Research Data Management for the PDB Archive: Unmerged Intensities, Structure Factors, Atomic Coordinates, and Metadata

Stephen K. Burley, M.D., D.Phil.
Director, RCSB Protein Data Bank

2017 ACA Meeting: Workshop
Outline

- History of the PDB Archive
- PDB Facts and Figures
- PDB Data Life Cycle/Unified OneDep System
- PDBx/mmCIF Working Group
- Evolution of the PDBx/mmCIF Data Dictionary
  - Structure Factors
  - Unmerged Intensities
  - Versioning
  - Updating of Atomic Coordinates
- Acknowledgements/wwPDB Foundation
Protein Data Bank Pioneered Open Access

- PDB 1st Open Access digital data resource in all of biology
- Founded 1971 with 7 X-ray structures
- Today, single global archive for experimental 3D macromolecular structure data
- Adhere to FAIR Principles

Some of the very first structures in the PDB
PDB: Massive Archive with Global Reach

- PDB safeguards >US$13 Billion worth of X-ray, NMR, and 3DEM data
- >130,000 structures freely available to Researchers, Educators/Students, and the Curious Public
- Archive growing at ~10%/year
- >1 Million Users worldwide served every year
- >1.5 Million structure data files downloaded/day
- Managed by worldwide collaboration (wwPDB)
- US PDB operations based at Rutgers/UCSD
PDB Data Life Cycle

Structure Biologists
Generate atomic model and data files
Assemble mandatory data items for deposition

Pre-deposition Validation
validate.wwpdb.org

Deposition
deposit.wwpdb.org

wwPDB OneDep
Unified global deposition, biocuration, and validation system

Biocuration

Archive Keeper
RCSB PDB packages and re-distributes data to wwPDB partners

Data Production

Public Release
http://rcsb.org
ftp.rcsb.org

Archive Update
ftp://ftp.wwpdb.org

Data Users
Enables research in various fields
Data accessed via ftp download or web service

Young et al. (2017) *Structure* 25, 536-545.
PDBx/mmCIF Dictionary

- wwPDB manages PDB data using the macromolecular Crystallographic Information Framework (mmCIF)
- PDBx/mmCIF Dictionary contains >4400 data items for PX, NMR, and 3DEM (see mmcif.wwpdb.org)
- wwPDB PDBx/mmCIF Working Group coordinates evolution of the standard and implementation within software packages

Working Group Roster

Paul Adams (Chair)
Jeff Bell
Gerard Bricogne
Paul Emsley
Rasmus Fogh
John Ionides
Eugene Krissinel
Nigel Moriarty
Garib Murshudov
Nicholas Sauter
Mike Word
Going Beyond $h, k, l, F_{hkl}, \Sigma(F_{hkl})$

- $h, k, l, F_{hkl}, \Sigma(F_{hkl})$ for refinement
- $h, k, l, F_{hkl}^{\text{map}}, \phi_{hkl}^{\text{map}}$ from Depositor interpretation
- $h, k, l, F_{hkl}^{\text{native}}, \phi_{hkl}^{\text{native}}$ from Depositor
- $h, k, l, F_{hkl}^{+}, \Sigma(F_{hkl})^{+}, F_{hkl}^{-}, \Sigma(F_{hkl})^{-}$ for phasing from Depositor
- $h, k, l, F_{hkl}^{\text{Der}#+}, \Sigma(F_{hkl}^{\text{Der}#+}), F_{hkl}^{\text{Der}#-}, \Sigma(F_{hkl}^{\text{Der}#-})$ for phasing from Depositor
Unmerged Intensities

- Crystal ID
- Sweep ID
- Frame ID
- $h, k, l, I_{hkl}, \sigma(I_{hkl}),$ partiality
- Identifier for raw diffraction data image with location of peak centroid
- Misset Angles for crystal/sweep/frame
- Rotation Range
Versioning of the PDB Archive

- A new versioned data file naming convention:
  - Semantic version numbers (e.g., major-minor)
  - Uniform identification of data content types
  - Extended accession codes grandfathering of current 4-character codes
    
```
    pdb_00001abc_xyz_v2-0.cif.gz
```

- Versioned files contain additional audit records describing revision details at the granularity of PDBx/mmCIF data categories.

- A new FTP server will be deployed to host the versioned data files in parallel to the current FTP service.
Updating of Atomic Coordinates

- Historical policy of obsoleting PDB entries upon coordinate data replacement has been revised.
- Future coordinate replacements by the Author of Record will retain the original PDB entry code, incrementing the entry major version number.
- Major versions of each entry (i.e., the latest minor revision) will be accessible from the primary wwPDB FTP archive.
Acknowledgements

John Westbrook, Paul Adams, wwPDB PDBx/mmCIF Working Group

The RCSB PDB is funded by a grant from the National Science Foundation, the National Institutes of Health, and the US Department of Energy.
Support PDB’s Spirit of Openness, Cooperation and Education

The wwPDB Foundation funds outreach activities of the wwPDB that are crucial to the future of the PDB archive, including workshops, symposia, and advisory meetings.

Visit foundation.wwpdb.org to make a donation.

2017 Industrial Sponsors

Platinum

FEI  JEOL  OpenEye  SCHRÖDINGER

Gold

DECTRIS

Silver

Anton Paar

Individual Sponsorships Available

Visit Booth 210 to support the wwPDB Foundation and receive this gift!