The archiving and dissemination of biological structure data

Helen M Berman¹, Stephen K Burley¹,², Gerard J Kleywegt³, John L Markley⁴, Haruki Nakamura⁵ and Sameer Velankar³

The global Protein Data Bank (PDB) was the first open-access digital archive in biology. The history and evolution of the PDB are described, together with the ways in which molecular structural biology data and information are collected, curated, validated, archived, and disseminated by the members of the Worldwide Protein Data Bank organization (wwPDB; http://wwpdb.org). Particular emphasis is placed on the role of community in establishing the standards and policies by which the PDB archive is managed day-to-day.

Addresses
¹ Research Collaboratory for Structural Bioinformatics Protein Data Bank, Department of Chemistry and Chemical Biology, Center for Integrative Proteomics Research, Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, 174 Frelinghuysen Road, Piscataway, NJ 08854, USA
² Research Collaboratory for Structural Bioinformatics Protein Data Bank, Skaggs School of Pharmacy and Pharmaceutical Sciences and San Diego Supercomputer Center, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92030, USA
³ Protein Data Bank in Europe, European Molecular Biology Laboratory — European Bioinformatics Institute, Wellcome Genome Campus, Cambridge CB10 1SD, UK
⁴ Biological Magnetic Resonance Bank, Department of Biochemistry, University of Wisconsin-Madison, Madison, WI 53706, USA
⁵ Protein Data Bank Japan, Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka, 565-0871, Japan

Corresponding author: Berman, Helen M (berman@rcsb.rutgers.edu)

The PDB as a community data resource

From its inception, the PDB has been a community effort that has evolved with changes in scientific culture. For example, when the PDB was first created, data submission was voluntary. However, in the 1980s, members of the community became outspoken about the need to enforce mandatory data deposition. Various committees were set up to define what data should be required and when to disseminate the data. These guidelines were published in 1989, and over time, adopted by virtually all of the scientific journals that now require PDB deposition(s) as a prerequisite for publication of structural studies [9]. In 2008, further shifts in community sentiment led to mandatory deposition of experimental data together with atomic coordinates. In the current decade, the importance of reproducibility has been highlighted. The PDB convened method-specific Validation Task Forces and Workshops [10*,11*,12*,13*] to define what data should be collected and how best to validate the structural models, the experimental data, and the fit of the models to the data. Now every structure in the PDB comes with a publicly available validation report, and
authors are strongly encouraged to include these reports with their manuscript submissions to journals.

The importance of global participation in data archiving was understood early in the creation of the PDB. Indeed, the announcement of the PDB in 1971 described the collaboration with the Cambridge Crystallographic Database Centre [7]. In 2003, a Memorandum of Understanding (MOU) among partners in the US (RCSB Protein Data Bank; http://www.rcsb.org), Japan (Protein Data Bank Japan or PDBj; http://www.pdbj.org), and Europe (Protein Data Bank in Europe or PDBs; http://pdbe.org) established the Worldwide Protein Data Bank (wwPDB) partnership, which is responsible for formalizing the procedures involved in collecting, standardizing, annotating and disseminating the data [14*]. Subsequently, a global NMR specialist data repository BioMagResBank, composed of deposition sites in the US (BMRB; http://www.bmrb.wisc.edu) and Japan (PDBj-BMRB; http://bmrbdep.pdbj.org), joined the wwPDB.

The X-ray crystallography community has led the biological sciences in the area of data sharing. While the sociological/anthropological underpinnings of this leadership role have not been fully explored, much of what has transpired in the creation and evolution of the PDB can be traced to J.D. Bernal, who, in addition to being a brilliant scientific innovator, was a prominent social activist, whose beliefs were consistent with the conduct of the PDB [15].

**Content of the PDB archive**

The PDB archive contains information about structural models that have been derived from experimental methods, including X-ray/neutron/electron crystallography, NMR spectroscopy, and 3D electron microscopy (3DEM). In addition to the 3D coordinates, the details of the chemistry of the polymers and small molecules are archived, as are metadata describing the experimental conditions, data-processing statistics and structural features such as the secondary and quaternary structure. The structure-factor amplitudes (or intensities) used to determine X-ray structures, and chemical shifts and restraints used in determining NMR structures are also archived. The electron density maps used to derive 3DEM models are archived in EMDB [16*], and the experimental data underpinning them can be archived in EMPIAR [17]. In collaboration with community experts, pertinent data items are defined for each experimental field, with requirements evolving over time. The PDB data dictionary, originally developed to describe macromolecular crystallography, contains more than 4400 data items. The dictionary combines data items common to all methods as well as those that are method specific. For example, the current dictionary contains 250 NMR-specific data and 1200 3DEM-specific data definitions.

Over time, the holdings of the PDB have increased dramatically as has the complexity of the structures being archived (Figure 1).

A workshop held in 2005 led to the policy that purely *in silic* models should not be part of the PDB [18*]. and, instead, a modeling portal should be created for these models. The Protein Modeling Portal was established in 2007 [19].

**Representation of PDB data**

The first data format used by the PDB was established in the early 1970s and was on the basis of the 80-column Hollerith format used for punched cards. The atom records included atom name, residue name and sequence number. A ‘header record’ contained some metadata. This format was readily accepted because it was simple and both human- and machine-readable. However, it had many serious drawbacks in that the size of the structural models was limited to 99 999 atoms and that relationships among the data items were implicit. These inherent weaknesses meant that significant domain knowledge was necessary in order to write software using this format.

In the 1990s, the IUCr chartered a committee to create a more formal data model. This committee proposed the Macromolecular Crystallographic Information File (mmCIF) [20*]. mmCIF is a self-defining format in which every data item has attributes describing its features including relationships to other data items. Most importantly, mmCIF has no limitations with respect to the size of the archived structural model. The dictionary and the data files are completely machine-readable, and no domain knowledge is required to read the files. The first dictionary contained over 3000 data items relevant to X-ray crystallography. Over time, terms specific to NMR and 3DEM were added, and the dictionary was renamed PDBx/mmCIF. In 2007, it was decided that PDBx would be the Master Format for data collected by the PDB. In 2011, major X-ray structure determination software developers agreed to adopt this data model so that all output from their programs would be in PDBx. In 2015, large structures archived in the PDB that had formerly been split into multiple entries were combined into single entries and mmCIF formatted files. Other structural biology communities are in the process of building on the PDBx/mmCIF framework to establish their own controlled vocabulary and specialist data items [19,21].

PDBML, an XML format on the basis of PDBx/mmCIF [22], and its RDF (Resource Description Framework) conversion were developed to facilitate the integration of structure data with other life sciences data resources could be facilitated [23*].

**The data pipeline**

Every data resource has a set of procedures for deposition, curation, validation, archiving and dissemination of data.
The pipeline currently used by the wwPDB to populate the PDB archive is illustrated schematically in Figure 2.

In the very early days of the PDB, structures were deposited to BNL on magnetic tapes containing atomic coordinates with paper forms listing other data items, all sent first by mail and then via a web-based system, called AutoDep, was created in the 1990s [24]. This system was later modified and used by PDBe [25] until very recently. The RCSB PDB and PDBe collected data using a system on the basis of mmCIF called ADIT [26], and the BMRB in the US and its affiliate in Japan adopted a similar system called ADIT-NMR [27]. Although these systems were distinct, since 2003, the wwPDB partners have determined jointly what data should be collected and which procedures and algorithms should be used for data processing. In 2007, it was agreed within the wwPDB to create a single deposition, Structures are made available to the public either immediately after they have been fully curated or—in most cases—when they are published in a journal. Usually, either the author or the journal informs wwPDB that the paper describing the structure is about to be published. PDB data are released in a two-stage process. Every Saturday at 03:00 UTC the polymer sequences, ligand SMILES strings, and crystallization pH for new structures designated for release are made public (http://wwpdb.org/download/downloads) as a courtesy to the protein structure modeling and computational chemistry communities to enable weekly blinded prediction challenge efforts (e.g., CAMEO [19] and D3R CELPP [28]). Every Wednesday at 00:00 UTC, all new structures designated for release are made publicly available through the wwPDB FTP sites. On average about 200 structures are released every week. As evidence

![Growth of the PDB archive.](image)

**Figure 1**

The pipeline currently used by the wwPDB to populate the PDB archive is illustrated schematically in Figure 2.

In the very early days of the PDB, structures were deposited to BNL on magnetic tapes containing atomic coordinates with paper forms listing other data items, all sent first by mail and then via a web-based system, called AutoDep, was created in the 1990s [24]. This system was later modified and used by PDBe [25] until very recently. The RCSB PDB and PDBe collected data using a system on the basis of mmCIF called ADIT [26], and the BMRB in the US and its affiliate in Japan adopted a similar system called ADIT-NMR [27]. Although these systems were distinct, since 2003, the wwPDB partners have determined jointly what data should be collected and which procedures and algorithms should be used for data processing. In 2007, it was agreed within the wwPDB to create a single deposition, Structures are made available to the public either immediately after they have been fully curated or—in most cases—when they are published in a journal. Usually, either the author or the journal informs wwPDB that the paper describing the structure is about to be published. PDB data are released in a two-stage process. Every Saturday at 03:00 UTC the polymer sequences, ligand SMILES strings, and crystallization pH for new structures designated for release are made public (http://wwpdb.org/download/downloads) as a courtesy to the protein structure modeling and computational chemistry communities to enable weekly blinded prediction challenge efforts (e.g., CAMEO [19] and D3R CELPP [28]). Every Wednesday at 00:00 UTC, all new structures designated for release are made publicly available through the wwPDB FTP sites. On average about 200 structures are released every week. As evidence
for the importance of this archive, in 2015, more than 500 million sets of atomic coordinates were downloaded from the wwPDB FTP sites.

**Value-added resources**

The wwPDB FTP sites provide the core data for many databases, services, and websites, including those run by the individual wwPDB partners. In the original wwPDB MOU, it was agreed that to best serve science, wwPDB partner websites would compete with one another and would offer many different kinds of services and features. The RCSB PDB has extensive search and reporting capabilities as well as an education portal called PDB-101 [26,29]. PDBj has multiple search and browse facilities as well as analysis and bioinformatics tools [30,31]. PDBj provides a variety of services and viewers and supports browsing in multiple Asian languages [23,32]. BMRB has many capabilities designed to serve the NMR community [33].

CATH [34] and SCOP [35,36] use the data in the PDB to classify the structural domains of proteins with an attempt to relate them to function. More recently, these two databases have agreed to work together and with other resources in the UK to provide predicted structural features under a unified system called Genome3D [37].

Additional specialty databases provide information on particular classes of macromolecules such as nucleic acids [38].

The Protein Structure Initiative (PSI) Structural Biology Knowledgebase (SBKB) [39] was an ambitious effort to unify information about protein sequence, structure and function. Unfortunately, the decision to discontinue funding the PSI means that this resource will cease to exist.

**Challenges going forward**

A review of the holdings of the PDB shows a steady growth (~10,000 new structures annually). More significantly, the complexity of the structural models continues to increase with more and more large heterogeneous assemblies entering the archive. Fortunately, there are no longer technical restrictions to receiving, annotating, validating, and disseminating these very large structures.

Historically, most structures were determined exclusively with the aid of a single experimental method: X-ray crystallography, NMR or 3DEM. In recent years, these traditional techniques are being combined with other methods to yield improved models. For example, it is now common practice to add data from small-angle scattering measurements to NMR-derived restraints to determine solution structures [40,41]. Similarly, NMR or X-ray data can be combined with cryoEM data in integrative modeling approaches [42]. Such integrative methods make it possible to combine data from different biophysical techniques with computational methods to create models of very large macromolecular machines [43]. However, hybrid approaches also present a variety of challenges including how to validate these structures and then how to archive them. As in the past, with the help and advice of an expert Task Force [44], this integrative challenge will be met by the wwPDB partners.
Acknowledgements

RCsb PDB is supported by NSF [DBI-1338415], NIH, DOE; PDBc by EMBL-EBI, Wellcome Trust [109489], BBSRC [BB/J007471/1, BB/K016970/1, BB/K020428/1], NIGMS [1R01 GM07929-01A1], EU [284209,675858] and MRC [MR/L007835/1]; PDBj by JST-NIBDC and BMRB by NIGMS [1R01 GM09046].

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest


This seminar meeting highlighted the key structures that had been determined and brought together the leading figures in structural biology. To quote David Phillips it was a ‘coming of age.’


This paper contained a thorough analysis of methods that could be used for validation of structures determined by X-ray crystallography. The recommendations were used to create the validation tools used by the wwPDB.


This paper was the first one to analyze what is needed to validate 3DEM maps and models. It is the basis for current research in this field.


This paper made recommendations for how to validate structures determined by NMR, and is the basis for current ongoing research.


The criteria for judging the quality of ligands in protein complexes are laid out and will form the basis for improved validation of these molecules.


This is the formal announcement for how the Protein Data Bank will be managed by an international consortium.


This paper describes the procedures for streamlining the deposition and distribution of 3DEM maps and models.


The recommendation to remove purely in silico models from the PDB is contained in this paper.


A complete description of the mmCIF data dictionary is contained here.


Descriptions of PDB services are given here as well as the RDF format.


This is the first complete description of the services provided by the RCSB PDB.


A summary of the services provided by BMRB is given here.


The services offered by PDBe are described.


This paper summarized the steps necessary to establish an archive of structure data from hybrid/integrative methods.