WORLDWIDE PDB PROTEIN DATA BANK

www.wwpdb.org • info@wwpdb.org

Annotation and Curation of the Protein Data Bank

Partner Directors: Kim Henrick*, John Markley***, Haruki Nakamura** and Helen M. Berman
wwPDB Team: Jasmine Young, Shuchismita Dutta, Zukang Feng, John Westbrook, Martha Quesada, Dimitris Dimitropoulos*, Miri Hirshberg*, Tom Oldfield*, Jawahar Swaminathan*, Sameer Velankar*, Takanori Matsuura**, Steve Mading***, Eldon Ulrich***

RCSB PDB, Rutgers University, New Jersey, USA.
*PDBe, European Bioinformatics Institute, Hinxton, UK.
**PDBj, Institute for Protein Research, Osaka University, Osaka, Japan.
***BMRB, BioMagResBank, University of Wisconsin, Wisconsin, USA.

Abstract

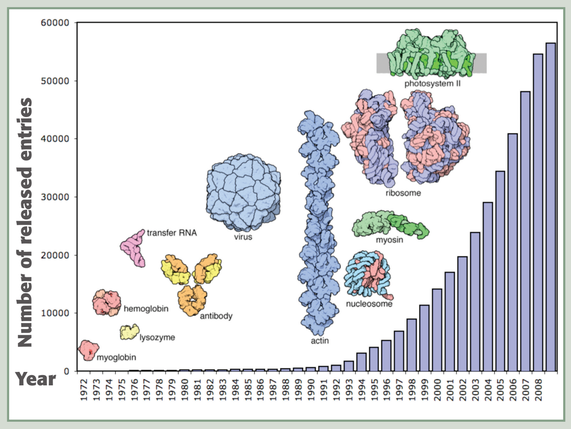
The Protein Data Bank (PDB) is the worldwide repository for experimentally determined 3D structures of biological macromolecules. Established in 1971 with just seven structures, it presently includes more than 56,000 entries. To maintain the highest standards in curation and processing, the members of the worldwide Protein Data Bank (wwPDB) collaborate in data annotation and the development of procedures, tools, and resources. Annotation-related issues, particularly those impacted by new developments in structural biology, are critically reviewed at in-person and virtual meetings regularly and frequently. Comprehensive documentation of the procedures, formats, and related data dictionaries used in data annotation are available at the wwPDB website (www.wwpdb.org).

Mindful of the impact that changes in annotation procedures or data format may have on users, changes are carefully managed and communicated in a timely fashion. In cases involving complex scientific or policy issues, input is sought from advisory committees, standing task forces, experimental method developers, and community experts. This is exemplified by creation of the recently-released version of the PDB archive which updates and further standardizes database references, small molecule chemistry, biological assemblies, and active sites.

PDB Archive Contents

- **Public archive**
  - More than 413,000 files (as of April 3, 2009)
  - Requires over 88 GB of storage
  - Data dictionaries
  - Derived data files
- **For each entry**
  - Atomic coordinates
  - Sequence information
  - Description of structure
  - Experimental data
  - Release status information
- **Internal archive**
  - Depositor correspondence
  - Depositor contact information
  - Paper records
  - Documentation
  - Historical records from Day One of deposition

Number of structures available in archive per year. Example structures are shown.



wwPDB

- Formalization of current working practice
- Members
  - RCSB PDB (Research Collaboratory for Structural Bioinformatics)
    - PDBj (Osaka University)
    - PDBe (EMBL-EBI)
    - BioMagResBank
- Memorandum Of Understanding signed July 1, 2003
- Announced in Nature Structural Biology, November 21, 2003

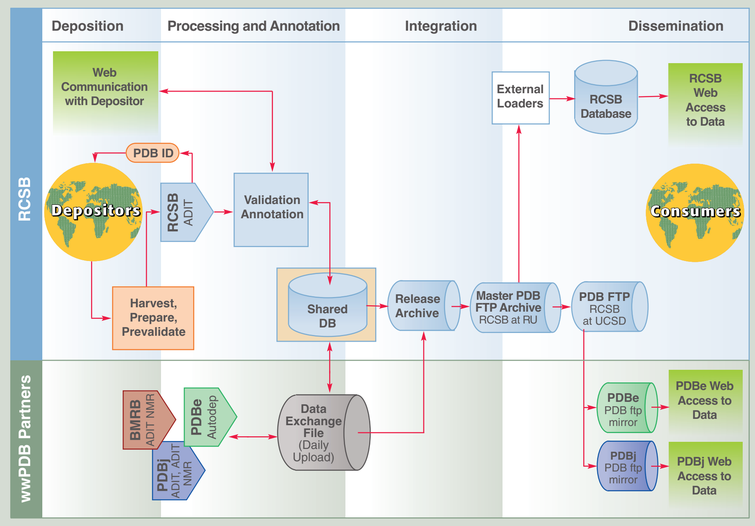
Guidelines and Responsibilities

- All members issue PDB IDs and serve as distribution sites for data
- One member is the archive keeper (RCSB PDB)
- All format documentation publicly available
- Strict rules for redistribution of PDB files
- All sites can create their own websites

www.wwpdb.org

wwPDB Data Sharing Logistics

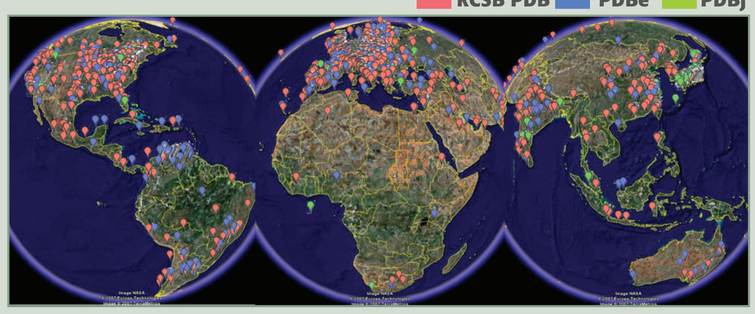
RCSB PDB and wwPDB Full Data Flow



The data sharing logistics among wwPDB partners.

Global Impact

3 months of PDB FTP Traffic



New Complete Documentation

Goal:

Clarify all format descriptions and procedures to ensure the most uniform archive possible.

Process

- Each PDB file format record was reviewed for scientific correctness and clarity by the wwPDB annotators
- Some PDB records and corresponding mmCIF items were added and others expanded
- Advisory Task Force members consulted for input
- Community input requested and taken into account

Deliverables

- A new PDB File Format Contents Guide Version 3.20 was developed and released to the public September 15, 2008
- Data files have been processed according to this specification since Nov. 15, 2008

www.wwpdb.org/docs.html

Improved Data Annotation and Curation

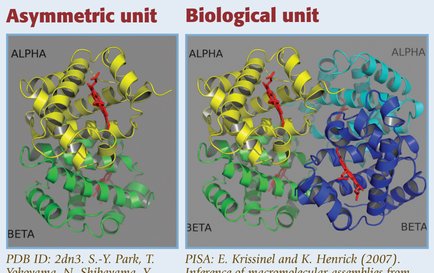
Database references

- **Taxonomy ID** - The source organism as listed in the NCBI Taxonomy database is indicated by the taxonomy id
- **PubMed IDs** - Available for the primary citations of entries in the the PDB, mmCIF, and XML formatted files
- **DOI** - Digital Object Identifiers (DOI) are also included in the PDB, mmCIF, and XML formatted files

Biological assemblies

The quaternary assembly is calculated using PISA/PQS and evaluated by the wwPDB biocurators, while the biological unit is provided by the author. Example here shows that the biological unit can be different from the asymmetric unit.

Asymmetric unit **Biological unit**



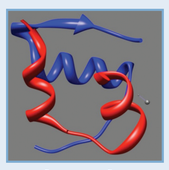
PDB ID: 2dn3. S.-Y. Park, T. Yokoyama, N. Shibayama, Y. Shiro, J.R. Tame (2006) 1.25 Å resolution crystal structures of human haemoglobin in the oxy, deoxy and carbonmonoxy forms. J.Mol.Biol. **360**: *690-701*

PISA: E. Krissinel and K. Henrick (2007). Inference of macromolecular assemblies from crystalline state. J. Mol. Biol. **372**, *774-797*

Example: Biological assemblies

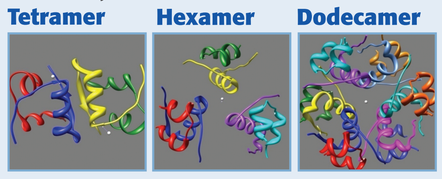
Multiple possible oligomeric states are provided by software, author provided assembly is also indicated in the PDB entry.

Dimer: Biological unit determined by both author and software



Other assemblies in crystal as determined by software

Tetramer **Hexamer** **Dodecamer**



PDB ID: 3e7y. V.I. Timofeev, A.N. Baidus, Y.A. Kislitsyn, I.P. Kuranova. Structure of human insulin. DOI: 10.2210/pdb3e7y/pdb

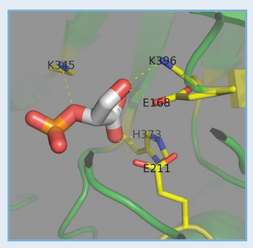
Example: Curation of Virus Biological Assemblies

C.L. Lawson, S. Dutta, J.D. Westbrook, K. Henrick, and H.M. Berman (2008) Representation of viruses in the remediated PDB archive. Acta Cryst. **D64**: *874-882.*

Binding site

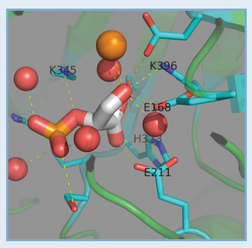
- In version 3.2 PDB entries, SITE records define any interacting residues, based on distance. An evidence code has been added to identify whether the SITE records are software calculated or author provided
- Additional SITE records may also be included upon author request to highlight biologically important residues in the protein (catalytic residues and metal binding site)

Example: Active site



Catalytic site provided by author (highlighted in yellow)

The ligand sits at catalytic site is 2-PHOSPHOGLYCERIC ACID



Binding site predicted by software (highlighted in cyan).

The orange sphere is calcium and red spheres are waters

PDB ID: 5enl. L. Lebioda, B. Stec, J.M. Brewer, E. Tykarska (1991) Inhibition of enolase: the crystal structures of enolase-Ca2(+)- 2-phosphoglycerate and enolase-Zn2(+)-phosphoglycolate complexes at 2.2 A resolution. Biochemistry **30**: *2823-2827*

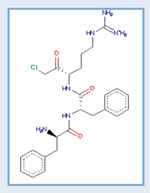
Small molecule chemistry

The Chemical Component Dictionary (www.wwpdb.org/ccd.html) is as an external reference file describing all residue and small molecule components found in PDB entries. This dictionary contains detailed chemical descriptions for standard and modified amino acids/nucleotides, small molecule ligands, and solvent molecules. Each chemical definition includes descriptions of chemical properties such as stereochemical assignments, aromatic bond assignments, idealized coordinates, chemical descriptors (SMILES & InChI), and systematic chemical names.

The Chemical Component Dictionary has been enhanced and unified by:

- Making the chemical name consistent with the systematic name
- Providing various software-generated SMILES strings
- Verifying the correctness of the chirality between coordinates and systematic names
- Capturing the sequence information (subcomponents) for peptide inhibitors
- Capturing author's nomenclature and residue names

Example: PDB entry curation on small molecule chemistry-PPACII inhibitor



Before curation: inconsistent sequence annotation

PDB entry	Sequence
1cvr	DPN F ACL
1dan	DPN F R CH2
1j9c	DPN F ARM
1qfk	DPN F R

After curation: all presented as single molecule 0Z6

Name: D-phenylalanyl-N-[(1S)-4{[amino(iminio)methyl]amino}-1-(chloroacetyl)butyl]-L-phenylalaninamide
Synonyms: FFRCK; PPACK II
Formula: C25 H34 Cl N6 O3
Formal Charge: 1
Subcomponents: DPN PHE ARG 0QE

Sequence information

Future:

Common Deposition and Annotation Processes and Tools for the wwPDB

Goal: To collaboratively develop the new processes and supporting systems that will support the wwPDB over the next 10 years.

The new systems will provide a high quality and dependable resource that will effectively

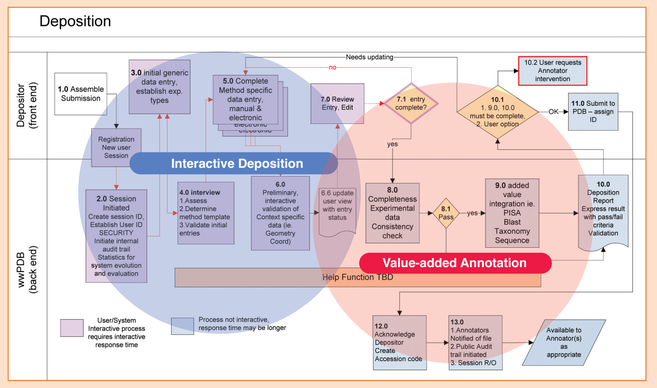
- Support the anticipated increase in deposition throughput
- Address the anticipated increase in complexity and experimental variety of submissions
- Focus on quality enhancement through the use of community-based validation tools

Ensure quality, consistency and efficiency of data deposition and processing

- Enhance the deposition process by providing:
  - System generated annotations, including both PDB internally calculated values and external links.
  - Interactive feedback through the implementation of recommendations from the NMR and X-ray Validation Task Forces

Leverage global resources:

- Load sharing of data processing
- Shared maintenance and tool updates



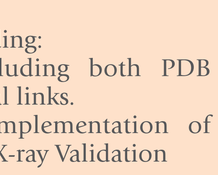
November 2007 Initiation of the wwPDB Common Deposition and Annotation Project

Scope:

X-ray, NMR, EM, and hybrid methods

Assumptions & Constraints:

Steering Committee



- wwPDB partners will adopt common tools and processes
- Must be able to handle all current, agreed upon, data entry formats
- All data elements in the PDB Annotation manual must be included

Project Strategy: Path Forward

- Core team of functional leaders from all sites will manage the project with advise from the Steering Committee
- Project Team made up of experts from all partner sites
  - Quarterly face to face meetings
  - Frequent video conferencing as needed
  - On-going teleconferences and email
- Final design and full requirements will be realized through incremental deliveries, using lessons learned along the way

Funding: RCSB PDB PROTEIN DATA BANK NSF, NIGMS, DOE, NLM, NCI, NINDS, NIDDK; RUTGERS; UCSD; PDBe PROTEIN DATA BANK EUROPE Wellcome Trust, EU, CCP4, BBSRC, MRC, EMBL; PDBj Protein Data Bank Japan BIRD-JST, MEXT; BMRB NLM

Nature Precedings : doi:10.1038/npre.2009.3379.1 : Posted 27 Jun 2009