2021 wwPDB AC Meeting

Tuesday, October 19, 2021

14:00 - 16:30 UK Time

Agenda

14:00 (UK Time)	Welcome and Introductions	Ardan Patwardhan
14:10	Executive Session No. 1	Breakout Room *via* Host
14:30	Discussion/Questions for AC	All
15:30	Executive Session No. 2	Breakout Room *via* Host
16:00	Feedback to wwPDB Leadership	All
16:20	Acknowledgements	Ardan Patwardhan
16:30	Group Photo/Meeting Close	All, Photo *via* Host

Welcome and Introductions

Ardan Patwardhan

Executive Session No. 1

Host will add Advisors into Zoom Breakout Room

Please rejoin us at the end of your discussion by selecting
Leave the breakroom (lower righthand corner)

Discussion/Questions for AC

Questions for Advisors Overview

1. Does the Advisory Committee have any questions regarding our responses to the 2020 AC report? (AP/GK)

2. Does the Advisory Committee have any questions/concerns regarding individual wwPDB member 2021 AC reports? (GK)

3. Does the Advisory Committee concur with our updated plans to engage with the relevant structure prediction and experimental data resources to accelerate development of the wwPDB data federation, employing community data standards and distributed archiving? (SV)

4. Does the Advisory Committee concur with our goal of engaging the community to develop general, statistically rigorous tools for validation of both experimental structures (*e.g.*, MX, 3DEM, NMR, I/HM) and computed structure models? (JCH)

5. Next Advisory Committee Meeting Hosted by RCSB PDB: format/scheduling? (SKB)

Advisor Question #1

Does the Advisory Committee have any questions or concerns regarding our responses to the 2020 AC report?

Advisor Question #2

Does the Advisory Committee have any questions/concerns about individual wwPDB member 2021 AC reports?

Advisor Question #3

Does the Advisory Committee concur with our updated plans to engage with the relevant structure prediction and experimental data resources to accelerate development of the wwPDB data federation, employing community data standards and distributed archiving?

Advisor Question #4

Does the Advisory Committee concur with our goal of engaging the community to develop general, statistically rigorous tools for validation of both experimental structures (e.g., MX, 3DEM, NMR, I/HM) and computed structure models?

Advisor Question #5

2022 wwPDB AC will be hosted by RCSB PDB

Would the Advisory Committee prefer a face-to-face meeting in Piscataway, NJ
(pandemic permitting)?

Which date would the Advisory Committee prefer?

   Friday October 14th 2022

   Friday October 21st 2022

Executive Session No. 2

Host will add Advisors into Zoom Breakout Room

Please rejoin us at the end of your discussion by selecting
Leave the breakroom (lower righthand corner)

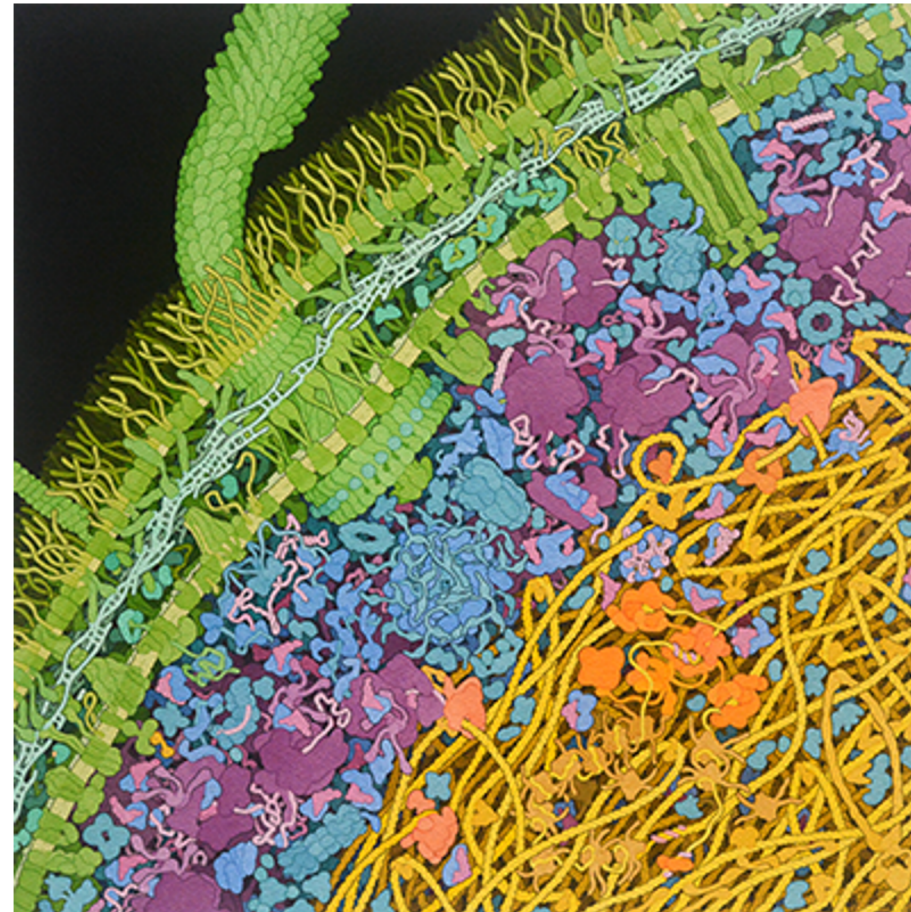# Feedback to wwPDB Leadership

Acknowledgements

Ardan Patwardhan

Group Photo

Meeting Close

Thank you

Molecular Landscapes: *E. coli*
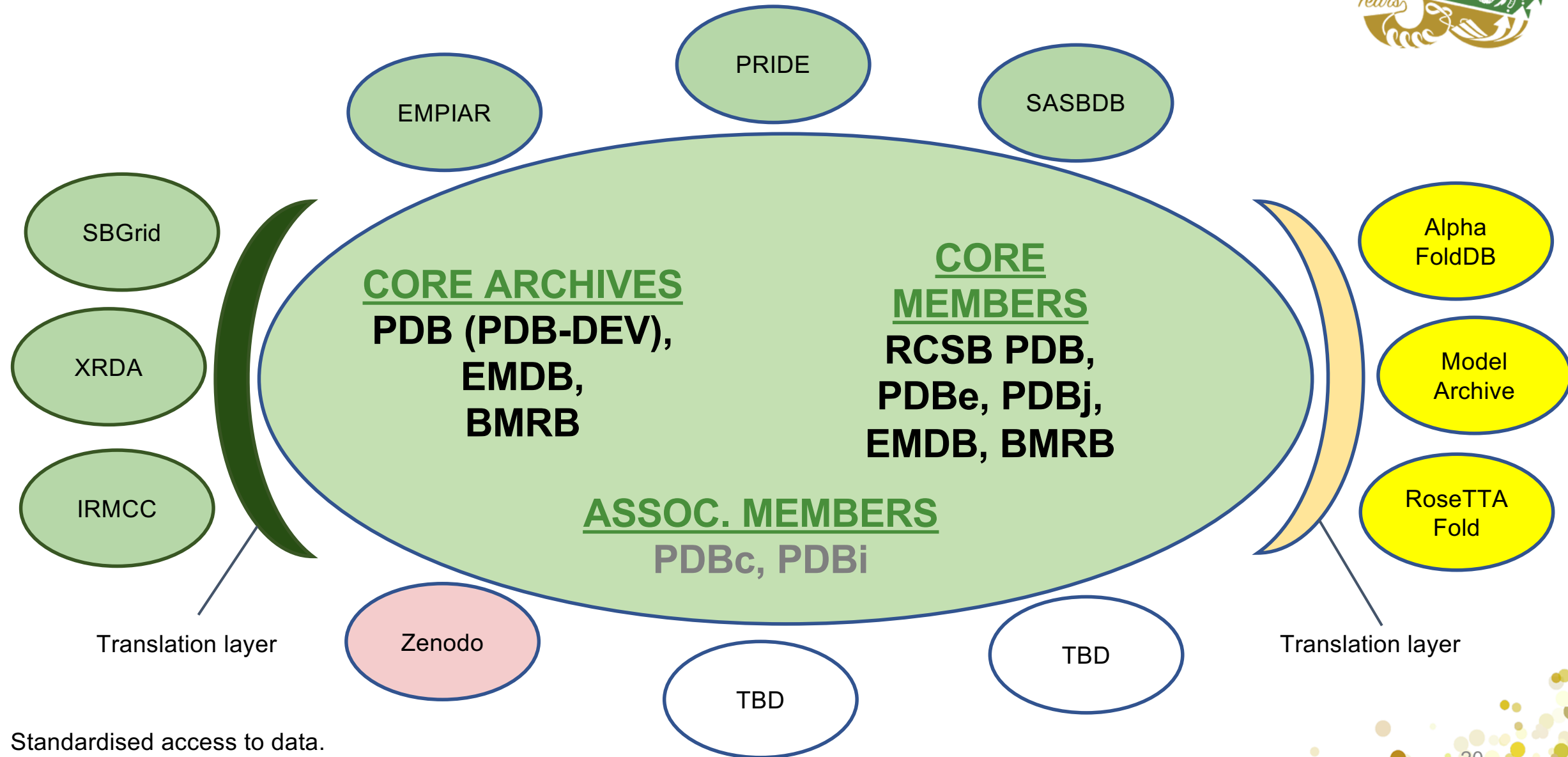
Question Briefing Slides

Advisor Question #3

Does the Advisory Committee concur with our updated plans to engage with the relevant structure prediction and experimental data resources to accelerate development of the wwPDB data federation, employing community data standards and distributed archiving?

wwPDB is Responding to a Changing BioData Ecosystem:

- Multiple data centers have emerged for various types of data repositories (*e.g.*, MX diffraction images, computed structure models).
  wwPDB Plan: Promote development of common standards and APIs for data exchange that support delivery of common data items.

- Stand alone servers are now supporting *de novo* protein structure prediction without archiving (*e.g.*, GoogleCollab, RoseTTAFold, …).
  wwPDB Plan: Promote and further develop existing ModelCIF standard for delivery of common data items, and encourage preservation of "published" computed structure models for reference and reuse, possibly using ModelArchive.

Structure Data EcoSystem

Structure Data EcoSystem

- Data resources related to experimental data archiving.
- Represent candidates for becoming federated resources

- Standardized access to data.
- Work with the individual resources to standardize data model

- A generic data resource used for archiving raw experimental data

PRIDE

EMPIAR

SASBDB

Alpha FoldDB

SBGrid

CORE ARCHIVES
PDB (PDB-DEV), EMDB, BMRB

CORE MEMBERS
RCSB PDB, PDBe, PDBj, EMDB, BMRB

Model Archive

XRDA

RoseTTA Fold

IRMCC

ASSOC. MEMBERS
PDBc, PDBi

SWISS MODEL

Translation layer

Zenodo

TBD

TBD

MODBase

Translation layer
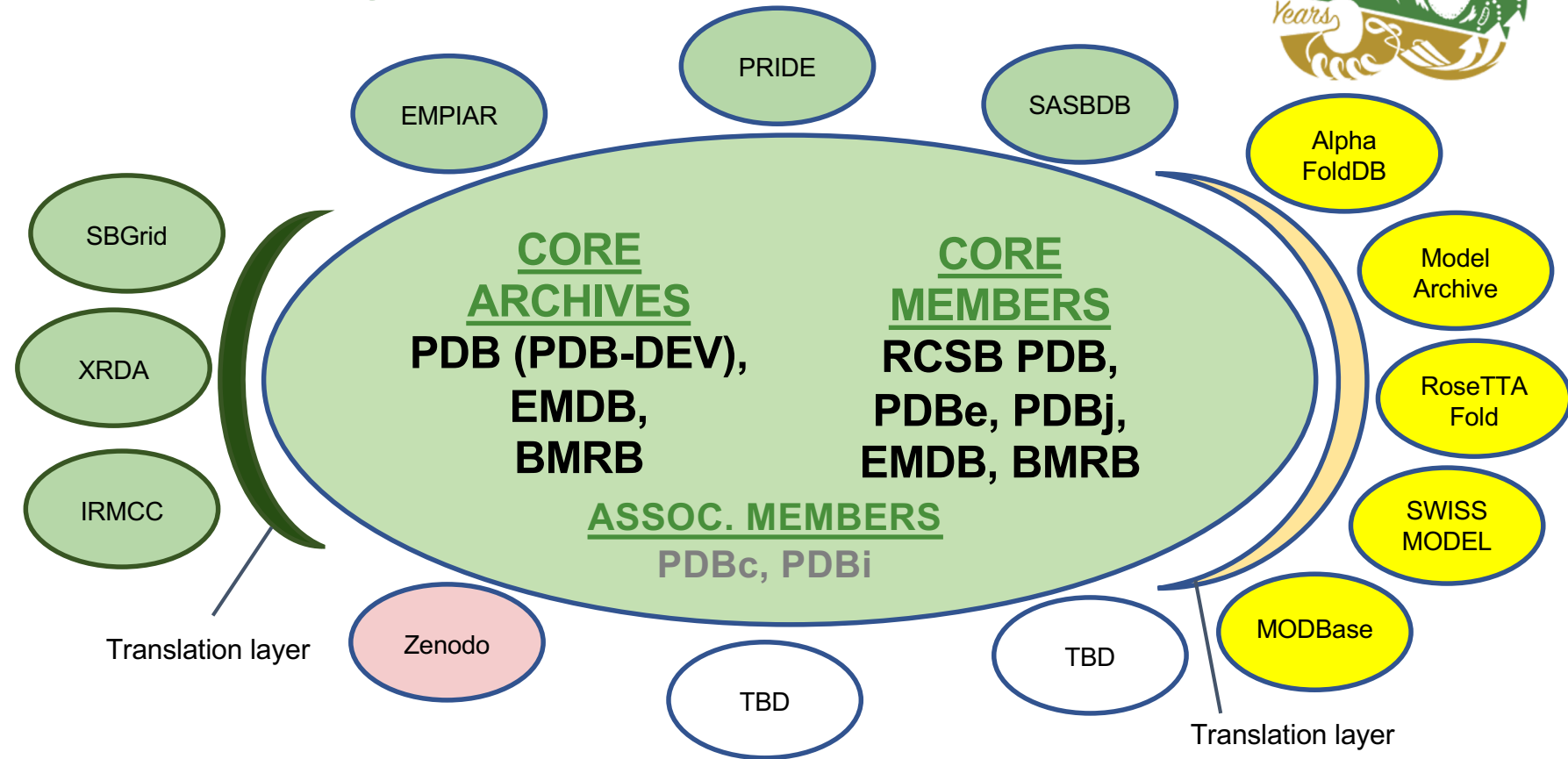
- SASBDB will be the first federated archive as wwPDB has already establish data exchange standards and implemented a transparent mechanism for data deposition *via* OneDep.
- PDBe a grant to establish a similar mechanism with PRIDE. wwPDB plans to work with EMPIAR to do the same.
- Where there are multiple resources for archiving same type of data (*e.g.*, SBGRID, IRMCC and XRDA archiving raw diffraction images,) we will work with these data resources to establish common data standards and data access mechanism.
- In the case of computed structure models, there are already efforts in EU to standardise data model and data access via 3d-beacons collaboration (3d-beacons.org), which may be considered in this context.

21

Advisor Question #4

Does the Advisory Committee concur with our goal of engaging the community to develop general, statistically rigorous tools for validation of both experimental structures (e.g., MX, 3DEM, NMR, I/HM) and computed structure models?

wwPDB seeks to advance the rigor and utility of 3D structure information:

- Use of Bayesian methods could put 3D structure validation on a statistically more rigorous footing (*i.e.*, current wwPDB validation system has various limitations due to nature of data distributions, *etc.*).
  wwPDB Plan: Promote the development of community standards for the application of Bayesian principles to validation of experimental structures/computed structure models.

- Emerging structure prediction resources present new challenges to interpretation and utilization similar to the challenges posed by I/HM (*e.g.*, low resolution 3DEM structure *versus* computed structure model).
  wwPDB Plan: Promote the development of common approaches for validating experimentally determined structures and computed structure models with community stakeholders.

Pre-Meeting Review Slides

Funding Update

Ardan Patwardhan

wwPDB Member Core Funding

- RCSB PDB: Joint NSF/NIH/DOE funding renewed: 2019-2023
- BMRB: NIH NIGMS funding: 2019-2023
  - Inadequate budget: still need to find additional support
  - UConn funding 25% administrative assistant and 25% project manager positions
  - NIH U24 submitted but not funded. Re-submitted 9/21
- PDBe: EMBL-EBI, Wellcome Trust: 2021-2025
- PDBj: NBDC-JST and AMED funding: 2019-2022
  - Additional budget from S. Korea (decision pending due to COVID-19)
- EMDB: EMBL-EBI, Wellcome Trust: 2019-2024
- RCSB PDB/PDBe: Joint NSF/BBCRC NextGen Archive funding: 2020-2022
- RCSB PDB/PDBe: Joint NSF/BBCRC Mol* Visualization funding: 2021-2024

Governance Update

Stephen K. Burley

New wwPDB Charter Signed

- Effective Date January 1 2021
- Admitted EMDB
- Established the framework for admission of PDB China and PDB India as Associate wwPDB Members (with roadmap for transition to Full Membership)
- Adopted Creative Commons CC0 1.0 Universal License for all wwPDB Core Archives

CHARTER OF THE WORLDWIDE PROTEIN DATA BANK

Agreement between

RESEARCH COLLABORATORY FOR STRUCTURAL BIOINFORMATICS PROTEIN DATA BANK (RCSB PDB Organization), Rutgers, The State University of New Jersey, Piscataway, New Jersey, United States

and

The EUROPEAN MOLECULAR BIOLOGY LABORATORY ("EMBL"), an intergovernmental institution established by treaty, headquartered at Meyerhofstrasse 1, 69117 Heidelberg, Germany acting through its UK Outstation the European Bioinformatics Institute ("EMBL-EBI"), located on the Wellcome Genome Campus in Hinxton, Cambridgeshire, UK, operating both the PROTEIN DATA BANK IN EUROPE (PDBe Organization) and the ELECTRON MICROSCOPY DATA BANK (EMDB Organization)

and

PROTEIN DATA BANK JAPAN (PDBj Organization), Osaka University, Osaka, Japan

and

BIOLOGICAL MAGNETIC RESONANCE DATA BANK (BMRB Organization), University of Connecticut, Farmington, Connecticut, United States

Effective from January 1, 2021

CHARTER OF THE WORLDWIDE PROTEIN DATA BANK (wwPDB)

1. Rationale

The Worldwide Protein Data Bank (**wwPDB**) was created to provide an enduring organizational framework for global management and dissemination of public-domain structural biology data.

It is essential for the progress of international science that structural biology data be maintained within a limited number of unfragmented archives, identified herein as *wwPDB Core Archives*.

At present, the **wwPDB** manages three *wwPDB Core Archives*, including the *Protein Data Bank* (*PDB*), the *Biological Magnetic Resonance Data Bank* (*BMRB*), and the *Electron Microscopy Data Bank* (*EMDB*).

It is paramount that facilities for deposition, validation, biocuration, remediation, and storage of public-domain structural biology data in the *wwPDB Core Archives* be managed jointly by the **wwPDB Core Members** with all services provided at no charge to **wwPDB Data Depositors**.

It is equally important that structural biology data stored in the *wwPDB Core Archives* be freely and publicly disseminated by the **wwPDB Core Members** and the **wwPDB Associate Members** without charge or limitations on usage.

wwPDB Core Members and **wwPDB Associate Members** are fully committed to the **FAIR** Principles of **F**indability-**A**ccessibility-**I**nteroperability-**R**eusability, emblematic of responsible stewardship of public domain information.

PDB China Update (PI/co-PI: Profs. Wenqing Xu, James Z. Liu, Quan Wang)

- The National Facility for Protein Science in Shanghai (NFPS), the primary host of PDB China, has purchased computational hardware dedicated to PDB China and performed related infrastructure set up;

- NFPS has recruited/formed the Data-in team (5 people) and the joint Data-out team (5 people, in collaboration with the ShanghaiTech University);

- PDB Japan obtained funding to purchase the computers for PDB China's first OneDep data-in system (to be sited in Osaka); equipment just arrived; remote data-in training will start soon;

- The first version of the PDB China data-out system, with primary search functions, has been developed, and is under test and further development;

- Dedicated future funding for protein structure databank research from MOST (Ministry of Science and Technology of China) can be expected for 2022-2026. More funding from other funding resources, including those from the Shanghai government, the ShanghaiTech University and NFPS, can be expected. No funding concerns.

PDB India Update (PI/Co-PI: Prof. M. Bansal, Prof. K. Sekar, Dr. D. Mohanty)

- First phase funding (Oct 2019-2022) received from NSM(India) for hiring staff, but planned second phase support for procurement of servers/storage has not materialized yet.

- PDBi PIs have made arrangements to utilize servers/storage of the Indian Biological Data Center (IBDC) which will be operational by December 2021. IBDC is a facility funded by DBT, India and efforts are underway to secure PDBi funding from DBT for 2022-2027.

- Six PDBi staff are currently working on development of structural bioinformatics software for comparative analysis of Protein-DNA complexes in PDB, structure-based analysis of PPI network of M.Tb./Plasmodium and integrating other structure analysis software with a local FTP mirror of PDB coordinate files.

- Discussions have been held with PDBe (Dr. Velankar) for online remote training of two PDBi staff on annotation/curation of structural data after securing required permissions from EBI/wwPDB PIs.

Outreach Update

Ardan Patwardhan

PDB50 Celebration Calendar*

- wwPDB/ASBMB: May 4-5
- RCSB PDB/ACA: July 30-31
- RCSB PDB/ACS: August 25
- RCSB PDB/BPS: October 6
- PDBe/EMBL: October 20-22
- RCSB PDB/Rutgers IQB: November 2
- PDBe/RCSB PDB/RSC: November 16 and 18
- PDBj/BSJ : November 24

* All events virtual

OneDep Update

Genji Kurisu, PDBj

wwPDB Collaboration Resource November 2020-October 2021

wwPDB Partner	Software Development	Production Maintenance / Management	Requirements Setting/ Testing	Core Archive Keeping*	Outreach	Biocuration/ Remediation	Total FTE Commitments
RCSB PDB	2.0**	1.3	0.35/0.35	2.0	0.3	6.3	12.6
PDBe	1.4**	1.0	0.35/0.35	-	0.2	5.0	8.3
PDBj	0.4	0.4	0.2/0.2	-	0.1	4.5	5.8
BMRB	0.95	-	0.1/0/1	0.5	-	0.2	1.85
EMDB	0.9	0.35	0.1/0.2	0.3	-	0.5	2.35
Total wwPDB	5.65	3.05	2.3	2.8	0.6	16.5	30.9

*RCSB PDB; EMDB; BMRB

**Excluding additional resource from BBSRC/NSF joint grant, 1.0 FTE at PDBe and 1.3 FTE at RCSB PDB

OneDep 2020/2021 Progress vs. Goals I

- Our familiar Table is here, and details are in Reference.

Ref. Appendix XXX

Delivered,
To be delivered,
Delayed

	Projects	Timeline			
		2020	2021		
		Q4	Q1	Q2	Q3
1. Validation	1.1 NMR restraint validation	←			
	1.2 mmCIF formatted validation reports				
	1.3 Improve chemical shift validation code				
	1.4 Highlight NMR violations				→
2. Backend Stabilization	2.1 Upgrade OneDep software to Python 3	←			
	2.2 Validation EDS software upgrade				
	2.3 Automated OneDep software packaging, update, and documentation				
	2.4 Remove PDB format dependent software				
	2.5 Streamline weekly update- data exchange via API			→	
3. Public facing (OneDep or wwPDB.ORG)	3.1 Present annotated assembly to depositors				
	3.2 Enable author-annotated assembly				
	3.3 Improve EM deposition				
	3.4 Improve NMR deposition				
	3.5 Present graphical analysis from NMR validation report				
4. Biocuration	4.1 Enable automated biocuration on Entity Transformation module				
	4.2 Establish global biocuration training				
	4.3 Integrate Mol* 3D view in annotation pipeline				
	4.4 Support depositor-annotated assembly				
	4.5 Improve biocuration efficiency via API software				→
5. Archive Improvements	5.1 Regenerate EM validation reports with map analysis	←			
	5.2 Planning for extension of CCD and PDB code				
	5.3 Protein Modification remediation planning	←			→

OneDep 2020/2021 Progress vs. Goals II

Ad hoc projects:
- Support EBI server upgrade (ongoing)
- Entry PI contact information made public
- Better support of extended mmCIF SF data from structure determination software

Bold: re-forecasted to 2021-2022
*PTM remediation re-forecasted to 2023

| Major Projects	Primary resource
Implement NMR restraint validation	BMRB
Improve EM map validation	EMDB
Provide mmCIF formatted validation report	RCSB PDB
Upgrade 3rd party EDS software in validation package	**PDBe**
Refactor NMR chemical shifts validation	**BMRB**
Improve NMR and EM depositions	PDBj/EMDB
Enable author-annotated assembly	**PDBe**
Establish global Biocuration training/quality assessments	RCSB PDB/PDBe/PDBj
Improve assembly review processes provided by biocurators	PDBe
Improve OneDep installation and documentation	PDBe/RCSB PDB
Planning for ligand ID and PDB ID extension	RCSB PDB
PTM remediation*	**PDBe**
3DEM validation reports recalculation	EMDB/PDBe
Present graphical outliers of NMR validation at deposition	**BMRBj/BMRB**

wwPDB Biocurator Productivity

- Record high on 15436 depositions in 2020

- Better automation for biocurating incoming depositions

  - Based on existing structure annotations (where possible)

  - Median processing time/entry reduced for large and/or complex entries (3-4 days to <2 days)

New Structures/wwPDB Biocurator
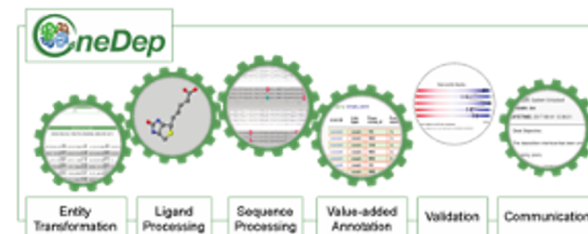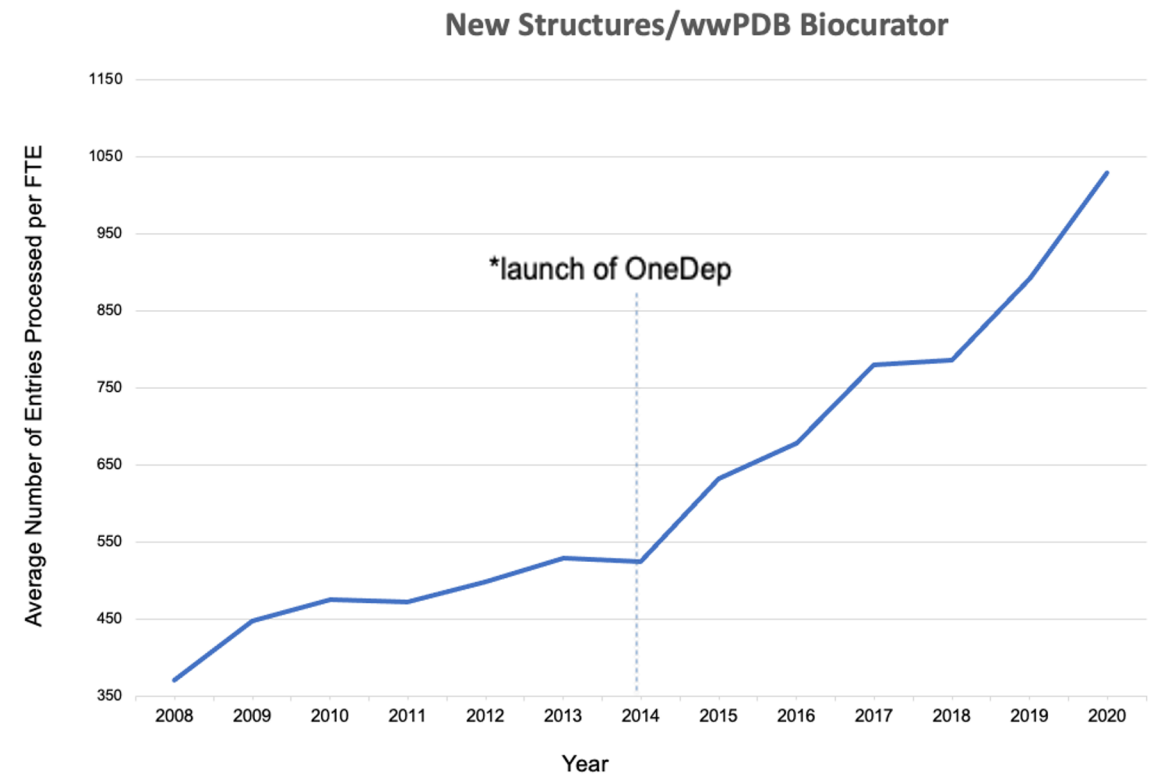
wwPDB Collaboration Resources November 2021-October 2022

wwPDB Partner	Software Development	Production Maintenance/ Project Management	Requirements Setting/ Testing	Core Archive Keeping*	Outreach	Biocuration/ Remediation	Total FTE Commitments
RCSB PDB	2.8**	1.3	0.35/0.35	2.0	0.3	6.3	13.4
PDBe	1.4**+0.8#	0.6	0.2/0.2	-	0.2	4	7.4
PDBj	0.9***	0.4	0.2/0.2	-	0.1	4.5	6.3
BMRB	0.95	-	0.1/0.1	0.5	-	0.2	1.85
EMDB	1.15	0.5	0.1/0.1	0.35	-	1.2	3.4
Total wwPDB	8.0	2.8	1.9	2.85	0.6	16.2	32.35

*RCSB PDB; EMDB; BMRB

**including additional resource from BBSRC/NSF joint grant, 1.0 FTE at PDBe and 1.3 FTE at RCSB PDB

***PDBj has a plan to add 0.5 FTE for OneDep development from April 2022

#New hire at PDBe to work on PTM remediation project; Will start in Jan/Feb 2022
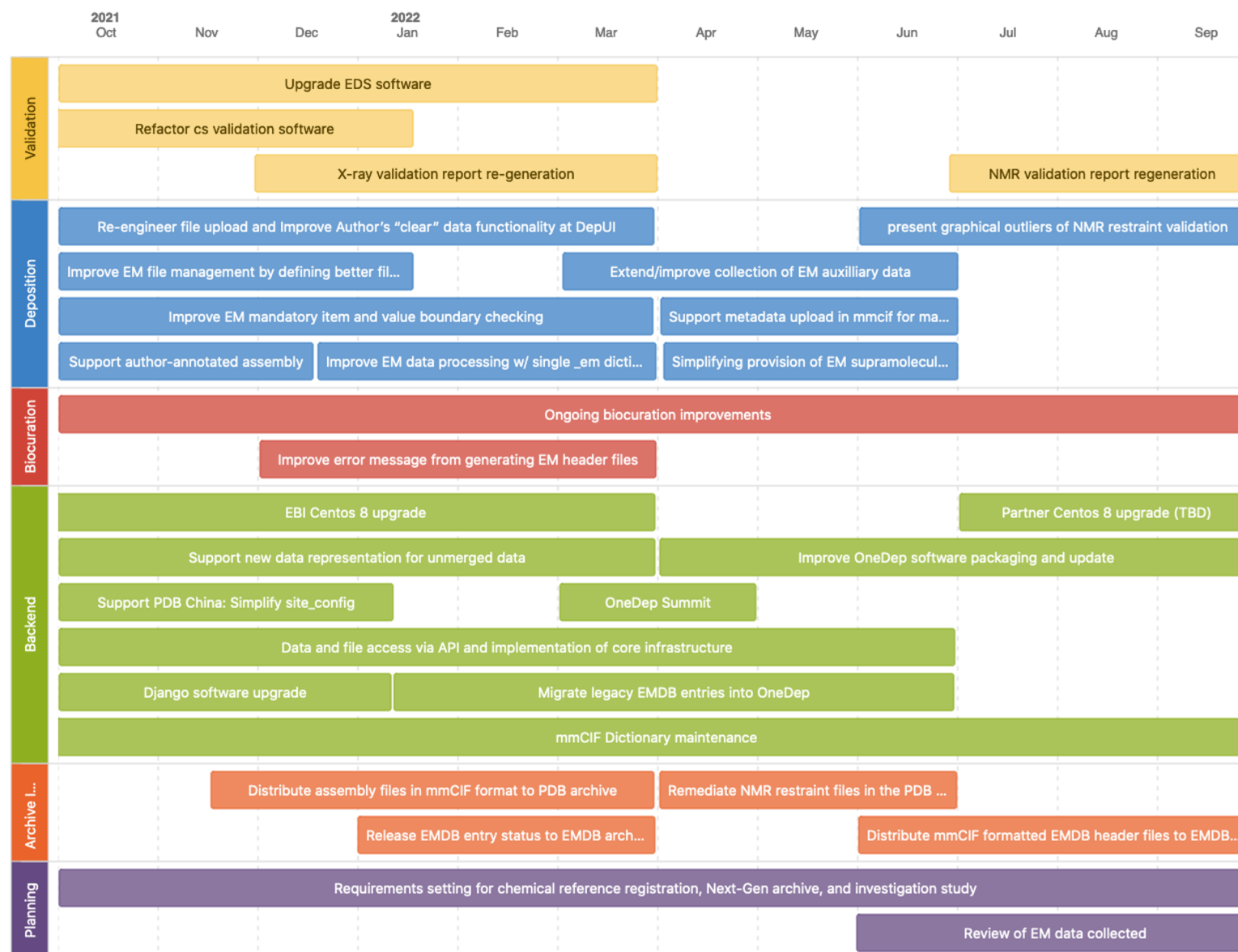
OneDep 2021/2022 Goal Setting

Bold: re-forecasted from 2020-2021

Components	Major Projects to be Completed	Primary resource
Validation	**Upgrade 3rd party EDS software**	**PDBe**
	Refactor NMR chemical shifts validation software	**BMRB**
	Validation report re-generation for X-ray and NMR entries	PDBe
Deposition	Extend/improve collection of EM auxiliary data	EMDB
	Streamline EM data model	EMDB
	Improve EM metadata checking/validation	EMDB
	Enable author-annotated assembly	**PDBe**
	Present graphical outliers of NMR restraint validation	**BMRBj/BMRB**
Infrastructure	Improve file and data access via API	RCSB PDB/PDBe
	Re-engineer file upload processes at deposition	PDBe
	Support new data representation for unmerged data for X-ray entries	RCSB PDB
PDB Archive	Distribute assembly files in mmCIF format	RCSB PDB
	Remediate NMR restraint files for validation report generation	BMRBj/BMRB

OneDep 2021/2022 roadmap

- will be further adjusted based on requirement setting

	2021 Oct	Nov	Dec	2022 Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep

Validation
- Upgrade EDS software
- Refactor cs validation software
- X-ray validation report re-generation
- NMR validation report regeneration

Deposition
- Re-engineer file upload and Improve Author's "clear" data functionality at DepUI
- present graphical outliers of NMR restraint validation
- Improve EM file management by defining better fil...
- Extend/improve collection of EM auxilliary data
- Improve EM mandatory item and value boundary checking
- Support metadata upload in mmcif for ma...
- Support author-annotated assembly
- Improve EM data processing w/ single _em dicti...
- Simplifying provision of EM supramolecul...

Biocuration
- Ongoing biocuration improvements
- Improve error message from generating EM header files

Backend
- EBI Centos 8 upgrade
- Partner Centos 8 upgrade (TBD)
- Support new data representation for unmerged data
- Improve OneDep software packaging and update
- Support PDB China: Simplify site_config
- OneDep Summit
- Data and file access via API and implementation of core infrastructure
- Django software upgrade
- Migrate legacy EMDB entries into OneDep
- mmCIF Dictionary maintenance

Archive I...
- Distribute assembly files in mmCIF format to PDB archive
- Remediate NMR restraint files in the PDB ...
- Release EMDB entry status to EMDB arch...
- Distribute mmCIF formatted EMDB header files to EMDB...

Planning
- Requirements setting for chemical reference registration, Next-Gen archive, and investigation study
- Review of EM data collected

PDB Archive Update

Stephen K. Burley

Current Archive Status

- Total Entries=182,418
  (as of 9/23/2021)

- Core Archive Storage
  - OneDep Sessions: ~26 TB
  - ftp(legacy + versioned): ~1.5TB
  - EMDB ftp: ~4 TB

- CoreTrustSeal certification renewed
  through April 2024 (CoreTrustSeal.org)

- Improved access of chemical references
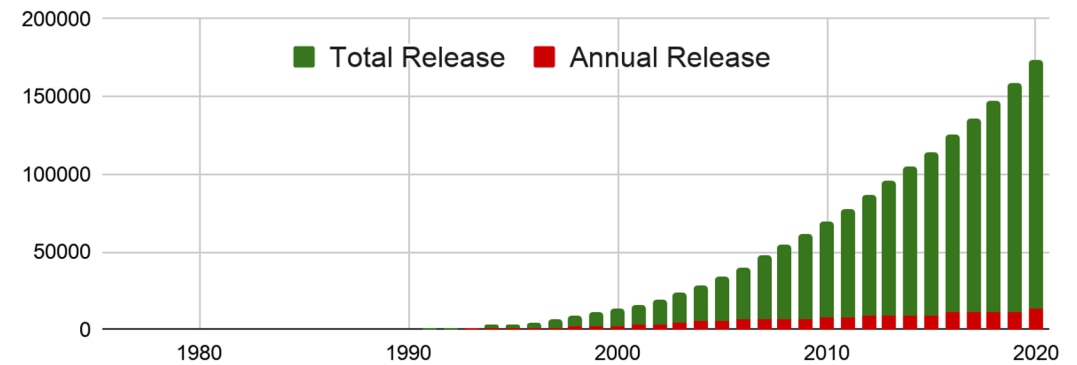 and information about PDB archive
  holdings
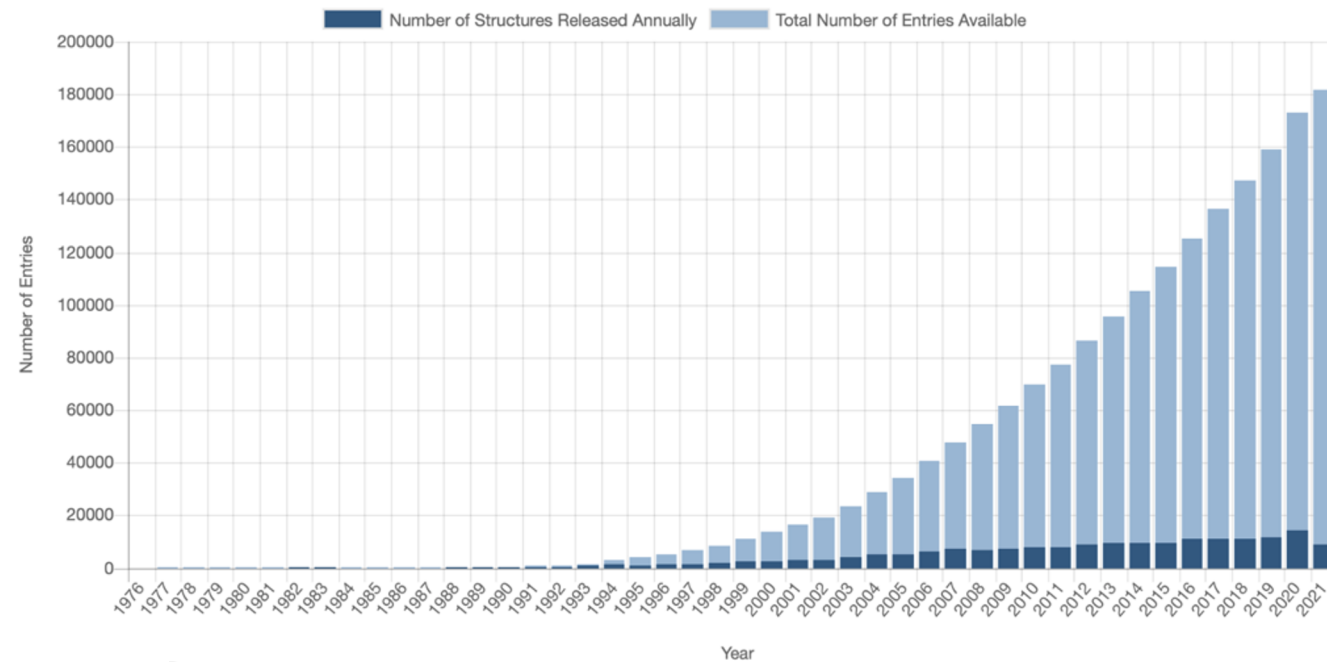
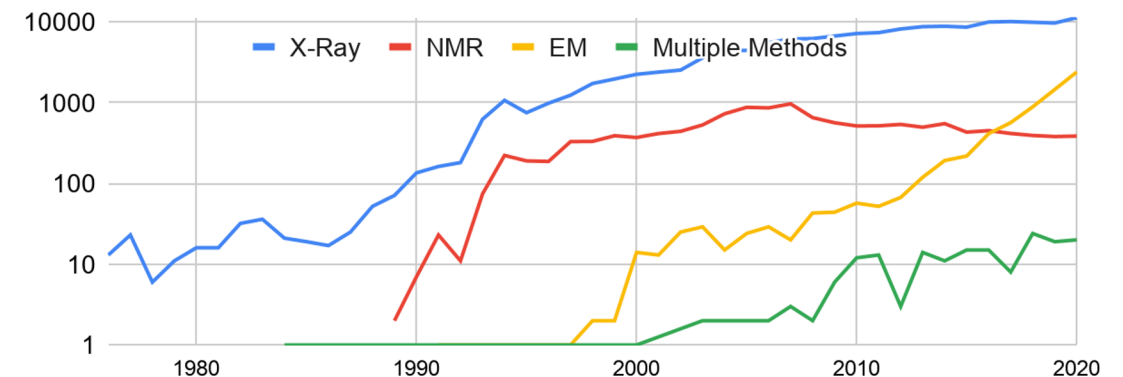Anticipate 200,000 structures in early 2023!

Archive Growth in 2020

- Year-end holdings >173,000

- Record 14,044 new entries released (~22% increase *versus* 2019)

- ~9% growth in the archive!

- Record 2,780 new EM entries released (~40% increase *versus* 2019; ~20% of new 2020 entries)

Total and Annual Release

Released Entries By Method/Year (log scale)

# PDB Chemical Reference Data

~34K Chemical Component Definitions
- 3118 new in 2020 (Up 29% over 2019)
- 6598 updated (Up 1.6x over 2019)

~1K Biologically Interesting molecule Reference Dictionary (BIRD) Definitions
- 174 new in 2020 (Up 9x over 2019)
- 130 updated (Up 60x over 2019)

Overlap with External Resources
- PubChem – 31603 (92% of CCD)
- ChEMBL – 12902 (38%)
- ChEBI – 4616 (14%)
- DrugBank – 6022 (18%)
- Pharos – 4277 (13%)

PRD_900012

MAN

PRD_900007
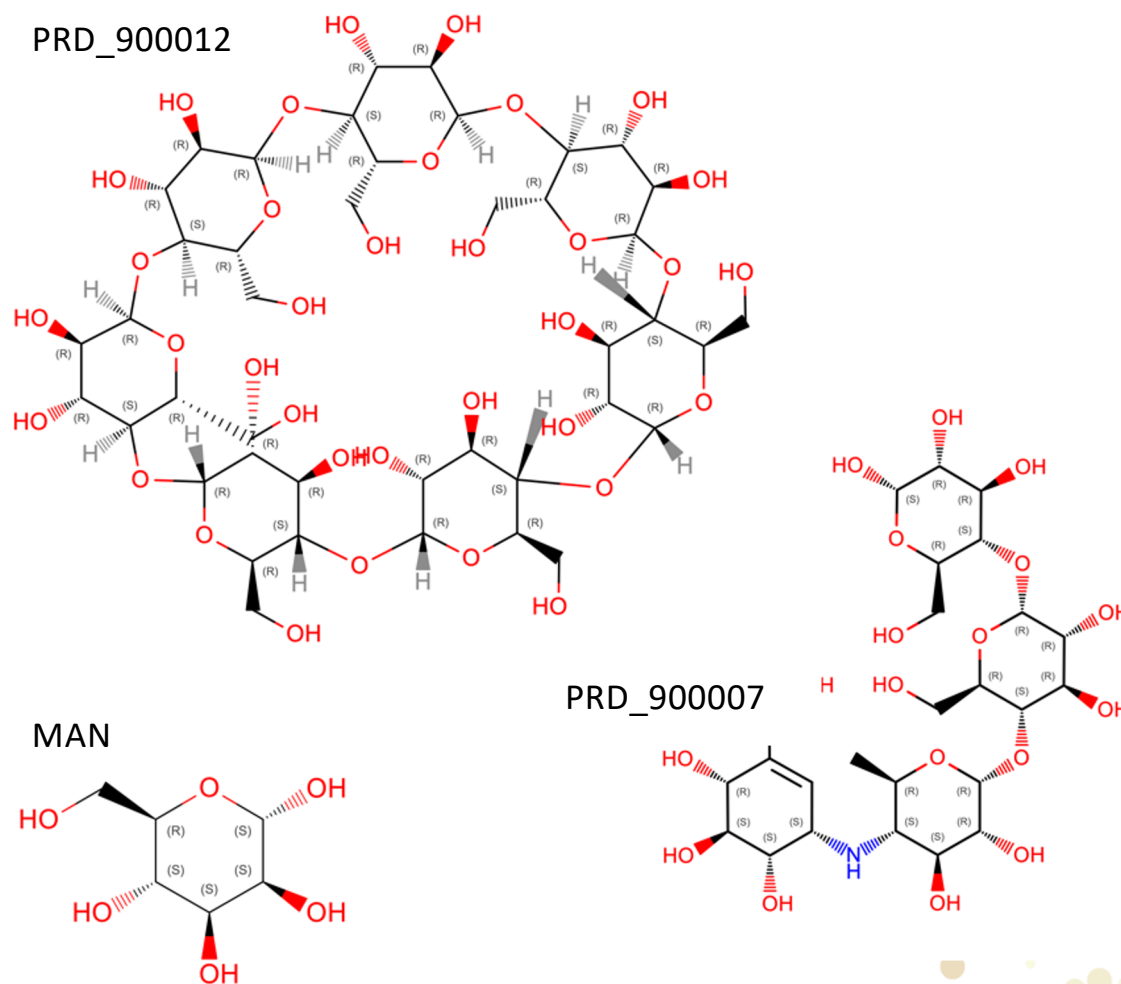
EMDB Update

Ardan Patwardhan, EMDB

Achievements in 2020

- 2020 EM-VTF meeting

- Improvements to EM validation reports

- Improvements to EM deposition

- 10,000[th] EMDB entry released

2020 EM-VTF meeting

- Face-to-face meeting held in January 2020 (over 40 participants)

- Productive and broad discussions on:

  - improvements to (meta-)data capture (EMDB/PDB)

  - EM validation strategy

  - improvement of wwPDB EM validation reports

- White paper in preparation (Kleywegt *et al.*)

Improvements to EM validation reports

- OneDep now generates EM map-only and map-model validation reports

- Reports generated upon deposition, annotation and release of an entry

- A number of EM-VTF recommendations for improvement already implemented

Improvements to EM validation reports

- Improvements to the FSC plot:
  a. FSC plot only shows '0.143', '0.5' and 'half-bit' criteria
  b. Author-provided and calculated FSC curves in one plot
  c. Resolution-estimate table re-designed for easier viewing

- Added:
  a. If half maps were deposited
     - Calculate raw map
     - Show various slices and projections of the raw map
     - Include orthogonal surface views of the raw map
     - Rotationally Averaged Power Spectrum of raw map added to RAPS plot
  b. Visualisation of map and full assembly (e.g., for viruses)

6.4.2   Raw map

X   Y   Z

These images show the 3D surface of the raw map. The raw map's contour level was selected so that its surface encloses the same volume as the primary map does at its recommended contour level.

Improvements to EM deposition

- Improved upload of large files

- Better EM-data-related integrity checks

- General improvements concerning how EM-related data is collected

- A large number of historical bugs fixed

Over 10,000 EMDB entries

- Achieved in February 2020 after 18 years

- Created a poster to celebrate this milestone

- Over 16,000 entries by mid-August 2021

- Prediction: 20,000 entries in first half of 2022

- 3,830 new entries released in 2020

- 2,660 released in 2021 (until mid-August)

CELEBRATING
10 000
EMDB ENTRIES
emdb-empiar.org
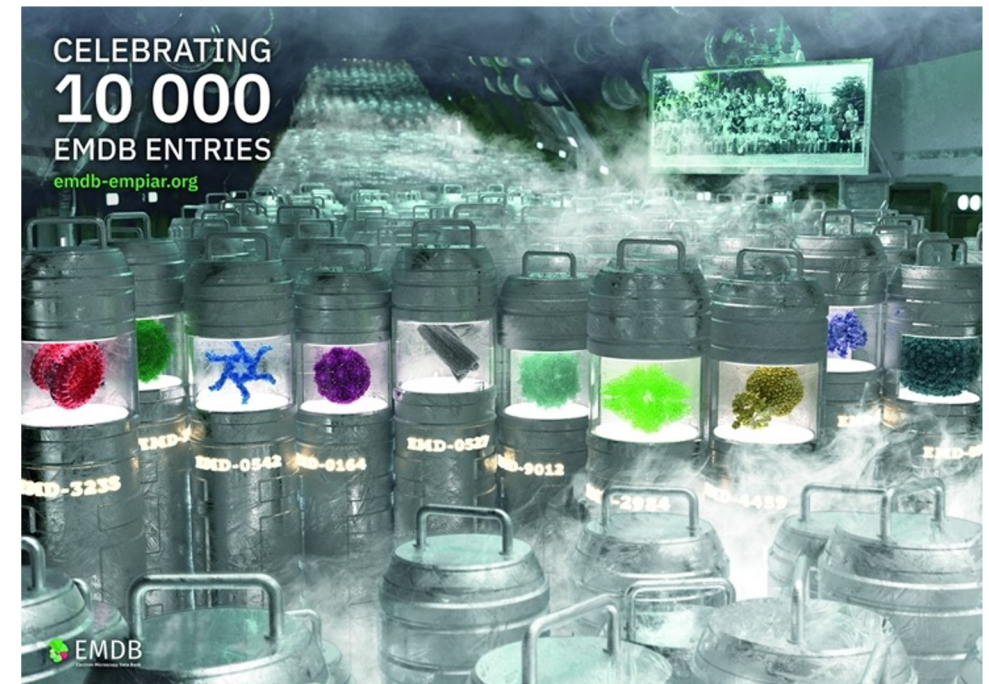
BRMB Update

Jeffrey C. Hoch, BMRB

Developments since 2020 Meeting

OneDep Activity

- <u>Completed</u>
- Support for single file upload for NMR data in with NMR-STAR/NEF in OneDep

- NMR distance and dihedral angle restraints analysis is now available through OneDep and wwPDB validation servers.

- <u>On Going</u>
- Installation of OneDep instance at BMRB
    - Docker method installed.
    - One script installation, PDBe working on simplifying config files.

- Chemical shift validation code refactoring

- Remediate restraints data available in BMRB restraints grid and copy to wwPDB FTP

- Generate restraints validation report for the whole archive.

FTP update

Developments since 2020 Meeting

BMRB Core Archive

- Completed migration of BMRB services to UConn

- Completed migration of BMRB-related services developed in tandem with NMRFAM (RUNER, GISSMO, etc.) to UConn

- Completed migration of BMRB CS-Rosetta server to CHTC

- Implemented new MailDB to archive annotator messages (within the UConn architecture we don't control our own mail server)

- Improvements made to many internal processes and workflows as part of the migration

- Substantial packaging, speed, and error message helpfulness improvements to PyNMR-STAR library

- Website implementation of new graphical design nearly complete – full-time web developed hired

- Progress towards re-implementation of BMRB dictionary management tools in Python (replacing decades-old Visual Basic code)

- Completed and deployed M2MTool to facilitate BMRB depositions directly from inside of NMRbox

- Partnered with NMRbox to add BMRB to the NMRbox "Reboxitory" data lake

BMRB Core Archive Growth

- BMRB has released 857 new entries so far in 2021 (180 via OneDep)
- Total released entries estimated to reach ~15050 by the end of 2021.
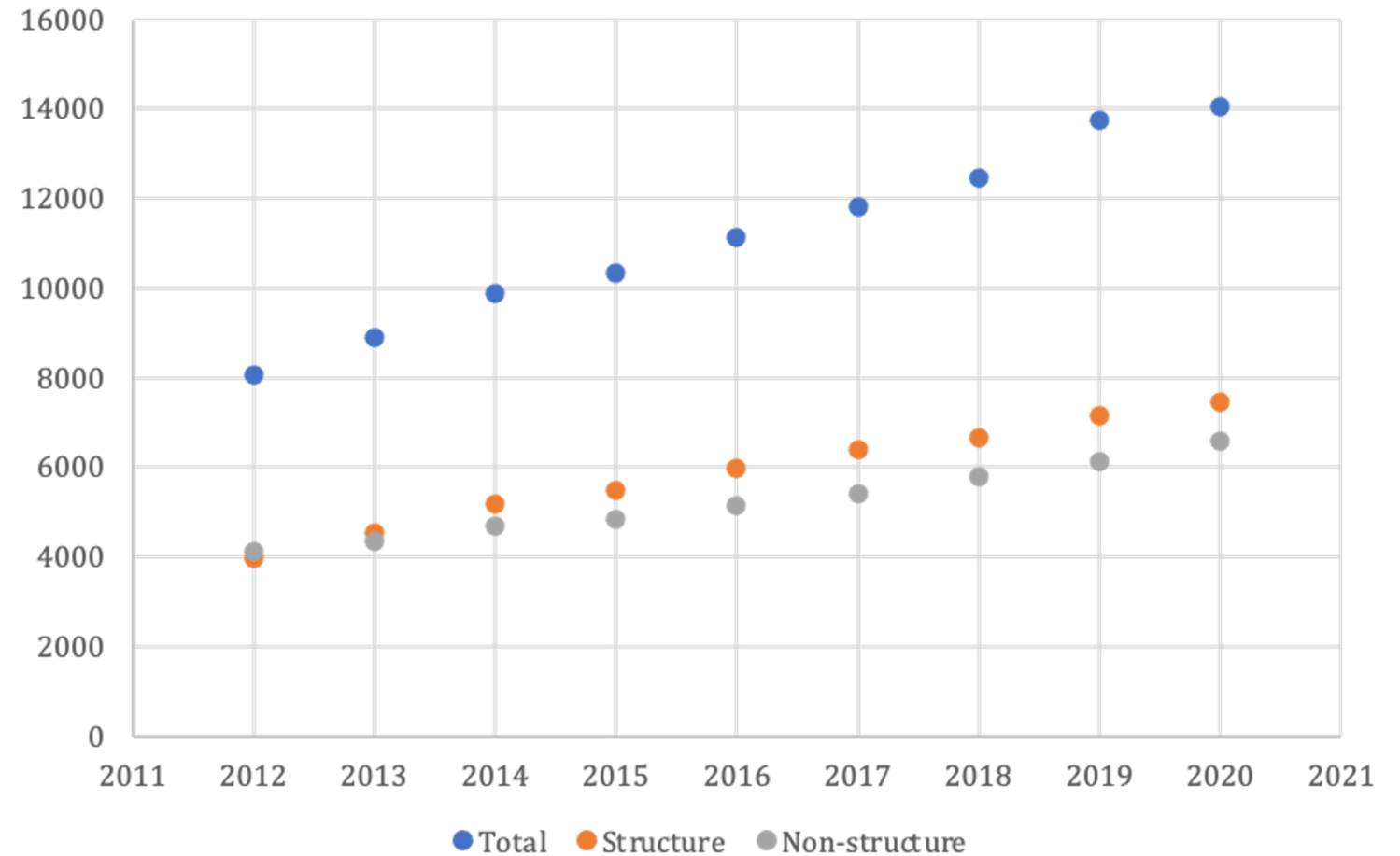
BMRB Core Archive Growth

Total Released Entries

Year	Total released	Yearly increase	Structures	Yearly increase	Non-structures	Yearly increase
2012	8068	814	3953	536	4115	278
2013	8886	818	4524	571	4362	247
2014	9867	981	5182	658	4685	323
2015	10322	455	5481	299	4841	156
2016	11112	790	5977	496	5135	294
2017	11803	691	6395	418	5408	273
2018	12438	635	6666	271	5772	364
2019	13728	867	7147	491	6131	376
2020	14031	792	7458	337	6573	455

BMRB Core Archive Growth

Internet Server Traffic (Website) – All Mirrors*

Year	Server requests	Page requests	File requests	Distinct hosts served	Total data transferred (TB)
2012	39,614,021	5,934,563	946,365	839,596	16.74
2013	42,154,225	6,236,267	864,391	987,792	17.99
2014	37,703,581	6,124,679	589,180	1,091,830	25.54
2015	46,899,762	7,011,339	664,415	1,225,468	32.68
2016	54,081,884	7,869,456	1,028,193	1,322,094	36.38
2017	56,209,400	8,537,562	1,433,036	1,026,426	18.37
2018	87,818,181	25,523,384	2,415,042	1,330,894	17.75
2019	87,989,534	33,885,684	1,173,443	1,779,867	27.55
2020	83,732,415	22,489,261	1,603,535	1,783,602	47.67

- ~300K/day server and page requests
- BMRB has mirror sites in Italy and Japan, and PDBj-BMRB branch for deposition
- Updates to accounting methods resulted in slight changes to historical data from previous reports

BMRB Core Archive Growth

Internet Server Traffic (FTP Servers) – All Mirrors*

Year	Server requests	Distinct files requested	Distinct hosts served	Total data transferred
2012	2,058,066	1,597,183	5,037	1.1 TB
2013	2,018,662	1,503,932	5,494	1.4 TB
2014	1,991,174	1,486,165	4,930	1.6 TB
2015	2,185,255	1,655,143	3,915	0.9 TB
2016	5,704,287	1,722,143	5,956	1.7 TB
2017	4,862,305	2,335,675	4,226	4.6 TB
2018	4,715,647	2,788,527	3,866	2.0 TB
2019	4,845,421	2,423,941	3,908	5.5 TB
2020	FTP deprecated GLOBUS endpoint deployed			

Network for Advanced NMR

NSF Mid-scale RI2 4-year grant to UConn, UW-Madison, UGA

2 x 1.1 GHz

Virtual network of 26 NMR spectrometers in total

"AirB$_0$" resource discovery

Knowledgebases:

> SSNMR for biomacromolecules
>
> SSNMR for materials
>
> Solution NMR for biomacromolecules
>
> Solution NMR metabolomics

Potential impact for BMRB:

> All connected spectrometers will stream experimental data to a central archive, with simple opt-in for public accessibility and curation

Joint Projects Update

Sameer Velankar, PDBe

Joint NSF/BBSRC projects

- ## Three year projects
  - The US funding is provided by NSF and the UK funding is provided by BBSRC

- ## Project 1 started in Oct 2020 at RCSB and Jan 2021 at PDBe
  - Provides additional resources for OneDep development
  - Coordination with OneDep team

- ## Project 2 was awarded in August 2021
  - PDBe will need to complete the administrative processes at BBSRC
  - Will provide resources for development of Mol*, web-based components for displaying annotations and efficient data delivery mechanism

Project 1: OneDep and NextGen PDB archive development

- SA1. To improve the fidelity and completeness of 3D structure data deposited into the PDB by
  - Harvesting data automatically from structure determination software packages; and
  - Streamlining the wwPDB data deposition, validation, and biocuration system.

- SA2. To improve "FAIRness" of PDB data for researchers, educators, and students by
  - Extending chemical metadata for small-molecule ligands (*e.g.*, bound cofactors and inhibitors);
  - Incorporating enhanced descriptions of macromolecular assemblies;
  - Grouping related PDB structures into investigations for more efficient, parallel data delivery; and
  - Creating a "Next Generation" PDB data repository with up-to-date metadata.

- SA3. To modernize the IT infrastructure to futureproof PDB data management and weekly PDB archive release to the public domain by
  - Developing new application programming interfaces (APIs) and microservices infrastructure; and
  - Updating existing mechanisms for synchronization of data and software across wwPDB data centers in the US, Europe, and Asia.

Protect 1: Major Year One Activities

- SA1: Gather stakeholder content and API requirements for deposition data preparation and chemical component pre-registration, and

- SA1: Implement new deposition data preparation API for experimental metadata
  - Convened a Data Processing Subgroup of the PDBx/mmCIF Deposition Working
  - Developed ~50 new and ~100 refined definitions
  - New content supported by the STARANISO, Dials, and CCP4 program systems
  - New content accepted in OneDep –

    www.wwpdb.org/news/news?year=2021#60638da1931d5660393084c3
  - Subgroup is arranging virtual meetings with ISPyB and SynchWeb to discuss developing API requirements for data exchange
  - PDBx/mmCIF Working Group developers representing Phenix, CCP4, and Global Phasing have also implemented the delivery of complete chemical component definitions for each constituent molecule in a deposited structure.
  - Working group and dictionary development activities now documented on GitHub github.com/pdbxmmcifwg and github.com/wwpdb-dictionaries

# Protect 1: Major Year One Activities

- SA2: Gather stakeholder content requirements for sequence, assembly, ligand, investigation semantics, and packaging
  - Developed representation of SIFTS data in PDBx/mmCIF including residue level mapping between PBD sequences and UniProt, Pfam, SCOP and CATH. These data to be initially distributed by PDBe.
  - Accelerated the plan for PDBx/mmCIF assembly delivery in the main archive in 2022
  - In a separately funded BBSRC project, PDBe has gathered requirements for harvesting data from ligand screening projects in collaboration with ESRF and Diamond facilities
  - Developing requirements for multi-structure investigations is a major task for the coming year

# Protect 1: Major Year One Activities

- SA3: Evaluate software choices for updating IT infrastructure, and
- SA3: Update IT infrastructure
  - Reached consensus on open source technology choices for container packaging (Docker), container orchestration (Kubernetes and Singularity), workflow management (Prefect), an API delivery framework (Python FastAPI) and log management (Elastic, Logstash, and Kibana (ELK)).
  - Streamlined OneDep installation and system build tooling
  - Migrated version control system from SVN to GitHub
  - Python packages published to the public PyPI packaging service
  - Prototyped scheduled task management a new workflow management system (Prefect)
  - Prototyped self-hosted deployments of Kubernetes automated using the Ansible provisioning tool
  - Initial containerized deployments of public-facing OneDep services (Docker)
  - Developed requirements for a File access API to support data exchange within and between wwPDB sites (under development)

Project 1: progress in 2021 (SA1)

- SA1 - Improve fidelity and completeness of 3D structure data deposited into the PDB
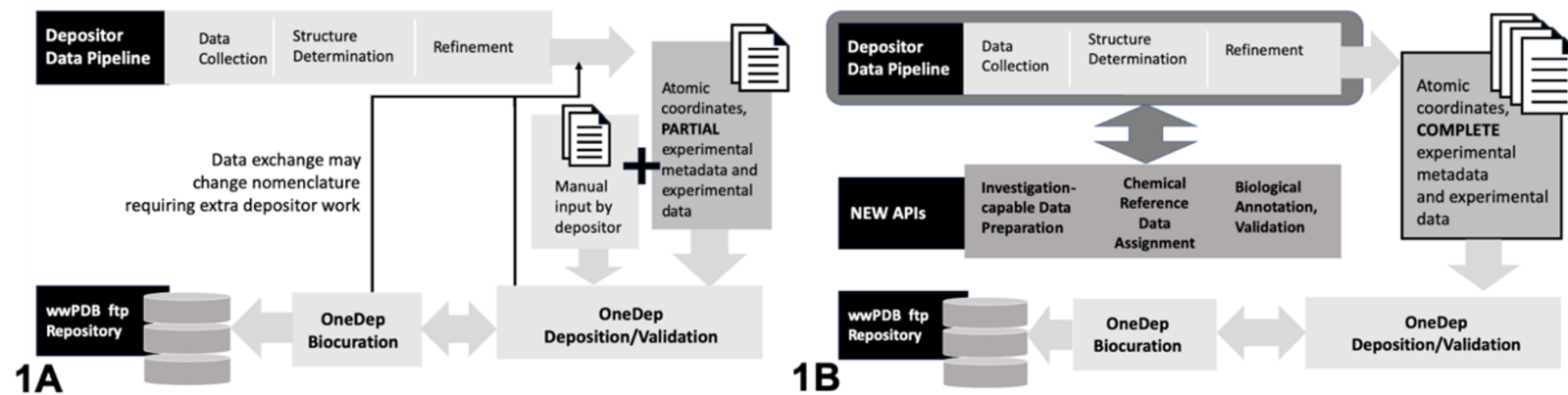

- Ongoing work towards enabling automated creation of and upload to deposition sessions

Authentication → API based creation of deposition session → API based upload files to deposition session
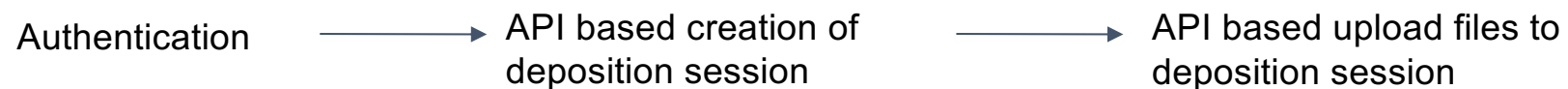
- Will make deposition of investigations possible *via* OneDep

Project 1: Progress in 2021 (SA2)

- Background
  - Each wwPDB site adds additional data to support their data out activities
  - Aggregating of PDB entries is only possible after the PDB weekly release
  - A NextGen FTP will be introduced which will make this additional data and add aggregated data available to users

- First example of additional data
  - Cross-reference to other data resources and residue level mapping to UniProt (based on the PDBe SIFTS data)

- Added new categories to PDBx/mmCIF dictionary for SIFTS data
- Includes
  - Per residue mapping to external resource (i.e. UniProt, Pfam, CATH, SCOP)
  - UniProt numbering of residues (where appropriate)
- mmCIF files with additional SIFTS data will initially be available from PDBe FTP area
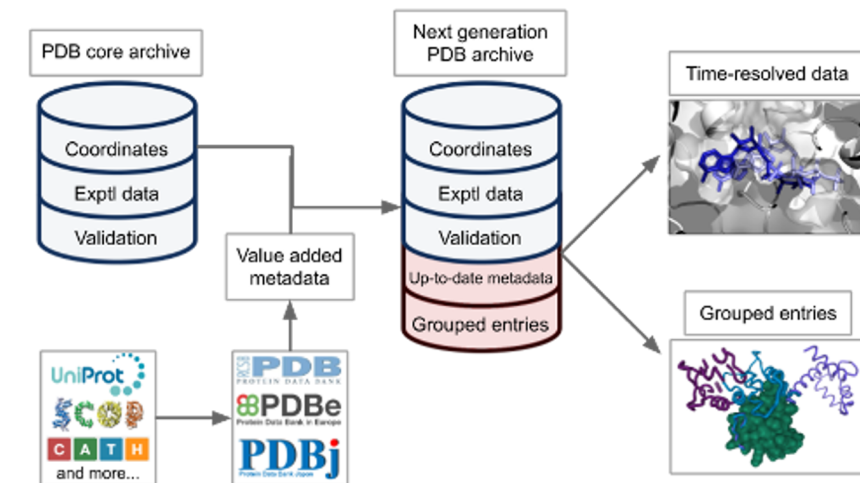
Project 1: Progress in 2021 (SA3)

- Installation procedure of wwPDB OneDep documented and simplified - from 133 manual steps to single installation script

- OneDep is comprised of the following components:
    - System configuration
    - Core utilities
    - Web application and software

- OneDep system configuration
    - On-going work on simplification
    - All OneDep apps changed to support a simpler configuration
    - Required to support additional wwPDB sites

- Core utilities published to PyPi Python repository

- Web applications and software
    - Docker containers created for all public facing modules
    - Docker containers automatically built using CI/CD
    - To be deployed on Kubernetes

Project 1: progress in 2021 (SA3)

- File handling API to provide
  - Single access point for data in OneDep
  - Data exchange mechanism for OneDep
  - Groundwork for data exchange for SA2
- Initial implementation scheduled for 2021

- OneDep workflow engine
  - Currently bespoke software which is tied to individual servers
  - Investigated open source replacements which are suitable for containerisation
  - Initial implementation developed to
    - Add resilience to regularly scheduled jobs (crons)
    - Monitor memory usage for OneDep processes
  - Will initially prototype deposition workflow and assess before making a decision about using the same approach for all of OneDep

Project 2: Visualization Infrastructure

- SA1 - To augment Mol-star for seamless operation across length scales ranging from atoms to cells with multiscale structure capability by
  a. Extending existing data standards to combine atomic coordinates with multiscale structure information from integrative/hybrid methods and in-cell molecular details from advanced electron or visible light/super-resolution imaging techniques;
  b. Building tools that enable superposition/simultaneous interactive display of multiscale structures;
  c. Enabling comparisons and analyses of multiscale structures.

- SA2 - To extend 3D visualization beyond structure to include display of associated biological and functional annotations by
  a. Developing a library of web-based components for displaying structural, biological and functional annotations;
  b. Integrating Mol-star 3D display tools with complementary web-based components, such as those displaying 2D topology and 1D sequence data.

- SA3 - To extend and strengthen existing IT infrastructure underpinning Mol-star by developing an agile data delivery system that supports high-speed interactive web-based visualization of 3D structure data for comparison and analyses of multiscale structures across length scales ranging Å units to microns.