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