CIF APPLICATIONS

Authors of any software that reads, writes or validates CIF data are invited to contribute to this series. Authors should state clearly when submitting a manuscript to a Co-editor that the paper should be included as part of the CIF Applications series. An appropriate series number will be assigned by the Editorial Office.

J. Appl. Cryst. (1997). **30**, 79–83

CIF Applications. VI. *CIFLIB*: an application-program interface to CIF dictionaries and data files

John D. Westbrook,[a] Shu-Hsin Hsieh[a] and Paula M. D. Fitzgerald[b] at [a]*Nucleic Acid Database Project, Department of Chemistry, Rutgers, The State University of New Jersey, USA, and* [b]*Merck Research Laboratories, Rahway, New Jersey, USA. E-mail: jwest@ndb.rutgers.edu*

(Received 23 January 1996; accepted 27 June 1996)

Abstract

A software library is described that provides simple and convenient access to information in Crystallographic Information File (CIF) dictionaries and data files. This library gives the application programmer a collection of high-level functions that can be used to process and check data stored in the CIF format using the full detail of the CIF dictionary of crystallographic terminology. Two applications are presented that demonstrate the features of the library: a CIF browser and a CIF dictionary-to-HTML converter.

1. Introduction

The Crystallographic Information File (CIF) (Hall, Allen & Brown, 1991) format has become a standard interchange format for the description, archiving and publication of crystallographic experiments on small organic and inorganic molecules. The dictionary-based approach employed in the CIF-format data description may well become a standard for the representation of many other kinds of structural experiments. CIF dictionaries of terminology are also under development for powder diffraction (Toby, 1993), macromolecular crystallography (Bourne *et al.*, 1997), and macromolecular nuclear magnetic resonance spectroscopy (Ulrich, 1995).

The CIF uses a subset of the features of the Self-defining Text Archive and Retrieval (STAR) format proposed by Hall (1991). It employs a simple syntax in which each data value or list of data values is accompanied by a data name. Corresponding to each data name is a definition that provides a very detailed description of the data item. These definitions are collected in dictionaries that are also represented in the CIF format. Each data definition consists of a set of components that are individually defined in a separate dictionary that is also represented in CIF format. Since this latter dictionary provides the framework for constructing dictionary definitions, it is referred to as a Dictionary Definition Language (DDL). Fig. 1 shows an example of a fragment of a CIF, a CIF dictionary entry and a DDL dictionary entry illustrating the uniform mode of expression used in each of the cases.

The regular appearance of CIFs and CIF dictionaries as collections of name and value pairs is dictated by STAR syntax rules; however, the underlying organization of the data and definitions is determined by the DDL. The core dictionary of crystallographic definitions was developed using the DDL proposed by Cook and Hall (Cook, 1991; Hall & Cook, 1995). In the development of the macromolecular CIF (mmCIF) dictionary (Fitzgerald, Berman, Bourne & Watenpaugh, 1993; Fitzgerald *et al.*, 1997), this DDL was extended in order to more rigorously express relationships among the macromolecular data items. Although the extended DDL uses different conventions for naming data, it provides a mechanism to reference alternative data names. The mmCIF dictionary uses this feature to show the correspondence between the mmCIF data items and the existing core CIF data items. Because the mmCIF dictionary incorporates all of the definitions in the core CIF dictionary, it is possible for software developed for the extended DDL to use the mmCIF dictionary to read, write and check data items derived from either dictionary.

CIFLIB (Berman & Westbrook, 1993) is a software library that was developed to provide an application interface to information in CIF format. *CIFLIB* is designed to completely

```
  _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
```

(*a*)

```
save__cell.length_a
  _item_description.description
;    Unit-cell length a corresponding to the structure
  reported.
;
  _item.name                          '_cell.length_a'
  _item.category_id                   cell
  _item.mandatory_code                no
  _item_sub_category.id               'cell_length'
  _item_aliases.alias_name            '_cell_length_a'
  _item_aliases.dictionary            'cifdic.c94'
  _item_aliases.version               '2.0'
  _item_related.related_name          '_cell.length_a_esd'
  _item_related.function_code         'associated_esd'
  _item_type.code                     float
  _item_type_conditions.code          esd
  _item_units.code                    'angstroms'
  save_
```

(*b*)

```
save__item_description.description
  _item_description.description
;    Text description of the defined data item.
;
  _item.name                    '_item_description.description'
  _item.category_id             item_description
  _item.mandatory_code          yes
  _item_type.code               text
  save_
```

(*c*)

Fig. 1. Abbreviated examples of (*a*) CIF data specifications, (*b*) CIF dictionary definition and (*c*) DDL definition.

Journal of Applied Crystallography
ISSN 0021-8898 © 1997

encapsulate all I/O and integrity-checking operations on CIF dictionaries and data files from a calling application. *CIFLIB* provides functions that perform the following types of operations:

(i) read and write operations on CIF format data files and dictionaries;

(ii) read, write and update operations on individual data items and dictionary-definition components;

(iii) detailed integrity checks of CIF data and dictionaries as defined by the Dictionary Definition Language (DDL) 2.1 (Berman & Westbrook, 1994; Westbrook & Hall, 1997);

(iv) efficient access to the CIF-dictionary data model;

(v) robust syntactic and semantic error handling.

Fig. 2 shows how this software library facilitates integration of the CIF interchange format with other applications. As the figure illustrates, *CIFLIB* provides complete access to the DDL, CIF dictionaries and CIF data files. This library can be used to build wrappers and filters around existing applications that need to access CIF data. Since *CIFLIB* provides complete access to the dictionary data model, the library can be conveniently used as an in-memory database or as a loader for an external database. The library can also be used as a vehicle to explore the contents of the CIF dictionary. Two such applications, a CIF browser and a CIF dictionary-to-HTML converter, are described in a later section.

2. The CIF data model

The smallest element of information in the CIF data model* is an individual data item such as a Cartesian coordinate. Collections of data items may be grouped together in subcategories. For instance, in the mmCIF dictionary the *x*, *y*, and *z* Cartesian components are assigned to the `cartesian_coordinate` subcategory.

A category is a stronger association among a group of data items requiring that the value(s) of one or more of the items in the group can be used to differentiate occurrences of the group. Thus `atom_site` is a category in the mmCIF dictionary that contains the subcategory `cartesian_coordinates` and uses the data item `_atom_id` as the unique identifier for the category.

An important feature of the CIF model is the ability to specify relationships between data items in different categories. Parent–child relationships arise frequently in the description of macromolecular structures. For instance, the definition of protein secondary structure includes the residue labels that define the limits of each structural feature. These residue labels are also a component of the definition of each atomic position. In some cases, it is important that a secondary structure description references only those residue labels for which positions have been determined. This requirement can be included in a dictionary definition by specification of a parent–child relationship between the residue labels in these two categories.

Broader associations are also provided by the combination of collections of related categories together in category groups. For example, in the mmCIF dictionary, all of the categories pertaining to refinement are assigned to a category group named `refine_group`. Finally, the highest level of association provided by a CIF is the data block. Each data block acts like an independent database. Typically, a data block is used to hold all of the information pertaining to a particular structural experiment.

3. *CIFLIB* C-language interface description

This section describes the C-language application-program interface to *CIFLIB*. The major areas of functionality provided by this interface are presented in summary form. A reference manual that describes each interface function in detail is available.

Accessing data in CIF format using *CIFLIB* is a multistep process. *CIFLIB* first reads a DDL dictionary. Although much of *CIFLIB* is necessarily hardwired for DDL 2.1, many DDL attributes act simply as placeholders for information, and these attributes can be extended without modification to the library. The DDL is also checked against itself using internally coded rules based on DDL 2.1. Once the DDL is read, a CIF dictionary that is based on this DDL can be read and checked. This process can be quite time consuming for large dictionaries, so a provision has been made to retain the state of any file that has been checked in an auxiliary file. This auxiliary file will be used in preference to the original file in subsequent file accesses if its modification date is more recent. Finally, CIF data files are read and checked with respect to the CIF dictionary. In any file access, *CIFLIB* provides complete access to the data blocks containing the DDL, the CIF dictionary, and any number of blocks containing user data.

Each CIF may be divided into data-block sections. *CIFLIB* treats each data block as an independent database loaded into the data model defined in its associated dictionary. The CIF

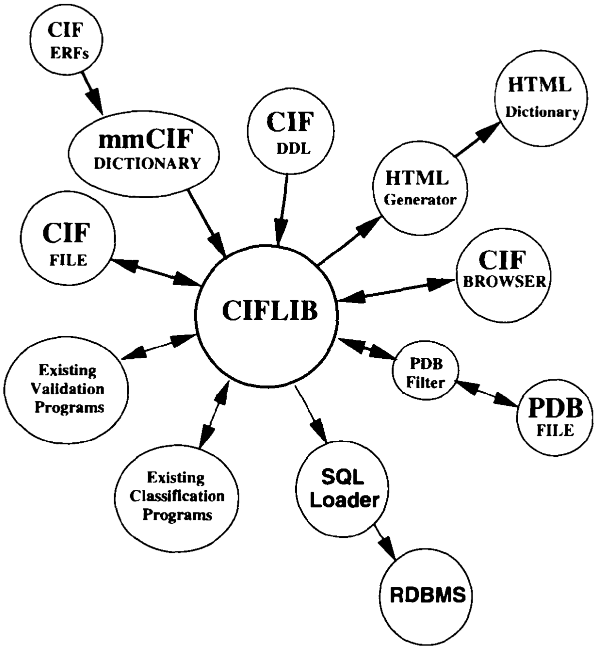


Fig. 2. Functional diagram of CIFLIB illustrating the intended use of the library in supporting CIF access for a variety of application-program types.

* The model described here is that defined in DDL 2.1 and is that used in the mmCIF dictionary. This model is a very similar to the recently published extension to the original core CIF DDL (Hall & Cook, 1995).

DDL is at the top of the chain and provides the data model for a CIF dictionary. The CIF dictionary in turn provides the data model for CIF data files. *CIFLIB* provides functions to read, write and merge data blocks. Any number of data blocks can be managed by the library.

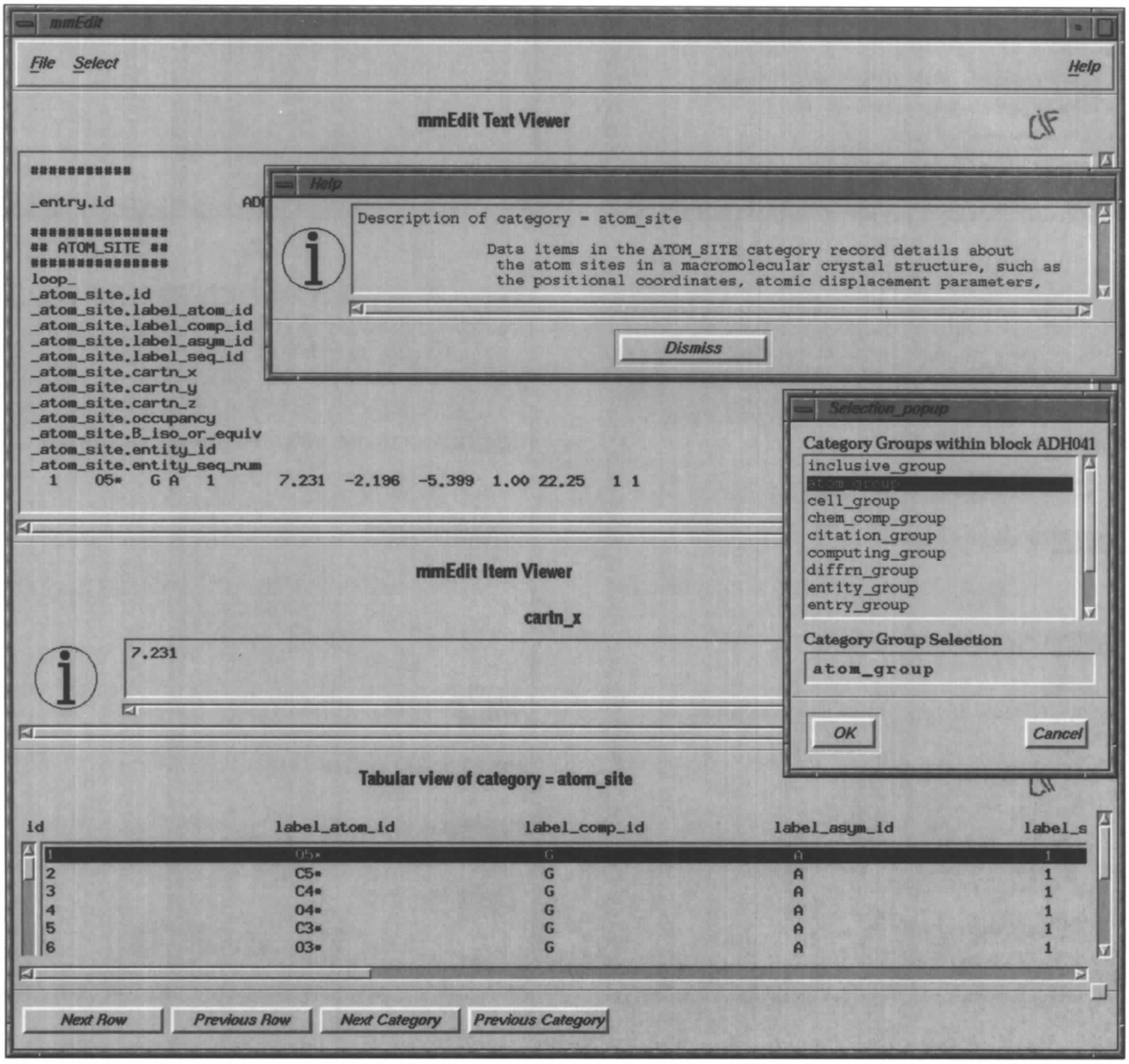
Within each individual data block, category groups provide a mechanism for organizing categories into conceptually meaningful collections. *CIFLIB* provides functions to obtain the list of category groups defined within a data block as well as the names of the member categories of each group.

The library provides a set of functions for accessing category-level features within a data block. These functions provide a complete list of the categories specified within a data block, the list of data items specified within each category, and the number of rows of data in a category. The attributes of a category defined in the CIF dictionary, such as the category description, category examples, member data items, key data items and member subcategories, can also be obtained.

Functions are provided to read, write and update individual data items, rows of data items and columns of data items. These functions also check the integrity of item values with respect to their dictionary definitions. Access to all of the item attributes defined in the CIF dictionary is provided, and convenience functions are provided for the most commonly used attributes such as alias names, data type, default value and enumeration.

CIFLIB provides a set of functions that give information about parent–child relationships and provide access to the parent and child item values. The parent and child relationships returned by the functions in this section span a single generation; however, complicated hierarchies of parentage can be easily traced.

CIFLIB provides a set of functions that access the error codes generated by those library functions that perform integrity checking. The *CIFLIB* functions that access and update individual item values return only a single error code. Functions providing read access return only the first error



```
###########

_entry.id                    AD

################
## ATOM_SITE ##
################
loop_
_atom_site.id
_atom_site.label_atom_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.label_seq_id
_atom_site.cartn_x
_atom_site.cartn_y
_atom_site.cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.entity_id
_atom_site.entity_seq_num
  1    05*   G A   1       7.231  -2.196  -5.399  1.00 22.25   1 1
```

```
Description of category = atom_site

             Data items in the ATOM_SITE category record details about
              the atom sites in a macromolecular crystal structure, such as
              the positional coordinates, atomic displacement parameters,
```

Category Groups within block ADH041

```
inclusive_group
atom_group
cell_group
chem_comp_group
citation_group
computing_group
diffrn_group
entity_group
entry_group
```

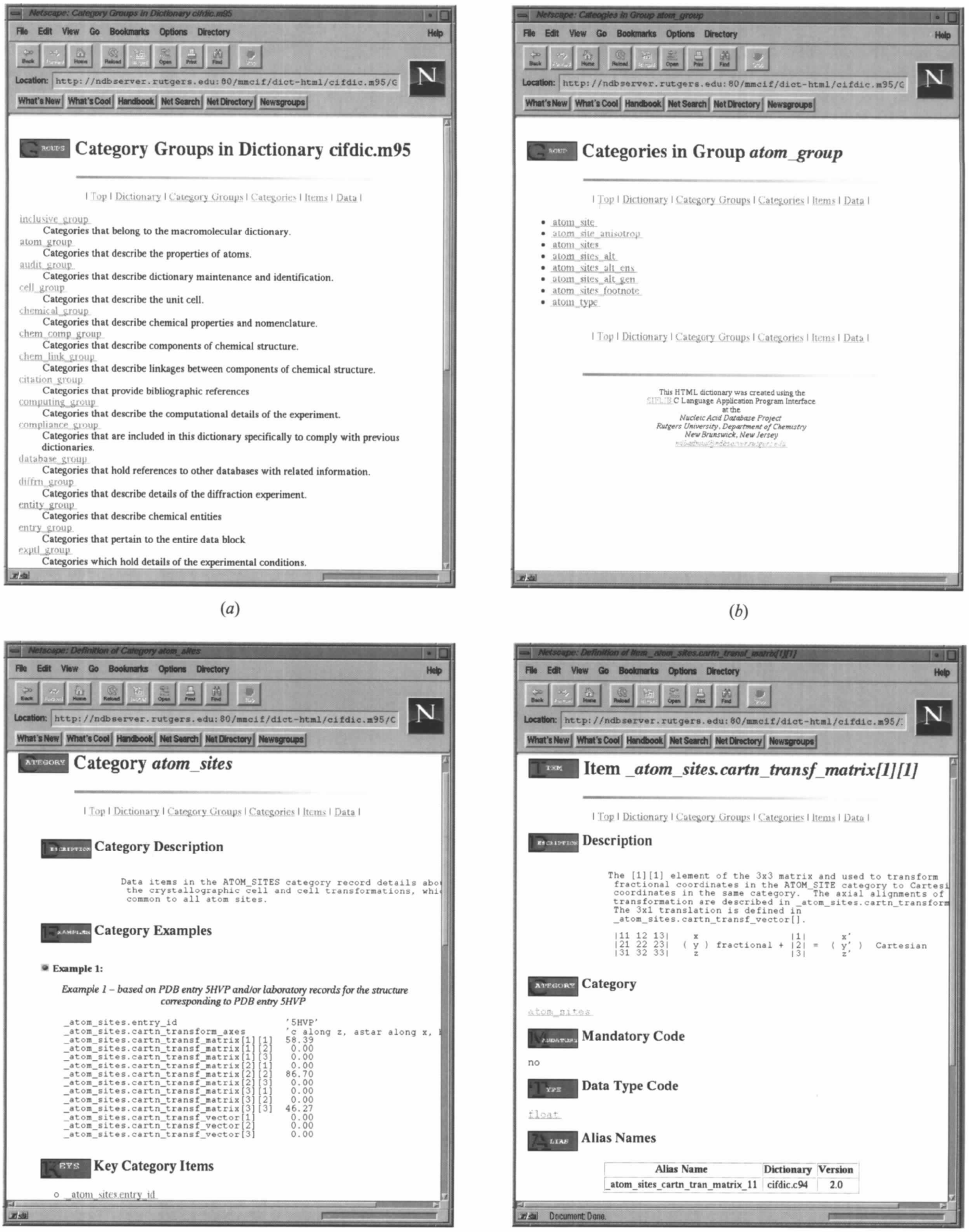
Category Group Selection: atom_group

cartn_x: 7.231

Tabular view of category = atom_site

id	label_atom_id	label_comp_id	label_asym_id	label_s
1	05*	G	A	1
2	C5*	G	A	1
3	C4*	G	A	1
4	04*	G	A	1
5	C3*	G	A	1
6	03*	G	A	1

Fig. 3. Main display window for the mmEdit browser application, the category group selection window and a portion of the help dialog for the `atom_site` category.



Netscape: Category Groups in Dictionary cifdic.m95
File Edit View Go Bookmarks Options Directory Help
Back Forward Home Reload Images Open Print Find Stop
Location: http://ndbserver.rutgers.edu:80/mmcif/dict-html/cifdic.m95/G
What's New What's Cool Handbook Net Search Net Directory Newsgroups
N
GROUPS

Category Groups in Dictionary cifdic.m95

| Top | Dictionary | Category Groups | Categories | Items | Data |

inclusive_group
Categories that belong to the macromolecular dictionary.
atom_group
Categories that describe the properties of atoms.
audit_group
Categories that describe dictionary maintenance and identification.
cell_group
Categories that describe the unit cell.
chemical_group
Categories that describe chemical properties and nomenclature.
chem_comp_group
Categories that describe components of chemical structure.
chem_link_group
Categories that describe linkages between components of chemical structure.
citation_group
Categories that provide bibliographic references
computing_group
Categories that describe the computational details of the experiment.
compliance_group
Categories that are included in this dictionary specifically to comply with previous dictionaries.
database_group
Categories that hold references to other databases with related information.
diffrn_group
Categories that describe details of the diffraction experiment.
entity_group
Categories that describe chemical entities
entry_group
Categories that pertain to the entire data block
exptl_group
Categories which hold details of the experimental conditions.

(*a*)

Netscape: Categories in Group atom_group
File Edit View Go Bookmarks Options Directory Help
Back Forward Home Reload Images Open Print Find Stop
Location: http://ndbserver.rutgers.edu:80/mmcif/dict-html/cifdic.m95/G
What's New What's Cool Handbook Net Search Net Directory Newsgroups
N
GROUP

Categories in Group *atom_group*

| Top | Dictionary | Category Groups | Categories | Items | Data |

- atom_site
- atom_site_anisotrop
- atom_sites
- atom_sites_alt
- atom_sites_alt_ens
- atom_sites_alt_gen
- atom_sites_footnote
- atom_type

| Top | Dictionary | Category Groups | Categories | Items | Data |

This HTML dictionary was created using the
CIFLIB C Language Application Program Interface
at the
Nucleic Acid Database Project
Rutgers University, Department of Chemistry
New Brunswick, New Jersey
ndbadmin@ndbserver.rutgers.edu

(*b*)

Netscape: Definition of Category atom_sites
File Edit View Go Bookmarks Options Directory Help
Back Forward Home Reload Images Open Print Find Stop
Location: http://ndbserver.rutgers.edu:80/mmcif/dict-html/cifdic.m95/C
What's New What's Cool Handbook Net Search Net Directory Newsgroups
N
CATEGORY
DESCRIPTION
EXAMPLES
KEYS

Category *atom_sites*

| Top | Dictionary | Category Groups | Categories | Items | Data |

Category Description

```
Data items in the ATOM_SITES category record details abo
 the crystallographic cell and cell transformations, whi
 common to all atom sites.
```

Category Examples

Example 1:

Example 1 – based on PDB entry 5HVP and/or laboratory records for the structure corresponding to PDB entry 5HVP

```
_atom_sites.entry_id                        '5HVP'
_atom_sites.cartn_transform_axes            'c along z, astar along x, 
_atom_sites.cartn_transf_matrix[1][1]  58.39
_atom_sites.cartn_transf_matrix[1][2]   0.00
_atom_sites.cartn_transf_matrix[1][3]   0.00
_atom_sites.cartn_transf_matrix[2][1]   0.00
_atom_sites.cartn_transf_matrix[2][2]  86.70
_atom_sites.cartn_transf_matrix[2][3]   0.00
_atom_sites.cartn_transf_matrix[3][1]   0.00
_atom_sites.cartn_transf_matrix[3][2]   0.00
_atom_sites.cartn_transf_matrix[3][3]  46.27
_atom_sites.cartn_transf_vector[1]      0.00
_atom_sites.cartn_transf_vector[2]      0.00
_atom_sites.cartn_transf_vector[3]      0.00
```

Key Category Items

○ _atom_sites.entry_id

(*c*)

Netscape: Definition of Item _atom_sites.cartn_transf_matrix[1][1]
File Edit View Go Bookmarks Options Directory Help
Back Forward Home Reload Images Open Print Find Stop
Location: http://ndbserver.rutgers.edu:80/mmcif/dict-html/cifdic.m95/
What's New What's Cool Handbook Net Search Net Directory Newsgroups
N
ITEM
DESCRIPTION
CATEGORY
MANDATORY
TYPE
ALIAS
Document: Done.

Item *_atom_sites.cartn_transf_matrix[1][1]*

| Top | Dictionary | Category Groups | Categories | Items | Data |

Description

```
The [1][1] element of the 3x3 matrix and used to transform
 fractional coordinates in the ATOM_SITE category to Cartesi
 coordinates in the same category.  The axial alignments of
 transformation are described in _atom_sites.cartn_transform
 The 3x1 translation is defined in
 _atom_sites.cartn_transf_vector[].

 |11 12 13|     x                 |1|       x'
 |21 22 23|   ( y ) fractional + |2| =  ( y' )  Cartesian
 |31 32 33|     z                 |3|       z'
```

Category

atom_sites

Mandatory Code

no

Data Type Code

float

Alias Names

Alias Name	Dictionary	Version
_atom_sites_cartn_tran_matrix_11	cifdic.c94	2.0

(*d*)

Fig. 4. Selected presentations from the CIF dictionary-to-HTML showing (*a*) the category-group organization of the dictionary, (*b*) the categories in the category group `atom_group`, (*c*) the description of the category `atom_sites` and (*d*) a portion of the description of an individual data item.

encountered in checking the target item. Similarly, functions providing update access return only the first error encountered in the checking process; however, all of the errors that may be detected during an I/O operation are appended to the warning or error lists maintained for each data block. Higher-level functions, which read and write files and data blocks, also append their diagnostic codes to internal error and warning lists. A set of functions has been provided to access and refresh these lists. Functions are also provided to translate individual error codes and to print the contents of an entire data block.

4. *CIFLIB* browser

The *mmEdit* browser application was developed as a general purpose browser and editor for macromolecular CIFs and CIF dictionaries. The browser was developed using OSF/Motif for the graphical user interface and using *CIFLIB* for all CIF access and navigation. The main *mmEdit* display window, shown in Fig. 3, is divided into three areas: a text viewer, an item edit area, and a table browser. The text viewer is just a scrollable window that allows the CIF to be scanned as an ordinary file. The item edit area divides a category row into data items that can be individually edited. The category browser can be used to select any row from the current category. Features such as data blocks, category groups, categories, subcategories and data items can be selected using pull-down menus. The category-group selection window is shown in Fig. 3. The help menu provides definitions and examples for the currently selected feature, and in Fig. 3 a portion of the help information is shown for the `atom_site` category. File access is also provided from a pull-down menu, and buttons are provided to navigate among rows and categories.

5. CIF dictionary-to-HTML converter

The dictionary-to-HTML converter application was developed to provide flexible access to the contents of the mmCIF and DDL dictionaries on the World Wide Web.*

The HTML mmCIF dictionary is organized so that a user can flexibly navigate through the hierarchy of the definitions and between all data-item relationships. The first level of presentation is a page of category groups and group descriptions. The contents of each category group can be explored and specific categories within each group can be selected. Each category is presented on a page that includes all of the DDL attributes pertaining to the category description. From within the category presentation, individual data items can be selected. The data item presentation includes all of the relevant DDL attributes and selections for all related data items. Each of these levels of presentation is illustrated in Fig. 4.

6. Language, documentation and availability

The C application-program interface described here is a library of wrapper functions for the *CIFLIB* class library. The class library was developed using the GNU C/C^{++} compiler and has been tested on a variety of UNIX platforms that support the 2.62 or later versions of the GNU compiler [*e.g.* Silicon Graphics (IRIX 5.3 & 6.2), SUN (SUN-OS 4.13), and Hewlett-Packard (HP-UX 9.05)]. The interface library package and example applications are available at http://ndbserver.rutgers.edu/software/CIFLIB or ftp://ndbserver.rutgers.edu/pub/programs/CIFLIB. Reference documentation is available in both PostScript and HTML format at this site.

For their contributions to the development of the mmCIF dictionary we gratefully acknowledge the members of the Macromolecular CIF working group: Phil Bourne at San Diego Supercomputer Center and Helen Berman at Rutgers. Thanks also go to Brian McMahon at the IUCr office in Chester for many valuable discussions on core dictionary compatibility. This work was supported by a grant from NSF (BIR9305135) for the Nucleic Acid Database.

* The HTML dictionaries are available at http://ndbserver.rutgers.edu/mmcif.

References

Berman, H. M. & Westbrook, J. D. (1993). *Proceedings of the First Macromolecular CIF Tools Workshop*, edited by P. E. Bourne, p. 65. Tarrytown, New York: National Science Foundation.

Berman, H. M. & Westbrook, J. D. (1994). *European Macromolecular Crystallographic Information (mmCIF) Workshop*, edited by S. D. Wodak. Free University of Brussels: European Commission.

Bourne, P. E., Berman, H. M., McMahon, B., Watenpaugh, K. D., Westbrook, J. & Fitzgerald, P. M. D. (1997). *Methods Enzymol.* Submitted.

Cook, A. F. P. (1991). *Dictionary Definition Language in STAR File format.* ORAC Report.

Fitzgerald, P., Berman, H. M., Bourne, P., McMahon, B., Watenpaugh, K. & Westbrook, J. D. (1997). *The Macromolecular Crystallographic Information File Dictionary,* http://ndbserver.rutgers.edu/mmcif.

Fitzgerald, P. M. D., Berman, H. M., Bourne, P. E., & Watenpaugh, K. (1993). *The Macromolecular CIF dictionary.* ACA Annual Meeting, Albuquerque, New Mexico, USA. No. D008.

Hall, S. R. (1991). *J. Chem. Inf. Comput. Sci.* **31**, 326–333.

Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *Acta Cryst.* A**47**, 655–685.

Hall, S. R. & Cook, A. F. P. (1995). *J. Chem. Inf. Comput. Sci.* **35**, 819–825.

Toby, B. (1993). *Proceedings of the First Macromolecular CIF Tools Workshop*, edited by P. E. Bourne, p. 49. Tarrytown, New York: National Science Foundation.

Ulrich, E. (1995). *The Nuclear Magnetic Resonance Information File Dictionary.* BioMagResBank Project at University of Wisconsin, http://nmrfam.wisc.edu.

Westbrook, J. D. & Hall, S. R. (1997). In preparation.