W O R L D W I D E PDB PROTEIN DATA BANK

wwPDB.org

The Life and Times of the PDB Format - Looking Towards the Future with mmCIF

Ezra Peisach[1], Sanja Abbott[5], Kumaran Baskaran[4], John Berrisford[2], Zukang Feng[1], Yasuyo Ikegawa[3], Catherine Lawson[1], John D. Westbrook[1], Masashi Yokochi[3], Jasmine Y. Young[1], Jeffrey Hoch[4], Genji Kurisu[3], Sameer Velankar[2], Stephen K. Burley[1]

Abstract

For 50 years, the PDB file format has been a standard used by many software packages. It is a fixed column format with character limits and cannot be extended.

In three to four years, wwPDB will need to extend the width of chemical component ids to four characters. When this happens, PDB files cannot be produced.

In addition, wwPDB also plans to implement extended PDB IDs beyond four characters. Once the four-character PDB IDs are all consumed, newly deposited PDB entries will only be available in PDBx/mmCIF format.

wwPDB is asking community and user software developers to review their code and ensure compatibility for the future.

PDBx/mmCIF is the Solution

- Since late 1990s, PDB archive has provided entries in PDBx/mmCIF format https://mmcif.wwpdb.org
- mmCIF is an extensible machine-readable dictionary-based format (Figure 3)
- Since 2014, entries that could not be produced in PDB format have been PDBx/mmCIF on;y
- The PDBx/mmCIF dictionary already has metadata not present in PDB file format. i.e., XFEL experiments

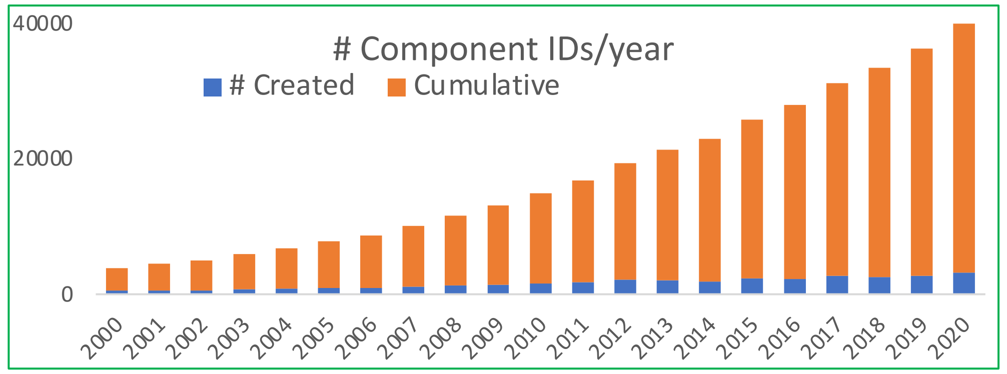

Figure 2: Number of chemical components created per year.

PDB Format and its Challenges

- PDB format was established in 1971
- See https://www.wwpdb.org/documentation/file-format
- Based on 72 column fixed width format (Figure 1)
- Last version of PDB format released in 2011 and frozen
- Limitations <=62 chains or <= 99999 atom records, so cannot accommodate “large” entries
- PDB format limits chemical components (residue names) to three alphanumeric characters
- PDB ID is limited to four characters
- The available three-characters chemical component names will be exhausted within the next three to four years. (Figure 2)

```
loop_
_database_2.database_id
_database_2.database_code
_database_2.pdbx_database_accession
_database_2.pdbx_DOI
PDB 1abc pdb_00001abc 10.2210/pdb1abc/pdb
WWPDB D_1xxxxxxxxx ? ?
#
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_alt_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.label_entity_id
_atom_site.label_seq_id
_atom_site.pdbx_PDB_ins_code
_atom_site.Cartn_x
_atom_site.auth_atom_id
. . .
ATOM 1 N N . SER A 1 1 ? 18.167 19.270 0.618 1.00 50.48  12
SER A N
. . .
```

Figure 3: Portion of a PDBx/mmCIF sample file.

```
COLUMNS     DATA TYPE     FIELD     DEFINITION
----------------------------------------------------------------------
 1 -  6     Record name   "ATOM  "
 7 - 11     Integer       serial     Atom serial number.
13 - 16     Atom          name       Atom name.
17          Character     altLoc     Alternate location indicator.
18 - 20     Residue name  resName    Residue name.
22          Character     chainID    Chain identifier.
23 - 26     Integer       resSeq     Residue sequence number.
27          AChar         iCode      Code for insertion of residues.
31 - 38     Real(8.3)     x          Coordinates for X in Angstroms.
. . .

ATOM      1  N   SER A  12      18.167  19.270   0.618  1.00 50.48           N
```

Figure 1: Partial column description of PDB file format and sample ATOM record.

Plan

- wwPDB will start issuing four-character codes for chemical components
- wwPDB will also expand PDB IDs to an eight-character format with prefix pdb_00001abc (see Figure 4)
- Once introduced, **PDB formatted files cannot be produced for these entries**

Action Needed by Users & Developers

- Support mmCIF format
- Correct the assumptions as to field width
- Achieve the above in two years

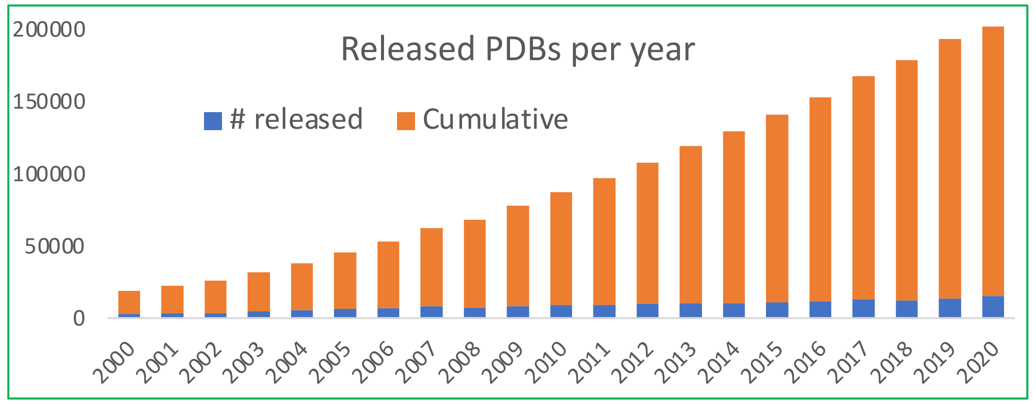

Figure 4: Growth in the number of assigned PDB IDs.

wwPDB members RCSB PDB BMRB PDBe PDBj EMDB

[1]RCSB Protein Data Bank, Rutgers, The State University of New Jersey, Piscataway, USA [2]Protein Data Bank in Europe, EMBL-EBI, Hinxton, UK.
[3]Protein Data Bank Japan, Institute for Protein Research, Osaka University, Japan [4]BMRB, University of Connecticut, USA, [5]EMDB, EMBL-EBI, Hinxton, UK

Presenter contact: ezra.peisach@rcsb.org