Title

Announcing mandatory submission of PDBx/mmCIF format files for crystallographic depositions to the Protein Data Bank (PDB)

Journal

Acta Crystallographica Section D, Structural Biology, 75(4)

ISSN

2059-7983

Authors

Adams, Paul D

Afonine, Pavel V

Baskaran, Kumaran

et al.

Publication Date

2019-04-01

DOI

10.1107/s2059798319004522

Peer reviewed

STRUCTURAL BIOLOGY
ISSN 2059-7983

Announcing mandatory submission of PDBx/mmCIF format files for crystallographic depositions to the Protein Data Bank (PDB)

Paul D. Adams,[a,b] Pavel V. Afonine,[a] Kumaran Baskaran,[c] Helen M. Berman,[d] John Berrisford,[e] Gerard Bricogne,[f] David G. Brown,[g] Stephen K. Burley,[d,h,i]* Minyu Chen,[j] Zukang Feng,[d] Claus Flensburg,[f] Aleksandras Gutmanas,[e] Jeffrey C. Hoch,[k]* Yasuyo Ikegawa,[j] Yumiko Kengaku,[j] Eugene Krissinel,[l] Genji Kurisu,[j]* Yuhe Liang,[d] Dorothee Liebschner,[a] Lora Mak,[e] John L. Markley,[c]* Nigel W. Moriarty,[a] Garib N. Murshudov,[m] Martin Noble,[n] Ezra Peisach,[d] Irina Persikova,[d] Billy K. Poon,[a] Oleg V. Sobolev,[a] Eldon L. Ulrich,[c] Sameer Velankar,[e]* Clemens Vonrhein,[f] John Westbrook,[d] Marcin Wojdyr,[f,l] Masashi Yokochi[j] and Jasmine Y. Young[d]

[a]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA, [b]Department of Bioengineering, University of California, Berkeley, CA 94720, USA, [c]BioMagResBank (BMRB), University of Wisconsin-Madison, Madison, WI 53706, USA, [d]Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA, [e]Protein Data Bank in Europe (PDBe), European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK, [f]Global Phasing Limited, Sheraton House, Castle Park, Cambridge, CB3 0AX, UK, [g]School of Biosciences, University of Kent, Canterbury, Kent CT2 7NJ, UK, [h]Rutgers Cancer Institute of New Jersey, Robert Wood Johnson Medical School, New Brunswick, NJ 08903, USA, [i]Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093, USA, [j]Protein Data Bank Japan (PDBj), Institute for Protein Research, Osaka University, Osaka, 565-0871, Japan, [k]BioMagResBank (BMRB), UConn Health, 263 Farmington Avenue, Farmington, CT 06030, USA, [l]CCP4, Research Complex at Harwell (RCaH), Rutherford Appleton Laboratory, Didcot, Oxon OX11 0FA, UK, [m]MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge, CB2 0QH, UK, and [n]Newcastle University, Framlington Place, Newcastle Upon Tyne, NE2 4HH, UK. *Correspondence e-mail: stephen.burley@rcsb.org, hoch@uchc.edu, gkurisu@protein.osaka-u.ac.jp, markley@biochem.wisc.edu, sameer@ebi.ac.uk

Received 21 February 2019
Accepted 3 April 2019

Edited by R. J. Read, University of Cambridge, England

Keywords: PDB; mmCIF; OneDep; wwPDB; data dictionary; data archiving; biocuration; validation; macromolecular crystallography; data standards; PDBx/mmCIF format; Protein Data Bank; Worldwide Protein Data Bank.

The Protein Data Bank (PDB) (wwPDB consortium, 2019) is the single global archive of experimentally determined three-dimensional (3D) structure data of biological macromolecules. The continuing growth in the numbers, size and complexity of macromolecular structures in the PDB archive, coupled with the rapid growth of evolving experimental methods such as 3D cryo-electron microscopy (3DEM) has made the traditional PDB format ('legacy PDB format') inadequate for fully representing these data. As described below, this format was based on a punched-card format that became obsolete long ago. In the following letter, we describe the changes necessary to address the challenges coming from the extraordinary success of structural biologists.

Since 2003, the PDB has been managed by the Worldwide Protein Data Bank (wwPDB; https://www.wwpdb.org/) (Berman *et al.*, 2003), an international partnership that collaboratively oversees deposition, validation, biocuration and open-access dissemination of 3D macromolecular structure data, adhering to the FAIR principles of Findability, Accessibility, Interoperability and Reusability (Wilkinson *et al.*, 2016). In 2007, the master file format for the archive was officially changed to PDB Exchange/Macromolecular Crystallographic Information File (PDBx/mmCIF), supported by the PDBx/mmCIF data dictionary, to address new challenges in structure archiving. Later, in 2012, the wwPDB terminated its support of the legacy PDB file format and froze its further development (https://wwpdb.org/documentation/file-formats-and-the-pdb).

We now announce that as of 1 July 2019, PDBx/mmCIF will be the only format allowed for deposition of the atomic coordinates for PDB structures resulting from macromolecular crystallography (MX), including X-ray, neutron, fiber and electron diffraction methods, *via OneDep* (Young *et al.*, 2017). This requirement will be extended to PDB structures resulting from nuclear magnetic resonance (NMR) spectroscopy and 3DEM methods at a later date to be determined. Elimination of the legacy PDB format will improve the efficiency of the deposition process and enhance validation through capture

of the more extensive experimental metadata supported by PDBx/mmCIF. By 2021, we anticipate that the PDB Chemical Component Identifier will need to be extended beyond three characters, which will result in full retirement of files in the PDB Core Archive that utilize the legacy PDB format.

The PDBx/mmCIF dictionary and format are machine-readable and fully extensible; they assure that the wwPDB partners can effectively represent, biocurate, validate and distribute structural biology data. The data dictionary is regularly updated to reflect evolution of the underlying science and technology. The current version of the dictionary (5.303), which includes >4300 data items, is organized as 553 interlinked categories, clustered into 41 groups that describe atomic coordinates, polymer and small-molecule chemistry, details of sample production, experimental setup, data collection and processing, and cross-references to other biological databases. To ensure consistency, more than 25% of the data items are subject to controlled vocabularies, which are periodically updated (Young *et al.*, 2018). Unlike the legacy PDB format, PDBx/mmCIF has no limitation on the size of structures that can be represented. The richer information content of PDBx/mmCIF also enables better validation of incoming data, which saves time for data depositors and wwPDB biocurators, and improves the consistency and quality of the PDB archive.

The wwPDB currently distributes atomic coordinate models from the PDB archive in both PDBx/mmCIF (http://mmcif.wwpdb.org) and XML/PDBML formats (http://pdbml.wwpdb.org). The biological assemblies are fully described in PDBx/mmCIF, but only partially described in the legacy PDB format (Lawson *et al.*, 2008). PDBx/mmCIF formatted files are available from *REFMAC* (Murshudov *et al.*, 2011), *PHENIX* (Adams *et al.*, 2010) and Global Phasing *BUSTER* (Blanc *et al.*, 2004; Bricogne *et al.*, 2009) – the three major MX software systems accounting for >90% of MX depositions. For users of other structure determination/refinement software packages, the wwPDB provides standalone and web-based tools to convert legacy PDB format files into PDBx/mmCIF format: *pdb_extract* (https://pdb-extract.wwpdb.org) (Yang *et al.*, 2004) and *MAXIT* (https://sw-tools.rcsb.org/apps/MAXIT/index.html). We encourage authors of all the software packages serving MX and other structural biology communities to produce PDBx/mmCIF files as default output for direct deposition to the PDB.

The legacy PDB format (based on the Hollerith IBM 80-column punched-card format) was first introduced by the PDB in 1970s (Protein Data Bank, 1971), wherein 'ATOM' records included atom and residue names and sequence numbers, and the 'HEADER' record contained a limited amount of metadata (https://wwpdb.org/documentation/file-format). The punched-card origin of the PDB format imposed hard limits on the maximum number of atoms (99 999) and polymer chains (62) that could be represented in a single PDB file. Advances in structural biology meant that by the mid-1990s some of the new structures deposited to the PDB archive (*e.g.* ribosomal subunits) could not be represented within a single legacy PDB file. As an interim measure, these structures were 'split' into multiple PDB entries, each with its own accession code, thereby rendering the data very difficult to use. wwPDB partners will continue to provide legacy PDB format files from the PDB Core Archive on a 'best-efforts' basis, with the caveat that many of the larger PDB structures cannot be distributed using the legacy format.

The mmCIF format was proposed as a replacement to the legacy PDB format in 2005 (Fitzgerald *et al.*, 2005*b*). The mmCIF format is based on a data dictionary, originally developed under the auspices of the International Union of Crystallography (IUCr; Hall *et al.*, 1991). Following its initial introduction, the mmCIF dictionary was extended to represent structure determination data from NMR spectroscopy and 3DEM, and renamed PDBx/mmCIF (Westbrook *et al.*, 2005). In 2007, this extended dictionary became the foundation for the master format for data used by the wwPDB to archive PDB data (Fitzgerald *et al.*, 2005*a*; Westbrook & Fitzgerald, 2009). For deposition of NMR data, the BioMagResBank (BMRB) utilizes the NMR-STAR format, which is fully compatible with PDBx/mmCIF (Ulrich *et al.*, 1996, 2018).

Every data item defined in the PDBx/mmCIF dictionary has attributes describing its features, including relationships to other data items. As outlined above, the mmCIF format has no limitations with respect to the size of archived structure, is fully machine-readable, and is extensible for new data content. In 2015, the formerly 'split' entry structures archived in the PDB were combined into single entries with PDBx/mmCIF formatted files, which enables users to visualize and analyze every one of the larger PDB structures in their entirety.

Atomic coordinate models of macromolecular structures determined by crystallographic methods are described using the PDBx/mmCIF dictionary (http://mmcif.wwpdb.org/), which specifies categories of information represented as tables and key-value pairs. Every data item found in the legacy PDB format has a corresponding data item(s) in the PDBx/mmCIF format, but the reverse is not true. Legacy PDB files should, therefore, no longer be used in bioinformatics or structural biology workflows. Chemical descriptions of all of the subunits and ligands in PDB entries are provided in the PDB Chemical Component Dictionary, which is also defined within the PDBx/mmCIF data dictionary.

The PDBx/mmCIF Working Group (https://wwpdb.org/task/mmcif), whose members include major crystallographic structure determination software developers, has committed to the PDBx/mmCIF data model. PDBx/mmCIF format files can be output from the following software packages: *CCP*4/*REFMAC* (Murshudov *et al.*, 2011), *PHENIX* (Adams *et al.*, 2010), Global Phasing *BUSTER* (Bricogne *et al.*, 2009), and dedicated PDB deposition tasks in *CCP*4*i*2 (Potterton *et al.*, 2018) and *CCP*4 *Cloud* (Krissinel *et al.*, 2018). PDBx/mmCIF is also supported by visualization software applications, including *CCP*4*mg* (Potterton *et al.*, 2004), *Coot* (Brown *et al.*, 2015), *Chimera* (Goddard *et al.*, 2018), *Jmol/JSMol* (Hanson, 2010; Hanson *et al.*, 2013), *LiteMole* (Sehnal *et al.*, 2017), *Molmil* (Bekker *et al.*, 2016), *NGL* (Rose *et al.*, 2018),

OpenRasMol (Bernstein, 2000), *PyMoL* (DeLano, 2002) and *VMD* (Humphrey *et al.*, 1996). In addition, other data resources, such as the Protein Model Portal (Haas *et al.*, 2013) and SASBDB (Malfois & Svergun, 2000; Kachala *et al.*, 2016), have adopted and extended the PDBx/mmCIF framework for data representation.

Funding information

Funding for the wwPDB consortium comes from multiple sources. The RCSB Protein Data Bank is funded jointly by the National Science Foundation, National Institute of General Medical Sciences, National Cancer Institute and Department of Energy. The Protein Data Bank in Europe is supported by the European Molecular Biology Laboratory-European Bioinformatics Institute, Wellcome Trust, Biotechnology and Biological Sciences Research Council, and the European Union. The Protein Data Bank Japan is supported by the Database Integration Coordination Program from the National Bioscience Database Centre (NBDC)-JST (Japan Science and Technology Agency) and the joint usage program of Institute for Protein Research, Osaka University. The BioMagResBank is supported by the US National Institutes of Health.

References

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. (2010). *Acta Cryst.* D**66**, 213–221.

Bekker, G. J., Nakamura, H. & Kinjo, A. R. (2016). *J. Cheminform*, **8**, 42, 1–5.

Berman, H. M., Henrick, K. & Nakamura, H. (2003). *Nat. Struct. Biol.* **10**, 980.

Bernstein, H. J. (2000). *Trends Biochem. Sci.* **25**, 453–455.

Blanc, E., Roversi, P., Vonrhein, C., Flensburg, C., Lea, S. M. & Bricogne, G. (2004). *Acta Cryst.* D**60**, 2210–2221.

Bricogne, G., Blanc, E., Brandl, M., Flensburg, C., Keller, P., Paciorek, W., Roversi, P., Sharff, A., Smart, O. S., Vonrhein, C. & Womack, T. O. (2009). *BUSTER*, Global Phasing Ltd., Cambridge, UK.

Brown, A., Long, F., Nicholls, R. A., Toots, J., Emsley, P. & Murshudov, G. (2015). *Acta Cryst.* D**71**, 136–153.

DeLano, W. (2002). *The pyMOL molecular graphics system*. http://www.pymol.org.

Fitzgerald, P. M. D., Westbrook, J. D., Bourne, P. E., McMahon, B., Watenpaugh, K. D. & Berman, H. M. (2005*a*). *International Tables for Crystallography*, Vol. G, edited by S. R. Hall & B. McMahon, pp. 144–198. Dordrecht, The Netherlands: Springer.

Fitzgerald, P. M. D., Westbrook, J. D., Bourne, P. E., McMahon, B., Watenpaugh, K. D. & Berman, H. M. (2005*b*). *International Tables for Crystallography*, Vol. G, edited by S. R. Hall & B. McMahon, pp. 295–443. Dordrecht, The Netherlands: Springer.

Goddard, T. D., Huang, C. C., Meng, E. C., Pettersen, E. F., Couch, G. S., Morris, J. H. & Ferrin, T. E. (2018). *Protein Sci.* **27**, 14–25.

Haas, J., Roth, S., Arnold, K., Kiefer, F., Schmidt, T., Bordoli, L. & Schwede, T. (2013). *Database*, **2013**, bat031.

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

Hanson, R. M. (2010). *J. Appl. Cryst.* **43**, 1250–1260.

Hanson, R. M., Prilusky, J., Renjian, Z., Nakane, T. & Sussman, J. L. (2013). *Isr. J. Chem.* **53**, 207–216.

Humphrey, W., Dalke, A. & Schulten, K. (1996). *J. Mol. Graph.* **14**, 33–38.

Kachala, M., Westbrook, J. & Svergun, D. (2016). *J. Appl. Cryst.* **49**, 302–310.

Krissinel, E., Uski, V., Lebedev, A., Winn, M. & Ballard, C. (2018). *Acta Cryst.* D**74**, 143–151.

Lawson, C. L., Dutta, S., Westbrook, J. D., Henrick, K. & Berman, H. M. (2008). *Acta Cryst.* D**64**, 874–882.

Malfois, M. & Svergun, D. I. (2000). *J. Appl. Cryst.* **33**, 812–816.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.

Potterton, L., Agirre, J., Ballard, C., Cowtan, K., Dodson, E., Evans, P. R., Jenkins, H. T., Keegan, R., Krissinel, E., Stevenson, K., Lebedev, A., McNicholas, S. J., Nicholls, R. A., Noble, M., Pannu, N. S., Roth, C., Sheldrick, G., Skubak, P., Turkenburg, J., Uski, V., von Delft, F., Waterman, D., Wilson, K., Winn, M. & Wojdyr, M. (2018). *Acta Cryst.* D**74**, 68–84.

Potterton, L., McNicholas, S., Krissinel, E., Gruber, J., Cowtan, K., Emsley, P., Murshudov, G. N., Cohen, S., Perrakis, A. & Noble, M. (2004). *Acta Cryst.* D**60**, 2288–2294.

Protein Data Bank (1971). *Nature New Biol.* **233**, 223.

Rose, A. S., Bradley, A. R., Valasatava, Y., Duarte, J. M., Prlić, A. & Rose, P. W. (2018). *Bioinformatics*, bty419.

Sehnal, D., Deshpande, M., Vařeková, R. S., Mir, S., Berka, K., Midlik, A., Pravda, L., Velankar, S. & Koča, J. (2017). *Nat. Methods*, **14**, 1121–1122.

Ulrich, E. L., Argentar, D., Klimowicz, A. & Markley, J. L. (1996). *Acta Cryst.* A**52**, C577–C577.

Ulrich, E. L., Baskaran, K., Dashti, H., Ioannidis, Y. E., Livny, M., Romero, P. R., Maziuk, D., Wedell, J. R., Yao, H., Eghbalnia, H. R., Hoch, J. C. & Markley, J. L. (2018). *J. Biomol. NMR*, pp. 1–5.

Westbrook, J., Henrick, K., Ulrich, E. L. & Berman, H. M. (2005). *International Tables for Crystallography*, Vol. G, edited by S. R. Hall & B. McMahon, pp. 195–198. Dordrecht, The Netherlands: Springer.

Westbrook, J. D. & Fitzgerald, P. M. D. (2009). *Structural Bioinformatics*, 2nd ed, edited by P. E. Bourne & J. Gu, pp. 271–291. Hoboken, NJ: John Wiley & Sons, Inc.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., t Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S. A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. (2016). *Sci Data*, **3**, 1–9.

wwPDB consortium (2019). *Nucleic Acids Res.* **47**, D520–D528.

Yang, H., Guranovic, V., Dutta, S., Feng, Z., Berman, H. M. & Westbrook, J. D. (2004). *Acta Cryst.* D**60**, 1833–1839.

Young, J. Y., Westbrook, J. D., Feng, Z., Peisach, E., Persikova, I., Sala, R., Sen, S., Berrisford, J. M., Swaminathan, G. J., Oldfield, T. J., Gutmanas, A., Igarashi, R., Armstrong, D. R., Baskaran, K., Chen, L., Chen, M., Clark, A. R., Di Costanzo, L., Dimitropoulos, D., Gao, G., Ghosh, S., Gore, S., Guranovic, V., Hendrickx, P. M. S., Hudson, B. P., Ikegawa, Y., Kengaku, Y., Lawson, C. L., Liang, Y., Mak, L., Mukhopadhyay, A., Narayanan, B., Nishiyama, K., Patwardhan, A., Sahni, G., Sanz-Garcia, E., Sato, J., Sekharan, M. R., Shao, C., Smart, O. S., Tan, L., van Ginkel, G., Yang, H., Zhuravleva, M. A., Markley, J. L., Nakamura, H., Kurisu, G., Kleywegt, G. J., Velankar, S., Berman, H. M. & Burley, S. K. (2018). *Database*, **2018**, bay002.

Young, J. Y., Westbrook, J. D., Feng, Z., Sala, R., Peisach, E., Oldfield, T. J., Sen, S., Gutmanas, A., Armstrong, D. R., Berrisford, J. M.,

Chen, L., Chen, M., Di Costanzo, L., Dimitropoulos, D., Gao, G., Ghosh, S., Gore, S., Guranovic, V., Hendrickx, P. M. S., Hudson, B. P., Igarashi, R., Ikegawa, Y., Kobayashi, N., Lawson, C. L., Liang, Y., Mading, S., Mak, L., Mir, M. S., Mukhopadhyay, A., Patwardhan, A., Persikova, I., Rinaldi, L., Sanz-Garcia, E., Sekharan, M. R., Shao, C., Swaminathan, G. J., Tan, L., Ulrich, E. L., van Ginkel, G., Yamashita, R., Yang, H., Zhuravleva, M. A., Quesada, M., Kleywegt, G. J., Berman, H. M., Markley, J. L., Nakamura, H., Velankar, S. & Burley, S. K. (2017). *Structure*, **25**, 536–545.