International Tables for Crystallography (2006). Vol. G, Section 3.6.7.3, pp. 174–176.

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

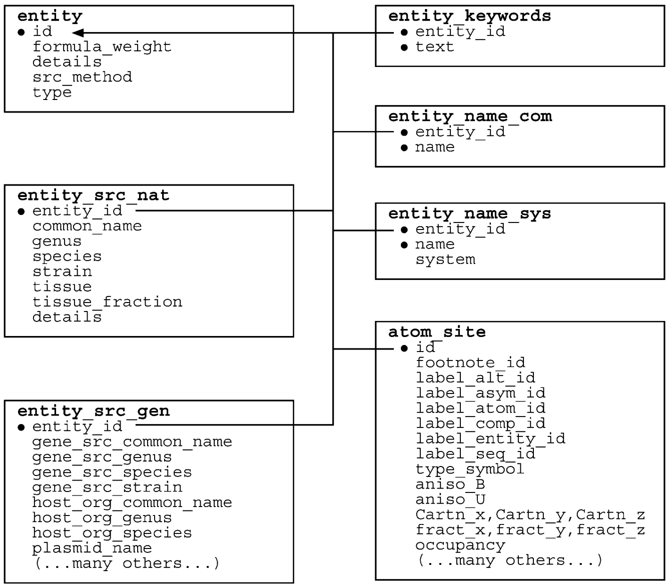

Fig. 3.6.7.5. The family of categories used to describe chemical entities. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (●). Lines show relationships between linked data items in different categories with arrows pointing at the parent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

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'
```

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.

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

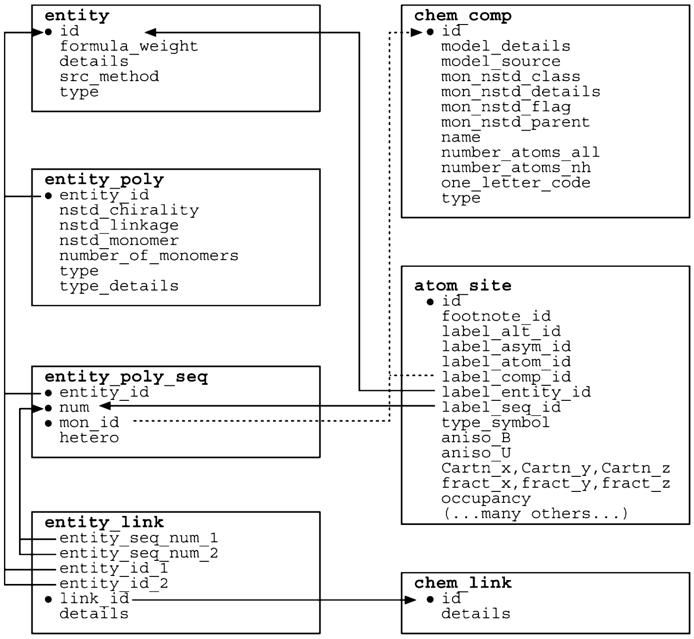

Fig. 3.6.7.6. The family of categories used to describe polymer chemical entities. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (●). Lines show relationships between linked data items in different categories with arrows pointing at the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