_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

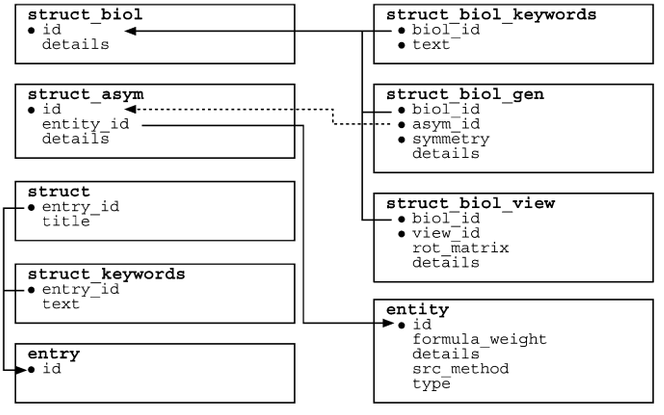

Fig. 3.6.7.7. The family of categories used to describe the higher-level macromolecular structure. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows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

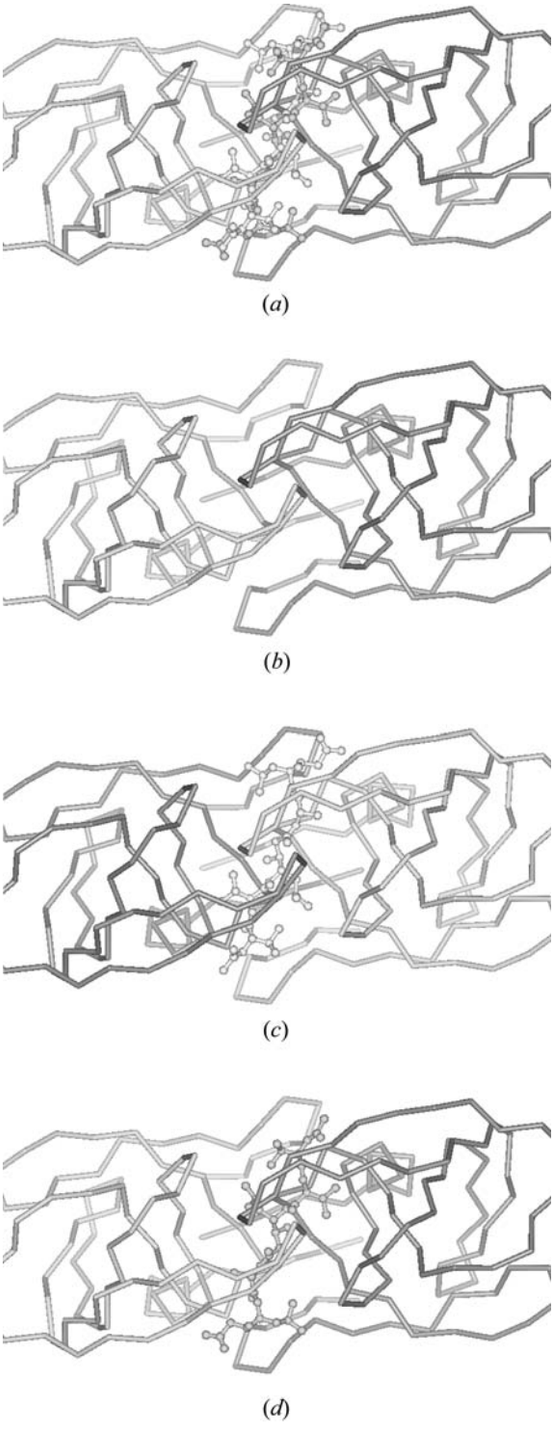

Fig. 3.6.7.8. The higher-level structure of the complex of HIV-1 protease with an inhibitor (PDB 5HVP) to be described with data items in the STRUCT_ASYM, STRUCT_BIOL, STRUCT_BIOL_KEYWORDS and STRUCT_BIOL_GEN categories. (*a*) Complete structure; (*b*), (*c*), (*d*) three different biological units.

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.

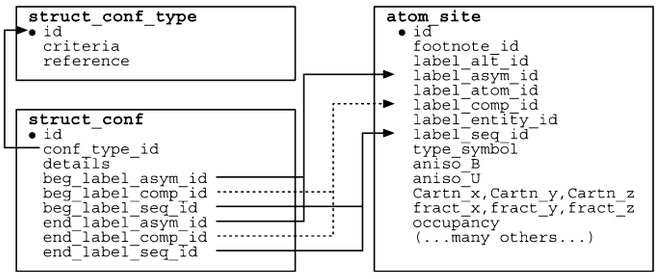

Fig. 3.6.7.9. The family of categories used to describe secondary structure. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent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

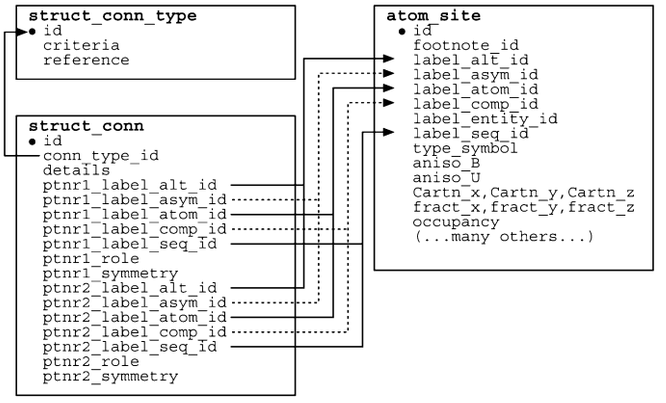

Fig. 3.6.7.10. The family of categories used to describe structural interactions such as hydrogen bonding, salt bridges and disulfide bridges. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked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

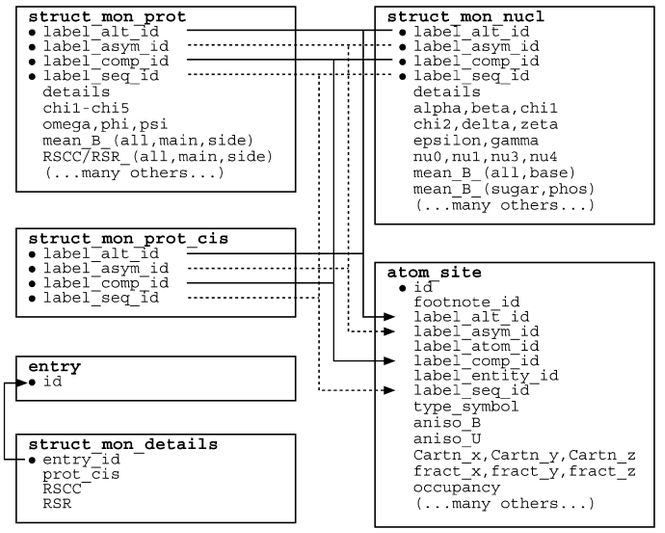

Fig. 3.6.7.11. The family of categories used to describe the structural features of monomers. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing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

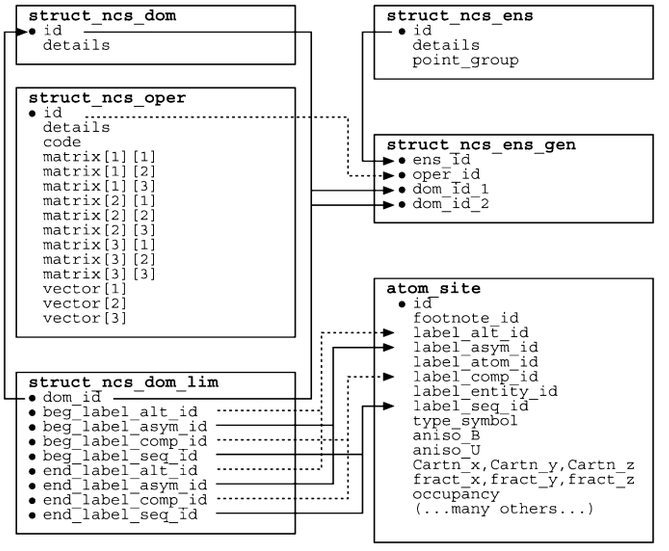

Fig. 3.6.7.12. The family of categories used to describe noncrystallographic symmetry. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing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

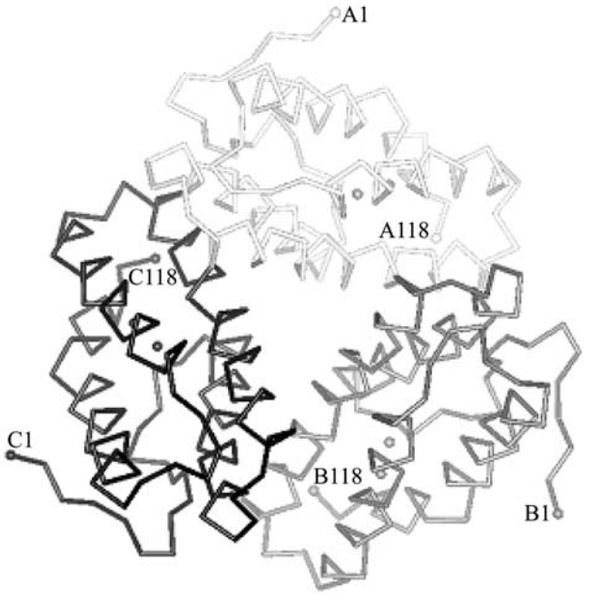

Fig. 3.6.7.13. Noncrystallographic symmetry in the structure of trimeric haemerythrin (PDB 1HR3) to be described with data items in the STRUCT_NCS_ENS, STRUCT_NCS_ENS_GEN, STRUCT_NCS_DOM and STRUCT_NCS_DOM_LIM categories.

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.

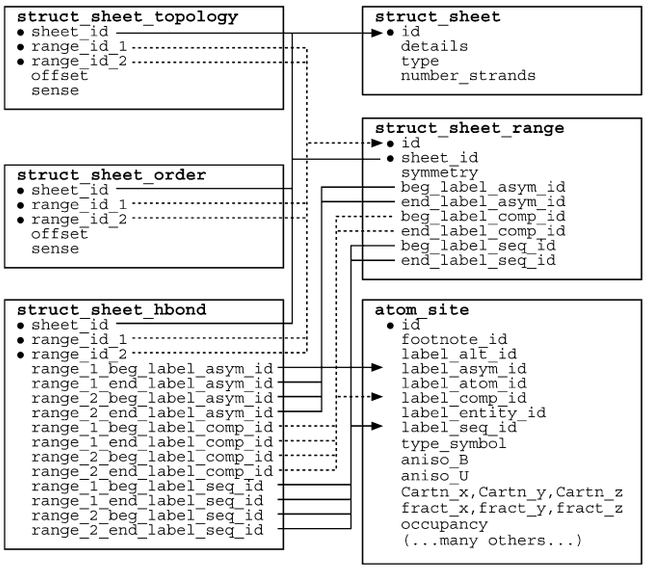

Fig. 3.6.7.14. The family of categories used to describe β-sheets. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

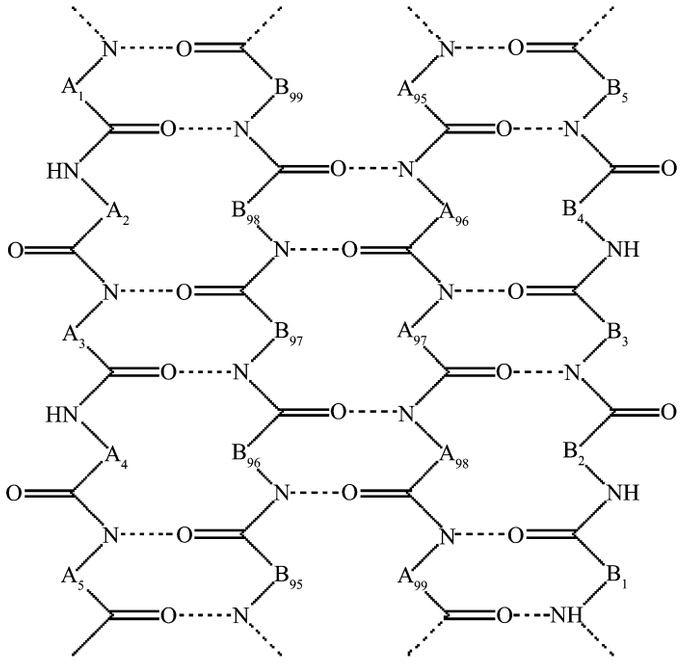

Fig. 3.6.7.15. A hypothetical β-sheet to be described with data items in the STRUCT_SHEET, STRUCT_SHEET_ORDER, STRUCT_SHEET_RANGE and STRUCT_SHEET_HBOND categories. Note that the strands come from two different polypeptides, labelled A and B.

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.

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.

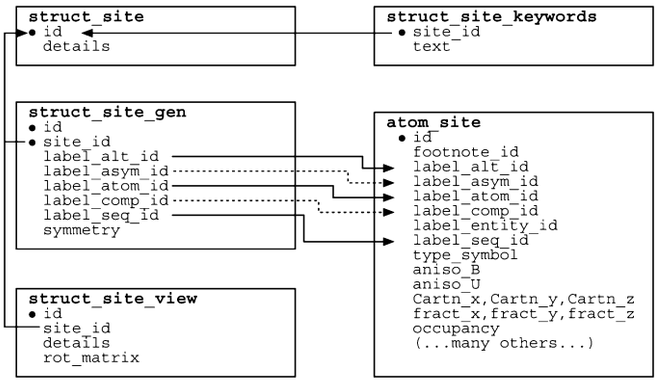

Fig. 3.6.7.16. The family of categories used to describe molecular sites. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent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

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
```