mon_nstd_details
  _chem_comp.mon_nstd_flag
  _chem_comp.mon_nstd_parent
  _chem_comp.mon_nstd_parent_comp_id
        → _chem_comp.id
  _chem_comp.name
  _chem_comp.number_atoms_all
  _chem_comp.number_atoms_nh
  _chem_comp.one_letter_code
  _chem_comp.three_letter_code
  _chem_comp.type
```

(*b*) CHEM_COMP_ANGLE

```
• _chem_comp_angle.atom_id_1
        → _chem_comp_atom.atom_id
• _chem_comp_angle.atom_id_2
        → _chem_comp_atom.atom_id
• _chem_comp_angle.atom_id_3
        → _chem_comp_atom.atom_id
• _chem_comp_angle.comp_id
        → _chem_comp.id
+ _chem_comp_angle.value_angle
+ _chem_comp_angle.value_dist
```

(*c*) CHEM_COMP_ATOM

```
• _chem_comp_atom.atom_id
• _chem_comp_atom.comp_id
        → _chem_comp.id
  _chem_comp_atom.alt_atom_id
  _chem_comp_atom.charge
+ _chem_comp_atom.model_Cartn_x
+ _chem_comp_atom.model_Cartn_y
+ _chem_comp_atom.model_Cartn_z
  _chem_comp_atom.partial_charge
  _chem_comp_atom.substruct_code
  _chem_comp_atom.type_symbol
        → _atom_type.symbol
```

(*d*) CHEM_COMP_BOND

```
• _chem_comp_bond.atom_id_1
        → _chem_comp_atom.atom_id
• _chem_comp_bond.atom_id_2
        → _chem_comp_atom.atom_id
• _chem_comp_bond.comp_id
        → _chem_comp.id
  _chem_comp_bond.value_order
+ _chem_comp_bond.value_dist
```

(*e*) CHEM_COMP_CHIR

```
• _chem_comp_chir.id
• _chem_comp_chir.comp_id
  _chem_comp_chir.atom_id
        → _chem_comp_atom.atom_id
  _chem_comp_chir.atom_config
        → _chem_comp.id
  _chem_comp_chir.number_atoms_all
  _chem_comp_chir.number_atoms_nh
  _chem_comp_chir.volume_flag
+ _chem_comp_chir.volume_three
```

(*f*) CHEM_COMP_CHIR_ATOM

```
• _chem_comp_chir_atom.atom_id
        → _chem_comp_atom.atom_id
• _chem_comp_chir_atom.chir_id
        → _chem_comp_chir.id
• _chem_comp_chir_atom.comp_id
        → _chem_comp.id
  _chem_comp_chir_atom.dev
```

(*g*) CHEM_COMP_LINK

```
• _chem_comp_link.link_id
        → _chem_link.id
  _chem_comp_link.details
  _chem_comp_link.type_comp_1
        → _chem_comp.type
  _chem_comp_link.type_comp_2
        → _chem_comp.type
```

(*h*) CHEM_COMP_PLANE

```
• _chem_comp_plane.id
• _chem_comp_plane.comp_id
        → _chem_comp.id
  _chem_comp_plane.number_atoms_all
  _chem_comp_plane.number_atoms_nh
```

(*i*) CHEM_COMP_PLANE_ATOM

```
• _chem_comp_plane_atom.atom_id
        → _chem_comp_atom.atom_id
• _chem_comp_plane_atom.comp_id
        → _chem_comp.id
• _chem_comp_plane_atom.plane_id
        → _chem_comp_plane.id
+ _chem_comp_plane_atom.dist
```

(*j*) CHEM_COMP_TOR

```
• _chem_comp_tor.id
• _chem_comp_tor.comp_id
        → _chem_comp.id
  _chem_comp_tor.atom_id_1
        → _chem_comp_atom.atom_id
  _chem_comp_tor.atom_id_2
        → _chem_comp_atom.atom_id
  _chem_comp_tor.atom_id_3
        → _chem_comp_atom.atom_id
  _chem_comp_tor.atom_id_4
        → _chem_comp_atom.atom_id
```

(*k*) CHEM_COMP_TOR_VALUE

```
• _chem_comp_tor_value.comp_id
• _chem_comp_tor_value.tor_id
+ _chem_comp_tor_value.angle
        → _chem_comp_atom.comp_id
+ _chem_comp_tor_value.dist
        → _chem_comp_tor.id
```

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. Data items marked with a plus (+) have companion data names for the standard uncertainty in the reported value, formed by appending the string **_esd** *to the data name listed.*

Data items in the CHEM_COMP and related categories allow the covalent geometry, stereochemistry and Cartesian coordinates for the chemical components of the structure to be specified. These components may be monomers, *e.g.* the amino acids that form proteins, the nucleotides that form nucleic acids or the sugars that form oligosaccharides, or they may be the small-molecule compounds, ions or water molecules that co-crystallize with the macromolecule(s).

In a small-molecule structure determination, the chemistry is often deduced from the electron density distribution. In contrast, in macromolecular crystallography, the chemistry of the monomers that form a polymeric macromolecule is usually known in advance and is used to interpret the electron density. In many cases, the chemistry of the monomers is so well determined that it is not worth storing a copy of the geometric restraints used in every mmCIF that uses the same set of data for the monomers. In these cases, the data item **_chem_comp.model_erf** can be used to identify an external reference file (e.r.f.) that contains standard chemical data for these monomers. Although the present version of the mmCIF dictionary does not specify the form that the file identifier might take, it is likely that users will specify the location of the file in their local file system or the URL of files of reference data accessible over the Internet. In the long term, it would be helpful to have a standard repository of reference data for monomers with a stable identifier that is independent of file names or access protocols.

The relationships between the categories used to describe chemical components are shown in Fig. 3.6.7.3.

The CHEM_COMP category provides data items for the chemical formula and formula weight of each component, the total number

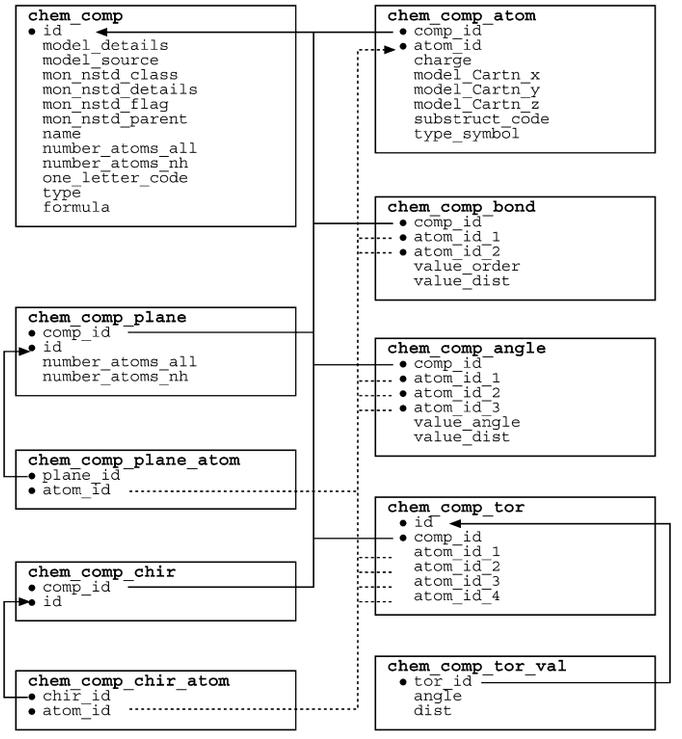

Fig. 3.6.7.3. The family of categories used to describe the chemical and structural features of the monomers and small molecules used to build a model of a structure. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

of atoms, the number of non-hydrogen atoms, and the name of the component. The name of the component will typically be a common name such as 'alanine' or 'valine'; it is recommended that the IUPAC name is used for components that are not among the usual monomers that make up proteins, nucleic acids or sugars.

The one-letter or three-letter code for a standard component may be given (using **_chem_comp.one_letter_code** and **_chem_comp.three_letter_code**, respectively). Values of X for the one-letter code or UNK for the three-letter code are used to indicate components that do not have a standard abbreviation. A component that has been formed by modification of a standard component can be indicated by prefixing the code with a plus sign. A value of '.', which means 'not applicable', should be used for components that are not monomers from which a polymeric macromolecule is built, for example co-crystallized small molecules, ions or water.

The data item **_chem_comp.type** can be used to describe the structural role of a monomer within a polymeric molecule. The types that are recognized are classified as linking monomers (for proteins, nucleic acids and sugars), monomers with an N-terminal or C-terminal cap (for proteins), and monomers with a 5′ or 3′ terminal cap (for nucleic acids). The specification of types for sugars is less complete than for proteins and nucleic acids and no types of terminal groups are currently specified for sugars. The values non-polymer and other are provided for types that have not been defined explicitly.

Information about the source of the model for the chemical component can be given using **_chem_comp.model_source** and **_chem_comp.model_details**. **_chem_comp.model_source** is a text field where the user might, for example, supply a reference to the Cambridge Structural Database or another small-molecule crystallographic database, or describe a molecular-modelling process. **_chem_comp.model_details** can be used to discuss any modification made to the model given in **_chem_comp.model_source**. As mentioned previously, **_chem_comp.model_erf** can be used to specify the location of an external reference file if the model is not described within the current data block.

Macromolecules often contain modifications of standard monomers, such as phosphorylated serines and threonines. In the mmCIF data model, a nonstandard monomer should be treated as a separate CHEM_COMP entry and described in full. However, it may be useful to refer to the standard monomer from which it was derived using the **_chem_comp.mon_nstd_*** data items. There are no fixed rules for what constitutes a 'standard' or 'nonstandard' monomer in this context, but any covalent modification of a standard amino acid or nucleotide would generally be considered nonstandard. Sometimes it is is difficult to decide whether a monomer is standard or nonstandard: selenomethionine is not one of the standard 20 amino acids, but it is so commonly used that geometric restraints for it are included in many standard packages for protein structure refinement.

Data items in the CHEM_COMP_ATOM category can be used to describe the atoms in a component. The position of each atom is given in orthogonal ångström coordinates. These coordinates correspond to the atom positions in the model of the component used in the refinement, not to the final set of refined atom positions recorded in the ATOM_SITE list.

Other CHEM_COMP_ATOM data items can be used to specify what element the atom is and its formal electronic charge, or partial charge. A code may also be assigned to the atom to indicate its role within a substructural classification of the component. The allowed codes are main and side for the main-chain and side-chain parts of amino acids, and base, phos and sugar for the base, phosphate and sugar parts of nucleotides. Atoms that do not belong to a substructure may be assigned the code none.

Data items in the CHEM_COMP_BOND category can be used to describe the intramolecular bonds between atoms in a component. Bond restraints may be described by the distance between the bonded atoms, the bond order, or both. The recognized bond types are the same as those for the core CIF dictionary data item **_chemical_conn_bond.type**, and they fulfil the same role: to characterize a model that could be used for database substructure searching, rather than to give a detailed description of unusual bond types.

In the CHEM_COMP_ANGLE category, atom 2 defines the vertex of the angle involving atoms 1, 2 and 3. The angle may be described as either an angle at the vertex atom or as a distance between atoms 1 and 3.

Data items in the CHEM_COMP_CHIR category can be used to describe the conformation of chiral centres within the component. The absolute configuration and the chiral volume may be specified, as well as the total number of atoms and the number of non-hydrogen atoms bonded to the chiral centre. There is also a flag to indicate whether a restrained chiral volume should match the target value in sign as well as in magnitude. Because chiral centres can involve a variable number of atoms, a separate list of the atoms should be given in CHEM_COMP_CHIR_ATOM.

Data items in the CHEM_COMP_PLANE category can be used to define planes within a component. The number of non-hydrogen atoms and the total number of atoms in each plane can be recorded. The atoms defining each plane should be listed separately in CHEM_COMP_PLANE_ATOM.

Data items in the CHEM_COMP_TOR category can be used to give details about the torsion angles in a component. A torsion angle may be described either as an angle or as a distance between the first and last atoms. (A torsion angle cannot be completely described by a distance, but sometimes a distance

Example 3.6.7.4. *The description of a component (adriamycin) of a macromolecule with data items in the* CHEM_COMP, CHEM_COMP_ATOM, CHEM_COMP_BOND, CHEM_COMP_TOR *and* CHEM_COMP_TOR_VALUE *categories (Leonard et al., 1993).*

```
_chem_comp.id                 'DM2'
_chem_comp.name               'adriamycin'
_chem_comp.type                non-polymer
_chem_comp.formula            'C27 H29 N1 O11'
_chem_comp.number_atoms_all  68
_chem_comp.number_atoms_nh   39
_chem_comp.formula_weight    543.51

loop_
_chem_comp_atom.comp_id
_chem_comp_atom.atom_id
_chem_comp_atom.type_symbol
_chem_comp_atom.model_Cartn_x
_chem_comp_atom.model_Cartn_y
_chem_comp_atom.model_Cartn_z
   DM2 'C1'  C  12.996  0.476  12.694
   DM2 'C2'  C  13.982 -0.225  13.183
   DM2 'C3'  C  12.482  0.165  11.515
# - - - abbreviated - - -

loop_
_chem_comp_bond.comp_id
_chem_comp_bond.atom_id_1
_chem_comp_bond.atom_id_2
_chem_comp_bond.value_order
_chem_comp_bond.value_dist
_chem_comp_bond.value_dist_esd
   DM2 'C1' 'C2' sing  1.517  0.0210
   DM2 'C2' 'C3' sing  1.445  0.0040
# - - - abbreviated - - -

loop_
  _chem_comp_tor.comp_id
  _chem_comp_tor.id
  _chem_comp_tor.atom_id_1
  _chem_comp_tor.atom_id_2
  _chem_comp_tor.atom_id_3
  _chem_comp_tor.atom_id_4
    phe  phe_chi1   N    CA   CB   CG
    phe  phe_chi2   CA   CB   CG   CD1
    phe  phe_ring1  CB   CG   CD1  CE1
    phe  phe_ring2  CB   CG   CD2  CE2
    phe  phe_ring3  CG   CD1  CE1  CZ
    phe  phe_ring4  CD1  CE1  CZ   CE2
    phe  phe_ring5  CE1  CZ   CE2  CD2

loop_
  _chem_comp_tor_value.tor_id
  _chem_comp_tor_value.comp_id
  _chem_comp_tor_value.angle
  _chem_comp_tor_value.dist
    phe_chi1   phe  -60.0  2.88
    phe_chi1   phe  180.0  3.72
    phe_chi1   phe   60.0  2.88
    phe_chi2   phe   90.0  3.34
    phe_chi2   phe  -90.0  3.34
    phe_ring1  phe  180.0  3.75
    phe_ring2  phe  180.0  3.75
    phe_ring3  phe    0.0  2.80
    phe_ring4  phe    0.0  2.80
    phe_ring5  phe    0.0  2.80
```

restraint is used in refinement, where the value of the angle is assumed to be close to the target value.) As torsion angles can have more than one target value, the target values are specified in the CHEM_COMP_TOR_VALUE category.

Data items in the CHEM_COMP_LINK category can be used to provide a table of links between the components of the structure. Each link is assigned an identifier (**_chem_comp_link.link_id**) and the types of monomer at each end of the link are stated. The types are those allowed for the parent data item **_chem_comp.type**.

The use of many of these data items to describe a typical component is shown in Example 3.6.7.4.

3.6.7.2.3. *Chemical links*

The data items in these categories are as follows:

(*a*) CHEM_LINK

```
• _chem_link.id
  _chem_link.details
```

(*b*) CHEM_LINK_ANGLE

```
• _chem_link_angle.atom_id_1
• _chem_link_angle.atom_id_2
• _chem_link_angle.atom_id_3
• _chem_link_angle.link_id
        → _chem_link.id
  _chem_link_angle.atom_1_comp_id
  _chem_link_angle.atom_2_comp_id
  _chem_link_angle.atom_3_comp_id
+ _chem_link_angle.value_angle
+ _chem_link_angle.value_dist
```

(*c*) CHEM_LINK_BOND

```
• _chem_link_bond.atom_id_1
• _chem_link_bond.atom_id_2
• _chem_link_bond.link_id
        → _chem_link.id
  _chem_link_bond.atom_1_comp_id
  _chem_link_bond.atom_2_comp_id
+ _chem_link_bond.value_dist
  _chem_link_bond.value_order
```

(*d*) CHEM_LINK_CHIR

```
• _chem_link_chir.id
• _chem_link_chir.link_id
        → _chem_link.id
  _chem_link_chir.atom_comp_id
  _chem_link_chir.atom_id
  _chem_link_chir.atom_config
  _chem_link_chir.number_atoms_all
  _chem_link_chir.number_atoms_nh
  _chem_link_chir.volume_flag
+ _chem_link_chir.volume_three
```

(*e*) CHEM_LINK_CHIR_ATOM

```
• _chem_link_chir_atom.atom_id
• _chem_link_chir_atom.chir_id
        → _chem_link_chir.id
  _chem_link_chir_atom.atom_comp_id
  _chem_link_chir_atom.dev
```

(*f*) CHEM_LINK_PLANE

```
• _chem_link_plane.id
• _chem_link_plane.link_id
        → _chem_link.id
  _chem_link_plane.number_atoms_all
  _chem_link_plane.number_atoms_nh
```

(*g*) CHEM_LINK_PLANE_ATOM

```
• _chem_link_plane_atom.atom_id
• _chem_link_plane_atom.plane_id
        → _chem_link_plane.id
  _chem_link_plane_atom.atom_comp_id
```

(*h*) CHEM_LINK_TOR

```
• _chem_link_tor.id
• _chem_link_tor.link_id
        → _chem_link.id
  _chem_link_tor.atom_1_comp_id
  _chem_link_tor.atom_2_comp_id
  _chem_link_tor.atom_3_comp_id
  _chem_link_tor.atom_4_comp_id
  _chem_link_tor.atom_id_1
  _chem_link_tor.atom_id_2
  _chem_link_tor.atom_id_3
  _chem_link_tor.atom_id_4
```

(*i*) CHEM_LINK_TOR_VALUE

```
• _chem_link_tor_value.tor_id
        → _chem_link_tor.id
+ _chem_link_tor_value.angle
+ _chem_link_tor_value.dist
```

(*j*) ENTITY_LINK

```
• _entity_link.link_id
        → _chem_link.id
  _entity_link.details
  _entity_link.entity_id_1
        → _entity.id
  _entity_link.entity_id_2
        → _entity.id
  _entity_link.entity_seq_num_1
        → _entity_poly_seq.num
  _entity_link.entity_seq_num_2
        → _entity_poly_seq.num
```

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. Data items marked with a plus (+) have companion data names for the standard uncertainty in the reported value, formed by appending the string ***_esd*** *to the data name listed.*

The geometry of the links between chemical components or entities can be described in the CHEM_LINK group of categories. Chemical components may be linked together according to the type of the component; defining the linking according to the type of the component rather than by each component in turn allows a type of polymer link for all the monomers in a polymer to be specified (*e.g.* L-peptide linking). The geometry of the links can be specified in the remaining CHEM_LINK categories. The relationships between categories used to describe links between chemical components are shown in Fig. 3.6.7.4, which also shows how information about the links is passed to the CHEM_COMP and CHEM_LINK categories. For simplicity, the categories CHEM_COMP_PLANE, CHEM_COMP_PLANE_ATOM, CHEM_COMP_CHIR, CHEM_COMP_CHIR_ATOM and ENTITY_LINK are not included in Fig. 3.6.7.4.

Note that this category group can be used to describe the links that connect the monomers within a macromolecular polymer (using the CHEM_LINK categories) and also the intramolecular links between separate molecules in the whole complex (using the ENTITY_LINK category). Intramolecular links, for example a covalent bond formed between a bound ligand and an amino-acid side chain, are usually discovered as a result of the structure determination, and it would therefore seem more appropriate to describe them in the STRUCT_CONN category. However, since one of the roles of the CHEM_LINK category group is to record target values used for restraints or constraints during the refinement of the model of the structure, ideal values for the geometry of any entity-to-entity links should be given here.

Data items in the CHEM_LINK category are used to assign a unique identifier to each link and allow the author to record any unusual aspects of each link. The other categories in the CHEM_LINK category group describe the geometric model of each link, and are closely analogous to the similarly named categories in the CHEM_COMP group.

The relationships among these categories are complex (see Fig. 3.6.7.4). Each atom that participates in an aspect of the link (for example, a bond, an angle, a chiral centre, a torsion angle or a plane) must be identified and it must also be specified whether the atom is in the first or second of the components that form the link.

Data items in the CHEM_LINK_BOND category describe the bonds between atoms participating in an intermolecular link between chemical components. Bond restraints may be described by the distance between the bonded atoms, the bond order or both.

An angle at a link may be described in the CHEM_LINK_ANGLE category as either an angle at the vertex atom or as a distance between the atoms attached to the vertex. For data items in both the CHEM_LINK_BOND and CHEM_LINK_ANGLE categories, a target value and its associated standard uncertainty may be specified (Example 3.6.7.5).

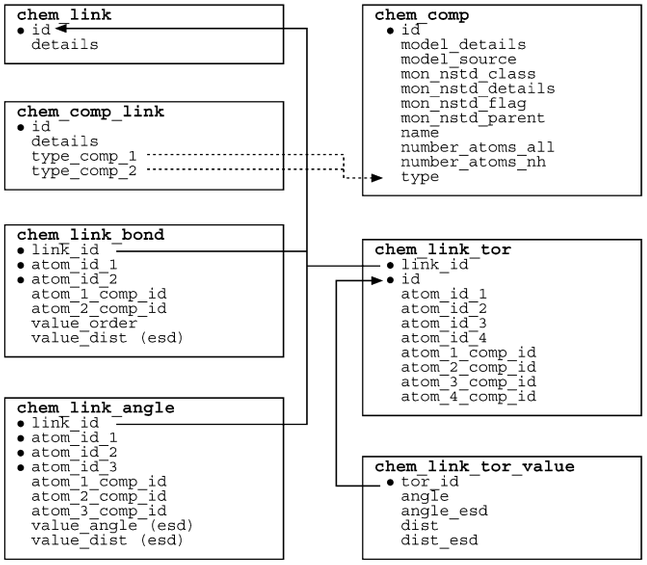

Fig. 3.6.7.4. The family of categories used to describe the links between chemical components. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

Data items in the CHEM_LINK_CHIR category can be used to describe the conformation of chiral centres in a link between two chemical components. The absolute configuration and the chiral volume may be specified, as well as the total number of atoms and the number of non-hydrogen atoms bonded to the chiral centre. There is also a flag to indicate whether a restrained chiral volume should match the target value in sign as well as in magnitude. Because chiral centres can involve a variable number of atoms, a separate list of the atoms should be given in CHEM_LINK_CHIR_ATOM.

Data items in the CHEM_LINK_PLANE category can be used to list planes defined across a link between two chemical components. Because planes can involve a variable number of atoms, a separate list of the atoms should be given in CHEM_LINK_PLANE_ATOM.

Data items in the CHEM_LINK_TOR category can be used to give details of the torsion angles across a link between two chemical

Example 3.6.7.5. *A peptide bond described with data items in the CHEM_LINK_BOND and CHEM_LINK_ATOM categories.*

```
loop_
  _chem_link_bond.link_id
  _chem_link_bond.value_dist
  _chem_link_bond.value_dist_esd
  _chem_link_bond.atom_id_1
  _chem_link_bond.atom_1_comp_id
  _chem_link_bond.atom_id_2
  _chem_link_bond.atom_2_comp_id
    PEPTIDE  1.329  0.014  C  1  N  2

loop_
  _chem_link_angle.link_id
  _chem_link_angle.value_angle
  _chem_link_angle.value_angle_esd
  _chem_link_angle.atom_id_1
  _chem_link_angle.atom_1_comp_id
  _chem_link_angle.atom_id_2
  _chem_link_angle.atom_2_comp_id
  _chem_link_angle.atom_id_3
  _chem_link_angle.atom_3_comp_id
    PEPTIDE  116.2  2.0  CA 1  C  1  N  2
    PEPTIDE  123.0  1.6  O  1  C  1  N  2
    PEPTIDE  121.7  1.8  C  1  N  2  CA 2
```

components. The torsion angle may be described either as an angle or as a distance between the first and last atoms. As torsion angles can have more than one target value, the target values are specified in the CHEM_LINK_TOR_VALUE category.

The ENTITY_LINK category is used to identify the participants in links between distinct molecular entities. A pointer to the details of the link is given in **_entity_link.link_id**, which matches a value of **_chem_link.id** in the CHEM_LINK category.

3.6.7.3. Distinct chemical species

The categories describing distinct chemical entities are as follows:

ENTITY group
Entities (§3.6.7.3.1)
ENTITY
ENTITY_KEYWORDS
ENTITY_NAME_COM
ENTITY_NAME_SYS
ENTITY_SRC_GEN
ENTITY_SRC_NAT
Polymer entities (§3.6.7.3.2)
ENTITY_POLY
ENTITY_POLY_SEQ

The ENTITY categories of the mmCIF dictionary should be used in preference to the CHEMICAL categories of the core CIF dictionary. In a typical small-molecule structure determination, for which the core CIF dictionary was designed, the substance being studied can be thought of as a single chemical species, even if it contains distinct ions or ligands. In a macromolecular structure, it is more often the case that separate descriptions are appropriate for each of the distinct chemical species that comprise the structural complex. The ENTITY categories allow the species present and their basic chemical properties to be specified. Their structures and connectivity are described in other categories.

It is important, therefore, to remember that the ENTITY data do not represent the result of the crystallographic experiment; those results are given using the ATOM_SITE data items and are discussed and described using data items in the STRUCT family of categories. The ENTITY categories describe the chemistry of the molecules under investigation and are most usefully considered as the ideal groups to which the structure is restrained or constrained during refinement.

It is also important to remember that entities do not correspond directly to the total contents of the asymmetric unit. Entities are described only once, even in structures in which the entity occurs several times. The STRUCT_ASYM data items, which reference the list of entities, describe and label the contents of the asymmetric unit.

The following discussion treats the data items used for entities in general (Section 3.6.7.3.1) and those used more specifically to describe polymeric entities (Section 3.6.7.3.2) separately.

3.6.7.3.1. *Description of entities*

The data items in these categories are as follows:

(*a*) ENTITY
- **_entity.id**
  _entity.details
  _entity.formula_weight
  _entity.src_method
  _entity.type

(*b*) ENTITY_KEYWORDS
- **_entity_keywords.entity_id**
  → **_entity.id**
- **_entity_keywords.text**

(*c*) ENTITY_NAME_COM
- **_entity_name_com.entity_id**
  → **_entity.id**
- **_entity_name_com.name**

(*d*) ENTITY_NAME_SYS
- **_entity_name_sys.entity_id**
  → **_entity.id**
- **_entity_name_sys.name**
  _entity_name_sys.system

(*e*) ENTITY_SRC_GEN
- **_entity_src_gen.entity_id**
  → **_entity.id**
  _entity_src_gen.gene_src_common_name
  _entity_src_gen.gene_src_details
  _entity_src_gen.gene_src_genus
  _entity_src_gen.gene_src_species
  _entity_src_gen.gene_src_strain
  _entity_src_gen.gene_src_tissue
  _entity_src_gen.gene_src_tissue_fraction
  _entity_src_gen.host_org_common_name
  _entity_src_gen.host_org_details
  _entity_src_gen.host_org_genus
  _entity_src_gen.host_org_species
  _entity_src_gen.host_org_strain
  _entity_src_gen.plasmid_details
  _entity_src_gen.plasmid_name

(*f*) ENTITY_SRC_NAT
- **_entity_src_nat.entity_id**
  → **_entity.id**
  _entity_src_nat.common_name
  _entity_src_nat.details
  _entity_src_nat.genus
  _entity_src_nat.species
  _entity_src_nat.strain
  _entity_src_nat.tissue
  _entity_src_nat.tissue_fraction

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.

An entity in mmCIF is a chemically distinct molecular component of the structural complex described in the mmCIF. The three possible types of molecular entities are polymer, non-polymer and water. Note that the 'water' entity is water, and only water. Any other well ordered solvent molecules or ions should be treated as non-polymer entities. The relationships between categories used to describe the features of entities are shown in Fig. 3.6.7.5, which also shows how the information describing the entity is linked to the coordinate list in the ATOM_SITE category.

Data items in the ENTITY category are used to label each distinct chemical molecule with a reference code (**_entity.id**), to give the formula weight in daltons (if available) and to define the type of the entity as one of polymer, non-polymer or water. The method by which the entity was produced may be indicated using the item **_entity.src_method**, whose allowed values are `nat` (indicating that the sample was isolated from a natural source), `man` (indicating a genetically manipulated source) or `syn` (indicating a chemical synthesis). A value of `nat` indicates that additional details should be given in the ENTITY_SRC_NAT category and a value of `man` indicates that additional details should be given in the ENTITY_SRC_GEN category. As these flags are only relevant to the macromolecular entities of a structural complex, a value of '.', indicating 'inapplicable', should be given to **_entity.src_method** for solvent or water molecules. The **_entity.details** field can be used for a free-text description of any special features of the entity.

Keywords characterizing the individual molecular species may be given using data items in the ENTITY_KEYWORD category. These keywords should only be used to record information that does not depend on knowledge of the molecular structure. Thus a polypeptide could be described as a polypeptide, or an enzyme, or

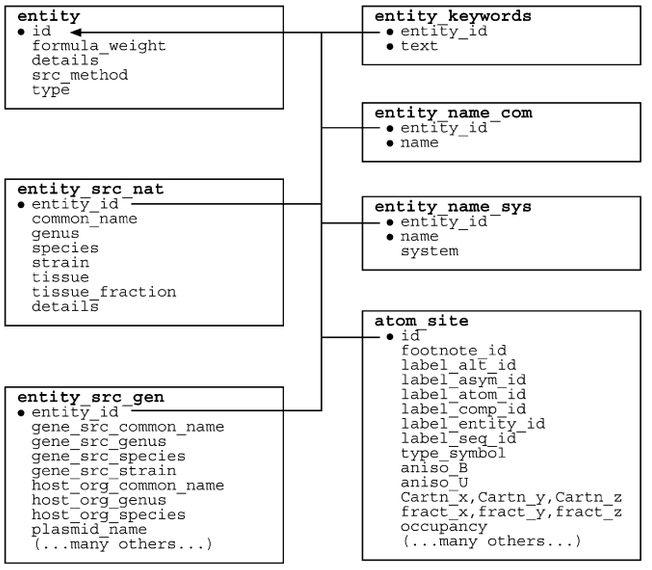

Fig. 3.6.7.5. The family of categories used to describe chemical entities. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data item.

a protease, but it should not be described as an $\alpha\beta$-barrel; a number of categories within the STRUCT family allow keywords specific to the structure of the macromolecule to be given.

Data items in the ENTITY_NAME_COM category may be used to give any common names for an entity. Several different names can be recorded for each entity if appropriate.

Similarly, data items in the ENTITY_NAME_SYS category may be used to give systematic names for each entity. Again, several different names can be recorded for each entity if appropriate. The data item **_entity_name_sys.system** can be used to record the system according to which the systematic name was generated.

The ENTITY_SRC_GEN category allows a description of the source of entities produced by genetic manipulation to be given. There are data items for describing the tissue from which the gene was obtained, the plasmid into which it was incorporated for expression, and the host organism in which the macromolecule was expressed (Example 3.6.7.6).

The ENTITY_SRC_NAT category allows a description of the source of entities obtained from a natural tissue to be given. Data items are provided for the common and systematic name (by genus, species and, where relevant, strain) of the organism from which the material was obtained. Other data items can be used to describe the tissue (and if necessary the subcellular fraction of the tissue) from which the entity was isolated.

Example 3.6.7.6. *An example of the description of the entities in an HIV-1 protease structure (PDB 5HVP), described using data items in the ENTITY, ENTITY_NAME_COM, ENTITY_NAME_SYS and ENTITY_SRC_GEN categories.*

```
loop_
_entity.id
_entity.type
_entity.formula_weight
_entity.details
   1  polymer       10916
; The enzymatically competent form of HIV protease is
  a dimer. This entity corresponds to one monomer of
  an active dimer.
;
   2  non-polymer   647.2    .
   3  water            18    .

loop_
_entity_name_com.entity_id
_entity_name_com.name
   1  'HIV-1 protease monomer'
   1  'HIV-1 PR monomer'
   2  'acetyl-pepstatin'
   2  'acetyl-Ile-Val-Asp-Statine-Ala-Ile-Statine'
   3  'water'

_entity_name_sys.entity_id       1
_entity_name_sys.name           'EC 2.1.1.1'
_entity_name_sys.system         'Enzyme convention'

loop_
_entity_src_gen.entity_id
_entity_src_gen.gene_src_common_name
_entity_src_gen.gene_src_strain
_entity_src_gen.host_org_common_name
_entity_src_gen.host_org_genus
_entity_src_gen.host_org_species
_entity_src_gen.plasmid_name
1 'HIV-1' 'NY-5' 'bacteria' 'Escherichia' 'coli'
'pB322'
```

3.6.7.3.2. *Polymer entities*

The data items in these categories are as follows:

(*a*) ENTITY_POLY

- **_entity_poly.entity_id**
  → **_entity.id**

_entity_poly.nstd_chirality
_entity_poly.nstd_linkage
_entity_poly.nstd_monomer
_entity_poly.number_of_monomers
_entity_poly.type
_entity_poly.type_details

(*b*) ENTITY_POLY_SEQ

- **_entity_poly_seq.entity_id**
  → **_entity.id**
- **_entity_poly_seq.mon_id**
  → **_chem_comp.id**
- **_entity_poly_seq.num**

_entity_poly_seq.hetero

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.

The polymer type, sequence length and information about any nonstandard features of the polymer may be specified using data items in the ENTITY_POLY category. The sequence of monomers in each polymer entity is given using data items in the ENTITY_POLY_SEQ category. The relationships between categories describing polymer entities are shown in Fig. 3.6.7.6, which also shows how the information describing the polymer is linked to the coordinate list in the ATOM_SITE category and to the full chemical description of each monomer or nonstandard monomer in the CHEM_COMP category.

Non-polymer entities are treated as individual chemical components, in the same way in which monomers within a polymer are treated as individual chemical components. They may be fully described in the CHEM_COMP group of categories (Example 3.6.7.7).

Data items in the ENTITY_POLY category can be used to give the number of monomers in the polymer and to assign the type of the polymer as one of the set of types `polypeptide(D)`, `polypeptide(L)`, `polydeoxyribonucleotide`, `polyribonucleotide`, `polysaccharide(D)`, `polysaccharide(L)` or `other`. Details of deviations from a standard type may be given in **_entity_poly.type_details**.

In some cases, the polymer is best described as one of the standard types even if it contains some nonstandard features. Flags are provided to indicate the presence of three types of nonstandard features. The presence of chiral centres other than those implied

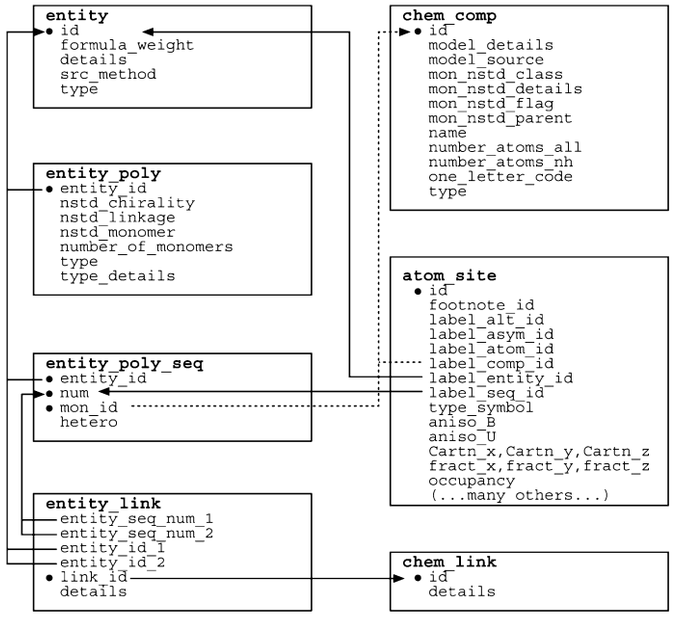

Fig. 3.6.7.6. The family of categories used to describe polymer chemical entities. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

Example 3.6.7.7. *An example of both polymer and non-polymer entities in a drug–DNA complex (NDB DDF040) described with data items in the* ENTITY, ENTITY_KEYWORDS, ENTITY_NAME_COM, ENTITY_POLY *and* ENTITY_POLY_SEQ *categories (Narayana et al., 1991).*

```
loop_
_entity.id
_entity.type
_entity.src_method
   1  polymer     man
   2  non-polymer man
   3  water        .

loop_
_entity_keywords.entity_id
_entity_keywords.text
   1  'nucleic acid'
   2  'drug'

loop_
_entity_name_com.entity_id
_entity_name_com.name
   2  adriamycin
   3  water

loop_
_entity_poly.entity_id
_entity_poly.number_of_monomers
_entity_poly.type
   1  8  'polydeoxyribonucleotide'

loop_
_entity_poly_seq.entity_id
_entity_poly_seq.mon_id
_entity_poly_seq.num
   1  T  1
   1  G  2
   1  G  3
   1  C  4
   1  C  5
   1  A  6
# - - - abbreviated - - -
```

by the assigned type is indicated by assigning a value of yes to the data item **_entity_poly.nstd_chirality**. A value of yes for **_entity_poly.nstd_linkage** indicates the presence of monomer-to-monomer links different from those implied by the assigned type and a value of yes for **_entity_poly.nstd_monomer** indicates the presence of one or more nonstandard monomer components.

Data items in the ENTITY_POLY_SEQ category describe the sequence of monomers in a polymer. By including **_entity_poly_seq.mon_id** in the category key, it is possible to allow for sequence heterogeneity by allowing a given sequence number to be correlated with more than one monomer ID. Sequence heterogeneity is shown in the example of crambin in Section 3.6.3.

3.6.7.4. Molecular or packing geometry

The categories describing geometry are as follows:

GEOM group
- GEOM
- GEOM_ANGLE
- GEOM_BOND
- GEOM_CONTACT
- GEOM_HBOND
- GEOM_TORSION

The categories within the GEOM group are used in the core CIF dictionary to describe the geometry of the model that results from the structure determination, and can be used to select values that will be published in a report describing the structure. The complexity of macromolecular structures means that a different approach to presenting the results of a structure determination is needed. The STRUCT family of categories was created to meet this need. The GEOM categories are retained in the mmCIF dictionary, but only for consistency with the core CIF dictionary.

The data items in the categories in the GEOM group are:

(*a*) GEOM

- **_geom.entry_id**
  → **_entry.id**
  _geom.details (∼ ***_geom_special_details***)

(*b*) GEOM_ANGLE

- ***_geom_angle.atom_site_id_1***
  (∼ ***_geom_angle_atom_site_label_1***)
- ***_geom_angle.atom_site_id_2***
  (∼ ***_geom_angle_atom_site_label_2***)
- ***_geom_angle.atom_site_id_3***
  (∼ ***_geom_angle_atom_site_label_3***)
- ***_geom_angle.site_symmetry_1***
- ***_geom_angle.site_symmetry_2***
- ***_geom_angle.site_symmetry_3***

_geom_angle.atom_site_auth_asym_id_1
→ **_atom_site.auth_asym_id**
_geom_angle.atom_site_auth_atom_id_1
→ **_atom_site.auth_atom_id**
_geom_angle.atom_site_auth_comp_id_1
→ **_atom_site.auth_comp_id**
_geom_angle.atom_site_auth_seq_id_1
→ **_atom_site.auth_seq_id**
_geom_angle.atom_site_auth_asym_id_2
→ **_atom_site.auth_asym_id**
_geom_angle.atom_site_auth_atom_id_2
→ **_atom_site.auth_atom_id**
_geom_angle.atom_site_auth_comp_id_2
→ **_atom_site.auth_comp_id**
_geom_angle.atom_site_auth_seq_id_2
→ **_atom_site.auth_seq_id**
_geom_angle.atom_site_auth_asym_id_3
→ **_atom_site.auth_asym_id**
_geom_angle.atom_site_auth_atom_id_3
→ **_atom_site.auth_atom_id**
_geom_angle.atom_site_auth_comp_id_3
→ **_atom_site.auth_comp_id**
_geom_angle.atom_site_auth_seq_id_3
→ **_atom_site.auth_seq_id**
→ **_atom_site.id**
_geom_angle.atom_site_label_alt_id_1
→ **_atom_site.label_alt_id**
_geom_angle.atom_site_label_asym_id_1
→ **_atom_site.label_asym_id**
_geom_angle.atom_site_label_atom_id_1
→ **_atom_site.label_atom_id**

_geom_angle.atom_site_label_comp_id_1
→ _atom_site.label_comp_id
_geom_angle.atom_site_label_seq_id_1
→ _atom_site.label_seq_id
→ _atom_site.id
_geom_angle.atom_site_label_alt_id_2
→ _atom_site.label_alt_id
_geom_angle.atom_site_label_asym_id_2
→ _atom_site.label_asym_id
_geom_angle.atom_site_label_atom_id_2
→ _atom_site.label_atom_id
_geom_angle.atom_site_label_comp_id_2
→ _atom_site.label_comp_id
_geom_angle.atom_site_label_seq_id_2
→ _atom_site.label_seq_id
→ _atom_site.id
_geom_angle.atom_site_label_alt_id_3
→ _atom_site.label_alt_id
_geom_angle.atom_site_label_asym_id_3
→ _atom_site.label_asym_id
_geom_angle.atom_site_label_atom_id_3
→ _atom_site.label_atom_id
_geom_angle.atom_site_label_comp_id_3
→ _atom_site.label_comp_id
_geom_angle.atom_site_label_seq_id_3
→ _atom_site.label_seq_id
_geom_angle.publ_flag
+ _geom_angle.value (∼ _geom_angle)

(*c*) GEOM_BOND
• _geom_bond.atom_site_id_1
(∼ _geom_bond_atom_site_label_1)
• _geom_bond.atom_site_id_2
(∼ _geom_bond_atom_site_label_2)
• _geom_bond.site_symmetry_1
• _geom_bond.site_symmetry_2
_geom_bond.atom_site_auth_asym_id_1
→ _atom_site.auth_asym_id
_geom_bond.atom_site_auth_atom_id_1
→ _atom_site.auth_atom_id
_geom_bond.atom_site_auth_comp_id_1
→ _atom_site.auth_comp_id
_geom_bond.atom_site_auth_seq_id_1
→ _atom_site.auth_seq_id
_geom_bond.atom_site_auth_asym_id_2
→ _atom_site.auth_asym_id
_geom_bond.atom_site_auth_atom_id_2
→ _atom_site.auth_atom_id
_geom_bond.atom_site_auth_comp_id_2
→ _atom_site.auth_comp_id
_geom_bond.atom_site_auth_seq_id_2
→ _atom_site.auth_seq_id
→ _atom_site.id
_geom_bond.atom_site_label_alt_id_1
→ _atom_site.label_alt_id
_geom_bond.atom_site_label_asym_id_1
→ _atom_site.label_asym_id
_geom_bond.atom_site_label_atom_id_1
→ _atom_site.label_atom_id
_geom_bond.atom_site_label_comp_id_1
→ _atom_site.label_comp_id
_geom_bond.atom_site_label_seq_id_1
→ _atom_site.label_seq_id
→ _atom_site.id
_geom_bond.atom_site_label_alt_id_2
→ _atom_site.label_alt_id
_geom_bond.atom_site_label_asym_id_2
→ _atom_site.label_asym_id
_geom_bond.atom_site_label_atom_id_2
→ _atom_site.label_atom_id
_geom_bond.atom_site_label_comp_id_2
→ _atom_site.label_comp_id
_geom_bond.atom_site_label_seq_id_2
→ _atom_site.label_seq_id
+ _geom_bond.dist (∼ _geom_bond_distance)
_geom_bond.publ_flag
_geom_bond.valence

(*d*) GEOM_CONTACT
• _geom_contact.atom_site_id_1
(∼ _geom_contact_atom_site_label_1)
• _geom_contact.atom_site_id_2
(∼ _geom_contact_atom_site_label_2)
• _geom_contact.site_symmetry_1
• _geom_contact.site_symmetry_2
_geom_contact.atom_site_auth_asym_id_1
→ _atom_site.auth_asym_id
_geom_contact.atom_site_auth_atom_id_1
→ _atom_site.auth_atom_id
_geom_contact.atom_site_auth_comp_id_1
→ _atom_site.auth_comp_id
_geom_contact.atom_site_auth_seq_id_1
→ _atom_site.auth_seq_id
_geom_contact.atom_site_auth_asym_id_2
→ _atom_site.auth_asym_id
_geom_contact.atom_site_auth_atom_id_2
→ _atom_site.auth_atom_id
_geom_contact.atom_site_auth_comp_id_2
→ _atom_site.auth_comp_id
_geom_contact.atom_site_auth_seq_id_2
→ _atom_site.auth_seq_id
→ _atom_site.id
_geom_contact.atom_site_label_alt_id_1
→ _atom_site.label_alt_id
_geom_contact.atom_site_label_asym_id_1
→ _atom_site.label_asym_id
_geom_contact.atom_site_label_atom_id_1
→ _atom_site.label_atom_id
_geom_contact.atom_site_label_comp_id_1
→ _atom_site.label_comp_id
_geom_contact.atom_site_label_seq_id_1
→ _atom_site.label_seq_id
→ _atom_site.id
_geom_contact.atom_site_label_alt_id_2
→ _atom_site.label_alt_id
_geom_contact.atom_site_label_asym_id_2
→ _atom_site.label_asym_id
_geom_contact.atom_site_label_atom_id_2
→ _atom_site.label_atom_id
_geom_contact.atom_site_label_comp_id_2
→ _atom_site.label_comp_id
_geom_contact.atom_site_label_seq_id_2
→ _atom_site.label_seq_id
+ _geom_contact.dist (∼ _geom_contact_distance)
_geom_contact.publ_flag

(*e*) GEOM_HBOND
• _geom_hbond.atom_site_id_A
→ _atom_site.id
• _geom_hbond.atom_site_id_D
→ _atom_site.id
• _geom_hbond.atom_site_id_H
→ _atom_site.id
• _geom_hbond.site_symmetry_A
• _geom_hbond.site_symmetry_D
• _geom_hbond.site_symmetry_H
+ _geom_hbond.angle_DHA
_geom_hbond.atom_site_auth_asym_id_A
→ _atom_site.auth_asym_id
_geom_hbond.atom_site_auth_atom_id_A
→ _atom_site.auth_atom_id
_geom_hbond.atom_site_auth_comp_id_A
→ _atom_site.auth_comp_id
_geom_hbond.atom_site_auth_seq_id_A
→ _atom_site.auth_seq_id
_geom_hbond.atom_site_auth_asym_id_D
→ _atom_site.auth_asym_id
_geom_hbond.atom_site_auth_atom_id_D
→ _atom_site.auth_atom_id
_geom_hbond.atom_site_auth_comp_id_D
→ _atom_site.auth_comp_id
_geom_hbond.atom_site_auth_seq_id_D
→ _atom_site.auth_seq_id
_geom_hbond.atom_site_auth_asym_id_H
→ _atom_site.auth_asym_id
_geom_hbond.atom_site_auth_atom_id_H
→ _atom_site.auth_atom_id
_geom_hbond.atom_site_auth_comp_id_H
→ _atom_site.auth_comp_id
_geom_hbond.atom_site_auth_seq_id_H
→ _atom_site.auth_seq_id
_geom_hbond.atom_site_label_alt_id_A
→ _atom_site.label_alt_id
_geom_hbond.atom_site_label_asym_id_A
→ _atom_site.label_asym_id

_geom_hbond.atom_site_label_atom_id_A
→ _atom_site.label_atom_id
_geom_hbond.atom_site_label_comp_id_A
→ _atom_site.label_comp_id
_geom_hbond.atom_site_label_seq_id_A
→ _atom_site.label_seq_id
_geom_hbond.atom_site_label_alt_id_D
→ _atom_site.label_alt_id
_geom_hbond.atom_site_label_asym_id_D
→ _atom_site.label_asym_id
_geom_hbond.atom_site_label_atom_id_D
→ _atom_site.label_atom_id
_geom_hbond.atom_site_label_comp_id_D
→ _atom_site.label_comp_id
_geom_hbond.atom_site_label_seq_id_D
→ _atom_site.label_seq_id
_geom_hbond.atom_site_label_alt_id_H
→ _atom_site.label_alt_id
_geom_hbond.atom_site_label_asym_id_H
→ _atom_site.label_asym_id
_geom_hbond.atom_site_label_atom_id_H
→ _atom_site.label_atom_id
_geom_hbond.atom_site_label_comp_id_H
→ _atom_site.label_comp_id
_geom_hbond.atom_site_label_seq_id_H
→ _atom_site.label_seq_id
\+ _geom_hbond.dist_DA (∼ _geom_hbond_distance_DA)
\+ _geom_hbond.dist_DH (∼ _geom_hbond_distance_DH)
\+ _geom_hbond.dist_HA (∼ _geom_hbond_distance_HA)
_geom_hbond.publ_flag

(*f*) GEOM_TORSION

• _geom_torsion.atom_site_id_1
(∼ _geom_torsion_atom_site_label_1)
• _geom_torsion.atom_site_id_2
(∼ _geom_torsion_atom_site_label_2)
• _geom_torsion.atom_site_id_3
(∼ _geom_torsion_atom_site_label_3)
• _geom_torsion.atom_site_id_4
(∼ _geom_torsion_atom_site_label_4)
• _geom_torsion.site_symmetry_1
• _geom_torsion.site_symmetry_2
• _geom_torsion.site_symmetry_3
• _geom_torsion.site_symmetry_4
_geom_torsion.atom_site_auth_asym_id_1
→ _atom_site.auth_asym_id
_geom_torsion.atom_site_auth_atom_id_1
→ _atom_site.auth_atom_id
_geom_torsion.atom_site_auth_comp_id_1
→ _atom_site.auth_comp_id
_geom_torsion.atom_site_auth_seq_id_1
→ _atom_site.auth_seq_id
_geom_torsion.atom_site_auth_asym_id_2
→ _atom_site.auth_asym_id
_geom_torsion.atom_site_auth_atom_id_2
→ _atom_site.auth_atom_id
_geom_torsion.atom_site_auth_comp_id_2
→ _atom_site.auth_comp_id
_geom_torsion.atom_site_auth_seq_id_2
→ _atom_site.auth_seq_id
_geom_torsion.atom_site_auth_asym_id_3
→ _atom_site.auth_asym_id
_geom_torsion.atom_site_auth_atom_id_3
→ _atom_site.auth_atom_id
_geom_torsion.atom_site_auth_comp_id_3
→ _atom_site.auth_comp_id
_geom_torsion.atom_site_auth_seq_id_3
→ _atom_site.auth_seq_id
_geom_torsion.atom_site_auth_asym_id_4
→ _atom_site.auth_asym_id
_geom_torsion.atom_site_auth_atom_id_4
→ _atom_site.auth_atom_id
_geom_torsion.atom_site_auth_comp_id_4
→ _atom_site.auth_comp_id
_geom_torsion.atom_site_auth_seq_id_4
→ _atom_site.auth_seq_id
→ _atom_site.id
_geom_torsion.atom_site_label_alt_id_1
→ _atom_site.label_alt_id
_geom_torsion.atom_site_label_asym_id_1
→ _atom_site.label_asym_id
_geom_torsion.atom_site_label_atom_id_1
→ _atom_site.label_atom_id
_geom_torsion.atom_site_label_comp_id_1
→ _atom_site.label_comp_id
_geom_torsion.atom_site_label_seq_id_1
→ _atom_site.label_seq_id
→ _atom_site.id
_geom_torsion.atom_site_label_alt_id_2
→ _atom_site.label_alt_id
_geom_torsion.atom_site_label_asym_id_2
→ _atom_site.label_asym_id
_geom_torsion.atom_site_label_atom_id_2
→ _atom_site.label_atom_id
_geom_torsion.atom_site_label_comp_id_2
→ _atom_site.label_comp_id
_geom_torsion.atom_site_label_seq_id_2
→ _atom_site.label_seq_id
→ _atom_site.id
_geom_torsion.atom_site_label_alt_id_3
→ _atom_site.label_alt_id
_geom_torsion.atom_site_label_asym_id_3
→ _atom_site.label_asym_id
_geom_torsion.atom_site_label_atom_id_3
→ _atom_site.label_atom_id
_geom_torsion.atom_site_label_comp_id_3
→ _atom_site.label_comp_id
_geom_torsion.atom_site_label_seq_id_3
→ _atom_site.label_seq_id
→ _atom_site.id
_geom_torsion.atom_site_label_alt_id_4
→ _atom_site.label_alt_id
_geom_torsion.atom_site_label_asym_id_4
→ _atom_site.label_asym_id
_geom_torsion.atom_site_label_atom_id_4
→ _atom_site.label_atom_id
_geom_torsion.atom_site_label_comp_id_4
→ _atom_site.label_comp_id
_geom_torsion.atom_site_label_seq_id_4
→ _atom_site.label_seq_id
_geom_torsion.publ_flag
\+ _geom_torsion.value (∼ _geom_torsion)

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* (_) *except where indicated by the* ∼ *symbol. Data items marked with a plus* (+) *have companion data names for the standard uncertainty in the reported value, formed by appending the string* _esd *to the data name listed.*

3.6.7.5. Molecular structure

The categories describing molecular structure are as follows:

STRUCT group

Higher-level macromolecular structure (§3.6.7.5.1)
- STRUCT
- STRUCT_ASYM
- STRUCT_BIOL
- STRUCT_BIOL_GEN
- STRUCT_BIOL_KEYWORDS
- STRUCT_BIOL_VIEW

Secondary structure (§3.6.7.5.2)
- STRUCT_CONF
- STRUCT_CONF_TYPE

Structural interactions (§3.6.7.5.3)
- STRUCT_CONN
- STRUCT_CONN_TYPE

Structural features of monomers (§3.6.7.5.4)
- STRUCT_MON_DETAILS
- STRUCT_MON_NUCL
- STRUCT_MON_PROT
- STRUCT_MON_PROT_CIS

Noncrystallographic symmetry (§3.6.7.5.5)
- STRUCT_NCS_DOM
- STRUCT_NCS_DOM_LIM
- STRUCT_NCS_ENS

STRUCT_NCS_ENS_GEN
STRUCT_NCS_OPER

External databases (§3.6.7.5.6)
STRUCT_REF
STRUCT_REF_SEQ
STRUCT_REF_SEQ_DIF

β-sheets (§3.6.7.5.7)
STRUCT_SHEET
STRUCT_SHEET_TOPOLOGY
STRUCT_SHEET_ORDER
STRUCT_SHEET_RANGE
STRUCT_SHEET_HBOND

Molecular sites (§3.6.7.5.8)
STRUCT_SITE_GEN
STRUCT_SITE_KEYWORDS
STRUCT_SITE_VIEW

The results of the determination of a structure can be described in mmCIF using data items in the categories contained in the STRUCT category group. This is a very large group of categories and it has been divided into eight groups of related categories for the discussions that follow: (1) those that describe the structure at the level of biologically relevant assemblies; (2) those that describe the secondary structure of the macromolecules present; (3) those that describe the structural interactions that determine the conformation of the macromolecules; (4) those that describe properties of the structure at the monomer level; (5) those that describe ensembles of identical domains related by noncrystallographic symmetry; (6) those that provide references to related entities in external databases; (7) those that describe the β-sheets present in the structure; and (8) those that provide detailed descriptions of the structure of biologically interesting molecular sites.

3.6.7.5.1. *Higher-level macromolecular structure*

The data items in these categories are as follows:

(*a*) STRUCT
- **_struct.entry_id**
  → **_entry.id**

_struct.title

(*b*) STRUCT_ASYM
- **_struct_asym.id**

_struct_asym.details
_struct_asym.entity_id
  → **_entity.id**

(*c*) STRUCT_BIOL
- **_struct_biol.id**

_struct_biol.details

(*d*) STRUCT_BIOL_GEN
- **_struct_biol_gen.asym_id**
  → **_struct_asym.id**
- **_struct_biol_gen.biol_id**
  → **_struct_biol.id**
- **_struct_biol_gen.symmetry**

_struct_biol_gen.details

(*e*) STRUCT_BIOL_KEYWORDS
- **_struct_biol_keywords.biol_id**
  → **_struct_biol.id**
- **_struct_biol_keywords.text**

(*f*) STRUCT_BIOL_VIEW
- **_struct_biol_view.biol_id**
  → **_struct_biol.id**
- **_struct_biol_view.id**

_struct_biol_view.details
_struct_biol_view.rot_matrix[1][1]
_struct_biol_view.rot_matrix[1][2]
_struct_biol_view.rot_matrix[1][3]
_struct_biol_view.rot_matrix[2][1]
_struct_biol_view.rot_matrix[2][2]
_struct_biol_view.rot_matrix[2][3]
_struct_biol_view.rot_matrix[3][1]
_struct_biol_view.rot_matrix[3][2]
_struct_biol_view.rot_matrix[3][3]

(*g*) STRUCT_KEYWORDS
- **_struct_keywords.entry_id**
  → **_entry.id**
- **_struct_keywords.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.

The data items in these categories serve two related but distinct purposes.

The first purpose is to label each of the entities in the asymmetric unit, using data items in the STRUCT_ASYM category. These labels become part of the category key that identifies each coordinate record and they are used extensively throughout the STRUCT family of categories, so care must be taken to select a labelling scheme that is concise and informative.

The second function is descriptive. The categories descending from STRUCT_BIOL allow the author of the mmCIF to identify and annotate the biologically relevant structural units found by the structure determination. What constitutes a biological unit can depend on the context. Take the case of a structure with two polymers related by noncrystallographic symmetry, each of which binds a small-molecule cofactor. If the author wishes to describe the dimer interface, the biological unit could be taken to be the two protein molecules. If the author wishes to highlight the cofactor binding mode, the biological unit could be taken to be one protein molecule and its bound cofactor. In this second case, there could be an additional biological unit of the second protein molecule and its bound cofactor, which may or may not be identical in conformation to the first.

The relationships between categories used to describe higher-level structure are illustrated in Fig. 3.6.7.7.

The STRUCT category serves to link the structure to the overall identifier for the data block, using **_struct.entry_id**, and to supply a title that describes the entire structure. The importance of this title as a succinct description of the structure should not be underestimated, and the author should express concisely but clearly in **_struct.title** the components of interest and the importance of this particular study. It is useful to think of this title as describing

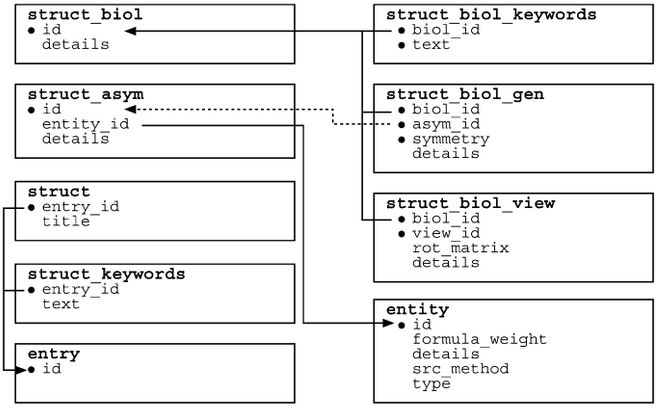

Fig. 3.6.7.7. The family of categories used to describe the higher-level macromolecular structure. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

Example 3.6.7.8. *The higher-level structure of the complex of HIV-1 protease with an inhibitor (PDB 5HVP) described with data items in the* STRUCT_ASYM, STRUCT_BIOL, STRUCT_BIOL_KEYWORDS *and* STRUCT_BIOL_GEN *categories.*

```
loop_
_struct_asym.id
_struct_asym.entity_id
_struct_asym.details
   A  1  'one monomer of the dimeric enzyme'
   B  1  'one monomer of the dimeric enzyme'
   C  2
 'one partially occupied position for the inhibitor'
   D  2
 'one partially occupied position for the inhibitor'

loop_
_struct_biol.id
_struct_biol.details
   1
; significant deviations from twofold symmetry exist
  in this dimeric enzyme
;
   2
; The drug binds to this enzyme in two roughly
  twofold symmetric modes.

  Hence this biological unit (2) is roughly twofold
  symmetric to biological unit (3). Disorder in the
  protein chain indicated with alternative ID 1
  should be used with this biological unit.
;
   3
; The drug binds to this enzyme in two roughly
  twofold symmetric modes.

  Hence this biological unit (3) is roughly twofold
  symmetric to biological unit (2). Disorder in the
  protein chain indicated with alternative ID 2
  should be used with this biological unit.
;

loop_
_struct_biol_gen.biol_id
_struct_biol_gen.asym_id
_struct_biol_gen.symmetry
   1  A  1_555     1  B  1_555
   2  A  1_555     2  B  1_555     2  C  1_555
   3  A  1_555     3  B  1_555     3  D  1_555
```

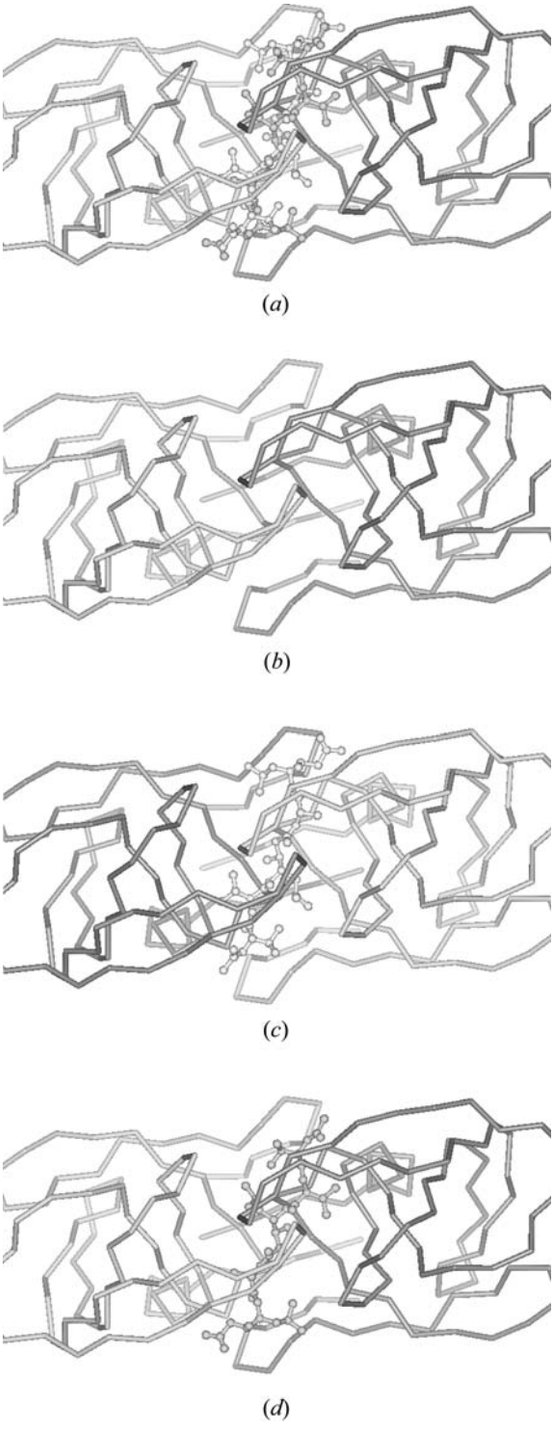

Fig. 3.6.7.8. The higher-level structure of the complex of HIV-1 protease with an inhibitor (PDB 5HVP) to be described with data items in the STRUCT_ASYM, STRUCT_BIOL, STRUCT_BIOL_KEYWORDS and STRUCT_BIOL_GEN categories. (*a*) Complete structure; (*b*), (*c*), (*d*) three different biological units.

the motivation for the structure determination, rather than the result. For instance, if the goal of the study was to determine the structure of enzyme A at pH 7.2 as part of a study of the mechanism of the reaction catalysed by the enzyme, an appropriate value for `_struct.title` would be 'Enzyme A at pH 7.2', even if the structure was found to contain two molecules per asymmetric unit, a bound calcium ion and a disordered loop between residues 47 and 52.

The STRUCT_KEYWORDS category allows an author to include keywords for the structure that has been determined. Other categories, such as STRUCT_BIOL_KEYWORDS and STRUCT_SITE_KEYWORDS, allow more specific keywords to be given, but the STRUCT_KEYWORDS category is the most likely category to be searched by simple information retrieval applications, so the author of an mmCIF might want to duplicate any keywords given elsewhere in the mmCIF in STRUCT_KEYWORDS as well.

The chemical entities that form the contents of the asymmetric unit are identified using data items in the ENTITY categories. The data items in the STRUCT_ASYM category link these entities to the structure itself. A unique identifier is attached to each occurrence of each entity in the asymmetric unit using `_struct_asym.id`. This identifier forms a part of the atom label in the ATOM_SITE category, which is used throughout the many categories in the STRUCT group in describing the structure. The identifier is also used in generating biological assemblies.

The usual reason for determining the structure of a biological macromolecule is to get information about the biologically relevant assemblies of the entities in the crystal structure. These assemblies take many forms and could encompass the complete contents of the asymmetric unit, a fraction of the contents of the asymmetric unit or the contents of more than one asymmetric unit. Each assembly, or 'biological unit', is given an identifier in the STRUCT_BIOL category and the author may annotate each biological unit using the data item `_struct_biol.details`. Key-

words for each biological unit can be given using data items in the STRUCT_BIOL_KEYWORD category.

The entities that comprise the biological unit are specified using data items in the STRUCT_BIOL_GEN category by reference to the appropriate values of **_struct_asym.id** and by specifying any symmetry transformation that must be applied to the entities to generate the biological unit.

Data items in the STRUCT_BIOL_VIEW category allow the author to specify an orientation of the biological unit that provides a useful view of the structure. The comments given in **_struct_biol_view.details** may be used as a figure caption if the view is intended to be a figure in a report describing the structure.

The example of crambin in Section 3.6.3 shows the relations between the categories defining higher-level structure for the straightforward case of a single protein molecule (with a small co-crystallization molecule and solvent) in the asymmetric unit. The structure of HIV-1 protease with a bound inhibitor (PDB 5HVP), shown in Example 3.6.7.8, is considerably more complex. There are two entities: the monomeric form of the enzyme and the small-molecule inhibitor. The asymmetric unit contains two copies of the enzyme monomer (both fully occupied) and two copies of the inhibitor (each of which is partially occupied) (Fig. 3.6.7.8). Three biological assemblies are constructed for this system. One biological unit contains only the dimeric enzyme (Fig. 3.6.7.8*b*), the second contains the dimeric enzyme with one partially occupied conformation of the inhibitor (Fig. 3.6.7.8*c*) and the third contains the dimeric enzyme with the second partially occupied conformation of the inhibitor (Fig. 3.6.7.8*d*). There are alternative conformations of the side chains in the enzyme that correlate with the binding mode of the inhibitor.

3.6.7.5.2. *Secondary structure*

The data items in these categories are as follows:

(*a*) STRUCT_CONF_TYPE

```
• _struct_conf_type.id
  _struct_conf_type.criteria
  _struct_conf_type.reference
```

(*b*) STRUCT_CONF

```
• _struct_conf.id
  _struct_conf.beg_label_asym_id
        → _atom_site.label_asym_id
  _struct_conf.beg_label_comp_id
        → _atom_site.label_comp_id
  _struct_conf.beg_label_seq_id
        → _atom_site.label_seq_id
  _struct_conf.beg_auth_asym_id
        → _atom_site.auth_asym_id
  _struct_conf.beg_auth_comp_id
        → _atom_site.auth_comp_id
  _struct_conf.beg_auth_seq_id
        → _atom_site.auth_seq_id
  _struct_conf.conf_type_id
        → _struct_conf_type.id
  _struct_conf.details
  _struct_conf.end_label_asym_id
        → _atom_site.label_asym_id
  _struct_conf.end_label_comp_id
        → _atom_site.label_comp_id
  _struct_conf.end_label_seq_id
        → _atom_site.label_seq_id
  _struct_conf.end_auth_asym_id
        → _atom_site.auth_asym_id
  _struct_conf.end_auth_comp_id
        → _atom_site.auth_comp_id
  _struct_conf.end_auth_seq_id
        → _atom_site.auth_seq_id
```

The bullet (•) indicates a category key. The arrow (→) is a reference to a parent data item.

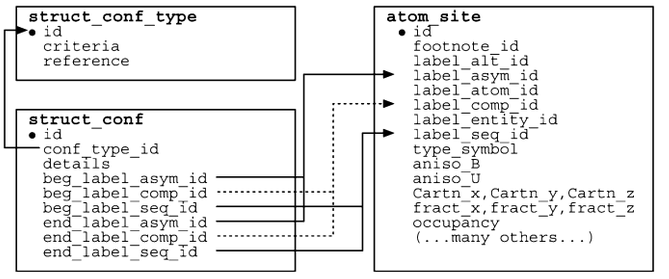

Fig. 3.6.7.9. The family of categories used to describe secondary structure. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

Example 3.6.7.9. *Secondary structure in an HIV-1 protease structure (PDB 5HVP) described with data items in the STRUCT_CONF_TYPE and STRUCT_CONF categories.*

```
loop_
_struct_conf_type.id
_struct_conf_type.criteria
   HELX_RH_AL_P   'author judgement'
   STRN           'author judgement'
   TURN_TY1_P     'author judgement'
   TURN_TY1P_P    'author judgement'
   TURN_TY2_P     'author judgement'
   TURN_TY2P_P    'author judgement'

loop_
_struct_conf.id
_struct_conf.conf_type_id
_struct_conf.beg_label_comp_id
_struct_conf.beg_label_asym_id
_struct_conf.beg_label_seq_id
_struct_conf.end_label_comp_id
_struct_conf.end_label_asym_id
_struct_conf.end_label_seq_id
   HELX1  HELX_RH_AL_P  ARG  A   87  GLN  A   92
   HELX2  HELX_RH_AL_P  ARG  B  287  GLN  B  292
   STRN1  STRN          PRO  A    1  LEU  A    5
   STRN2  STRN          CYS  B  295  PHE  B  299
   STRN3  STRN          CYS  A   95  PHE  A  299
   STRN4  STRN          PRO  B  201  LEU  B  205
   TURN1  TURN_TY1P_P   ILE  A   15  GLN  A   18
   TURN2  TURN_TY2_P    GLY  A   49  GLY  A   52
   TURN3  TURN_TY1P_P   ILE  A   55  HIS  A   69
   TURN4  TURN_TY1_P    THR  A   91  GLY  A   94
```

The primary structure of a macromolecule is defined by the sequence of the components (amino acids, nucleic acids or sugars) in the polymer chain. The polymer chains assume conformations based on the torsion angles adopted by the rotatable bonds in the polymer backbone; the resulting conformations are referred to as the secondary structure of the polymer. Several patterns of values of backbone torsion angles have been described and given names, such as α-helix, β-strand, turn and coil for proteins, and A-, B- and Z-helix for nucleic acids.

In the mmCIF dictionary, these secondary structures are described in the STRUCT_CONF and STRUCT_CONF_TYPE categories. Note that the data items in these categories describe only the secondary structure; the tertiary organization of β-strands into β-sheets is described in the STRUCT_SHEET_* categories. There are no data items for describing the tertiary organization of α-helices or nucleic acids in the current version of the mmCIF dictionary.

The relationships between categories used to describe secondary structure are shown in Fig. 3.6.7.9.

The type of the secondary structure is specified in the STRUCT_CONF_TYPE category, along with the criteria used to identify it. The range of monomers assigned to each secondary-structure element is given in the STRUCT_CONF category.

The allowed values for the data item **_struct_conf_type.id** cover most types of protein and nucleic acid secondary structure (Example 3.6.7.9). The criteria that define the secondary structure may be given using the data item **_struct_conf_type.criteria**. **_struct_conf_type.reference** can be used to specify a reference to the literature in which the criteria are explained in more detail.

The residues that define the beginning and end of each region of secondary structure are identified with the appropriate ***_asym**, ***_comp** and ***_seq** identifiers. The standard labelling system or the author's alternative labelling system may be used. The identification of the residues assigned to each region of secondary structure is linked to the labelling information in the ATOM_SITE category. Unusual features of a conformation may be described using **_struct_conf.details**.

3.6.7.5.3. *Structural interactions*

The data items in these categories are as follows:

(*a*) STRUCT_CONN_TYPE

```
• _struct_conn_type.id
  _struct_conn_type.criteria
  _struct_conn_type.reference
```

(*b*) STRUCT_CONN

```
• _struct_conn.id
  _struct_conn.conn_type_id
        → _struct_conn_type.id
  _struct_conn.details
  _struct_conn.ptnr1_label_alt_id
        → _atom_sites_alt.id
  _struct_conn.ptnr1_label_asym_id
        → _atom_site.label_asym_id
  _struct_conn.ptnr1_label_atom_id
        → _chem_comp_atom.atom_id
  _struct_conn.ptnr1_label_comp_id
        → _atom_site.label_comp_id
  _struct_conn.ptnr1_label_seq_id
        → _atom_site.label_seq_id
  _struct_conn.ptnr1_auth_asym_id
        → _atom_site.auth_asym_id
  _struct_conn.ptnr1_auth_atom_id
        → _atom_site.auth_atom_id
  _struct_conn.ptnr1_auth_comp_id
        → _atom_site.auth_comp_id
  _struct_conn.ptnr1_auth_seq_id
        → _atom_site.auth_seq_id
  _struct_conn.ptnr1_role
  _struct_conn.ptnr1_symmetry
  _struct_conn.ptnr2_label_alt_id
        → _atom_sites_alt.id
  _struct_conn.ptnr2_label_asym_id
        → _atom_site.label_asym_id
  _struct_conn.ptnr2_label_atom_id
        → _chem_comp_atom.atom_id
  _struct_conn.ptnr2_label_comp_id
        → _atom_site.label_comp_id
  _struct_conn.ptnr2_label_seq_id
        → _atom_site.label_seq_id
  _struct_conn.ptnr2_auth_asym_id
        → _atom_site.auth_asym_id
  _struct_conn.ptnr2_auth_atom_id
        → _atom_site.auth_atom_id
  _struct_conn.ptnr2_auth_comp_id
        → _atom_site.auth_comp_id
  _struct_conn.ptnr2_auth_seq_id
        → _atom_site.auth_seq_id
  _struct_conn.ptnr2_role
  _struct_conn.ptnr2_symmetry
```

The bullet (•) indicates a category key. The arrow (→) is a reference to a parent data item.

The structural interactions that are described with data items in the STRUCT_CONN family of categories are the tertiary result of a structure determination, not the chemical connectivity of the components of the structure. In general, the interactions described using the STRUCT_CONN data items are noncovalent, such as hydrogen bonds, salt bridges and metal coordination.

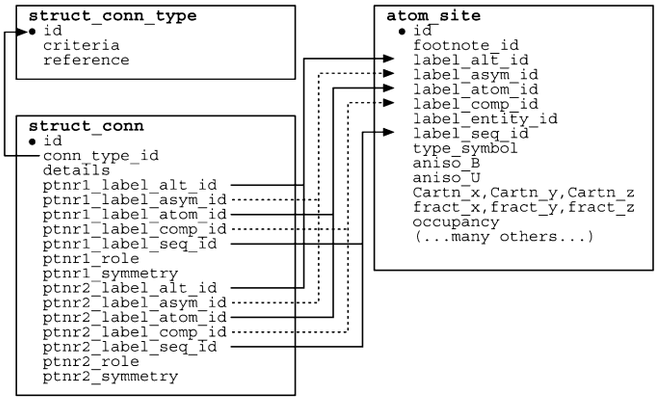

Fig. 3.6.7.10. The family of categories used to describe structural interactions such as hydrogen bonding, salt bridges and disulfide bridges. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

It is useful to think of the structure interactions given in CHEM_COMP_BOND, CHEM_LINK and ENTITY_LINK as the covalent interactions that are known in advance of the structure determination because the chemistry of the components is well defined. Literature or calculated values for these interactions are often used as restraints during the refinement. In contrast, the structural interactions described in the STRUCT_CONN family of categories are not known in advance and are part of the results of the structure determination.

This distinction only holds approximately, as there are clearly bonds, such as disulfide links, that are covalent and usually restrained during the refinement but that are also a result of the folding of the protein revealed by the structure determination, and thus should be described using STRUCT_CONN data items.

In general, the STRUCT_CONN data items would not be used to list all the structure interactions. Instead, the author of the mmCIF would use the STRUCT_CONN data items to identify and annotate only the structural interactions worthy of discussion. The relationships between categories used to describe structural interactions are shown in Fig. 3.6.7.10.

Structural interactions such as hydrogen bonds, salt bridges and disulfide bridges can be described in the STRUCT_CONN category. The type of each interaction and the criteria used to identify the interaction can be specified in the STRUCT_CONN_TYPE category (Example 3.6.7.10).

The atoms participating in each interaction are arbitrarily labelled as 'partner 1' and 'partner 2'. Each is identified by the ***_alt**, ***_asym**, ***_atom**, ***_comp** and ***_seq** constituents of the corresponding atom-site label. The role of each partner in the interaction (*e.g.* donor, acceptor) may be specified, and any crystallographic symmetry operation needed to transform the atom from the position given in the ATOM_SITE list to the position where the interaction occurs can be given. The atoms participating in the interaction may also be identified using an alternative labelling scheme if the author has supplied one.

Unusual aspects of the interaction may be discussed in **_struct_conn.details**. The general type of an interaction can be indicated using **_struct_conn.conn_type_id**, which references one of the standard types described using data items in the STRUCT_CONN_TYPE category.

The specific types of structural connection that may be recorded are those allowed for **_struct_conn_type.id**, namely covalent and hydrogen bonds, ionic (salt-bridge) interactions, disulfide

Example 3.6.7.10. *A hypothetical salt bridge and hydrogen bond described with data items in the* STRUCT_CONN_TYPE *and* STRUCT_CONN *categories.*

```
loop_
_struct_conn_type.id
_struct_conn_type.criteria
   saltbr
; negative to positive distance > 2.5 Angstroms,
  < 3.2 Angstroms
;
   hydrog
; N-O distance > 2.5 Angstroms, < 3.5 Angstroms,
  N-O-C  angle < 120 degrees
;

loop_
_struct_conn.id
_struct_conn.conn_type_id
_struct_conn.ptnr1_label_comp_id
_struct_conn.ptnr1_label_asym_id
_struct_conn.ptnr1_label_seq_id
_struct_conn.ptnr1_label_atom_id
_struct_conn.ptnr1_role
_struct_conn.ptnr1_symmetry
_struct_conn.ptnr2_label_comp_id
_struct_conn.ptnr2_label_asym_id
_struct_conn.ptnr2_label_seq_id
_struct_conn.ptnr2_label_atom_id
_struct_conn.ptnr2_role
_struct_conn.ptnr2_symmetry
  C1 saltbr ARG  A  87 NZ1 positive 1_555
            GLU  A  92 OE1 negative 1_555
  C2 hydrog ARG  B 287 N   donor    1_555
            GLY  B 292 O   acceptor 1_555
```

links, metal coordination, mismatched base pairs, covalent residue modifications and covalent modifications of nucleotide bases, sugars or phosphates. The criteria used to define each interaction may be described in detail using `_struct_conn_type.criteria` or a literature reference to the criteria can be given in `_struct_conn_type.reference`.

3.6.7.5.4. *Structural features of monomers*

The data items in these categories are as follows:

(*a*) STRUCT_MON_DETAILS

- `_struct_mon_details.entry_id`
  → `_entry.id`

`_struct_mon_details.prot_cis`
`_struct_mon_details.RSCC`
`_struct_mon_details.RSR`

(*b*) STRUCT_MON_NUCL

- `_struct_mon_nucl.label_alt_id`
  → `_atom_sites_alt.id`
- `_struct_mon_nucl.label_asym_id`
  → `_atom_site.label_asym_id`
- `_struct_mon_nucl.label_comp_id`
  → `_atom_site.label_comp_id`
- `_struct_mon_nucl.label_seq_id`
  → `_atom_site.label_seq_id`

`_struct_mon_nucl.alpha`
`_struct_mon_nucl.auth_asym_id`
  → `_atom_site.auth_asym_id`
`_struct_mon_nucl.auth_comp_id`
  → `_atom_site.auth_comp_id`
`_struct_mon_nucl.auth_seq_id`
  → `_atom_site.auth_seq_id`
`_struct_mon_nucl.beta`
`_struct_mon_nucl.chi1`
`_struct_mon_nucl.chi2`
`_struct_mon_nucl.delta`
`_struct_mon_nucl.details`
`_struct_mon_nucl.epsilon`
`_struct_mon_nucl.gamma`
`_struct_mon_nucl.mean_B_all`
`_struct_mon_nucl.mean_B_base`
`_struct_mon_nucl.mean_B_phos`
`_struct_mon_nucl.mean_B_sugar`
`_struct_mon_nucl.nu0`
`_struct_mon_nucl.nu1`
`_struct_mon_nucl.nu2`
`_struct_mon_nucl.nu3`
`_struct_mon_nucl.nu4`
`_struct_mon_nucl.P`
`_struct_mon_nucl.RSCC_all`
`_struct_mon_nucl.RSCC_base`
`_struct_mon_nucl.RSCC_phos`
`_struct_mon_nucl.RSCC_sugar`
`_struct_mon_nucl.RSR_all`
`_struct_mon_nucl.RSR_base`
`_struct_mon_nucl.RSR_phos`
`_struct_mon_nucl.RSR_sugar`
`_struct_mon_nucl.tau0`
`_struct_mon_nucl.tau1`
`_struct_mon_nucl.tau2`
`_struct_mon_nucl.tau3`
`_struct_mon_nucl.tau4`
`_struct_mon_nucl.taum`
`_struct_mon_nucl.zeta`

(*c*) STRUCT_MON_PROT

- `_struct_mon_prot.label_alt_id`
  → `_atom_sites_alt.id`
- `_struct_mon_prot.label_asym_id`
  → `_atom_site.label_asym_id`
- `_struct_mon_prot.label_comp_id`
  → `_atom_site.label_comp_id`
- `_struct_mon_prot.label_seq_id`
  → `_atom_site.label_seq_id`

`_struct_mon_prot.auth_asym_id`
  → `_atom_site.auth_asym_id`
`_struct_mon_prot.auth_comp_id`
  → `_atom_site.auth_comp_id`
`_struct_mon_prot.auth_seq_id`
  → `_atom_site.auth_seq_id`
`_struct_mon_prot.chi1`
`_struct_mon_prot.chi2`
`_struct_mon_prot.chi3`
`_struct_mon_prot.chi4`
`_struct_mon_prot.chi5`
`_struct_mon_prot.details`
`_struct_mon_prot.RSCC_all`
`_struct_mon_prot.RSCC_main`
`_struct_mon_prot.RSCC_side`
`_struct_mon_prot.RSR_all`
`_struct_mon_prot.RSR_main`
`_struct_mon_prot.RSR_side`
`_struct_mon_prot.mean_B_all`
`_struct_mon_prot.mean_B_main`
`_struct_mon_prot.mean_B_side`
`_struct_mon_prot.omega`
`_struct_mon_prot.phi`
`_struct_mon_prot.psi`

(*d*) STRUCT_MON_PROT_CIS

- `_struct_mon_prot_cis.label_alt_id`
  → `_atom_sites_alt.id`
- `_struct_mon_prot_cis.label_asym_id`
  → `_atom_site.label_asym_id`
- `_struct_mon_prot_cis.label_comp_id`
  → `_atom_site.label_comp_id`
- `_struct_mon_prot_cis.label_seq_id`
  → `_atom_site.label_seq_id`

`_struct_mon_prot_cis.auth_asym_id`
  → `_atom_site.auth_asym_id`
`_struct_mon_prot_cis.auth_comp_id`
  → `_atom_site.auth_comp_id`
`_struct_mon_prot_cis.auth_seq_id`
  → `_atom_site.auth_seq_id`

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.

Most macromolecules have complex structures which contain regions of well defined structure and flexible regions that are difficult to model accurately. Overall measures of the quality of a model, such as the standard crystallographic *R* factors, do not represent the local quality of the model. During the development of

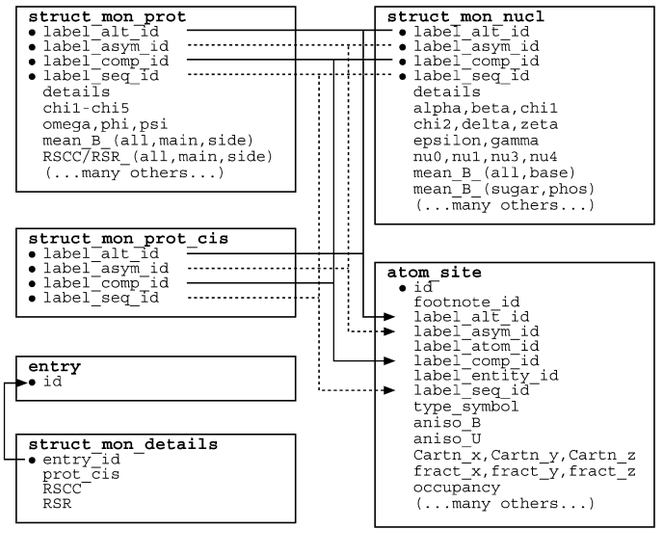

Fig. 3.6.7.11. The family of categories used to describe the structural features of monomers. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

Example 3.6.7.11. *A hypothetical example of the structural features of a single protein residue described with data items in the* STRUCT_MON_PROT *category.*

```
_struct_mon_prot.label_comp_id          ARG
_struct_mon_prot.label_seq_id           35
_struct_mon_prot.label_asym_id          A
_struct_mon_prot.label_alt_id           .
_struct_mon_prot.chi1                 -67.9
_struct_mon_prot.chi2                -174.7
_struct_mon_prot.chi3                 -67.7
_struct_mon_prot.chi4                 -86.3
_struct_mon_prot.chi5                   4.2
_struct_mon_prot.RSCC_all               0.90
_struct_mon_prot.RSR_all                0.18
_struct_mon_prot.mean_B_all            30.0
_struct_mon_prot.mean_B_main           25.0
_struct_mon_prot.mean_B_side           35.1
_struct_mon_prot.omega                180.1
_struct_mon_prot.phi                  -60.3
_struct_mon_prot.psi                  -46.0
```

the mmCIF dictionary, it was found that the biological crystallography community felt that mmCIF should contain data items that allowed the local quality of the model to be recorded: these data items are found in the categories STRUCT_MON_DETAILS, STRUCT_MON_NUCL (for nucleotides), and STRUCT_MON_PROT and STRUCT_MON_PROT_CIS (for proteins). Using these categories, quantities that reflect the local quality of the structure, such as isotropic displacement factors, real-space *R* factors and real-space correlation coefficients, can be given at the monomer and submonomer levels.

In addition, these categories can be used to record the conformation of the structure at the monomer level by listing side-chain torsion angles. These values can be derived from the atom coordinate list, so it would not be common practice to include them in an mmCIF for archiving a structure unless it was to highlight conformations that deviate significantly from expected values (Engh & Huber, 1991). However, there are applications, such as comparative studies across a number of independent determinations of the same structure, where it would be useful to store torsion-angle information without having to recalculate it each time it is needed.

The relationships between the categories used to describe the structural features of monomers are shown in Fig. 3.6.7.11.

Three indicators of the quality of a structure at the local level are included in this version of the dictionary: the mean displacement (*B*) factor, the real-space correlation coefficient (Jones *et al.*, 1991) and the real-space *R* factor (Brändén & Jones, 1990). Other indicators are likely to be added as they become available. In the current version of the dictionary, these metrics can be given at the monomer level, or at the levels of main- and side-chain for proteins, or base, phosphate and sugar for nucleic acids (Altona & Sundaralingam, 1972).

The variables used when calculating real-space correlation coefficients and real-space *R* factors, such as the coefficients used to calculate the map being evaluated or the radii used for including points in a calculation, can be recorded using the data items `_struct_mon_details.RSC` and `_struct_mon_details.RSR`.

These data items are also provided for recording the full conformation of the macromolecule, using a full set of data items for the torsion angles of both proteins and nucleic acids. Although one could use these data items to describe the whole macromolecule, it is more likely that they would be used to highlight regions of the structure that deviate from expected values (Example 3.6.7.11). Deviations from expected values could imply inaccuracies in the model in poorly defined parts of the structure, but in some cases nonstandard torsion angles are found in very well defined regions and are essential to the proper configurations of active sites or ligand binding pockets.

A special case of nonstandard conformation is the occurrence of *cis* peptides in proteins. As the *cis* conformation occurs quite often, the category STRUCT_MON_PROT_CIS is provided so that an explicit list can be made of *cis* peptides. The related data item `_struct_mon_details.prot_cis` allows an author to specify how far a peptide torsion angle can deviate from the expected value of 0.0 and still be considered to be *cis*.

In these categories, properties are listed by residue rather than by individual atom. The only label components needed to identify the residue are `*_alt`, `*_asym`, `*_comp` and `*_seq`. If the author has provided an alternative labelling system, this can also be used. Since the analysis is by individual residue, there is no need to specify symmetry operations that might be needed to move one residue so that it is next to another.

3.6.7.5.5. *Noncrystallographic symmetry*

Data items in these categories are as follows:

(*a*) STRUCT_NCS_ENS

- `_struct_ncs_ens.id`
  `_struct_ncs_ens.details`
  `_struct_ncs_ens.point_group`

(*b*) STRUCT_NCS_ENS_GEN

- `_struct_ncs_ens_gen.dom_id_1`
  → `_struct_ncs_dom.id`
- `_struct_ncs_ens_gen.dom_id_2`
  → `_struct_ncs_dom.id`
- `_struct_ncs_ens_gen.ens_id`
  → `_struct_ncs_ens.id`
- `_struct_ncs_ens_gen.oper_id`
  → `_struct_ncs_oper.id`

(*c*) STRUCT_NCS_DOM

- `_struct_ncs_dom.id`
  `_struct_ncs_dom.details`

(*d*) STRUCT_NCS_DOM_LIM

- `_struct_ncs_dom_lim.beg_label_alt_id`
  → `_atom_sites_alt.id`
- `_struct_ncs_dom_lim.beg_label_asym_id`
  → `_atom_site.label_asym_id`
- `_struct_ncs_dom_lim.beg_label_comp_id`
  → `_atom_site.label_comp_id`

- **_struct_ncs_dom_lim.beg_label_seq_id**
  → **_atom_site.label_seq_id**
- **_struct_ncs_dom_lim.dom_id**
- **_struct_ncs_dom_lim.end_label_alt_id**
  → **_atom_sites_alt.id**
- **_struct_ncs_dom_lim.end_label_asym_id**
  → **_atom_site.label_asym_id**
- **_struct_ncs_dom_lim.end_label_comp_id**
  → **_atom_site.label_comp_id**
- **_struct_ncs_dom_lim.end_label_seq_id**
  → **_atom_site.label_seq_id**

_struct_ncs_dom_lim.beg_auth_asym_id
→ **_atom_site.auth_asym_id**
_struct_ncs_dom_lim.beg_auth_comp_id
→ **_atom_site.auth_comp_id**
_struct_ncs_dom_lim.beg_auth_seq_id
→ **_atom_site.auth_seq_id**
_struct_ncs_dom_lim.end_auth_asym_id
→ **_atom_site.auth_asym_id**
_struct_ncs_dom_lim.end_auth_comp_id
→ **_atom_site.auth_comp_id**
_struct_ncs_dom_lim.end_auth_seq_id
→ **_atom_site.auth_seq_id**

(*e*) STRUCT_NCS_OPER

- **_struct_ncs_oper.id**

_struct_ncs_oper.code
_struct_ncs_oper.details
_struct_ncs_oper.matrix[1][1]
_struct_ncs_oper.matrix[1][2]
_struct_ncs_oper.matrix[1][3]
_struct_ncs_oper.matrix[2][1]
_struct_ncs_oper.matrix[2][2]
_struct_ncs_oper.matrix[2][3]
_struct_ncs_oper.matrix[3][1]
_struct_ncs_oper.matrix[3][2]
_struct_ncs_oper.matrix[3][3]
_struct_ncs_oper.vector[1]
_struct_ncs_oper.vector[2]
_struct_ncs_oper.vector[3]

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.

Biological macromolecular complexes may be built from domains related by symmetry transformations other than those arising from the crystal lattice symmetry. These domains are not necessarily discrete molecular entities: they may be composed of one or more segments of a single polypeptide or nucleic acid chain, of segments from more than one chain, or of small-molecule components of the structure. The categories above allow the distinct domains that participate in ensembles of structural elements related by noncrystallographic symmetry to be listed and described in detail. The relationships between categories used to describe noncrystallographic symmetry are shown in Fig. 3.6.7.12.

In the mmCIF model of noncrystallographic symmetry, the highest level of organization is the ensemble, which corresponds to the complete symmetry-related aggregate (*e.g.* tetramer, icosahedron). An identifier is given to the ensemble using the data item **_struct_ncs_ens.id**.

The symmetry-related elements within the ensemble are referred to as domains. The elements of structure that are to be considered part of the domain are specified using the data items in the STRUCT_NCS_DOM and STRUCT_NCS_DOM_LIM categories. By using the STRUCT_NCS_DOM_LIM data items appropriately, domains can be defined to include ranges of polypeptide chain or nucleic acid strand, bound ligands or cofactors, or even bound solvent molecules. Note that the category keys for STRUCT_NCS_DOM_LIM include the domain ID and the range specifiers. Thus a single domain may be composed of any number of ranges of elements.

Finally, the ensemble is generated from the domains using the rotation matrix and translation vector specified by data items in

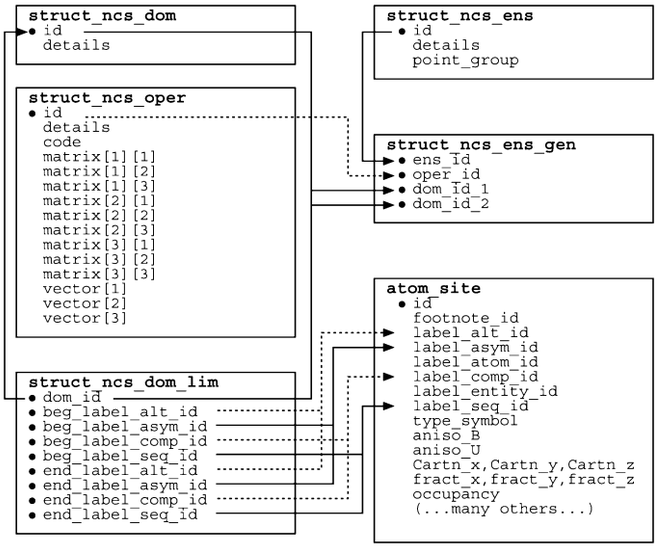

Fig. 3.6.7.12. The family of categories used to describe noncrystallographic symmetry. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

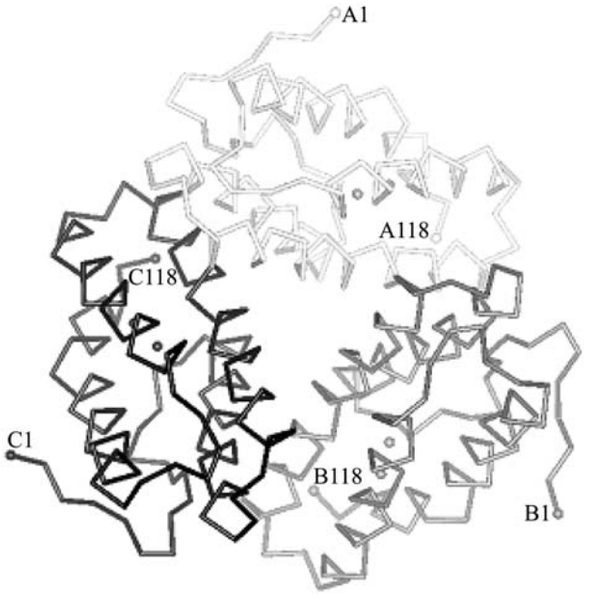

Fig. 3.6.7.13. Noncrystallographic symmetry in the structure of trimeric haemerythrin (PDB 1HR3) to be described with data items in the STRUCT_NCS_ENS, STRUCT_NCS_ENS_GEN, STRUCT_NCS_DOM and STRUCT_NCS_DOM_LIM categories.

the STRUCT_NCS_OPER category, which are referenced by the data items in the STRUCT_NCS_ENS_GEN category. There are data items appropriate for two common methods of describing noncrystallographic symmetry:

(1) In the first method, the coordinate list includes all copies of domains related by noncrystallographic symmetry and the aim is to describe the relationships between domains in the ensemble; in this case the data items in STRUCT_NCS_ENS_GEN specify a pair of domains and reference the appropriate operator in STRUCT_NCS_OPER. This method is indicated by giving the data item **_struct_ncs_oper.code** the value `given`.

(2) In the second method, the coordinate list contains only one copy of the domain and the aim is to generate the entire ensemble; in this case the data items in STRUCT_NCS_ENS_GEN

Example 3.6.7.12. *Noncrystallographic symmetry in the structure of trimeric haemerythrin (PDB 1HR3) described with data items in the STRUCT_NCS_ENS, STRUCT_NCS_ENS_GEN, STRUCT_NCS_DOM and STRUCT_NCS_DOM_LIM categories. For brevity, the data items in the STRUCT_NCS_OPER category are not shown.*

```
_struct_ncs_ens.id                    trimer
_struct_ncs_ens.point_group           3

loop_
_struct_ncs_ens_gen.ens_id
_struct_ncs_ens_gen.dom_id_1
_struct_ncs_ens_gen.dom_id_2
_struct_ncs_ens_gen.oper_id
 trimer  chain_A  chain_B  1
 trimer  chain_A  chain_C  2

loop_
_struct_ncs_dom.id
 chain_A    chain_B    chain_C

loop_
_struct_ncs_dom_lim.dom_id
_struct_ncs_dom_lim.beg_label_asym_id
_struct_ncs_dom_lim.beg_label_comp_id
_struct_ncs_dom_lim.beg_label_seq_id
_struct_ncs_dom_lim.beg_label_alt_id
_struct_ncs_dom_lim.end_label_asym_id
_struct_ncs_dom_lim.end_label_comp_id
_struct_ncs_dom_lim.end_label_seq_id
_struct_ncs_dom_lim.end_label_alt_id
 chain_A  A  ala  1  .  A  ala  118  .
 chain_B  B  ala  1  .  B  ala  118  .
 chain_C  C  ala  1  .  C  ala  118  .
```

specify a pair of domains and reference the appropriate operator in STRUCT_NCS_OPER, but now the data item **_struct_ncs_oper.code** is given the value `generate`.

Noncrystallographic symmetry in a trimeric molecule is shown in Fig. 3.6.7.13 and described in Example 3.6.7.12.

3.6.7.5.6. *External databases*

The data items in these categories are as follows:

(*a*) STRUCT_REF

```
• _struct_ref.id
  _struct_ref.biol_id
        → _struct_biol.id
  _struct_ref.db_code
  _struct_ref.db_name
  _struct_ref.details
  _struct_ref.entity_id
        → _entity.id
  _struct_ref.seq_align
  _struct_ref.seq_dif
```

(*b*) STRUCT_REF_SEQ

```
• _struct_ref_seq.align_id
  _struct_ref_seq.db_align_beg
  _struct_ref_seq.db_align_end
  _struct_ref_seq.details
  _struct_ref_seq.ref_id
        → _struct_ref.id
  _struct_ref_seq.seq_align_beg
        → _entity_poly_seq.num
  _struct_ref_seq.seq_align_end
        → _entity_poly_seq.num
```

(*c*) STRUCT_REF_SEQ_DIF

```
• _struct_ref_seq_dif.align_id
        → _struct_ref_seq.align_id
• _struct_ref_seq_dif.seq_num
        → _entity_poly_seq.num
  _struct_ref_seq_dif.db_mon_id
        → _chem_comp.id
  _struct_ref_seq_dif.details
  _struct_ref_seq_dif.mon_id
        → _chem_comp.id
```

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.

Data items in the STRUCT_REF category allow the author of an mmCIF to provide references to information in external databases that is relevant to the entities or biological units described in the mmCIF. For example, the database entry for a protein or nucleic acid sequence could be referenced and any differences between the sequence of the macromolecule whose structure is reported in the mmCIF and the sequence of the related entry in the external database can be recorded. Alternatively, references to external database entries can be used to record the relationship of the structure reported in the mmCIF to structures already reported in the literature, for example by referring to previously determined structures of the same or a similar protein, or to a small-molecule structure determination of a bound inhibitor or cofactor. STRUCT_REF data items are not intended to be used to reference a database entry for the structure in the mmCIF itself (this would be the role of data items in the DATABASE_2 category), but it would not be formally incorrect to do so.

When the data items in these categories are used to provide references to external database entries describing the sequence of a polymer, data items from all three categories could be used. The value of the data item **_struct_ref.seq_align** is used to indicate whether the correspondence between the sequence of the entity or biological unit in the mmCIF and the sequence in the related external database entry is `complete` or `partial`. If the value is `partial`, the region (or regions) of the alignment may be identified using data items in the STRUCT_REF_SEQ category. Comments on the alignment may be given in **_struct_ref_seq.details** (Example 3.6.7.13).

The value of the data item **_struct_ref.seq_dif** is used to indicate whether the two sequences contain point differences. If the value is `yes`, the differences may be identified and annotated using data items in the STRUCT_REF_SEQ_DIF category. Comments on specific point differences may be recorded in **_struct_ref_seq_dif.details**.

Example 3.6.7.13. *The relationship of the sequence of the protein PDB 5HVP to a sequence in an external database described with data items in the STRUCT_REF and STRUCT_REF_SEQ categories.*

```
loop_
  _struct_ref.id
  _struct_ref.biol_id
  _struct_ref.entity_id
  _struct_ref.db_name
  _struct_ref.db_code
  _struct_ref.seq_align
  _struct_ref.seq_dif
 seq_pdb     1  .  PDB      5HVP      .         .
 seq_genbank .  1  GenBank  AAG30358  complete  yes

loop_
  _struct_ref_seq.align_id
  _struct_ref_seq.ref_id
  _struct_ref_seq.seq_align_beg
  _struct_ref_seq.seq_align_end
  _struct_ref_seq.db_align_beg
  _struct_ref_seq.db_align_end
  _struct_ref_seq.details
    align_seq_pdb_genbank seq_genbank 1 99 24 122
; The genbank reference is to the sequence of
  residues 1-376 of the viral pol 1 polypeptide;
  the protease is proteolytically released from
  this precursor during viral maturation.
;
```

References do not have to be to entries in databases of sequences: any external database can be referenced. For other kinds of databases, only the data items in the STRUCT_REF category would usually be used. The element of the structure that is referenced could be either an entity or a biological unit, that is, either a building block of the structure or a structurally meaningful assembly of those building blocks. Since the identification of the part of the structure being linked to an entry in an external database can be made using either `_struct_ref.biol_id` or `_struct_ref.entity_id`, and since any part of the structure could be linked to any number of entries in external databases, the data item `_struct_ref.id` was introduced as the category key.

3.6.7.5.7. *β-sheets*

Data items in these categories are as follows:

(*a*) STRUCT_SHEET

```
• _struct_sheet.id
  _struct_sheet.details
  _struct_sheet.number_strands
  _struct_sheet.type
```

(*b*) STRUCT_SHEET_TOPOLOGY

```
• _struct_sheet_topology.range_id_1
        → _struct_sheet_range.id
• _struct_sheet_topology.range_id_2
        → _struct_sheet_range.id
• _struct_sheet_topology.sheet_id
        → _struct_sheet.id
  _struct_sheet_topology.offset
  _struct_sheet_topology.sense
```

(*c*) STRUCT_SHEET_RANGE

```
• _struct_sheet_range.id
• _struct_sheet_range.sheet_id
        → _struct_sheet.id
  _struct_sheet_range.beg_label_asym_id
        → _struct_asym.id
  _struct_sheet_range.beg_label_comp_id
        → _chem_comp.id
  _struct_sheet_range.beg_label_seq_id
        → _atom_site.label_seq_id
  _struct_sheet_range.end_label_asym_id
        → _struct_asym.id
  _struct_sheet_range.end_label_comp_id
        → _chem_comp.id
  _struct_sheet_range.end_label_seq_id
        → _atom_site.label_seq_id
  _struct_sheet_range.beg_auth_asym_id
        → _atom_site.auth_atom_id
  _struct_sheet_range.beg_auth_comp_id
        → _atom_site.auth_comp_id
  _struct_sheet_range.beg_auth_seq_id
        → _atom_site.auth_seq_id
  _struct_sheet_range.end_auth_asym_id
        → _atom_site.auth_atom_id
  _struct_sheet_range.end_auth_comp_id
        → _atom_site.auth_comp_id
  _struct_sheet_range.end_auth_seq_id
        → _atom_site.auth_seq_id
  _struct_sheet_range.symmetry
```

(*d*) STRUCT_SHEET_ORDER

```
• _struct_sheet_order.range_id_1
        → _struct_sheet_range.id
• _struct_sheet_order.range_id_2
        → _struct_sheet_range.id
• _struct_sheet_order.sheet_id
        → _struct_sheet.id
  _struct_sheet_order.offset
  _struct_sheet_order.sense
```

(*e*) STRUCT_SHEET_HBOND

```
• _struct_sheet_hbond.range_id_1
        → _struct_sheet_range.id
• _struct_sheet_hbond.range_id_2
        → _struct_sheet_range.id
• _struct_sheet_hbond.sheet_id
        → _struct_sheet.id
  _struct_sheet_hbond.range_1_beg_label_atom_id
        → _atom_site.label_atom_id
  _struct_sheet_hbond.range_1_beg_label_seq_id
        → _atom_site.label_seq_id
  _struct_sheet_hbond.range_1_end_label_atom_id
        → _atom_site.label_atom_id
  _struct_sheet_hbond.range_1_end_label_seq_id
        → _atom_site.label_seq_id
  _struct_sheet_hbond.range_2_beg_label_atom_id
        → _atom_site.label_atom_id
  _struct_sheet_hbond.range_2_beg_label_seq_id
        → _atom_site.label_seq_id
  _struct_sheet_hbond.range_2_end_label_atom_id
        → _atom_site.label_atom_id
  _struct_sheet_hbond.range_2_end_label_seq_id
        → _atom_site.label_seq_id
  _struct_sheet_hbond.range_1_beg_auth_atom_id
        → _atom_site.auth_atom_id
  _struct_sheet_hbond.range_1_beg_auth_seq_id
        → _atom_site.auth_seq_id
  _struct_sheet_hbond.range_1_end_auth_atom_id
        → _atom_site.auth_atom_id
  _struct_sheet_hbond.range_1_end_auth_seq_id
        → _atom_site.auth_seq_id
  _struct_sheet_hbond.range_2_beg_auth_atom_id
        → _atom_site.auth_atom_id
  _struct_sheet_hbond.range_2_beg_auth_seq_id
        → _atom_site.auth_seq_id
  _struct_sheet_hbond.range_2_end_auth_atom_id
        → _atom_site.auth_atom_id
  _struct_sheet_hbond.range_2_end_auth_seq_id
        → _atom_site.auth_seq_id
```

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.

Different methods of describing β-sheets are in widespread use. The mmCIF dictionary provides data items for two methods and it is anticipated that future versions of the dictionary could cover others. The model used in the STRUCT_SHEET_TOPOLOGY category is the simpler of the two. It is a convenient shorthand for describing the topology, but it does not provide details about strand registration and it is not suitable for describing sheets that contain strands from more than one polypeptide. A more general model is provided by the linked data items in the STRUCT_SHEET_RANGE, STRUCT_SHEET_ORDER and STRUCT_SHEET_HBOND categories. For both methods of representing β-sheets, data items in the parent category STRUCT_SHEET can be used to provide an identifier for each sheet, a free-text description of its type, the number of participating strands and a free-text description of any peculiar aspects of the sheet. The relationships between categories used to describe β-sheets are shown in Fig. 3.6.7.14.

In the description of β-sheet topology based on the STRUCT_SHEET_TOPOLOGY category, the strand that occurs first in the polypeptide chain is numbered 1. Subsequent strands are described by their position in the sheet relative to the previous strand (+1, −3 *etc.*) and by their orientation relative to the previous strand (parallel or antiparallel).

While writing this chapter, a few errors in the mmCIF dictionary were discovered. The use of `_struct_sheet_topology.range_id_1` and `*_2` as pointers to the residues participating in β-sheets is one; the correct data items should be `_struct_sheet_topology.comp_id_1` and `*_2`, and these data items should be pointers to `_atom_site.label_comp_id`. This error will be corrected in future versions of the dictionary. As the data model encoded in the current version of the dictionary is incorrect, no example of its use is given.

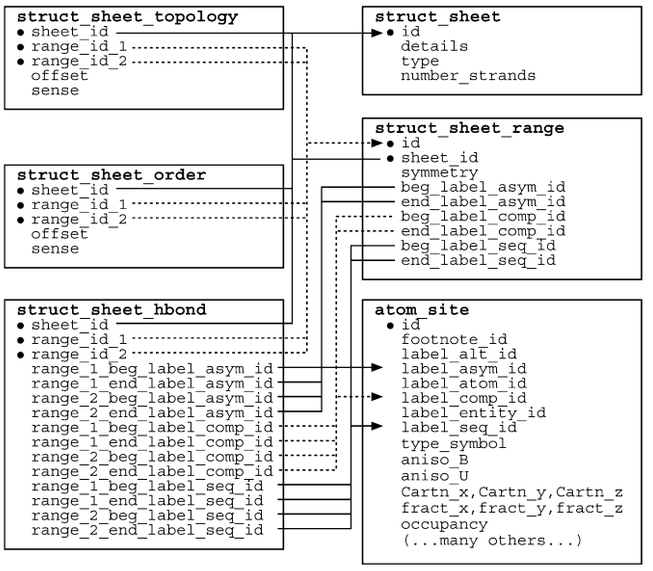

Fig. 3.6.7.14. The family of categories used to describe β-sheets. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

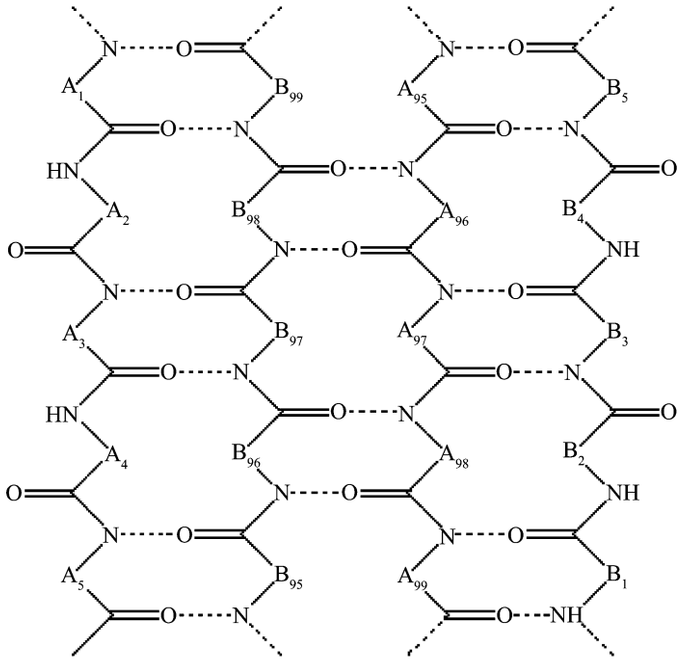

Fig. 3.6.7.15. A hypothetical β-sheet to be described with data items in the STRUCT_SHEET, STRUCT_SHEET_ORDER, STRUCT_SHEET_RANGE and STRUCT_SHEET_HBOND categories. Note that the strands come from two different polypeptides, labelled A and B.

In the more detailed and more general method for describing β-sheets, data items in the STRUCT_SHEET_RANGE category specify the range of residues that form strands in the sheet, data items in the STRUCT_SHEET_ORDER category specify the relative pairwise orientation of strands and data items in the STRUCT_SHEET_HBOND category provide details of specific hydrogen-bonding interactions between strands (see Fig. 3.6.7.15 and Example 3.6.7.14). Note that the specifiers for the strand ranges include the amino acid (**`*_comp_id`** and **`*_seq_id`**), the chain (**`*_asym_id`**) and a symmetry code (**`_struct_sheet_range.symmetry`**). Thus sheets that are composed of strands from more than one polypeptide chain

Example 3.6.7.14. *A hypothetical β-sheet described with data items in the STRUCT_SHEET, STRUCT_SHEET_ORDER, STRUCT_SHEET_RANGE and STRUCT_SHEET_HBOND categories.*

```
loop_
  _struct_sheet.id
  _struct_sheet.number_strands
    S1   4

loop_
  _struct_sheet_order.sheet_id
  _struct_sheet_order.range_id_1
  _struct_sheet_order.range_id_2
  _struct_sheet_order.sense
    S1 1 2 anti-parallel
    S1 2 3 anti-parallel
    S1 3 4 anti-parallel
    S2 1 2 anti-parallel

loop_
  _struct_sheet_range.sheet_id
  _struct_sheet_range.id
  _struct_sheet_range.beg_label_comp_id
  _struct_sheet_range.beg_label_asym_id
  _struct_sheet_range.beg_label_seq_id
  _struct_sheet_range.end_label_comp_id
  _struct_sheet_range.end_label_asym_id
  _struct_sheet_range.end_label_seq_id
    S1 1 PRO A 1  LEU A 5
    S1 2 CYS B 95 PHE B 99
    S1 3 CYS A 95 PHE A 99
    S1 4 PRO B 1  LEU B 5

loop_
  _struct_sheet_hbond.sheet_id
  _struct_sheet_hbond.range_id_1
  _struct_sheet_hbond.range_id_2
  _struct_sheet_hbond.range_1_beg_label_atom_id
  _struct_sheet_hbond.range_1_beg_label_seq_id
  _struct_sheet_hbond.range_2_beg_label_atom_id
  _struct_sheet_hbond.range_2_beg_label_seq_id
    S1 1 2 A 3  O 97
    S1 2 3 B 98 O 96
    S1 3 4 A 97 O 3
```

or from polypeptides in more than one asymmetric unit can be described.

It is conventional to assign the number 1 to an outermost strand. The choice of which outermost strand to number as 1 is arbitrary, but would usually be the strand encountered first in the amino-acid sequence. The remaining strands are then numbered sequentially across the sheet.

In some simple cases, the complete hydrogen bonding of the sheet could be inferred from the strand-range pairings and the relationship between the strands (parallel or antiparallel). However, in most cases it is necessary to specify at least one hydrogen bond between adjacent strands in order to establish the registration. The data items in the STRUCT_SHEET_HBOND category can be used to do this. Hydrogen bonds also need to be specified precisely when a sheet contains a nonstandard feature such as a β-bulge. This is a case where it is sufficient to specify a single hydrogen-bonding interaction to establish the registration; here only the **`*_beg_*`** or **`*_end_*`** data items need to be used to reference the atom-label components. However, it is preferable, wherever possible, to specify the initial and final atoms of the two ranges participating in the hydrogen bonding.

3.6.7.5.8. *Molecular sites*

The data items in these categories are as follows:

(*a*) STRUCT_SITE

- **`_struct_site.id`**
  `_struct_site.details`

(*b*) STRUCT_SITE_KEYWORDS

```
• _struct_site_keywords.site_id
        → _struct_site.id
• _struct_site_keywords.text
```

(*c*) STRUCT_SITE_GEN

```
• _struct_site_gen.id
• _struct_site_gen.site_id
        → _struct_site.id
  _struct_site_gen.details
  _struct_site_gen.label_alt_id
        → _atom_sites_alt.id
  _struct_site_gen.label_asym_id
        → _atom_site.label_asym_id
  _struct_site_gen.label_atom_id
        → _chem_comp_atom.atom_id
  _struct_site_gen.label_comp_id
        → _atom_site.label_atom_id
  _struct_site_gen.label_seq_id
        → _atom_site.label_seq_id
  _struct_site_gen.auth_asym_id
        → _atom_site.auth_asym_id
  _struct_site_gen.auth_atom_id
        → _atom_site.auth_atom_id
  _struct_site_gen.auth_comp_id
        → _atom_site.auth_comp_id
  _struct_site_gen.auth_seq_id
        → _atom_site.auth_seq_id
  _struct_site_gen.symmetry
```

(*d*) STRUCT_SITE_VIEW

```
• _struct_site_view.id
  _struct_site_view.details
  _struct_site_view.rot_matrix[1][1]
  _struct_site_view.rot_matrix[1][2]
  _struct_site_view.rot_matrix[1][3]
  _struct_site_view.rot_matrix[2][1]
  _struct_site_view.rot_matrix[2][2]
  _struct_site_view.rot_matrix[2][3]
  _struct_site_view.rot_matrix[3][1]
  _struct_site_view.rot_matrix[3][2]
  _struct_site_view.rot_matrix[3][3]
  _struct_site_view.site_id
        → _struct_site.id
```

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.

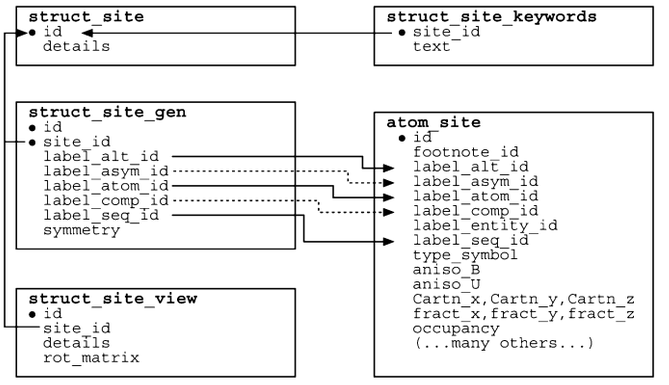

Fig. 3.6.7.16. The family of categories used to describe molecular sites. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

Substrate-binding sites, active sites, metal coordination sites and any other sites of interest may be described using data items in a collection of categories descending from STRUCT_SITE. These categories are intended to enable the author to generate views of molecular sites that could be used as figures in a report describing the structure or to enable a database to store standard views of common molecular sites (*e.g.* ATP-binding sites or the coordination of a calcium atom). The relationships between categories used to describe structural sites are shown in Fig. 3.6.7.16.

An identifier for each site that an author wishes to describe is given using `_struct_site.id` and the site can be described using `_struct_site.details`.

Keywords can be given for each site using data items in the STRUCT_SITE_KEYWORD category. Because keywords can be given at many levels of the mmCIF description of a structure, it may be worth duplicating the most significant higher-level keywords at this level to ensure that the site is detected in all search strategies.

The structural elements that generate each molecular site can be specified using data items in the STRUCT_SITE_GEN category. 'Structural elements' in this sense may be at any level of detail in the structure: single atoms, complete amino acids or nucleotides, or elements of secondary, tertiary or quaternary structure. Therefore the labels for each element may include, as required, the relevant `*_alt`, `*_asym`, `*_atom`, `*_comp` or `*_seq` parts of atom or residue identifiers. If the author has used an alternative labelling scheme, this can also be used. Noteworthy features of a structural element that forms part of the site can be described using the data item `_struct_site_gen.details`. Any crystallographic symmetry operations that are needed to form the site can be given using `_struct_site_gen.symmetry`.

Data items in the STRUCT_SITE_VIEW category allow the author to specify an orientation of the molecular site that gives a useful view of the components. The comments given in `_struct_site_view.details` could be used as a figure caption if the view is intended for use as a figure in a report.

Example 3.6.7.15 illustrates the use of these categories for describing a DNA binding site.

Example 3.6.7.15. *A DNA binding site with an intercalated drug (NDB DDF040) described with data items in the STRUCT_SITE, STRUCT_SITE_KEYWORDS, STRUCT_SITE_GEN and STRUCT_SITE_VIEW categories.*

```
loop_
_struct_site.id
_struct_site.details
   B1  'Binding at TG/AC Step 1'

loop_
_struct_site_keywords.site_id
_struct_site_keywords.text
   B1  'Intercalation complex'

loop_
_struct_site_gen.id
_struct_site_gen.site_id
_struct_site_gen.label_asym_id
_struct_site_gen.label_comp_id
_struct_site_gen.label_seq_id
_struct_site_gen.symmetry
   1  B1  A  T    1  1_555
   2  B1  A  G    2  1_555
   3  B1  A  C    5  8_555
   4  B1  A  A    6  8_555
   5  B1  D  DM2  .  8_555

loop_
_struct_site_view.id
_struct_site_view.site_id
_struct_site_view.details
_struct_site_view.rot_matrix[1][1]
_struct_site_view.rot_matrix[1][2]
# - - - abbreviated - - -
_struct_site_view.rot_matrix[3][3]
   View1  B1
  'View along the base-pair plane'
   0.133  0.922  . . . . . -0.172
```

3.6.7.6. Crystal symmetry

The categories describing symmetry are as follows:

SYMMETRY group
- SYMMETRY
- SYMMETRY_EQUIV
- SPACE_GROUP
- SPACE_GROUP_SYMOP

Data items in the SYMMETRY category are used to give details about the crystallographic symmetry. The equivalent positions for the space group are listed using data items in the SYMMETRY_EQUIV category. These categories are used in the same way in the core CIF and mmCIF dictionaries, and Section 3.2.4.4 can be consulted for details.

The current version of the mmCIF dictionary includes the SPACE_GROUP categories that were derived from the symmetry CIF dictionary (Chapter 3.8) and included in version 2.3 of the core CIF dictionary. At the time of writing, macromolecular applications have not yet begun to make use of these new categories.

Data items in these categories are as follows:

(*a*) SYMMETRY
- **_symmetry.entry_id**
  → **_entry.id**
- **_symmetry.cell_setting**
- **_symmetry.Int_Tables_number**
- **_symmetry.space_group_name_Hall**
- **_symmetry.space_group_name_H-M**

(*b*) SYMMETRY_EQUIV
- **_symmetry_equiv.id** (∼ **_symmetry_equiv_pos_site_id**)
- **_symmetry_equiv.pos_as_xyz**

(*c*) SPACE_GROUP
- **_space_group.id**
- **_space_group.crystal_system**
- **_space_group.IT_number**
- **_space_group.name_H-M_alt**
- **_space_group.name_Hall**

(*d*) SPACE_GROUP_SYMOP
- **_space_group_symop.id**
- **_space_group_symop.operation_xyz**
- **_space_group_symop.sg_id**

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 (_) except where indicated by the ∼ symbol.

The data item **_symmetry.entry_id** has been added to the SYMMETRY category to provide the formal category key required by the DDL2 data model.

3.6.7.7. Bond-valence information

The categories describing bond valences are as follows:

VALENCE group
- VALENCE_PARAM
- VALENCE_REF

These categories were introduced into version 2.2 of the core CIF dictionary to provide the information about bond valences required in inorganic crystallography. They appear in the mmCIF dictionary only for full compatibility with the core dictionary.

Data items in these categories are as follows:

(*a*) VALENCE_PARAM
- **_valence_param.atom_1**
- **_valence_param.atom_1_valence**
- **_valence_param.atom_2**
- **_valence_param.atom_2_valence**
- **_valence_param.B**
- **_valence_param.details**
- **_valence_param.id**
- **_valence_param.ref_id**
  → **_valence_ref.id**
- **_valence_param.Ro**

(*b*) VALENCE_REF
- **_valence_ref.id**
- **_valence_ref.reference**

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 (_).

Information about the use of these data items in the core CIF dictionary is given in Section 3.2.4.5.

3.6.8. Publication

The results of the determination of the crystal structure of a biological macromolecule might be published in an academic journal and/or deposited in a structural database. The data items in the core CIF dictionary cover most of the requirements for constructing an article for publication from an mmCIF and the many well defined data fields in mmCIF allow an extensively annotated record of the structure to be deposited in a database. However, the formalism of two of the core CIF categories for publication did not fit the relational database model of mmCIF, so new categories were required. The core CIF category COMPUTING, which is used to list the programs used to determine the structure, is replaced by the mmCIF category SOFTWARE, and the core CIF category DATABASE, which is used to identify the records associated with the structure in various databases, is replaced by the mmCIF category DATABASE_2.

The category groups discussed here are: the CITATION group, which is used to give citations to the literature (Section 3.6.8.1); the COMPUTING group, which is used to cite software (Section 3.6.8.2); the DATABASE group for citing related database entries (Section 3.6.8.3), which includes a group of categories used to ensure compatibility with specific database records in the Protein Data Bank (Section 3.6.8.3.2); journal administration categories that might be used by a publisher (Section 3.6.8.4.1); and the PUBL family of categories used to store the text of an article for publication (Section 3.6.8.4.2).

3.6.8.1. Literature citations

The categories describing literature citations are as follows:

CITATION group
- CITATION
- CITATION_AUTHOR
- CITATION_EDITOR

Data items in these categories are as follows:

(*a*) CITATION
- **_citation.id**
- **_citation.abstract**
- **_citation.abstract_id_CAS**
- **_citation.book_id_ISBN**
- **_citation.book_publisher**
- **_citation.book_publisher_city**
- **_citation.book_title**
- **_citation.coordinate_linkage**
- **_citation.country**
- **_citation.database_id_CSD**
- **_citation.database_id_Medline**
- **_citation.journal_abbrev**
- **_citation.journal_full**
- **_citation.journal_id_ASTM**
- **_citation.journal_id_CSD**
- **_citation.journal_id_ISSN**
- **_citation.journal_issue**
- **_citation.journal_volume**
- **_citation.language**
- **_citation.page_first**

_citation.page_last
_citation.details (∼ *_citation_special_details*)
_citation.title
_citation.year

(*b*) CITATION_AUTHOR
• *_citation_author.citation_id*
→ _citation.id
_citation_author.name
_citation_author.ordinal

(*c*) CITATION_EDITOR
• *_citation_editor.citation_id*
→ _citation.id
_citation_editor.name
_citation_editor.ordinal

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* (_).

The original core CIF dictionary contained the data item **_publ_section_references** for citations of journal articles, book chapters and monographs. The authors of the mmCIF dictionary felt that a more detailed and structured approach to literature citations was required. This is provided by the mmCIF categories CITATION, CITATION_AUTHOR and CITATION_EDITOR. These categories were subsequently included in the core CIF dictionary and are used in the same way in both dictionaries. Section 3.2.5.1 may be consulted for details. Although **_publ.section_references** remains a valid mmCIF data item, it is expected that the CITATION, CITATION_AUTHOR and CITATION_EDITOR categories will be used for literature citations in mmCIFs.

3.6.8.2. Citation of software packages

The categories describing software citations are as follows:
COMPUTING group
COMPUTING
SOFTWARE

It is expected that citations of software packages in an mmCIF will be made using data items in the SOFTWARE category. However, in some cases, a particular publisher or database may require that this information is given using data items in the COMPUTING category instead (see Section 3.2.5.2 for details).

Data items in these categories are as follows:

(*a*) COMPUTING
• **_computing.entry_id**
→ _entry.id
_computing.cell_refinement
_computing.data_collection
_computing.data_reduction
_computing.molecular_graphics
_computing.publication_material
_computing.structure_refinement
_computing.structure_solution

(*b*) SOFTWARE
• **_software.name**
• **_software.version**
_software.citation_id
→ _citation.id
_software.classification
_software.compiler_name
_software.compiler_version
_software.contact_author
_software.contact_author_email
_software.date
_software.dependencies
_software.description
_software.hardware
_software.language
_software.location
_software.mods
_software.os
_software.os_version
_software.type

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 data item **_computing.entry_id** has been added to the COMPUTING category to provide the formal category key required by the DDL2 data model.

The data items in the SOFTWARE category are used to cite the software packages used in the structure analysis. The software can be described in great detail if necessary. However, for most applications a small subset of these data items, for example just **_software.name** and **_software.version**, could be used (see Example 3.6.8.1).

Most data items in the SOFTWARE category are self-explanatory, but a few require further comment. The data item **_software.citation_id** provides a way to link the details of a program to the citation of an article in the literature that describes the program; this data item must match a value of **_citation.id** in the CITATION category. The name and e-mail address of the author of the software can also be given using **_software.contact_author** and **_software.contact_author_email**, respectively. (This may be the original author or someone who subsequently modifies or maintains the software; these data items would generally refer to the person most closely associated with the maintenance of the code at the time it was used.) The release date of the software may be recorded in **_software.date**. As far as possible, the date should be that of the version recorded in **_software.version**. The data item **_software.location** may be used to supply a URL from which the software may be downloaded or where it is described in detail.

3.6.8.3. Citation of related database entries

Categories describing related database entries are as follows:
DATABASE group
Related database entries (§3.6.8.3.1)
DATABASE
DATABASE_2

Example 3.6.8.1. *The refinement program Prolsq described with data items in the SOFTWARE category.*

```
_software.name
_software.version
_software.date
_software.type
_software.contact_author
_software.contact_author_email
_software.location
_software.classification
_software.citation_id
_software.language
_software.compiler_name
_software.compiler_version
_software.hardware
_software.os
_software.os_version
_software.dependencies
_software.mods
_software.description
   Prolsq    unknown  .    program
    'Wayne A. Hendrickson'  ?
    'ftp://rosebud.sdsc.edu/pub/sdsc/xtal/CCP4/ccp4/'
    refinement ref5  Fortran
  'Convex Fortran' v8.0 'Convex C220' ConvexOS  v10.1
    'Requires that Protin be run first'    optimized
    'restrained least-squares refinement'
```

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

freely used by other publishers who wish to handle articles submitted in CIF format. The JOURNAL and JOURNAL_INDEX categories are used in the same way in the core CIF and mmCIF dictionaries, and Section 3.2.5.4 can be consulted for details.

3.6.8.4.2. *Contents of a publication*

Data items in these categories are as follows:

(*a*) PUBL

• **_publ.entry_id**
  → **_entry.id**
  _publ.contact_author
  _publ.contact_author_address
  _publ.contact_author_email
  _publ.contact_author_fax
  _publ.contact_author_name
  _publ.contact_author_phone
  _publ.contact_letter
  _publ.manuscript_creation
  _publ.manuscript_processed
  _publ.manuscript_text
  _publ.requested_category
  _publ.requested_coeditor_name
  _publ.requested_journal
  _publ.section_abstract
  _publ.section_acknowledgements
  _publ.section_comment
  _publ.section_discussion
  _publ.section_experimental
  _publ.section_exptl_prep
  _publ.section_exptl_refinement
  _publ.section_exptl_solution
  _publ.section_figure_captions
  _publ.section_introduction
  _publ.section_references
  _publ.section_synopsis
  _publ.section_table_legends
  _publ.section_title
  _publ.section_title_footnote

(*b*) PUBL_AUTHOR
  _publ_author.address
  _publ_author.email
  _publ_author.footnote
  _publ_author.id_iucr
  _publ_author.name

(*c*) PUBL_BODY
  _publ_body.contents
  _publ_body.element
  _publ_body.format
  _publ_body.label
  _publ_body.title

(*d*) PUBL_MANUSCRIPT_INCL

• **_publ_manuscript_incl.entry_id**
  → **_entry.id**
  _publ_manuscript_incl.extra_defn
  _publ_manuscript_incl.extra_info
  _publ_manuscript_incl.extra_item

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* (_).

The categories PUBL, PUBL_AUTHOR, PUBL_BODY and PUBL_MANUSCRIPT_INCL are also members of the IUCR group in the mmCIF dictionary. They are used in the same way in the core CIF and mmCIF dictionaries, and Section 3.2.5.5 can be consulted for details.

3.6.9. File metadata

As in the core CIF dictionary, information about the source and the revision history of an mmCIF may be given in the AUDIT group of categories: AUDIT, AUDIT_AUTHOR, AUDIT_CONTACT_AUTHOR and AUDIT_CONFORM (Section 3.6.9.1). However, the mmCIF dictionary differs from the core CIF dictionary in the way it expresses relationships between data blocks: instead of the core AUDIT_LINK category, mmCIF has two categories, ENTRY and ENTRY_LINK, that essentially fulfil the same role but are classified in a distinct category group (Section 3.6.9.2).

3.6.9.1. History of a data block

The categories describing the history of a data block are as follows:

AUDIT group
  AUDIT
  AUDIT_AUTHOR
  AUDIT_CONFORM
  AUDIT_CONTACT_AUTHOR

Data items in these categories are as follows:

(*a*) AUDIT

• **_audit.revision_id**
  _audit.creation_date
  _audit.creation_method
  _audit.update_record

(*b*) AUDIT_AUTHOR

• ***_audit_author.name***
  _audit_author.address

(*c*) AUDIT_CONFORM

• ***_audit_conform.dict_name***
• ***_audit_conform.dict_version***
  _audit_conform.dict_location

(*d*) AUDIT_CONTACT_AUTHOR

• ***_audit_contact_author.name***
  _audit_contact_author.address
  _audit_contact_author.email
  _audit_contact_author.fax
  _audit_contact_author.phone

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. Items in italics have aliases in the core CIF dictionary formed by changing the full stop* (**.**) *to an underscore* (_).

The data items in these categories are used in the same way in the mmCIF dictionary as in the core CIF dictionary (see Section 3.2.6). The data item **_audit.revision_id** has been added to the AUDIT category to provide the formal category key required by the DDL2 data model. The core data item **_audit_block_code** has been replaced by **_entry.id** (see Section 3.6.9.2).

3.6.9.2. Links between data blocks

The categories describing links between data blocks are as follows:

ENTRY group
  ENTRY
  ENTRY_LINK

AUDIT group
  AUDIT_LINK

Data items in these categories are as follows:

(*a*) ENTRY

• **_entry.id**

(*b*) ENTRY_LINK

• **_entry_link.entry_id**
  → **_entry.id**
• **_entry_link.id**
  _entry_link.details

(*c*) AUDIT_LINK
- **_audit_link.block_code**
- **_audit_link.block_description**

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 sole data item in the category ENTRY, **_entry.id**, is a label that identifies the current data block. This label is used as the formal key in several categories that record information that is relevant to the entire data block (*e.g.* **_cell.entry_id**, **_geom.entry_id**), so care should be taken to select a label that is informative and unique.

Data items in the ENTRY_LINK category record the relationships between the current data block and other data blocks within the current file which may be referenced in the current data block. Since there are no formal constraints on the value of **_entry.id** assigned to each data block, authors must take care to ensure that an mmCIF comprised of several distinct data blocks uses a different value for **_entry.id** in each block.

As mentioned in the introductory paragraph of Section 3.6.9, the ENTRY_LINK category is used in mmCIF applications instead of the core category AUDIT_LINK. The latter is retained formally in the mmCIF dictionary for strict compatibility with the core dictionary, and the data items in this category, **_audit_link.blockcode** and **_audit_link.block_description**, are aliased to corresponding core data names (see Section 3.2.6.1). Their use is not recommended in mmCIF applications.

3.6.9.3. Other category classifications

The following categories, already described elsewhere in this chapter, are included in other formal category groups:

Compliance with earlier dictionaries

COMPLIANCE group
- DATABASE

Compatibility with PDB format files

PDB group
- DATABASE_PDB_CAVEAT
- DATABASE_PDB_MATRIX
- DATABASE_PDB_REMARK
- DATABASE_PDB_REV
- DATABASE_PDB_REV_RECORD
- DATABASE_PDB_TVECT

The COMPLIANCE group includes categories that appear in the mmCIF dictionary for the sole purpose of ensuring compliance with earlier dictionaries. They are not intended for use in the creation of new mmCIFs. As was discussed in Section 3.6.8.3, the DATABASE category of the core CIF is replaced in mmCIF by the more structured DATABASE_2 category. Thus the core CIF DATABASE category appears in the mmCIF COMPLIANCE group. At the time of writing (2005), DATABASE is the only category in the COMPLIANCE group.

The PDB group includes a number of categories that record unstructured information imported from various records in Protein Data Bank (PDB) format files. These categories are also part of the DATABASE group and were discussed in Section 3.6.8.3.2.

Appendix 3.6.1
Category structure of the mmCIF dictionary

Table A3.6.1.1 provides an overview of the structure of the mmCIF dictionary by category group and member categories.

Appendix 3.6.2
The Protein Data Bank exchange data dictionary

BY J. D. WESTBROOK, K. HENRICK, E. L. ULRICH AND H. M. BERMAN

In developing a data-management infrastructure, the Protein Data Bank (PDB; Berman *et al.*, 2000) has chosen the mmCIF dictionary technology for describing the data that it collects and disseminates. To accommodate the growth in the PDB's activities, data collection, processing and annotation now occur at three sites worldwide: the Research Collaboratory for Structural Bioinformatics (RCSB/PDB), the Macromolecular Structural Database (MSD) at the European Bioinformatics Institute (EBI) and the Protein Data Bank Japan (PDBj) at Osaka. Together these facilities form the Worldwide PDB (wwPDB) (Berman *et al.*, 2003). In order to maintain the fidelity of the single archive of three-dimensional macromolecular structure, a precise content description is required to support the accurate exchange of data among the different sites and the exchange of information between different file formats.

A key strength of the mmCIF technology is the extensibility afforded by a framework based on a software-accessible data dictionary. The PDB has exploited this functionality by using the mmCIF dictionary as a foundation and supplementing it with extensions in order to describe all aspects of data processing and database operations.

These extensions include content required to support reversible format translation, noncrystallographic structure determination methods and the details of protein production. They also support recommendations by the International Union of Crystallography (IUCr) and the International Structural Genomics Organization (ISGO) as to which data should be deposited. In the following sections, the extensions to the mmCIF data dictionary developed by the PDB (http://mmcif.pdb.org/) are described.

A3.6.2.1. Data exchange and format translation

The majority of crystallographic and structural concepts embodied in the PDB are already well described in the mmCIF data dictionary. However, while there is a conceptual description of most crystallographic information in PDB-format files within the mmCIF dictionary, the precise representation of this information can differ subtly. To guarantee accurate data exchange and to facilitate reversible format translation between PDB and mmCIF formats, all such differences in representation must be resolved.

To accommodate content and semantic differences between formats, extensions to the dictionary have been created. These extensions take one of two forms: the addition of new definitions to existing categories or the creation of new categories. Where possible, extensions are added to existing categories. This is done when the new definition supplements the content of the category without changing the category definition or its fundamental organization. However, if a new definition cannot be added to an existing category, a new category is created to hold the extension. All new data items and categories include the prefix **pdbx_** in their names.

For example, the level of detail in the PDB description of the biological source exceeds the description provided by mmCIF. In this case, dictionary extensions have been added to the existing categories ENTITY_SRC_NAT and ENTITY_SRC_GEN (where 'nat' and 'gen' stand for naturally occurring and genetically engineered, respectively). The PDB description of atomic coordinates includes two items that are not described in mmCIF: the insertion code

Table A3.6.1.1. *Categories in the mmCIF dictionary*

Numbers in parentheses refer to the section of this chapter in which each category is described in detail.

ATOM group (§3.6.7.1)
- ATOM_SITE (§3.6.7.1.1(*a*))
- ATOM_SITE_ANISOTROP (§3.6.7.1.1(*b*))
- ATOM_SITES (§3.6.7.1.2(*a*))
- ATOM_SITES_ALT (§3.6.7.1.4(*a*))
- ATOM_SITES_ALT_ENS (§3.6.7.1.4(*b*))
- ATOM_SITES_ALT_GEN (§3.6.7.1.4(*c*))
- ATOM_SITES_FOOTNOTE (§3.6.7.1.2(*b*))
- ATOM_TYPE (§3.6.7.1.3)

AUDIT group (§3.6.9.1)
- AUDIT (§3.6.9.1(*a*))
- AUDIT_AUTHOR (§3.6.9.1(*b*))
- AUDIT_CONFORM (§3.6.9.1(*c*))
- AUDIT_CONTACT_AUTHOR (§3.6.9.1(*d*))
- AUDIT_LINK (§3.6.9.2(*c*))

CELL group (§3.6.5.1)
- CELL (§3.6.5.1(*a*))
- CELL_MEASUREMENT (§3.6.5.1(*b*))
- CELL_MEASUREMENT_REFLN (§3.6.5.1(*c*))

CHEM_COMP group (§3.6.7.2.2)
- CHEM_COMP (§3.6.7.2.2(*a*))
- CHEM_COMP_ANGLE (§3.6.7.2.2(*b*))
- CHEM_COMP_ATOM (§3.6.7.2.2(*c*))
- CHEM_COMP_BOND (§3.6.7.2.2(*d*))
- CHEM_COMP_CHIR (§3.6.7.2.2(*e*))
- CHEM_COMP_CHIR_ATOM (§3.6.7.2.2(*f*))
- *CHEM_COMP_LINK* (*see CHEM_LINK group*)
- CHEM_COMP_PLANE (§3.6.7.2.2(*h*))
- CHEM_COMP_PLANE_ATOM (§3.6.7.2.2(*i*))
- CHEM_COMP_TOR (§3.6.7.2.2(*j*))
- CHEM_COMP_TOR_VALUE (§3.6.7.2.2(*k*))

CHEM_LINK group (§3.6.7.2.3)
- CHEM_COMP_LINK (§3.6.7.2.2(*g*))
- CHEM_LINK (§3.6.7.2.3(*a*))
- CHEM_LINK_ANGLE (§3.6.7.2.3(*b*))
- CHEM_LINK_BOND (§3.6.7.2.3(*c*))
- CHEM_LINK_CHIR (§3.6.7.2.3(*d*))
- CHEM_LINK_CHIR_ATOM (§3.6.7.2.3(*e*))
- CHEM_LINK_PLANE (§3.6.7.2.3(*f*))
- CHEM_LINK_PLANE_ATOM (§3.6.7.2.3(*g*))
- CHEM_LINK_TOR (§3.6.7.2.3(*h*))
- CHEM_LINK_TOR_VALUE (§3.6.7.2.3(*i*))
- ENTITY_LINK (§3.6.7.2.3(*j*))

CHEMICAL group (§3.6.7.2)
- CHEMICAL (§3.6.7.2.1(*a*))
- CHEMICAL_CONN_ATOM (§3.6.7.2.1(*b*))
- CHEMICAL_CONN_BOND (§3.6.7.2.1(*c*))
- CHEMICAL_FORMULA (§3.6.7.2.1(*d*))

CITATION group (§3.6.8.1)
- CITATION (§3.6.8.1(*a*))
- CITATION_AUTHOR (§3.6.8.1(*b*))
- CITATION_EDITOR (§3.6.8.1(*c*))

COMPUTING group (§3.6.8.2)
- COMPUTING (§3.6.8.2(*a*))
- SOFTWARE (§3.6.8.2(*b*))

DATABASE group (§§3.6.8.3, 3.6.9.3)
- DATABASE (§3.6.8.3.1(*a*))
- DATABASE_2 (§3.6.8.3.1(*b*))
- *The following also belong to the PDB group*
- DATABASE_PDB_CAVEAT (§3.6.8.3.2(*e*))
- DATABASE_PDB_MATRIX (§3.6.8.3.2(*c*))
- DATABASE_PDB_REMARK (§3.6.8.3.2(*f*))
- DATABASE_PDB_REV (§3.6.8.3.2(*a*))
- DATABASE_PDB_REV_RECORD (§3.6.8.3.2(*b*))
- DATABASE_PDB_TVECT (§3.6.8.3.2(*d*))

DIFFRN group (§3.6.5.2)
- DIFFRN (§3.6.5.2(*a*))
- DIFFRN_ATTENUATOR (§3.6.5.2(*b*))
- DIFFRN_DETECTOR (§3.6.5.2(*c*))
- DIFFRN_MEASUREMENT (§3.6.5.2(*d*))
- DIFFRN_ORIENT_MATRIX (§3.6.5.2(*e*))
- DIFFRN_ORIENT_REFLN (§3.6.5.2(*f*))
- DIFFRN_RADIATION (§3.6.5.2(*g*))
- DIFFRN_RADIATION_WAVELENGTH (§3.6.5.2(*h*))
- DIFFRN_REFLN (§3.6.5.2(*i*))
- DIFFRN_REFLNS (§3.6.5.2(*j*))
- DIFFRN_REFLNS_CLASS (§3.6.5.2(*k*))
- DIFFRN_SCALE_GROUP (§3.6.5.2(*l*))
- DIFFRN_SOURCE (§3.6.5.2(*m*))
- DIFFRN_STANDARD_REFLN (§3.6.5.2(*n*))
- DIFFRN_STANDARDS (§3.6.5.2(*o*))

ENTITY group (§3.6.7.3)
- ENTITY (§3.6.7.3.1(*a*))
- ENTITY_KEYWORDS (§3.6.7.3.1(*b*))
- *ENTITY_LINK* (*see CHEM_LINK group*)
- ENTITY_NAME_COM (§3.6.7.3.1(*c*))
- ENTITY_NAME_SYS (§3.6.7.3.1(*d*))
- ENTITY_POLY (§3.6.7.3.2(*a*))
- ENTITY_POLY_SEQ (§3.6.7.3.2(*b*))
- ENTITY_SRC_GEN (§3.6.7.3.1(*e*))
- ENTITY_SRC_NAT (§3.6.7.3.1(*f*))

ENTRY group (§3.6.9.2)
- ENTRY (§3.6.9.2(*a*))
- ENTRY_LINK (§3.6.9.2(*b*))

EXPTL group (§3.6.5.3)
- EXPTL (§3.6.5.3.1(*a*))
- EXPTL_CRYSTAL (§3.6.5.3.1(*b*))
- EXPTL_CRYSTAL_FACE (§3.6.5.3.1(*c*))
- EXPTL_CRYSTAL_GROW (§3.6.5.3.2(*a*))
- EXPTL_CRYSTAL_GROW_COMP (§3.6.5.3.2(*b*))

GEOM group (§3.6.7.4)
- GEOM (§3.6.7.4(*a*))
- GEOM_ANGLE (§3.6.7.4(*b*))
- GEOM_BOND (§3.6.7.4(*c*))
- GEOM_CONTACT (§3.6.7.4(*d*))
- GEOM_HBOND (§3.6.7.4(*e*))
- GEOM_TORSION (§3.6.7.4(*f*))

IUCR group (§3.6.8.4)
- JOURNAL (§3.6.8.4.1(*a*))
- JOURNAL_INDEX (§3.6.8.4.1(*b*))
- PUBL (§3.6.8.4.2(*a*))
- PUBL_AUTHOR (§3.6.8.4.2(*b*))
- PUBL_BODY (§3.6.8.4.2(*c*))
- PUBL_MANUSCRIPT_INCL (§3.6.8.4.2(*d*))

PHASING group (§3.6.6.1)
- PHASING (§3.6.6.1.1)
- PHASING_AVERAGING (§3.6.6.1.2)
- PHASING_ISOMORPHOUS (§3.6.6.1.3)
- PHASING_MAD (§3.6.6.1.4(*a*))
- PHASING_MAD_CLUST (§3.6.6.1.4(*b*))
- PHASING_MAD_EXPT (§3.6.6.1.4(*c*))
- PHASING_MAD_RATIO (§3.6.6.1.4(*d*))
- PHASING_MAD_SET (§3.6.6.1.4(*e*))
- PHASING_MIR (§3.6.6.1.5(*a*))
- PHASING_MIR_DER (§3.6.6.1.5(*c*))
- PHASING_MIR_DER_REFLN (§3.6.6.1.5(*d*))
- PHASING_MIR_DER_SHELL (§3.6.6.1.5(*e*))
- PHASING_MIR_DER_SITE (§3.6.6.1.5(*f*))
- PHASING_MIR_SHELL (§3.6.6.1.5(*b*))
- PHASING_SET (§3.6.6.1.6(*a*))
- PHASING_SET_REFLN (§3.6.6.1.6(*b*))
- PUBL (*see IUCR group*)
- PUBL_AUTHOR (*see IUCR group*)
- PUBL_BODY (*see IUCR group*)
- PUBL_MANUSCRIPT_INCL (*see IUCR group*)

REFINE group (§3.6.6.2)
- REFINE (§3.6.6.2.1(*a*))
- REFINE_ANALYZE (§3.6.6.2.2)
- REFINE_B_ISO (§3.6.6.2.4(*a*))
- REFINE_FUNCT_MINIMIZED (§3.6.6.2.1(*b*))
- REFINE_HIST (§3.6.6.2.5)
- REFINE_LS_RESTR (§3.6.6.2.3(*a*))
- REFINE_LS_RESTR_NCS (§3.6.6.2.3(*b*))
- REFINE_LS_CLASS (§3.6.2.2.3(*e*))
- REFINE_LS_RESTR_TYPE (§3.6.6.2.3(*c*))
- REFINE_LS_SHELL (§3.6.6.2.3(*d*))
- REFINE_OCCUPANCY (§3.6.6.2.4(*b*))

REFLN group (§3.6.6.3)
- REFLN (§3.6.6.3.1(*a*))
- REFLN_SYS_ABS (§3.6.6.3.1(*b*))
- REFLNS (§3.6.6.3.2(*a*))
- REFLNS_CLASS (§3.6.6.3.2(*d*))
- REFLNS_SCALE (§3.6.6.3.2(*b*))
- REFLNS_SHELL (§3.6.6.3.2(*c*))

- *SOFTWARE* (*see COMPUTING group*)
- *SPACE_GROUP* (*see SYMMETRY group*)
- *SPACE_GROUP_SYMOP* (*see SYMMETRY group*)

STRUCT group (§3.6.7.5)
- STRUCT (§3.6.7.5.1(*a*))
- STRUCT_ASYM (§3.6.7.5.1(*b*))
- STRUCT_BIOL (§3.6.7.5.1(*c*))
- STRUCT_BIOL_GEN (§3.6.7.5.1(*d*))
- STRUCT_BIOL_KEYWORDS (§3.6.7.5.1(*e*))
- STRUCT_BIOL_VIEW (§3.6.7.5.1(*f*))
- STRUCT_CONF (§3.6.7.5.2(*b*))
- STRUCT_CONF_TYPE (§3.6.7.5.2(*a*))
- STRUCT_CONN (§3.6.7.5.3(*b*))
- STRUCT_CONN_TYPE (§3.6.7.5.3(*a*))
- STRUCT_KEYWORDS (§3.6.7.5.1(*g*))
- STRUCT_MON_DETAILS (§3.6.7.5.4(*a*))
- STRUCT_MON_NUCL (§3.6.7.5.4(*b*))
- STRUCT_MON_PROT (§3.6.7.5.4(*c*))
- STRUCT_MON_PROT_CIS (§3.6.7.5.4(*d*))
- STRUCT_NCS_DOM (§3.6.7.5.5(*c*))
- STRUCT_NCS_DOM_LIM (§3.6.7.5.5(*d*))
- STRUCT_NCS_ENS (§3.6.7.5.5(*a*))
- STRUCT_NCS_ENS_GEN (§3.6.7.5.5(*b*))
- STRUCT_NCS_OPER (§3.6.7.5.5(*e*))
- STRUCT_REF (§3.6.7.5.6(*a*))
- STRUCT_REF_SEQ (§3.6.7.5.6(*b*))
- STRUCT_REF_SEQ_DIF (§3.6.7.5.6(*c*))
- STRUCT_SHEET (§3.6.7.5.7(*a*))
- STRUCT_SHEET_HBOND (§3.6.7.5.7(*e*))
- STRUCT_SHEET_ORDER (§3.6.7.5.7(*d*))
- STRUCT_SHEET_RANGE (§3.6.7.5.7(*c*))
- STRUCT_SHEET_TOPOLOGY (§3.6.7.5.7(*b*))
- STRUCT_SITE (§3.6.7.5.8(*a*))
- STRUCT_SITE_GEN (§3.6.7.5.8(*c*))
- STRUCT_SITE_KEYWORDS (§3.6.7.5.8(*b*))
- STRUCT_SITE_VIEW (§3.6.7.5.8(*d*))

SYMMETRY group (§3.6.7.6)
- SPACE_GROUP (§3.6.7.6(*c*))
- SPACE_GROUP_SYMOP (§3.6.7.6(*d*))
- SYMMETRY (§3.6.7.6(*a*))
- SYMMETRY_EQUIV (§3.6.7.6(*b*))

VALENCE group (§3.6.7.7)
- VALENCE_PARAM group (§3.6.7.7(*a*))
- VALENCE_REF group (§3.6.7.7(*b*))

and the model number. These have been added to the mmCIF category ATOM_SITE (as `_atom_site.pdbx_PDB_ins_code` and `_atom_site.pdbx_PDB_model_num`) and to all related categories that include atom nomenclature.

The convention for defining the hydrogen bonding in β-sheets differs between the PDB and mmCIF representations. Because the PDB model is fundamentally different from that found in mmCIF, a new category was created to hold the PDB data: PDBX_STRUCT_SHEET_HBOND. The correspondence between the PDB and mmCIF formats is tabulated at http://deposit.pdb.org/mmCIF/dictionaries/pdb-correspondence/pdb2mmcif.html.

A3.6.2.2. Extensions for structural genomics

An International Task Force on Deposition, Archiving, and Curation of Primary Information for Structural Genomics was formed under the auspices of the International Structural Genomics Organization (ISGO) in 2001 (Berman, 2001) and was asked to develop specifications for data from structural genomics projects to be deposited with the PDB. The recommendations from this working group are summarized at http://deposit.pdb.org/mmcif/sg-data/xstal.html and http://deposit.pdb.org/mmcif/sg-data/nmr.html. For data from crystallography-based projects, the content extensions are largely focused on a more detailed description of phasing, tracing and density modification. All of the ISGO recommendations have been incorporated into the PDB exchange dictionary.

A3.6.2.3. Noncrystallographic methods

The IUCr-sponsored development of data dictionaries has been focused exclusively on crystallographic methods. As the repository for all three-dimensional macromolecular structure data, the PDB accepts structures determined using noncrystallographic techniques such as NMR and cryo-electron microscopy. The description of noncrystallographic methods is beyond the remit of the IUCr, so the PDB has worked with the NMR and cryo-electron microscopy communities to develop data dictionaries that describe these techniques within the mmCIF framework.

A3.6.2.3.1. *NMR*

The PDB exchange dictionary includes a description of NMR sample preparation, structure solution methodology, refinement and refinement metrics. These extensions were developed in collaboration with the BioMagResBank (BMRB; Ulrich *et al.*, 1989). The BMRB is the archive for experimental NMR data for biological macromolecules and has played an active role in the development of the mmCIF data dictionary. In selecting a format for archiving NMR data, the BMRB opted to use the STAR syntax (Hall, 1991) rather than the more restrictive CIF syntax. Despite this difference in syntax, the conceptual representation of macromolecular structure in the NMR dictionary (NMRStar) has remained semantically very close to the mmCIF representation. This has facilitated the exchange of data and dictionaries between the BMRB and the PDB, the sharing of software tools, and the development of a common platform for depositing data.

A3.6.2.3.2. *Cryo-electron microscopy*

Cryo-electron microscopy (as a technique for the determination of the structure of large molecular assemblies) is also described in the PDB exchange dictionary. The data extensions for cryo-electron microscopy include a description of the sample preparation, raw volume data (Henrick *et al.*, 2003), structure solution and refinement. These extensions have a prefix of `em_` (http://mmcif.pdb.org/dictionaries/mmcif_iims.dic/Index/).

A3.6.2.3.3. *Protein production*

The International Task Force on Deposition, Archiving, and Curation of Primary Information for Structural Genomics (Section A3.6.2.2) has also provided recommendations for the deposition of information about protein production. These recommendations are summarized at http://deposit.pdb.org/mmcif/sg-data/protprod.html. These data extensions have been used as the foundation for the Protein Expression Purification and Crystallization database (PEPCdb, http://pepcdb.pdb.org/) and for the protein production process model developed to support the Structural Proteomics in Europe initiative (SPINE; http://www.spineurope.org/).

A3.6.2.4. Supporting software

The RCSB/PDB has developed a set of software tools which support the PDB exchange dictionary framework (Chapter 5.5). These include *PDB_EXTRACT*, a tool to extract data from the output files of structure determination applications; *ADIT*, a web-based editor for data files based on the PDB exchange dictionary; and *CIFTr*, a translator from mmCIF to PDB format. These applications and other supporting utilities can be downloaded from http://sw-tools.pdb.org/.

The development of the mmCIF dictionary and DDL2 has been an enormous task, and any list of contributors to the effort will certainly be incomplete. Still, we must try. We have so appreciated the people that have taken the time to think carefully and constructively about all of this, and we would like to recognize their efforts. We begin by recognizing Syd Hall, David Brown and Frank Allen, who began the entire CIF effort and who recruited us to do the extensions for macromolecular structure.

Chapter 1.1 describes the formation of the original mmCIF working group, chaired by Paula Fitzgerald and including Enrique Abola, Helen Berman, Phil Bourne, Eleanor Dodson, Art Olson, Wolfgang Steigemann, Lynn Ten Eyck and Keith Watenpaugh. However, the number of people who contributed to the original design of the mmCIF data structure is much larger. We would like to thank Steve Bryant, Vivian Stojanoff, Jean Richelle, Eldon Ulrich and Brian Toby.

There are also the people who realized the shortcomings of the original DDL and worked hard to convince us that a more rigorous underpinning for the dictionary would be needed. Among them are Michael Scharf, Peter Grey, Peter Murray-Rust, Dave Stampf and Jan Zelinka.

Writing the dictionary and developing the new DDL were just the starting points for evaluation and critique, and this effort has been greatly aided by the input from COMCIFS, the IUCr committee with oversight over this process (David Brown, Chair). But the real process of review, after the dictionary was released to the public for comment in August 1995, has involved a much larger number of people. We cannot say enough about the valuable input we have received from Frances Bernstein, Herbert Bernstein, Dale Tronrud and Peter Keller.

Our efforts have been greatly enabled by the staff of the Nucleic Acid Database at Rutgers University, who have dealt with many of the technical issues of the implementation of mmCIF with real data. So we would also like to thank Anke Gelbin, Shu-Hsin Hsieh and Christine Zardecki.

Without the three CIF workshops described in Chapter 1.1, this effort would never have taken the shape and focus it now has, and we are eternally gratefully to Eleanor Dodson (York), Phil Bourne (Tarrytown) and Shoshana Wodak (Brussels), who organized the workshops, and also to Helen Berman and John Westbrook for hosting the subsequent workshop at Rutgers following the publication of the mmCIF dictionary. We thank the European Science Foundation (ESF), the European Union (EU), the National Science Foundation (NSF) and the US Department of Energy (DOE), who provided the funding.

The RCSB/PDB is operated by Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the Center for Advanced Research in Biotechnology/UMBI/NIST. RCSB/PDB is supported by funds

from the National Science Foundation (NSF), the National Institute of General Medical Sciences (NIGMS), the Office of Science, Department of Energy (DOE), the National Library of Medicine (NLM), the National Cancer Institute (NCI), the National Center for Research Resources (NCRR), the National Institute of Biomedical Imaging and Bioengineering (NIBIB) and the National Institute of Neurological Disorders and Stroke (NINDS).

References

Altona, C. & Sundaralingam, M. (1972). *Conformational analysis of the sugar ring in nucleosides and nucleotides. New description using the concept of pseudorotation. J. Am. Chem. Soc.* **94**, 8205–8212.

Berman, H. M. (2001). Chair. *Report of the task force on the deposition, archiving, and curation of the primary information.* Task Force Reports from the Second International Structural Genomics Meeting, Airlie, Virginia, USA. http://www.nigms.nih.gov/news/reports/airlie_tasks.html.

Berman, H. M., Henrick, K. & Nakamura, H. (2003). *Announcing the worldwide Protein Data Bank. Nature Struct. Biol.* **10**, 980.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *The Protein Data Bank. Nucleic Acids Res.* **28**, 235–242.

Bourne, P., Berman, H. M., McMahon, B., Watenpaugh, K. D., Westbrook, J. D. & Fitzgerald, P. M. D. (1997). *Macromolecular Crystallographic Information File. Methods Enzymol.* **277**, 571–590.

Brändén C.-I. & Jones, T. A. (1990). *Between objectivity and subjectivity. Nature (London)*, **343**, 687–689.

Brünger, A. T. (1997). *Free R value: cross-validation in crystallography. Methods Enzymol.* **277**, 366–396.

Cruickshank, D. W. J. (1999). *Remarks about protein structure precision. Acta Cryst.* D**55**, 583–601.

Driessen, H., Haneef, M. I. J., Harris, G. W., Howlin, B., Khan, G. & Moss, D. S. (1989). *RESTRAIN: restrained structure-factor least-squares refinement program for macromolecular structures. J. Appl. Cryst.* **22**, 510–516.

Engh, R. A. & Huber, R. (1991). *Accurate bond and angle parameters for X-ray protein structure refinement. Acta Cryst.* A**47**, 392–400.

Fitzgerald, P. M. D., Berman, H., Bourne, P., McMahon, B., Watenpaugh, K. & Westbrook, J. (1996). *The mmCIF dictionary: community review and final approval. Acta Cryst.* A**52** (*Suppl.*), C575.

Fitzgerald, P. M. D., McKeever, B. M., VanMiddlesworth, J. F., Springer, J. P., Heimbach, J. C., Leu, C.-T., Kerber, W. K., Dixon, R. A. F. & Darke, P. L. (1990). *Crystallographic analysis of a complex between human immunodeficiency virus type 1 protease and acetyl-pepstatin at 2.0-A resolution. J. Biol. Chem.* **265**, 14209–14219.

Hall, S. R. (1991). *The STAR file: a new format for electronic data transfer and archiving. J. Chem. Inf. Comput. Sci.* **31**, 326–333.

Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *The crystallographic information file (CIF): a new standard archive file for crystallography. Acta Cryst.* A**47**, 655–685.

Hamilton, W. C. (1965). *Significance tests on the crystallographic R factor. Acta Cryst.* **18**, 502–510.

Hendrickson, W. A. & Konnert, J. H. (1979). *Stereochemically restrained crystallographic least-squares refinement of macromolecule structures.* In *Biomolecular structure, conformation, function and evolution*, edited by R. Srinavisan, Vol. I, pp. 43–57. New York: Pergamon Press.

Hendrickson, W. A. & Lattman, E. E. (1970). *Representation of phase probability distributions for simplified combination of independent phase information. Acta Cryst.* B**26**, 136–143.

Henrick, K., Newman, R., Tagari, M. & Chagoyen, M. (2003). *EMDep: a web-based system for the deposition and validation of high-resolution electron microscopy macromolecular structural information. J. Struct. Biol.* **144**, 228–237.

Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Cryst.* A**47**, 110–119.

Leonard, G. A., Hambley, T. W., McAuley-Hecht, K., Brown, T. & Hunter, W. N. (1993). *Anthracycline–DNA interactions at unfavourable base-pair triplet-binding sites: structures of d(CGGCCG)/daunomycin and d(TGGCCA)/adriamycin complexes. Acta Cryst.* D**49**, 458–467.

Luzzati, V. (1952). *Traitement statistique des erreurs dans la determination des structures cristallines. Acta Cryst.* **5**, 802–810.

Narayana, N., Ginell, S. L., Russu, I. M. & Berman, H. M. (1991). *Crystal and molecular structure of a DNA fragment: d(CGTGAATTCACG). Biochemistry*, **30**, 4449–4455.

Shapiro, L., Fannon, A. M., Kwong, P. D., Thompson, A., Lehmann, M. S., Grubel, G., Legrand, J. F., Als-Nielsen, J., Colman, D. R. & Hendrickson, W. A. (1995). *Structural basis of cell–cell adhesion by cadherins. Nature (London)*, **374**, 327–337.

Tickle, I. J., Laskowski, R. A. & Moss, D. S. (1998). *R_{free} and the R_{free} ratio. I. Derivation of expected values of cross-validation residuals used in macromolecular least-squares refinement. Acta Cryst.* D**54**, 547–557.

Ulrich, E. L., Markley, J. L. & Kyogoku, Y. (1989). *Creation of a nuclear magnetic resonance data repository and literature database. Protein Seq. Data Anal.* **2**, 23–37.

Zanotti, G., Berni, R. & Monaco, H. L. (1993). *Crystal structure of liganded and unliganded forms of bovine plasma retinol-binding protein. J. Biol. Chem.* **268**, 10728–10738.