How can PDB submission be made more automatic

Sameer Velankar
Protein Data Bank in Europe

wwpdb.org

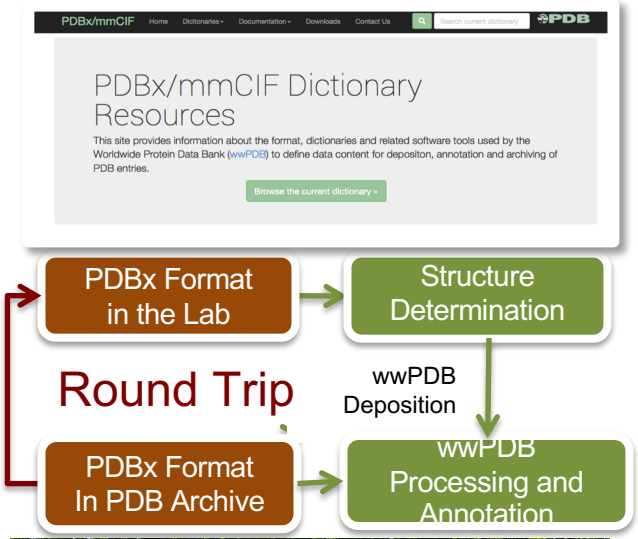
Worldwide Protein Data Bank (wwPDB)

Sustain freely accessible, interoperating Core Archives of structure data and metadata for biological macromolecules as an enduring public good to promote basic and applied research and education across the sciences.

- Manage the wwPDB Core Archives as a public good according to the FAIR Principles.
- Provide expert deposition, validation, biocuration, and remediation services at no charge to Data Depositors worldwide.
- Ensure universal open access to public domain structural biology data with no limitations on usage.
- Develop and promote community-endorsed data standards for archiving and exchange of global structural biology data.

PDBx/mmCIF Working Group

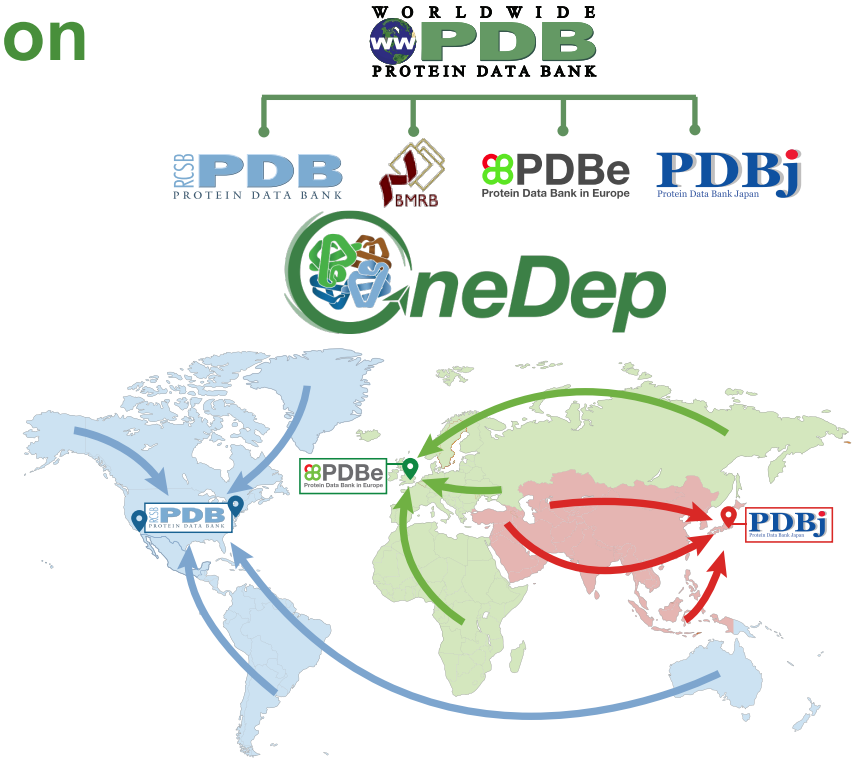
- PDBx/mmCIF is the archival data standard for the repository
- wwPDB together with the PDBx/mmCIF Working Group of community experts and methods developers oversee the evolution of the standard
- Working Group ensures that the standard is well supported by key community software tools.
- PDB hosts community workshops and maintains mmcif.wwpdb.org serving PDBx/mmCIF data dictionaries, schema and software tools



PDBx/mmCIF Workshop Participants, July 2017

Protein Data Bank Deposition

- Open Access to single global archive
  - Weekly release
- Regional sharing of data deposition efforts
  - RCSB PDB (US)
  - PDB Japan (PDBj)
  - PDB in Europe (PDBe)
- Partner-hosted websites offer complementary services and views of data

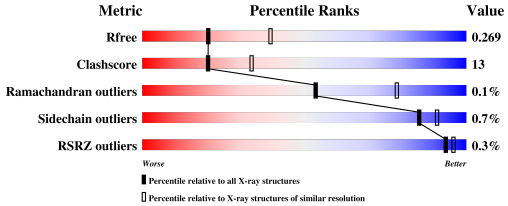


Protein Data Bank: the single global archive for 3D macromolecular structure data
wwPDB Consortium (2019). *Nucleic Acids Res.* 47, D520–D528.

Protein Data Bank Deposition

- Deposition
  - deposit.wwpdb.org

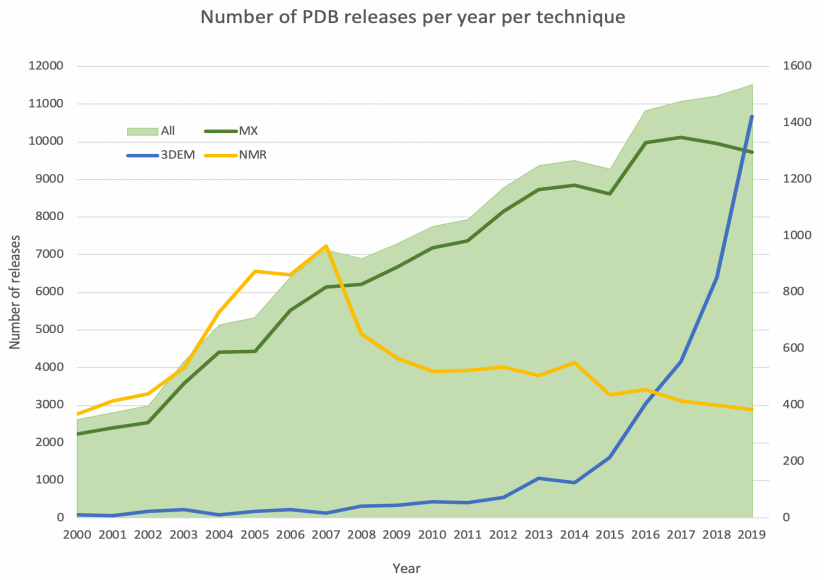
- Validation – validate your structures before deposition
  - validate.wwpdb.org



- Validation API
  - https://www.wwpdb.org/validation/onedep-validation-web-service-interface

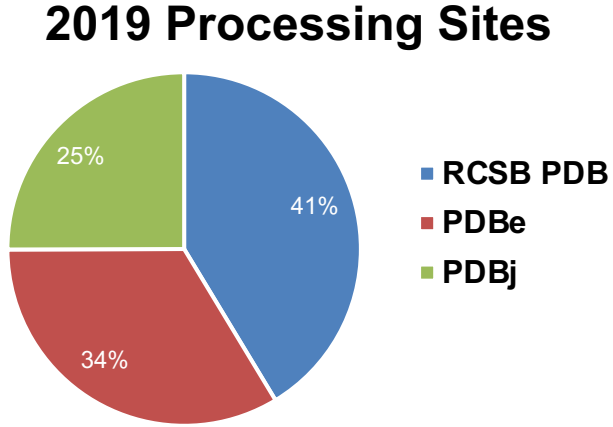
Growth/Complexity

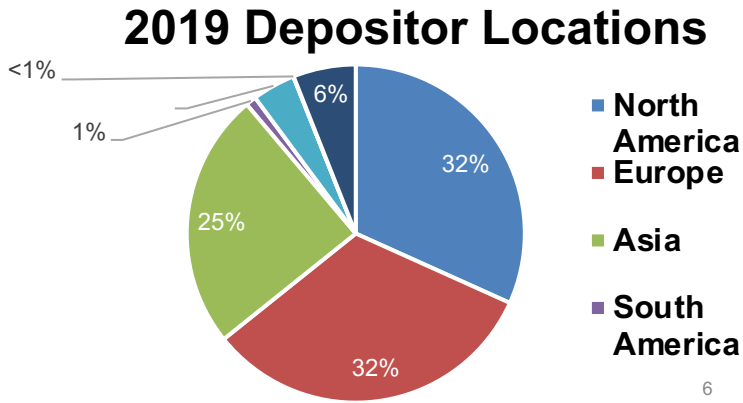
- 13,377 depositions in 2019
- Rapid growth in 3DEM
- More structures per Depositor
- Increasing data complexity
  - Increase in large complexes resolved by EM



All, MX

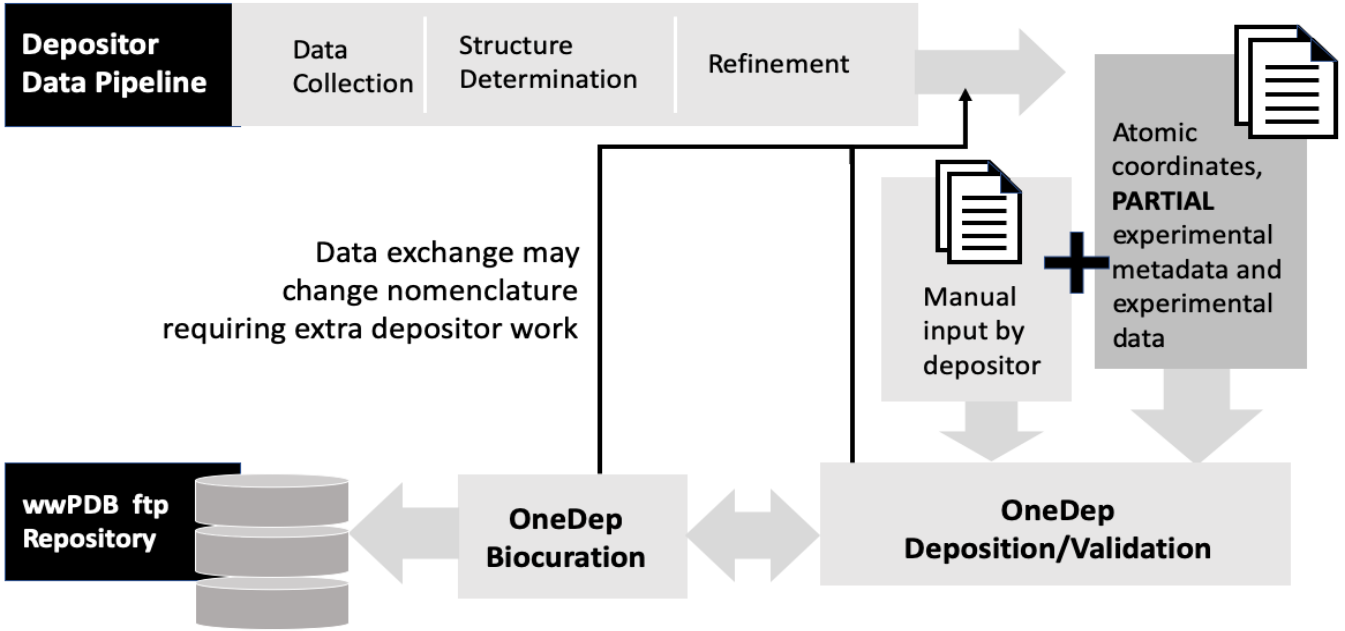
3DEM, NMR





Protein Data Bank Deposition

- How does deposition work now?



- Manual curation review within OneDep

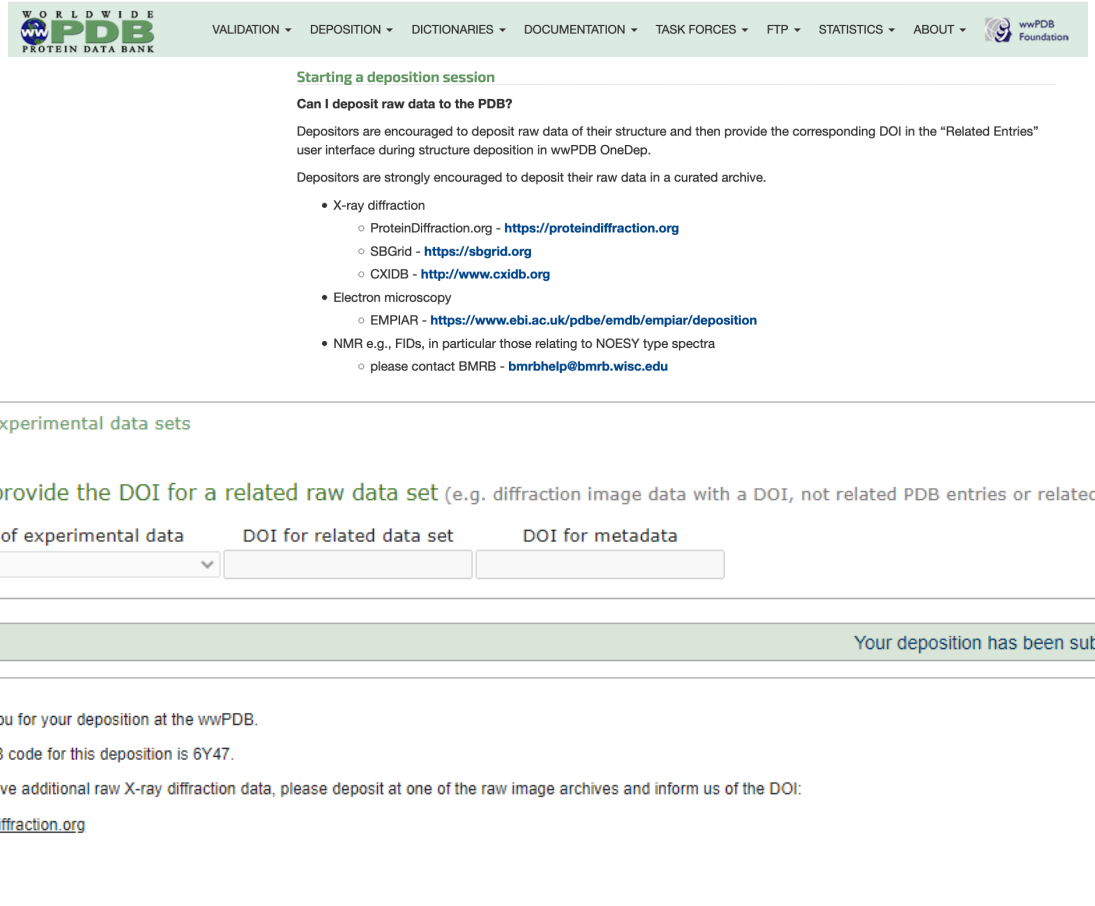
What data is captured?

Deposition type	Database accession codes issued	Mandatory file uploads	Optional file uploads
X-ray and neutron crystallography	PDB	Atomic coordinates Structure factor data Unmerged intensity data	Ligand definition file or image Auxiliary files
Solution and solid-state NMR	PDB BMRB	Atomic coordinates Assigned chemical shifts Restraints used in refinement Auxiliary sequence file from AMBER	Spectral peak lists Ligand definition file or image Auxiliary files
Electron crystallography	PDB EMDB	Atomic coordinates Structure factor data or Mass density map volume	Ligand definition file or image Auxiliary files
3DEM (map and model)	PDB EMDB	Atomic coordinates Mass density map volume Entry image for public display (EMDB)	Any number of additional maps Any number of masks Two half maps Fourier shell correlation (FSC) curve Ligand definition file or image
3DEM (map only)	EMDB	Mass density map volume Entry image for public display	As above

What data is captured?

- Administrative information (e.g., author release instructions).
- Description of each distinct macromolecule present in the sample.
- Description of the experimental setup (e.g., sample preparation and data collection).
- Description of experimental data, refinement (e.g., crystallographic refinement statistics), and software used.
- Description and matching of ligands and modified polymer residues to the PDB Chemical Component Dictionary (CCD).
- Information on the quaternary structure and, whenever possible, experimental support for the biologically relevant assembly.

Deposition of raw data

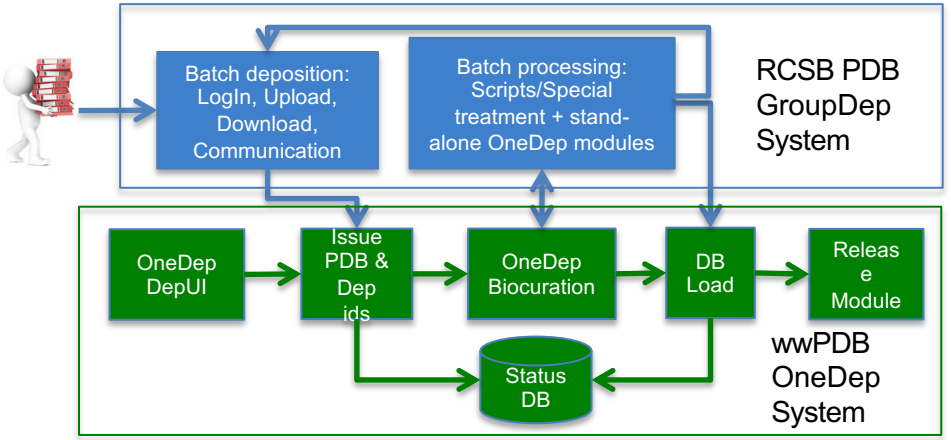


Group deposition

https://deposit-group.rcsb.rutgers.edu/groupdeposit/

- Provided support for D3R Blind Challenges
- Early Adopters: Roche, Merck Serono, U. Marburg, U. Essex

- Requires a complete PDBx/mmCIF file with all mandatory data
- Each group needs to implement the necessary infrastructure to generate complete PDBx/mmCIF file



Mandatory submission of PDBx/mmCIF – June ’19

- Increase in the amount of meta data collected in the entries during deposition
  - Work with Software community to develop automated process for deposition
- Improved consistency in biocuration between related entries
- Increased biocuration efficiency
  - Curating an increasing number of entries with the same resources

STRUCTURAL BIOLOGY
ISSN 2059-7983

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.

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

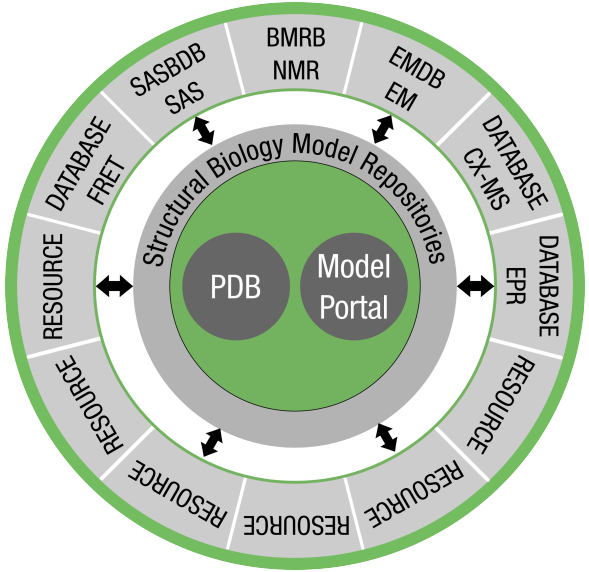
The Protein Data Bank (PDB) (wwPDB consortium, 2019) is the single global archive of experimentally determined three-dimensional (3D) structure data of biological macromolecules. The continuing growth in the numbers, size and complexity of macromolecular structures in the PDB archive, coupled with the 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.

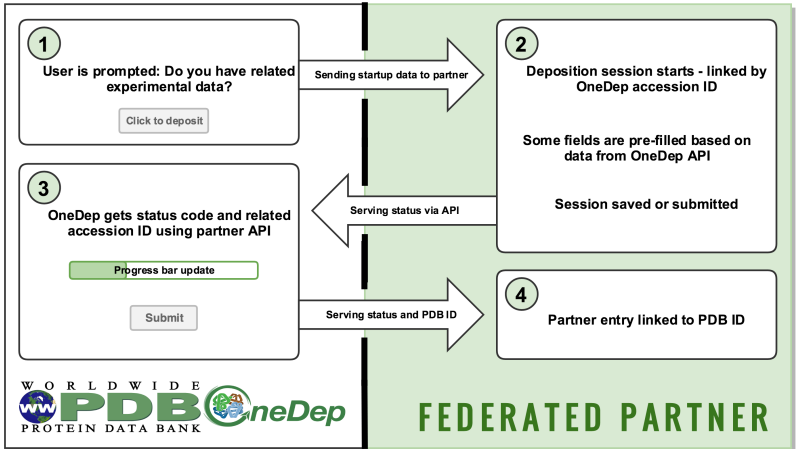
Since 2003, the PDB has been managed by the Worldwide Protein Data Bank (wwPDB; https://www.wwpdb.org/) (Berman *et al.*, 2003), an international partnership that collaboratively oversees deposition, validation, biocuration and open-access dissemination of 3D macromolecular structure data, adhering to 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

Deposition of associated experimental data

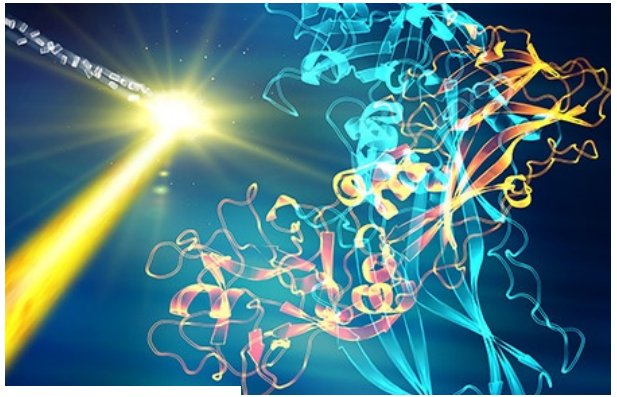
- Experience of API's within OneDep
- Interaction with SASBDB during deposition
- API services to interact with federated deposition systems



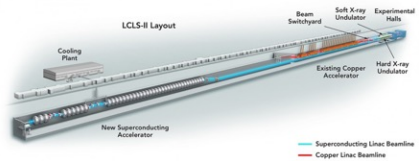


SFX/XFEL are Revolutionizing MX

- SFX: Serial Femtosecond X-ray crystallography
- XFEL: X-ray Free Electron Laser
- Time-resolved studies
  - Conformational changes in enzyme active sites
  - Response to photoexcitation
  - Membrane protein dynamics
- Rapid advances in data collection and sample handling technologies
- Linking of multiple depositions (“investigations”)



BioXFEL



PAUL SCHERRER INSTITUT
PSI

SwissFEL

LCLS

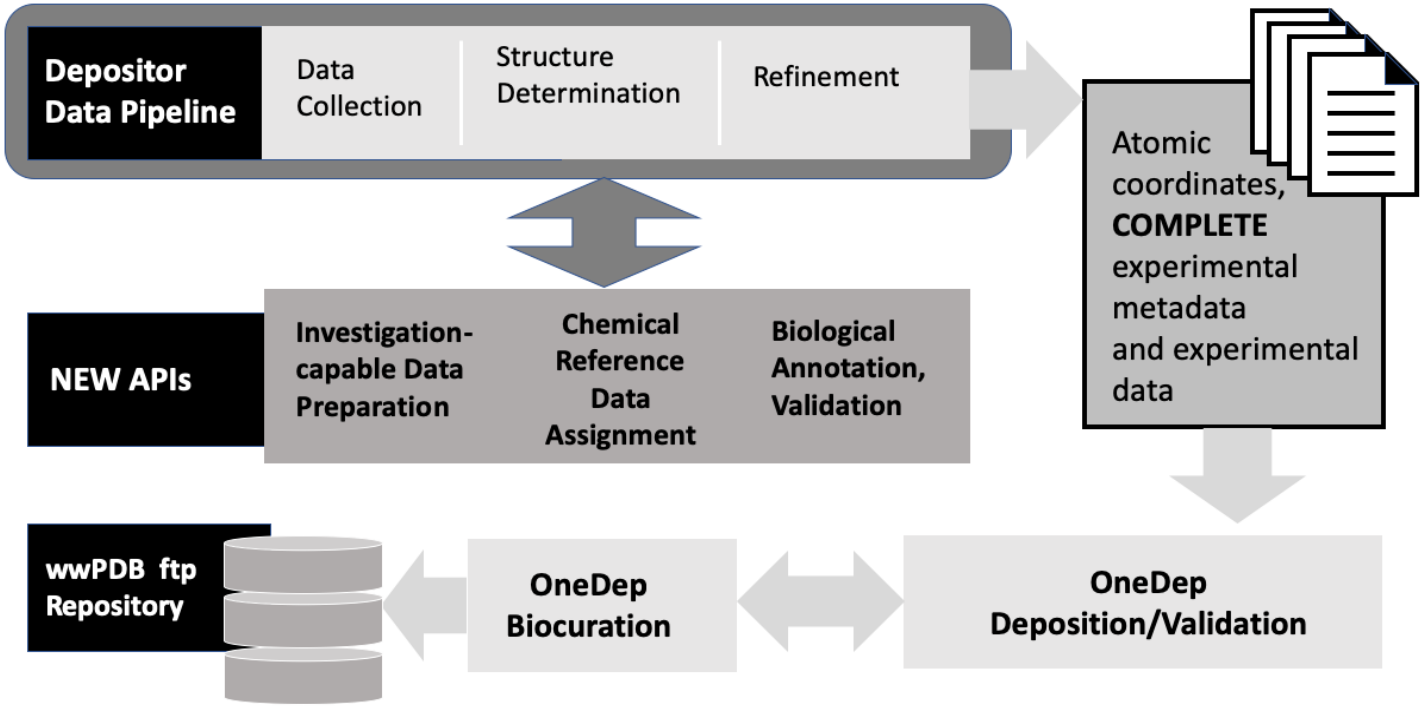
SACLA

European XFEL

Automated deposition

- Improved fidelity and completeness of 3D structure data deposited into the PDB
- Streamlining the wwPDB data deposition, validation, and biocuration system.
- Provide easy mechanism for deposition of multiple structures (SF/XFEL, Fragment screening)
- Improved efficiency for depositors and biocuration
- Consistent biocuration

How can PDB deposition be made more automatic



APIs to expose wwPDB deposition/annotation pipeline

New three-year project (PDBe and RCSB PDB)

- Funded by BBSRC and NSF
- Start date – July 2020 (RCSB PDB); Jan 2021 (PDBe)

- Gather requirements on Deposition API requirements
  - Automated deposition
  - Multi-structure deposition/annotation
    - “investigation” based deposition
    - “investigation” based annotation – Annotators to provide requirements

New three-year project (PDBe and RCSB PDB)

- Data Content Extension:
  - “Investigation” level semantics
  - Integrate validation data in mmCIF

- New Deposition Data Preparation APIs:
  - pre-deposition registration service for chemical reference data
    - Better validation of ligands before and during deposition

Questions and discussion