2023 wwPDB AC Meeting

Friday, October 27, 2023

13:00 - 16:00 UK (UTC+1)

Agenda

Time Zones								Session	Participants
Europe UTC+3	UK UTC+1	US NJ/CT UTC-4	US CA UTC-7	Australia Sydney UTC+10	Japan UTC+9	China UTC+8	India UTC+5: 30		
15:00	13:00	8:00	5:00	23:00	21:00	20:00	17:30	Welcome and Introductions (10 min)	EMDB
15:10	13:10	8:10	5:10	23:10	21:10	20:10	17:40	Discuss questions/answers/any other issues (90 min)	All
16:40	14:40	9:40	6:40	00:40 15th Oct.	22:40	21:40	19:10	Executive Session (35 min)	wwPDB-AC
17:15	15:15	10:15	7:15	01:15	23:15	22:15	19:45	Summarise wwPDB-AC feedback and further discussion (35 min)	All
17:50	15:50	10:50	7:50	01:50	23:50	22:50	20:20	Acknowledgements and photo (10 min)	All
18:00	16:00	11:00	8:00	02:00	24:00	23:00	20:30	Meeting End	

Welcome and Introductions

Kyle Morris, EMDB

Discuss questions/answers/ any other issues

Executive Session

Host will add Advisors into Zoom Breakout Room

Please re-join us at the end of your discussion by selecting
Leave the breakroom (lower right hand corner)

Questions for Advisor Feedback

1. Threats to the wwPDB Open Access Model (Slide 10)
2. Future of OneDep Development (Slide 16)
3. BMRB Funding Update (Slide 19)
4. Future of Data Archiving (Slide 22)
5. Next Advisory Committee Meeting (to discuss dates)

1. Threats to the wwPDB Open Access Model

Increasing political and economic polarization among major producers of structural biology information could lead to adoption and/or enforcement of local regulations in countries that interfere with wwPDB open access operations.

- Data downloads to select nations (*e.g.*, North Korea, Iran, Russia) could be prohibited by wwPDB member host countries.
- Data deposition to global wwPDB Core Archives (PDB, EMDB, BMRB) from select nations could also be prohibited.
  (N.B.: Many countries already ban "export" of genomic data.)

Current Status

- All data stored in the three wwPDB Core Archives are made available worldwide under the most permissive Creative Commons CC0 License.
- CC0 licensing means that anyone, anywhere can download information from the PDB, EMDB, and BMRB archives at no cost and with no restrictions on data utilization.
- There are no limitations on deposition of data to the PDB, EMDB, and BMRB archives in any geography.
- No charges are levied on wwPDB *Data Depositors* from any country.
- An NMR-related service provided by Jeff Hoch was asked to block access from an institute in Russia, but as of now there are no restrictions on access to any of the wwPDB Core Archives.

Feedback requested on: Opportunities to Strengthen Open Access

1. wwPDB Leader engagement of their national/regional funders to explain the potential impact of open access limitations on global research efforts across fundamental biology, biomedicine, and energy sciences.
2. wwPDB Leadership engagement with the Global Biodata Coalition.
3. wwPDB Leadership and the wwPDB AC could together contribute to the debate by publishing a “position paper” in a high-profile journal.

Discussion/Feedback/Questions

Questions for the wwPDB AC re Slides 7-9:

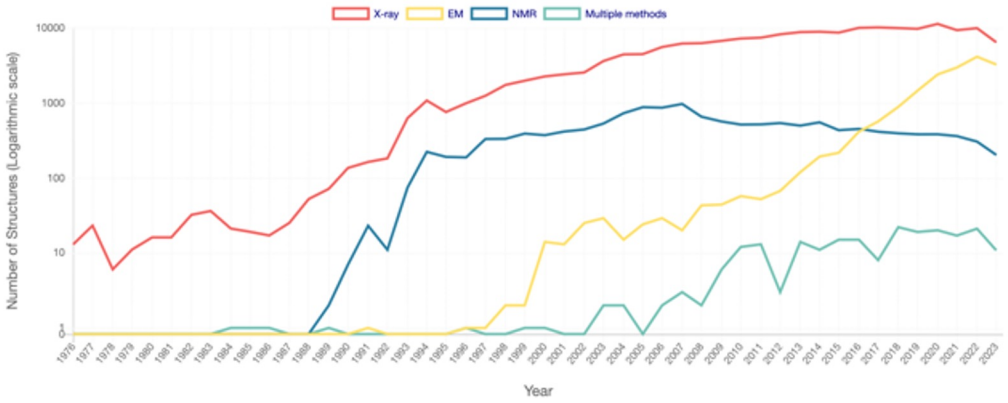
1. **Does the AC concur with 1 ?**
2. **Does the AC concur with 2?**
3. **Would the AC be willing to join an effort outlined in 3?**

2. Future of OneDep Development

Automate the deposition process by working with the key community software developers to establish a shared software package to prepare deposition-ready files and use the deposition API to automate deposition to the PDB and EMDB archive whilst improving the data quality and completeness.

Scientific Drivers

- Rapid advances in structure determination techniques have accelerated the number of submissions to the PDB and EMDB archives.
- Increasingly, individual experiments (EM, time-resolved, fragment screening, *etc*.) yield ensembles of 3D structures of molecular systems.
- The fraction of data captured and archived has not kept pace with the rapid advances, resulting in a loss of opportunity to take advantage of AI/ML methods to improve experimental protocols.
- Significant enhancements of the PDBx/mmCIF framework and the wwPDB OneDep system are required to improve semantic representation and simplify the Depositor experience whilst maintaining high-quality archiving to support the development of AI/ML tools for biocuration.



Current Status

- wwPDB partners have collaborated with key community software developers to build the fully extensible and human/machine-readable PDBx/mmCIF data standard (PDB exchange/ macromolecular crystallographic information framework).
- PDBx/mmCIF is now the standard for accepting data during deposition to the PDB archive.
- Each structure determination software tool has developed a process to prepare PDBx/mmCIF-compliant data files for starting PDB deposition, resulting in multiple implementations and add-ons.
- wwPDB partners have established a deposition API that facilitates automated data transfer in PDBx/mmCIF format from structure determination software tools to the OneDep system.
- wwPDB partners are working with the software developer community so that they can use the API to initiate structure depositions and transfer all relevant files.
- Once requisite files are uploaded, and OneDep automatically generates a deposition, structural biologists can add missing information *via* the OneDep web interface and submit a complete deposition for rigorous validation and expert biocuration before final structure review by the Depositor and eventual public release.

Feedback requested on: Future Plans

- Improve sustainability and efficiency of the deposition process by working with key community software developers on developing a common software framework that uses APIs to enable and manage data/metadata capture and prepare PDB/EMDB deposition-ready data/metadata files.
- Software developers and the wwPDB will work collaboratively to maintain these APIs, reducing the overall effort in software maintenance and providing a more efficient solution for automated 3D biostructure data/metadata deposition.
- Improve the Depositor experience by providing a standard user-facing functionality, making it easier for inexperienced Depositors to prepare data/metadata files necessary for automated deposition.
- Enhance metadata capture using the PDBx/mmCIF framework and develop additional semantics for missing metadata.
- Improve metadata validation based on the PDBx/mmCIF framework to improve data quality.

Feedback requested on: Future Plans

- Enrich the current set of deposition APIs to expose additional data validation and cross-reference methods, such as ligand matching to the wwPDB Chemical Component Dictionary (CCD) and polymer sample sequence matching to reference protein sequence data resources (e.g., UniProt, NCBI).
- Integration of additional APIs during deposition file preparation will allow the identification of any issues requiring resolution prior to deposition, making the entire process more efficient and transparent for both Depositors and wwPDB Biocurators.
- Enhance APIs to support deposition of multiple entries in an automated fashion using “Investigations” that can group multiple datasets with associated semantic descriptions of the relationship between different PDB and EMDB archival entries (e.g., different conformations providing information on macromolecular dynamics).

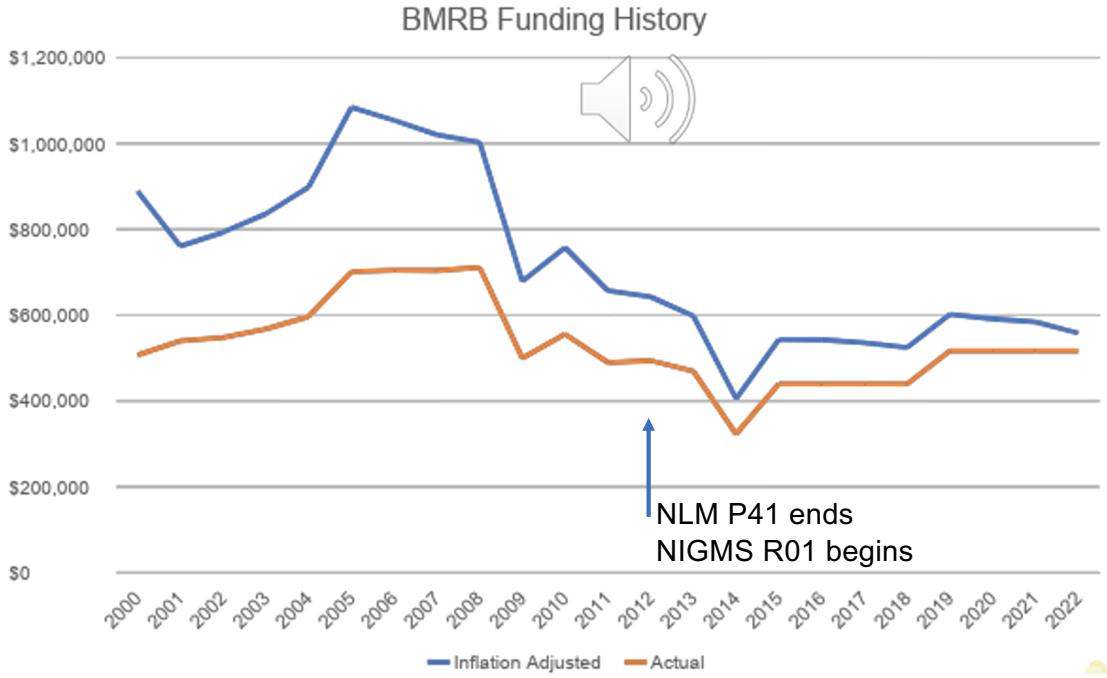
Discussion/Feedback

Questions re Slides 11-15:

1. **Does the AC concur with the plans for automating the deposition process as the future direction for OneDep development?**

3. BMRB Funding Status

- Prior R01 ended 3/31/23
- R24 NOA effective 9/20/23 **$542,792K/yr**
  - 44% cut to request, 5% increase over R01



BMRB Budget Breakdown

Minimal requirements for operations

- Biocuration of >1000 depositions/yr
  - 100% staff scientist $117,057
- Maintenance of deposition/validation pipeline, OneDep commitment
  - 100 % staff scientist $168,105
  - 25% software engineer $44,265
  - 100 hrs/yr consulting developer of BMRB software $20,000
- T & C requirement
  - 20% PI $54,383
- Misc (AC, DOIs, software licenses, community outreach) $45,000
- **Total: $448,810**
- <$100K insufficient for remediation, expanding curated data types, retiring technical debt, enhancing user experience, improving sustainability

Discussion/Feedback/Questions

Questions for the wwPDB AC re Slides 17-18:

1. **Does the AC have suggestions for alternative funding mechanisms to sustain BMRB?**

Feedback requested on:
4. Future of Data Archiving

Well-intentioned mandates from funders and publishers, but have unintended detrimental impacts

- Emphasis on generalist data repositories (because they are free!)
- Prioritizing novelty in data management plans (*vs*. standardization)
- The result will be Scientific Libraries of Babel, full of everything but nothing can be trusted and many files will never be found

Creating and maintaining trusted repositories is expensive and labor-intensive. And costs are growing.

- Short term: Additional funding is needed to sustain operations
- Long term: Research is needed to develop technologies that reduce skilled labor costs (AI/ML)

Lessons learned from 50 years of operating the PDB

- **Storage is cheap. Curation is expensive.** The capacity of data storage devices continues to increase, and unit costs continue to go down. Curation, on the other hand, requires work by domain experts, and labor costs continue to rise.
- **Trust is indispensable.** Storage of uncurated information does little to help the next generation of researchers. Most of the time, they do not know what to trust. Data curation by recognized domain experts adds trust, both in format and content. Trusted data sources are also essential for realizing the benefits of advances in artificial intelligence/machine learning (4).
- **Data curation is most accurate and efficient when done close to the source.** The higher the level of expertise and the more dialogue between curators and depositors resolving ambiguities and reducing errors, the better the final data product. Remediation (data cleaning) conducted long after data have been deposited is more challenging, very time-consuming, and inherently less accurate.
- **Broad consensus on data standards is essential and builds slowly.** Community standards for data formats and common definitions of terms (Ontologies) are essential for making stored information "Interoperable" and "Reusable". Achieving community consensus to adopt the mmCIF data standard utilized by wwPDB required two decades of constructive dialog.
- **Novelty in design of data management plans is not a virtue.** An overabundance of practices is the antithesis of standards. Innovations for data deposition, validation, and curation are needed to improve accuracy, efficiency, and sustainability, but not at the expense of community standards.
- **Data lakes too easily turn into digital swamps.** Generalist data repositories represent low-barrier pathways to making data "Accessible", but data standards are needed for robust discovery and federation. They contribute little or nothing to the imperatives of making data "Interoperable" or "Reusable". Nor do they bring communities together to establish trust.
- **Left to their own devices, scientists are unreliable narrators.** Scientists use different terms to describe the same thing, and frequently mistype manually entered data. Controlled vocabularies and limited options (e.g., pull-down menus) improve the accuracy of data capture and promote "Reusability".
- **Quotidian operations of data resources are considered uninspiring.** Funding agencies and scientific publishers routinely value innovation over consistency and awarding monies for new research over critical infrastructure. Consistency and staying power, both essential for effective data stewardship, are often misconstrued as being stodgy and boring.
- **Sustaining data resources is challenging.** Data generated with public research funding should be "Accessible". Earmarking a minuscule percentage of public research funding (5%) for sustaining FAIR stewardship of the information has been advocated by thought leaders and deserves serious consideration.

Discussion/Feedback

Questions re Slides 20-21:

1. **How can we engage the funders and the community to inform the discussion on the well-intentioned mandates from funders and publishers for archiving data in generalist repositories, but which will have an unintended detrimental impact on funding of specialist repositories and data FAIRness?**

1. **Would the AC be willing to join a wwPDB effort publishing a "position paper" in a high-profile journal?**

For information

wwPDB Plan for Transitioning to Extended PDB IDs and PDBx/mmCIF Format

Advisory Committee Meeting 2024

2024 wwPDB AC Meeting (format virtual) will be hosted by PDBj

Which date would the Advisory Committee prefer?

Friday October 18th 2024

Friday October 25th 2024

Summarise wwPDB-AC feedback and further discussion

Acknowledgements

Kyle Morris, EMDB

Group Photo

Meeting Close

Thank you

Poliovirus Neutralization (Artist: David S. Goodsell)

Pre-Meeting Review Slides

Table of Contents

Funding Update

Stephen K. Burley, RCSB PDB

wwPDB Full Member Core Funding

- RCSB PDB: Joint NSF/NIH/DOE funding (2019-2023)
  - On track. Renewal process for 2024-2028 to be completed before end 2023
- BMRB: NIH NIGMS funding (2023-2028)
  - Inadequate budget: still need to find additional support
  - UConn funding 25% administrative assistant and 25% project manager positions
  - NIH U24 funded as 5-year R24 with 44% reduction from the submitted budget request.
- PDBe: EMBL-EBI, Wellcome Trust (2021-2025)
- PDBj: NBDC-JST and AMED funding (2022-2027)
  - 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)

wwPDB Associate Member Funding

Basic support of PDB China (basic PDBc data-in and data-out operations) can be provided by guaranteed funds for NFPS. “Data-base and Computation” for protein science, with approved support for 9 FTEs (personnel), is one of these nine technical systems in NFPS.

For the current funding period (years 2020-2022), this basic yearly grant is CNY 57,130,000 (~USD $9,021,000/year).

On top of this basic grant, we also receive 0-15% merit-based bonus operation fund every year, and have supplementary salary support for personnel from the Chinese Academy of Sciences (CAS).

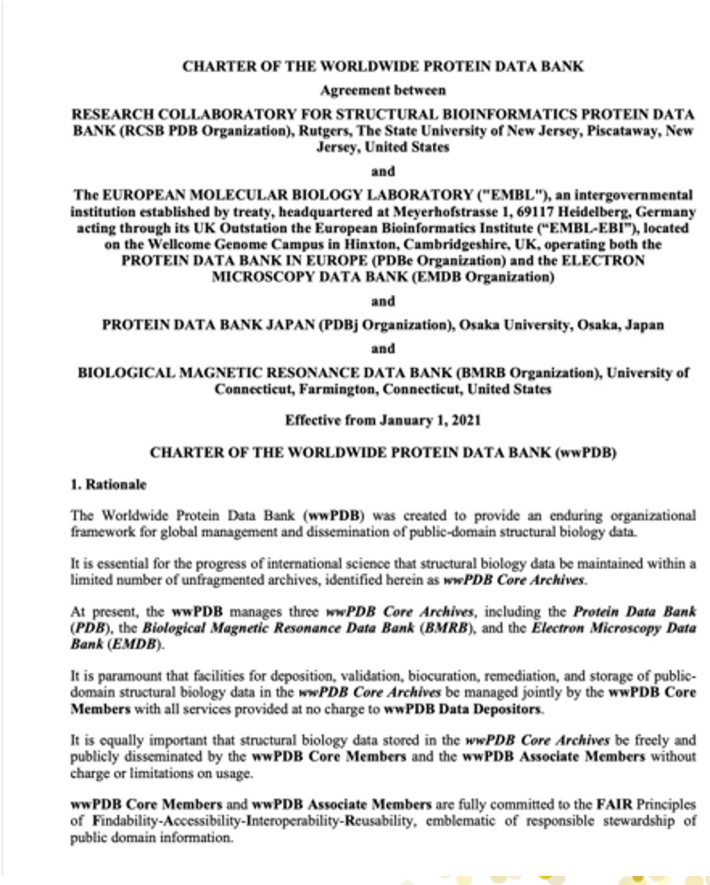
Governance Update

Stephen K. Burley, RCSB PDB

wwPDB Charter Updated

- PDBc admission as the first Associate Member announced in print

- Preparing to admit PDBi as an Associate Member



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.

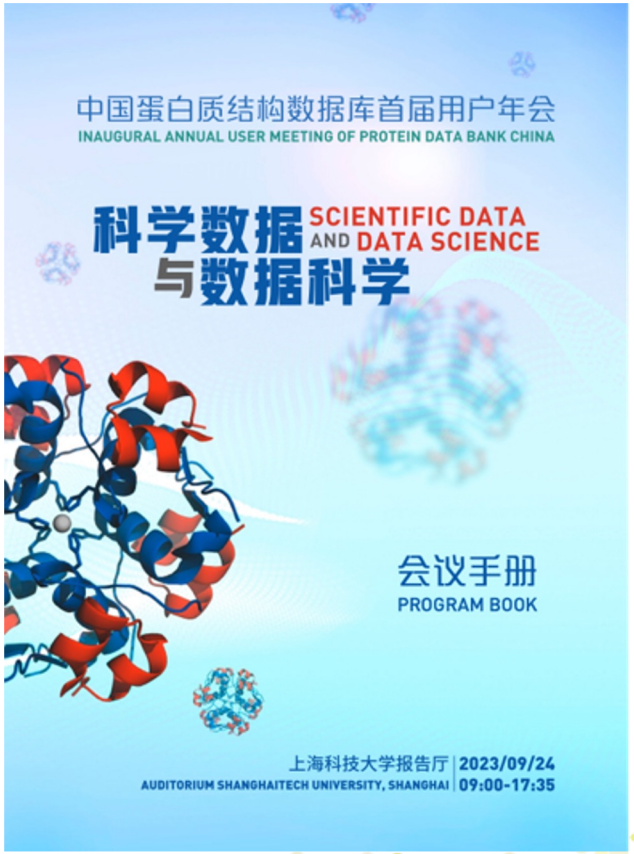
wwPDB Core Members and **wwPDB Associate Members** are fully committed to the **FAIR** Principles of Findability-Accessibility-Interoperability-Reusability, emblematic of responsible stewardship of public domain information.

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

•As of the end of 2022, the two PDBc biocurators had processed 297 new PDB depositions, representing ~10% of the 3D biostructure depositions from Mainland China for the year (~7% of all Asian depositions). During 2023, it is expected that PDBc will process most, possibly all, depositions to all three wwPDB Core Archives from Mainland China.

•Announcement of the launch of PDB China as an Associate Member of the wwPDB partnership was published in Acta Cryst D72, 792-795 (2023).

•Inaugural annual user meeting of Protein Data Bank China was held in Shanghai on Sep 24th, 2023.



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) was used for hiring PDBi staff, currently working at IISc, Bangalore & NII, New Delhi. NSM (India) has extended funding for salaries of PDBi staff till March 2024. Efforts are on to secure funding from DBT after 2024 for 5 years.
- 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.
- PDBi PIs have made arrangements to utilize servers/storage of the Indian Biological Data Center (IBDC), a facility funded by DBT, India. A prototype data-out server (http://pdbi.nii.ac.in) with very basic search interface has been set up by PDBi staff.
- After training of 3 PDBi staff (data curators) by PDBe at EBI, one of the PDBi staff (Dr. Vetriselvi Rangannan) has annotated 573 (357 X-ray + 216 EM) structures for PDBe in the span of 1 year while working remotely from India.
- PDBi PIs have approached two of the funding agencies (NSM & DBT) for support beyond March 2024. National Advisory Committee for PDB-India will be constituted in consultation with NSM & DBT after sanction is received for funding.

Outreach Update

Genji Kurisu, PDBj

wwPDB Outreach

- wwPDB collaboration celebrated 20th anniversary
- IUCr exhibit stand
- wwPDB Foundation sponsored poster prizes
  - IUCr
  - American Society for Biochemistry and Molecular Biology
  - Protein Society
- News and social media posts
- ORCiD Tutorial (PDF)

PDBe Outreach

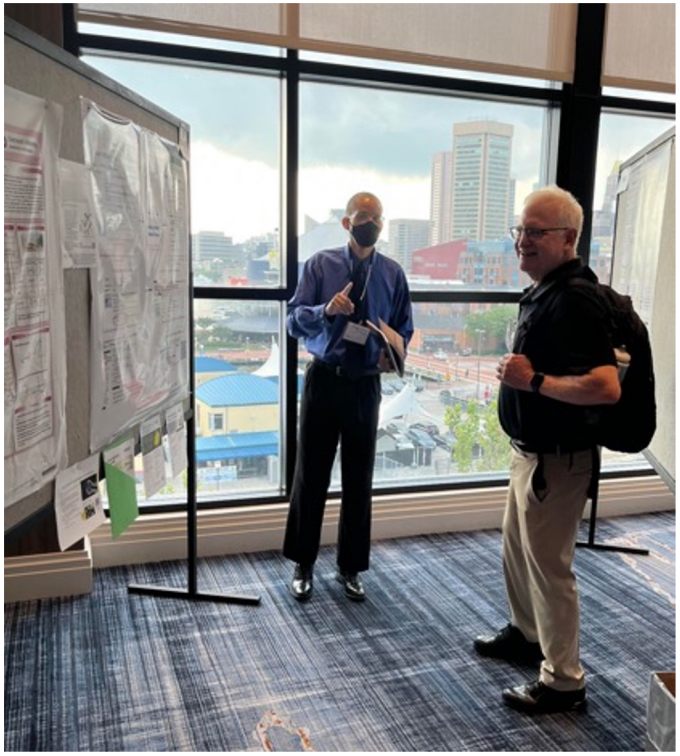
- PDBe workshops with a specific focus on data deposition and access:
  - Hot Topics in Contemporary Crystallography (HTCC5) - Dubrovnik, Croatia (April 16-19 2023)
  - Crystallography Course - Oulu, Finland (virtual, May 29-June 2, 2023)
  - European Crystallographic Meeting 8 - Berlin, Germany (June 18-24, 2023)
  - CCP4 Summer School - York, UK (Aug 23-25 2023)

PDBj Outreach

- PDBj Talk/AsCA: October 30-November 2
- PDBj Talk/Biophysical Society of Taiwan: May 17-20
- PDBj Talk/PSSJ2023: July 5-7
- PDBj booth: July 8 OsakaU@Expocity2023
- PDBJ Talk/IIBMP2023: September 8
- PDBj Talk&booth/CrSJ2023: October 26-28
- PDBj Talk/Biophysical Society of Japan: November 14-16

RCSB PDB Outreach

- Virtual Crash Course *Use PDB data to their full extent: Understanding PDBx/mmCIF*
- American Crystallographic Association (ACA)
  - *Growing the PDB Core Archive Using ORCiD Login for Depositing 3D Biostructures with OneDep* (poster)
  - *Growth of the PDB Archive Requires Transition to PDBx/mmCIF Format Files* (talk)
  - *Validation and Quality Assessment for Small-Molecule Ligands in the Protein Data Bank* (talk)
- Organometallic Chemistry Gordon Research Conference
  - *Chemical and Data Representation of Metalloproteins in the Protein Data Bank* (poster)
- CCP4/APS Crystallography School
  - *wwPDB OneDep Tools for Deposition and Validation* (talk)
- International Biocuration Conference
  - *wwPDB Biocuration: Supporting Advances in Science and Technology* (talk)



Poster presentation at ACA July 2023

EMDB Outreach

- EMDB Talk/Cold-Spring-Harbour Cryo-EM Course: March 30-April 1
- EMDB Talk/NCCAT Tomography Course: April 13
- EMDB Talk and Poster/CCPEM April 25 - 27
- EMDB Talk/CryoNET Single Particle Course: June 4-6
- EMDB Poster/3DEM GRC: June 11-16
- EMDB Talks/Sao Paulo Cryo-EM School July 16 - 22
- EMDB Talk/EMBO Image Processing for cryo-EM Course September 8

BMRB Outreach

- Biophysical Society, Booth, February 18-22, 2023
- ICBEB Course “NMR for combating diseases: from cancer to SARS-CoV-2”, Florence, Italy, March 27-31, 2023
- Experimental NMR Conference, Poster and Booth, April 16-20, 2023
- University of Georgia CCRC Metabolomics Data Workshop, May 23- 26, 2023
- Metabolomics Society of North America, Workshop and Poster, June 18-22, 2023
- Gordon Research Conference on Computational Aspects of Biomolecular NMR, Poster, June 25-30, 2023

OneDep Update

Stephen K. Burley, RCSB PDB

wwPDB Collaboration Resources November 2022-October 2023

wwPDB Partner	Software Development	Production Maintenance/ Project Management	Requirements Setting/ Testing	Core Archive Keeping*	Outreach	Biocuration/ Remediation	Total FTE Commitments
RCSB PDB	2.3**	1.3	0.35/0.35	2.0	0.3	6.2	12.8
PDBe	2.7**	0.6	0.2/0.2	-	0.2	5.5#	9.4
PDBj	1.1	0.5	0.2/0.2	-	0.1	4.6	6.7
BMRB	0.85	0.1	0.1/0.1	0.15	-	0.6	1.9
EMDB	2.1	0.65	0.25	0.15	-	0.9	4.05
Total wwPDB	9.15	3.05	1.95	2.65	0.6	17.4	34.8

*RCSB PDB; EMDB; BMRB

**including additional resource from BBSRC/NSF joint grant from PDBe and RCSB PDB and 0.8 FTE on PTM project from PDBe

Including 2.0 FTE PDB-India annotators in training

OneDep 2022/2023 Progress *versus* Goals I

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

Ref. Appendix No. 2

Delivered,
To be delivered,
Delayed,
Project change
(on hold)

	Projects	Timeline			
		2022	2023		
		Q4	Q1	Q2	Q3
1. Validation	1.1 Improve chemical shift validation code				
	1.2 EDS software upgrade				
	1.3 MolProbity software upgrade				
	1.4 Provide outlier detection and quality slider for 3DEM				
	1.5 Remove PDB format dependency in NMRClust software				
	1.6 Improve validation performance via parallel calculations				
2. Backend Stabilization	2.1 Support CCD ID extension				
	2.2 Improve process of validation report generation at release time				
	2.3 OS system upgrade at RCSB PDB				
	2.4 Upgrade 3rd party software packages/libraries				
	2.5 Migrate legacy EMDB entries into OneDep				
3. Public facing (OneDep or wwPDB.ORG)	3.1 Improve collection of starting model				
	3.2 Support mmCIF file upload of EM metadata for map-only entries				
	3.3 Provide ORCiD login and deposition API				
	3.4 Enable author-annotated assembly				
	3.5 Improve file re-upload process and logging via API				
	3.6 Improve 3DEM composite map deposition				
	3.7 Support more automated process on re-submission				
	3.8 Enable geographic redirection to PDB China				
	3.9 Provide a status review page on author-provided metadata items				
	3.10 Provide EM auxiliary data upload post release				
4. Biocuration	4.1 Provide a sequence summary page for review and actions				
	4.2 Make use of depositor-provided UNP ID				
	4.3 Make use of depositor-provided ligand restraints				
	4.4 Enable simultaneous editing of multiple related entries				
	4.5 Enable more automated process on re-submission				
5. Archive Improvements	5.1 Remediate X-ray entries with nonstandard crystal frame				
	5.2 Regenerate method specific validation reports (PDB)				
	5.3 NMR data remediation (PDB & BMRB archives)				
	5.4 Create next generation (NextGen) PDB archive				
	5.5 Distribute mmCIF header files at EMDB archive				
	5.6 PTM remediation (PDB archive)				
	5.7 Ongoing PDBx/mmCIF dictionary development (All archives)				

OneDep 2022/2023 Progress *versus* Goals II

Components	Major Projects to be Completed	Primary resource
Validation	**Upgrade 3rd party EDS software**	**PDBe**
	Refactor NMR chemical shifts validation software	BMRB
	Improve performance on validation calculations within OneDep	RCSB PDB
	Upgrade 3rd party MolProbity software	**PDBj**
	Enhance 3DEM map validation: False color images and symmetry analysis	EMDB
	Provide outlier detection and quality slider for 3DEM map-model fitness	**EMDB w/Greg&Chenghua**
Deposition	Improve composite map support for 3DEM	EMDB
	Support metadata upload in mmCIF for map-only 3DEM depositions	EMDB
	Improve EM metadata checking/validation	EMDB
	Enable author-annotated assembly*,#	**PDBe**
	Enable geographic redirection to PDBc*	**PDBj**
	Provide a Review page on the summary of author provided metadata at DepUI*	**PDBe**
	Improve collection of starting models	PDBe/EMDB
	Improve file re-upload process using file access API	**PDBe**

X-ray assembly remained

Bold: re-forecasted to 2023-2024
* put on hold as stretch goals

OneDep 2022/2023 Progress *versus* Goals III

Ad hoc projects:

- Developed data model to support MicroED depositions
- Set up OneDepPDBc instance at Osaka and provide biocuration training
- Created 5-years plan on transitioning to extended PDB IDs
- Provided intra-molecular connectivity to support retirement of PDB format files

Components	Major Projects to be Completed	Primary resource
Infrastructure	Centos upgrade	RCSB PDB
	Support CCD ID extension	ALL sites
	Migrate EMDB legacy entries into OneDep	**EMDB**
	Session login via ORCiD	PDBe
	Upgrade OneDep Third party packages/libraries	EMDB/RCSB PDB/PDBe
	Increase efficiency on validation report generation at release time	RCSB PDB
Core Archives	Remediate NMR restraint files for BMRB and PDB archives	BMRBj
	Archival re-calculation of validation reports for PDB and EMDB archives	**PDBe**
	Remediate entries with non-crystal frame (PDB)	RCSB PDB
	Create Next Generation PDB Archive (NextGen)	RCSB PDB/PDBe
	Distribute mmCIF formatted header files to the EMDB archive	EMDB
	PTM remediation (PDB)	PDBe

Bold: re-forecasted to 2023-2024

wwPDB Collaboration Resource November 2023-October 2024

wwPDB Partner	Software Development	Production Maintenance/ Project Management	Requirements Setting/ Testing	Core Archive Keeping*	Outreach	Biocuration/ Remediation	Total FTE Commitments
RCSB PDB	2.3	1.4	0.2/0.2	2.0	0.3	6.0	12.4
PDBe	0.8**+0.8#	0.6	0.1/0.1**	-	0.1	4.2##	6.7
PDBj	1.3	0.6	0.2/0.2	-	0.1	3.2	5.6
BMRB	0.85	0.1	0.1/0.1	0.15	-	0.6	1.9
EMDB	1.825	0.4	0.15/0.05	0.4	0.15	0.625	3.6
Total wwPDB	7.875	3.1	1.4	2.55	0.65	14.625	30.20

*RCSB PDB; EMDB; BMRB

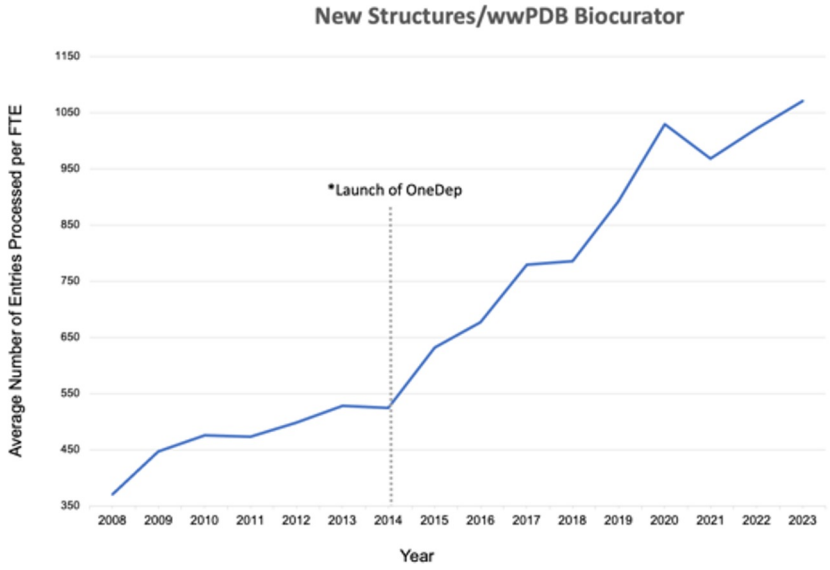
** If PDBe receives new BBSRC/NSF grant, will have additional 0.8 and 0.1/0.1 FTE

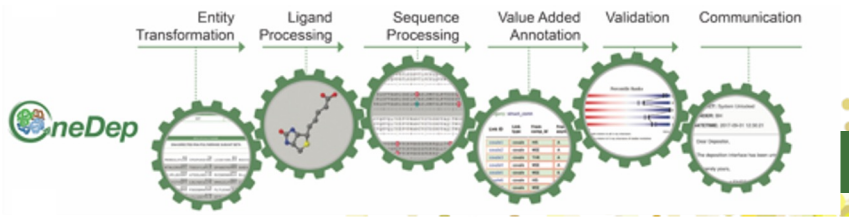
BBSRC/BBR grant for the PTM remediation project started in April 2022

Including 0.2 FTE PDB-India annotator in training

wwPDB Biocurator Productivity

- 16,344 depositions in 2022
- Improved efficiency for biocurating incoming depositions
  - Automated sequence annotation when depositor-provided UNP reference 100% matches to BLAST results
  - Auto-approval of processed entries when depositors do not respond and there are no major issues
  - Auto-completion of Annotation Module for map-only EM entries
  - Automated ligand assignment for well-known close matched ligand pairs
- Use of refinement software-provided ligand restraints for creating new chemical component definition





OneDep 2023/2024 Goal Setting I

Components	Major Projects to be Completed	Primary resource
Validation	Ongoing system improvements & bug fixing	All sites
	Upgrade 3rd party EDS software	**PDBe**
	Modularize validation software for easy maintenance and parallel execution	RCSB PDB
	Upgrade 3rd party MolProbity software	**PDBj**
	Improve EM map-model fitting in the validation with overall Qscore slider	**EMDB**
Deposition	Ongoing system improvements & bug fixing	PDBe/PDBj/EMDB
	Improve 3DEM composite map deposition	EMDB
	Improve EM cross-file checks at file upload	EMDB
	Show EM maps with 3D Mol* at the depUI	EMDB
	Capture metadata for MicroED	PDBe
	Improve file re-upload process with better tracking	**PDBe**

Bold: re-forecasted from 2022-2023

OneDep 2023/2024 Goal Setting II

Components	Major Projects to be Completed	Primary resource
Biocuration	Ongoing system improvements & bug fixing	RCSB PDB
	Annotation improvement- automation	RCSB PDB
	Annotation improvement- workflow for large structures	RCSB PDB
	Annotation improvement - sequence search	RCSB PDB
	Provide revision history for EMDB files	RCSB PDB
Infrastructure	Third party S/W updates - OpenBabel	RCSB PDB
	Third party S/W updates - Phenix	RCSB PDB
	Migrate EMDB legacy entries into OneDep	**EMDB**
	wwPDB data transfer changes due to PDBe and EMDB moving to Globus service	RCSB PDB/PDBj
	Remove dependency on email handler at RCSB PDB	RCSB PDB

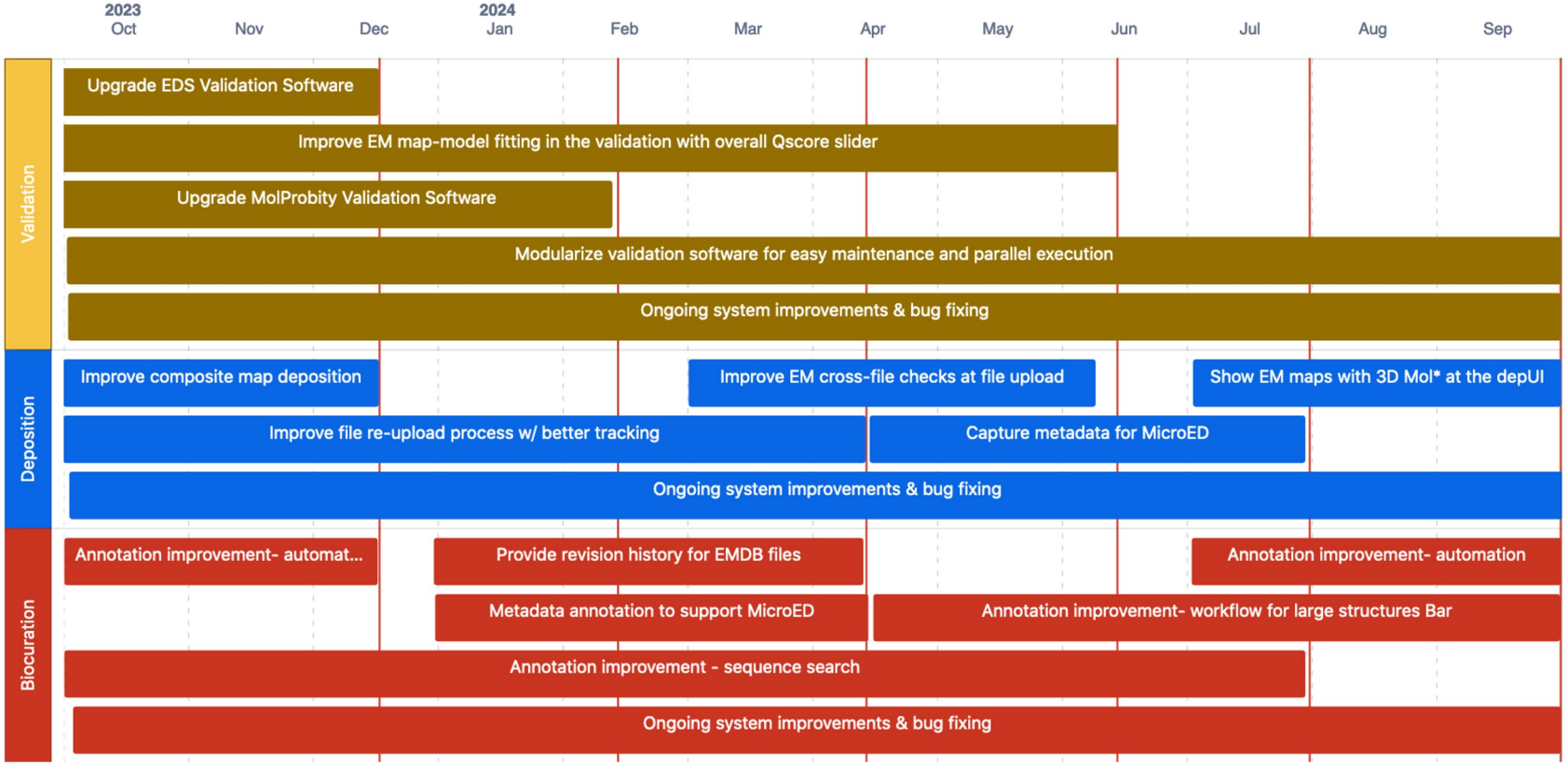
Bold: re-forecasted from 2022-2023

OneDep 2023/2024 Goal Setting III

Components	Major Projects to be Completed	Primary resource
Core Archives	mmCIF extension to support MicroED	RCSB PDB
	Archival re-calculation of validation reports for PDB and EMDB archives	**PDBe**
	Metadata remediation at EMDB archive	EMDB
	NMR data remediation with peak lists at PDB and BMRB archives	PDBj
	Provide holdings files at EMDB Archive	EMDB
	PDB archive-wide remediation	RCSB PDB/PDBe/PDBj
	Unifying PDB-dev into PDB Archive	RCSB PDB
	PTM remediation (PDB)	PDBe
	Metalloprotein remediation planning	RCSB PDB
	Ongoing mmCIF dictionary development and maintenance	RCSB PDB
	Community outreach activities to promote mmCIF and transition to extended PDB ID	All sites

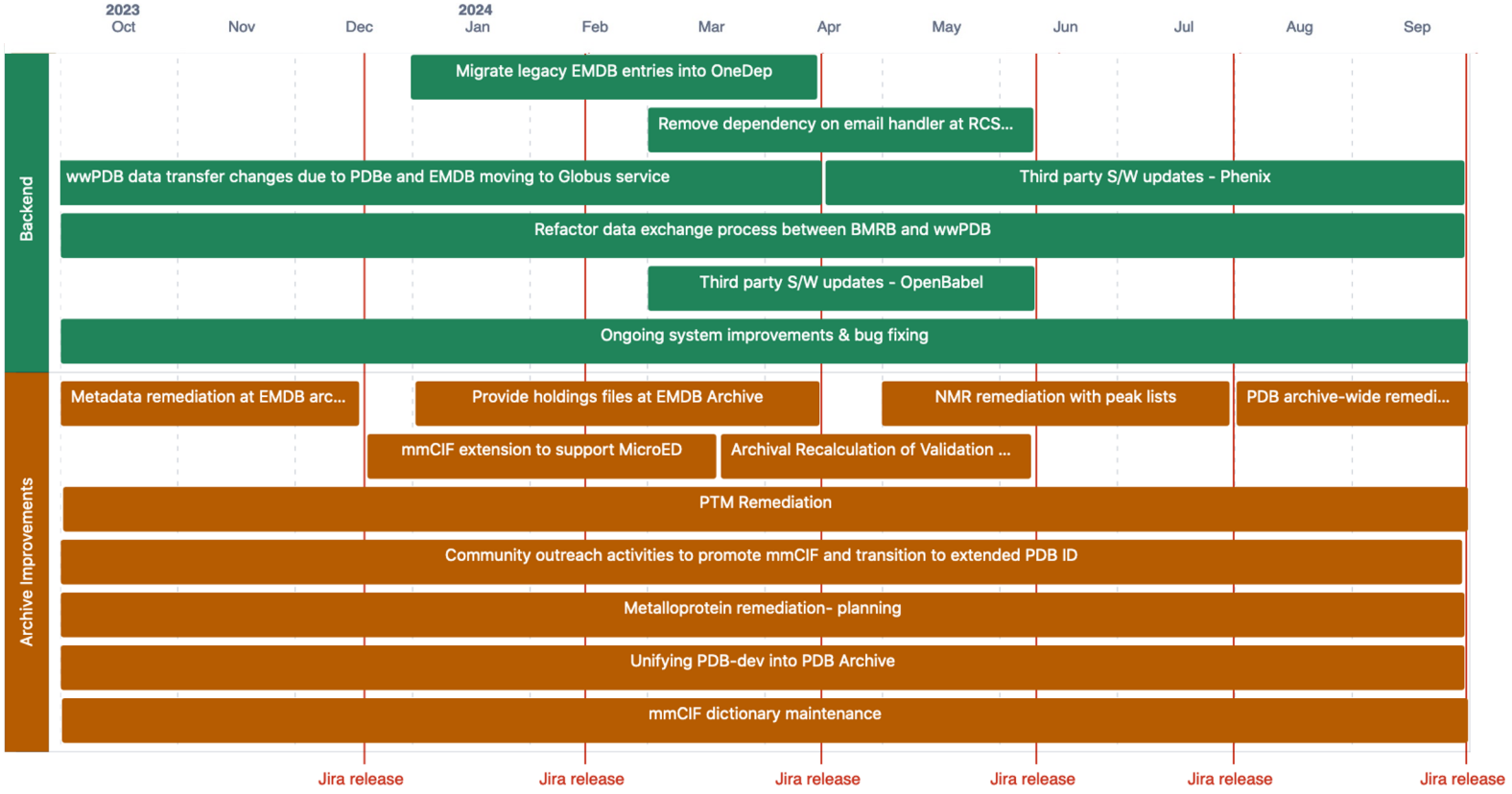
Bold: re-forecasted from 2022-2023

OneDep 2023/2024 Roadmap I



- will be further adjusted based on requirement setting

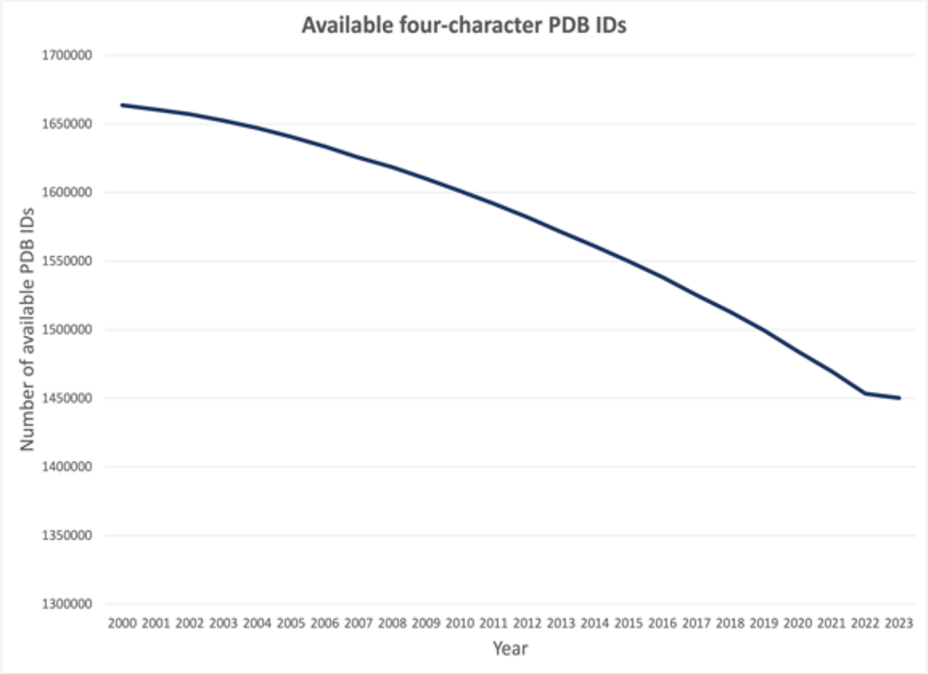
OneDep 2023/2024 Roadmap II



- will be further adjusted based on requirement setting

Five-year Plan for Transitioning to New PDB ID Format

- Anticipating change required in the 6 years
- Community outreach on transitioning to new PDB ID and PDBx/mmCIF formats
  - 2023: provide an FAQ page on new PDB IDs
  - 2024: create a PDBx/mmCIF documentation for user training
  - 2024-2025: update archival files to include new PDB IDs
  - 2026: create a PDB “beta” archive with new PDB IDs in file naming
  - 2028: modify OneDep software to issue extended PDB IDs with new file naming
  - 2029: switch PDB beta archive to PDB archive when four-character PDB IDs are consumed
  - Ongoing: provide PDBx/mmCIF training courses at regional data centers
  - Ongoing: advertise extended PDB ID format and transition plan at professional society meetings and social media



PDB Archive Update

Stephen K. Burley, RCSB PDB

Current PDB Archive Status

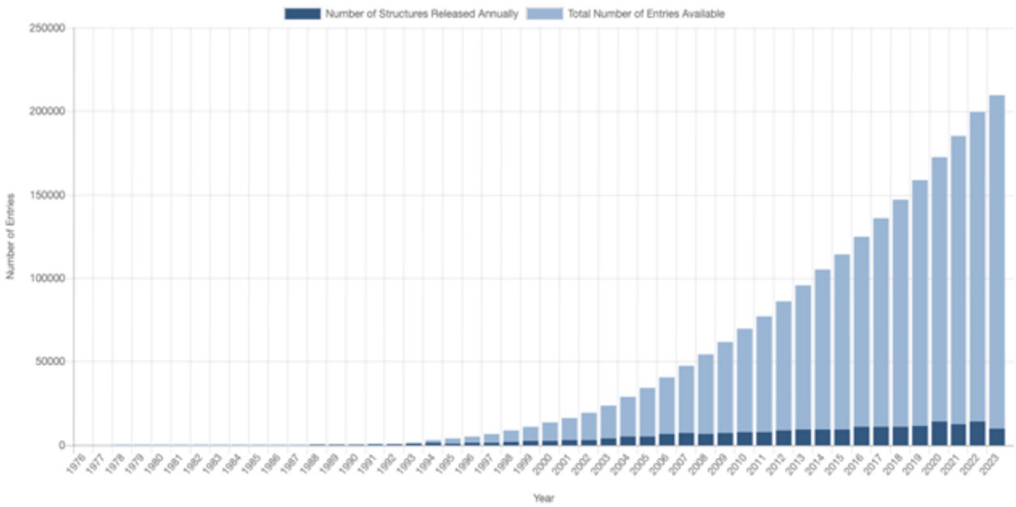
- Total Entries=209,957
  (as of 9/20/2023)

- Core Archive Storage
  - OneDep Sessions: ~67.1 TB
  - ftp(legacy + versioned): ~1.45TB
  - EMDB ftp: ~12 TB

- CoreTrustSeal certification due for renewal in April 2024 (CoreTrustSeal.org)

- PDB archival data now being housed and delivered by Amazon Web Services (AWS) with no storage or egress fees

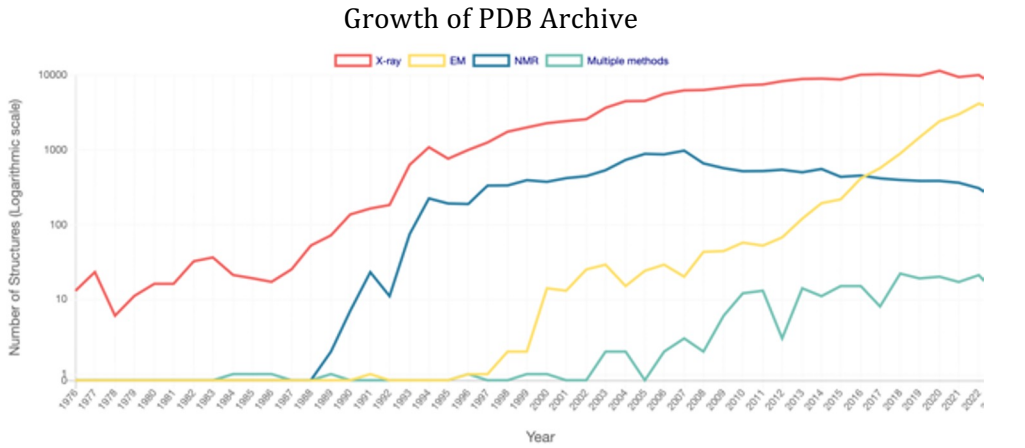
- NextGen archive now serving enriched annotation in the atomic coordinate files (https://files-nextgen.wwpdb.org)

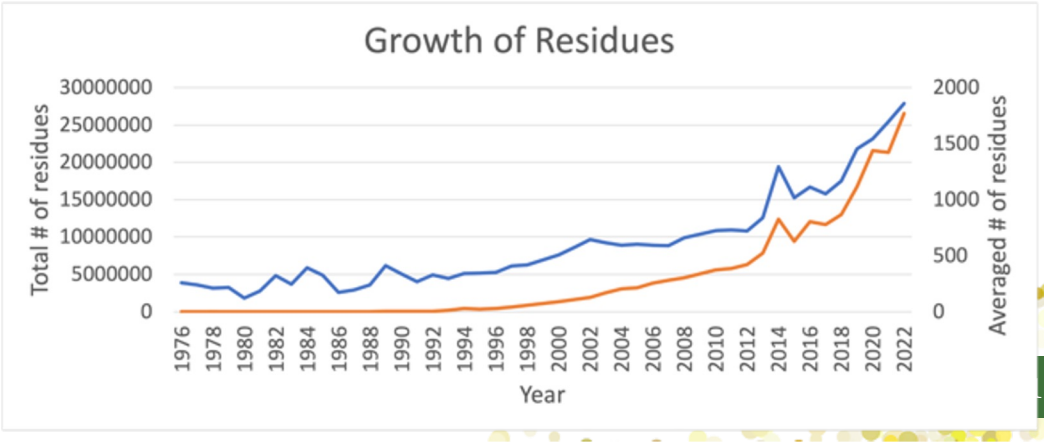


Reached 200,000 structures in January 2023!

PDB Archive Growth in 2022

- Year-end holdings 199,653
- 14,277 new entries released
  - ~13.5% increase *versus* 2021
  - ~7.7% growth in the archive
- Archival entries growing in both size and complexity
- Record 4,111 new 3DEM entries released
  - ~39% increase *versus* 2021!

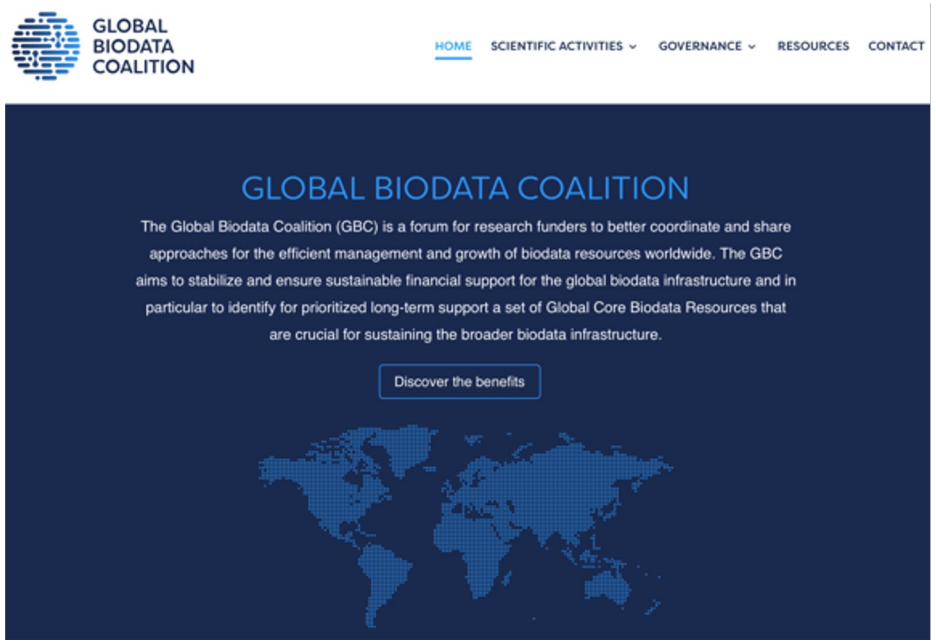




Global Biodata Coalition (GBC)

- PDB has been formally recognized as a Global Core Biodata Resource
  - Went through a rigorous two-stage process
  - Assessed by a panel of more than 50 independent expert reviewers

- wwPDB plans to seek formal recognition for EMDB and BMRB in response to subsequent GBC calls

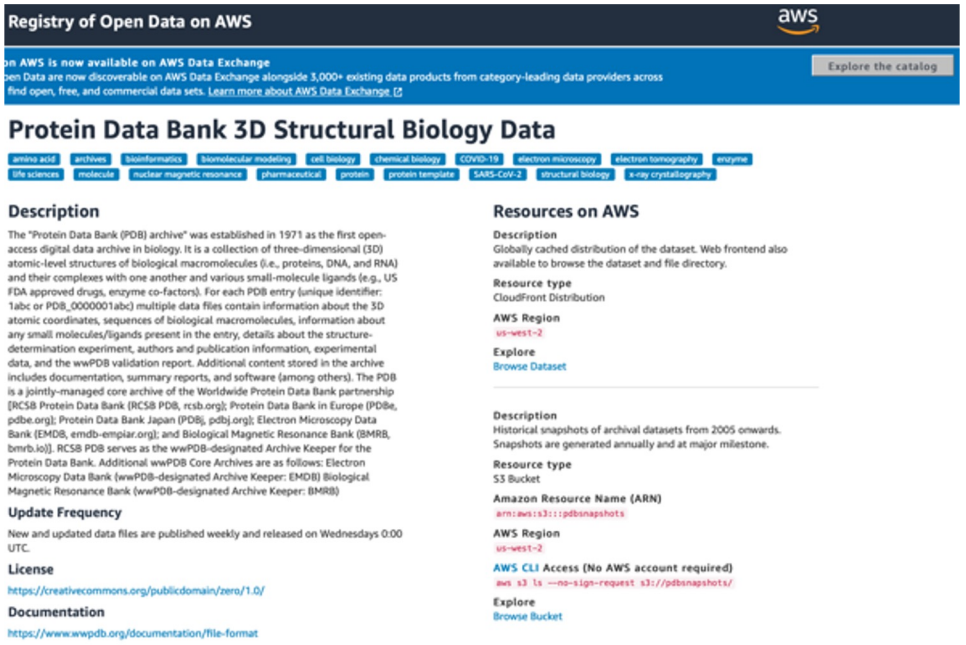


PDB Data Delivery by AWS

- Amazon Web Services (AWS) Open Data Sponsorship Program now housing and delivering PDB data
  - No storage fees charged to RCSB PDB
  - No egress fees charged to PDB users

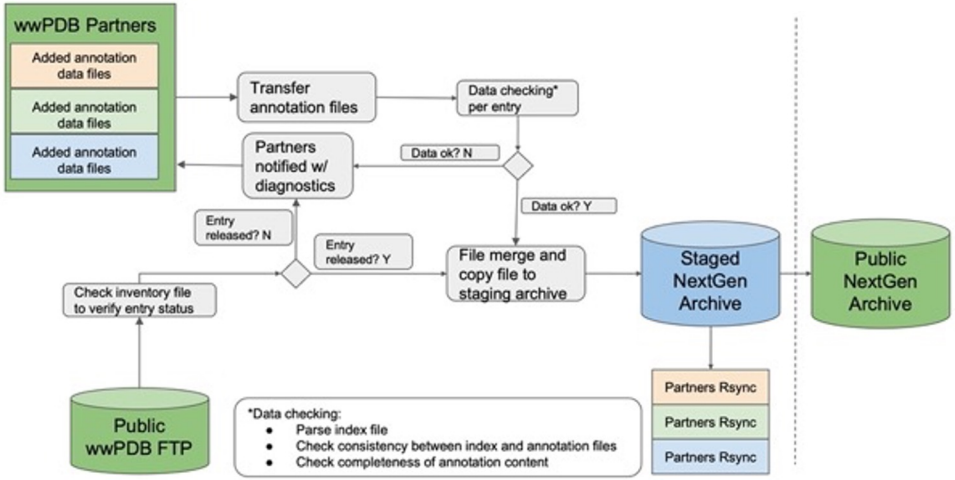
- Current AWS holdings include:
  - Annual PDB Archive Snapshots
  - Current PDB FTP Archive (updated weekly)

- AWS can deliver PDB data faster than RCSB PDB, PDBe, or PDBj!
  - https://s3.rcsb.org



PDB NextGen Archive

- NextGen Archive V1.0 launched in February 2023
  - https://files-nextgen.wwpdb.org, rsync://rsync-nextgen.wwpdb.org
- Provides enriched annotation from external database resources in the PDBx/mmCIF files
  - Sequence annotation such as UniProt, CATH, SCOP2 and Pfam from SIFTS
  - Intramolecular connectivity for each residue (atom pairs, bond order, aromatic flag, and stereochemistry)
- Automated monthly update process
  - Updated every month on the 1st Wednesday at 00:00 UTC
- Work is described in the manuscript which has been submitted to peer-reviewed journal and bioRxiv

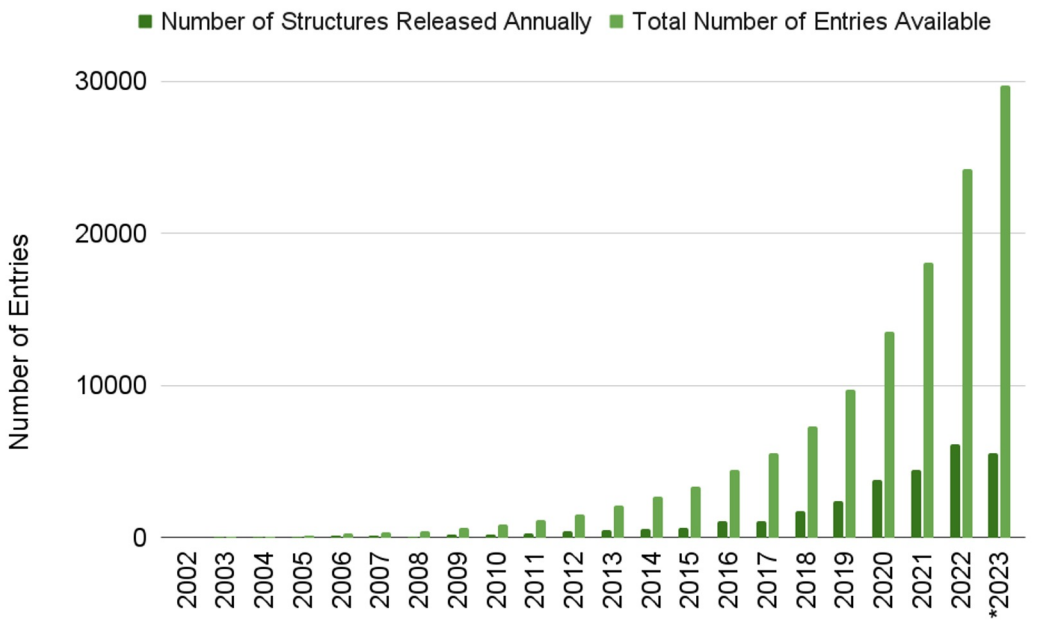


EMDB Archive Update

Kyle Morris, EMDB

Current EMDB Archive Status

- Total Entries = 28,685
  - (as of 2023/09/20)
- EMDB Archive Storage: 13 TB
  - (as of 2023/09/20)
- 30,000 entries achieved in early October 2023

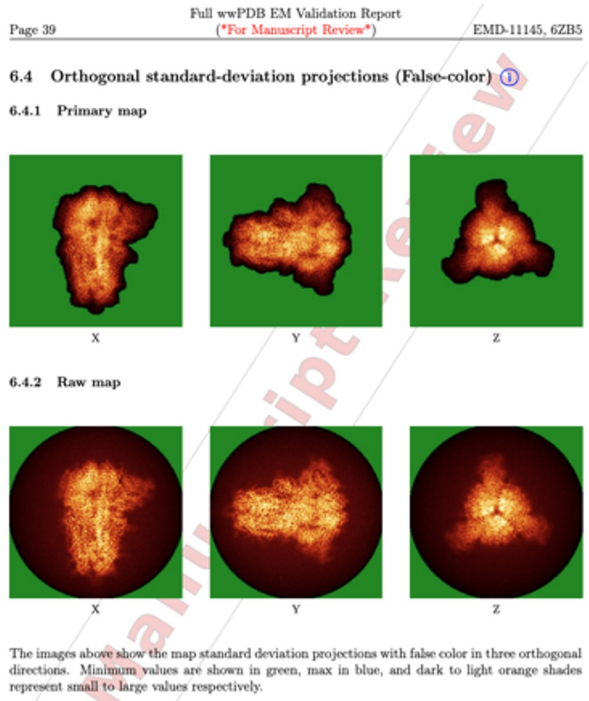


2022-2023 Achievements

- Improvements to 3DEM validation reports
- Improvements to 3DEM deposition
- Improvements to 3DEM annotation
- 30,000th EMDB entry released

Improvements to 3DEM Validation Reports

- Various 2D slices are being trialed on EMDB's Tier 1 validation pages in an effort to improve qualitative validation of structures
- False colour images provide a quick assessment for the presence and shape of masks in an EM volume
  - These have now been added to the wwPDB validation reports



Improvements to 3DEM Deposition

- Improved composite map support
  - Labelling of composite maps and the related entries now supported
- Additional metadata cross-checks implemented in the deposition interface
- The first of several planned inter-file checks has been implemented during the file upload process
  - An entry will now be halted if any atoms from a coordinate model are outside the bounding box of the EM map.
- Uploaded entry image now shown to depositor allowing it to be checked before submission
- metadata mmCIF file can be provided during EM deposition with or without a model present
- Experimenting with deep-learning to detect the EM method from the uploaded map

Improvements to 3DEM Annotation

- Improved efficiency of map-only entry annotation
- mmCIF formatted metadata files now available for EM entries in the archive
  - Currently available for entries released after 2023/10/04
  - Remediation planned to update older entries

Awaiting Community Feedback

- New Content Types
  - Planned and some implementation complete
  - Awaiting community feedback on names used for content types

Social Media

- Twitter/X
  - Group of EMDB and EMPIAR employees responsible for weekly tweets
    - 6200+ followers
    - Several tweets a week

- YouTube
  - Youtube channel hosting tutorials, talks and structure shorts
    - 3100+ views
    - 52 watch hours
    - 84 Subscribers

30,000 EMDB Entries

- Reached in early October 2023
- On track for number of EM releases to overtake X-ray in 2024
- 6,147 new entries release in 2022
- 4,973 new entries so far in 2023 (until late August)

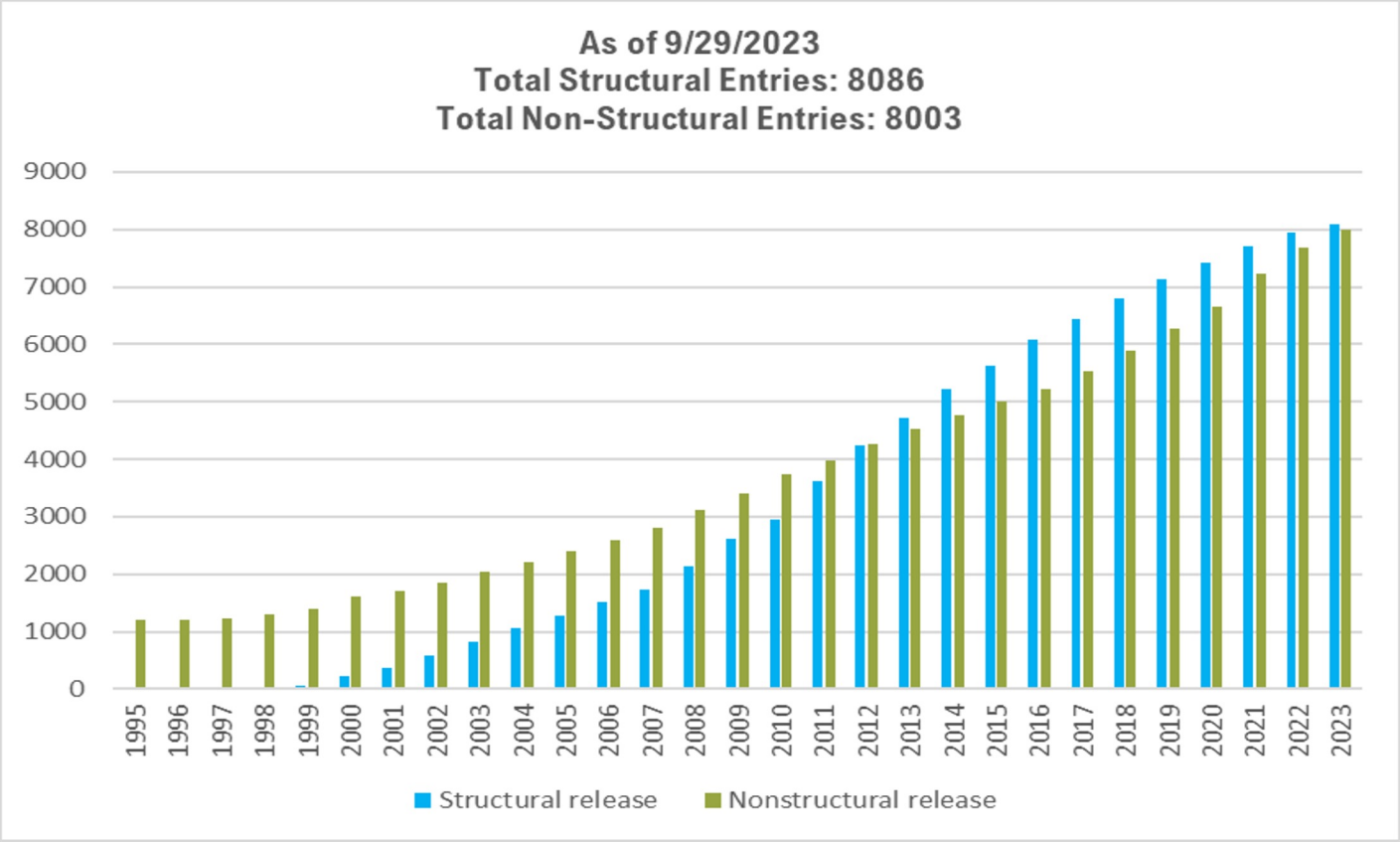
BMRB Archive Update

Jeffrey C. Hoch, BMRB

Developments since 2022 Meeting

- The chemical shift validation code has been refactored.
- Working with PDBj on restraints remediation project.
- Developed new tooling to facilitate deposition of large fragment screening data sets to BMRB small molecule database.
- Bug fixes and minor improvements to core BMRB software: PyNMR-STAR, BMRB-API, BMRBdep, and CS-Rosetta server.

BMRB Core Archive Growth



BMRB Core Archive Growth

Total Released Entries

Year	Yearly Totals					Cumulative Totals				
	Structural release	Nonstructural release	Original release	Total release	Withdrawn during year	First release	Total release	Structural release	Nonstructural release	Eventually withdrawn from this year
2012	631	294	925	1168	5	8510	10144	4240	4270	113
2013	468	248	716	1007	1	9226	11151	4708	4518	39
2014	522	242	764	1068	1	9990	12219	5230	4760	20
2015	402	234	636	696	0	10626	12915	5632	4994	75
2016	441	234	675	918	1	11301	13833	6073	5228	22
2017	377	307	684	839	5	11985	14672	6450	5535	70
2018	348	361	709	855	3	12694	15527	6798	5896	33
2019	323	375	698	968	4	13392	16495	7121	6271	37
2020	293	388	681	836	5	14073	17331	7414	6659	69
2021	280	573	853	1137	2	14926	18468	7694	7232	62
2022	243	460	703	1192	0	15629	19660	7937	7692	37
2023	149	311	460	684	0	16089	20344	8086	8003	59

BMRB Core Archive Growth

Internet Server Traffic (Website) – All Mirrors*

Year	Server requests	Page requests	File requests	Distinct hosts served	US hosts served	Total data transferred
2012	42,555,193	6,395,316	959,834	854,837	279,922	17.62 TB
2013	43,503,533	6,489,163	873,911	1,006,097	374,274	19.32 TB
2014	39,144,507	6,407,405	606,332	1,117,340	426,682	28.21 TB
2015	46,899,762	7,011,339	664,415	1,227,112	500,589	32.68 TB
2016	54,081,884	7,869,456	1,028,193	1,321,909	531,196	36.38 TB
2017	56,209,400	8,537,562	1,433,036	1,026,426	387,809	18.37 TB
2018	87,818,181	25,523,384	2,415,042	1,330,889	439,600	17.75 TB
2019	87,989,534	33,885,684	1,173,443	1,779,863	638,348	27.55 TB
2020	83,732,415	22,489,261	1,603,539	1,783,604	608,285	47.67 TB
2021	82,585,307	25,655,514	737,829	2,609,272	627,506	44.85 TB
2022	127,891,445	25,774,500	1,628,260	2,730,172	1,045,469	40.41 TB
2023	102,842,615	19,735,184	682,634	2,841,707	1,081,334	36.98 TB

- 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

Network for Advanced NMR (NAN): update

- In less than a full year of operation, NAN has harvested >10,000 data sets
- Two 1.1 GHz instruments in US are scheduled for delivery fall 2023.
- Experiment data harvesting capabilities are operational.
- Data browser functionalities are released to production.
- U Delaware (Tatyana Polenova, PI; Jeff Hoch, co-PI; Angela Gronenborn, co-PI) is invited for full proposal for a Midscale-RI2 proposal.
- Combined target between the 2 Midscale RI-2 projects is deploying 5 1.0/1.1 GHz NMR spectrometers, distributed geographically across the US, and connected to the NAN cyberinfrastructure for automatically harvesting experimental data.

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 and concluded in July 2023
  - Provides additional resources for OneDep development
  - Coordination with mmCIF Working Group and OneDep team

- Project 2 was started in December 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

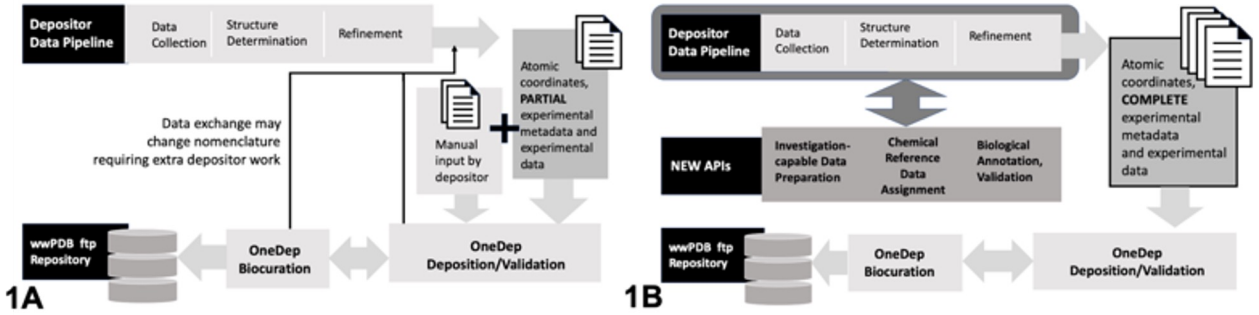
- SA1. To improve the fidelity and completeness of 3D structure data deposited into the PDB by
  - Harvesting data automatically from structure determination software packages; and
  - Streamlining the wwPDB data deposition, validation, and biocuration system.

- 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.

Project 1: 2021-23 (SA1)

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



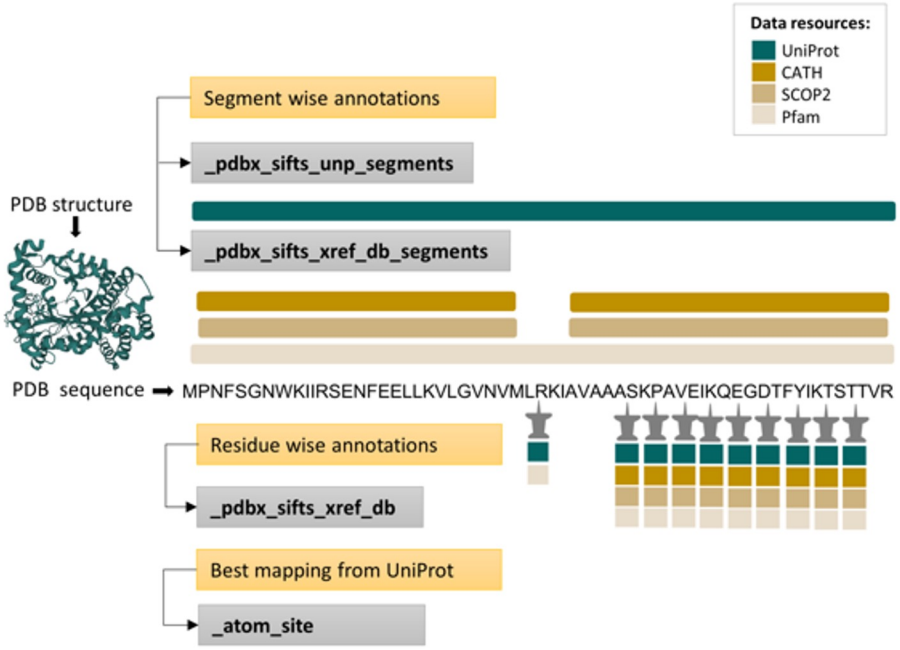
- Enabling automated creation of and upload to deposition sessions

Authentication → API based creation of deposition session → API based upload files to deposition session

  - Implemented deposition API via ORCiD login authentication
  - Held community hackathon workshop on API based deposition training
  - API based deposition currently under community testing and adoption

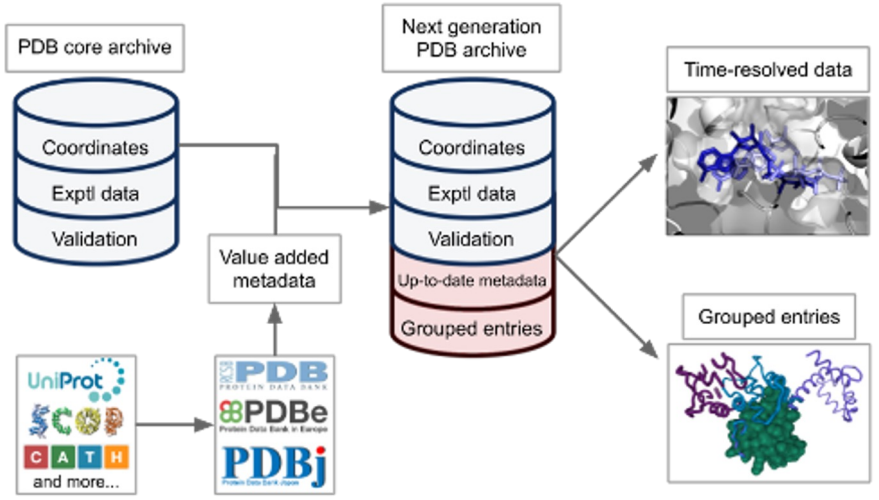
Project 1: 2021-23 (SA2)

- 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)



Project 1: 2021-23 (SA2)

- NextGen archive was launched in February 2023
  - https://files-nextgen.wwpdb.org, rsync://rsync-nextgen.wwpdb.org
- Incorporated enriched annotation
  - Sequence annotation such as UniProt, SCOP2 and Pfam from SIFTS
  - Intramolecular connectivity for each residue (atom pairs, bond order, aromatic flag, and stereochemistry)
- Automated monthly update process
  - Updated every month on the 1st Wednesday at 00:00 UTC
- Work is described in the manuscript which has been submitted to peer-reviewed journal and bioRxiv



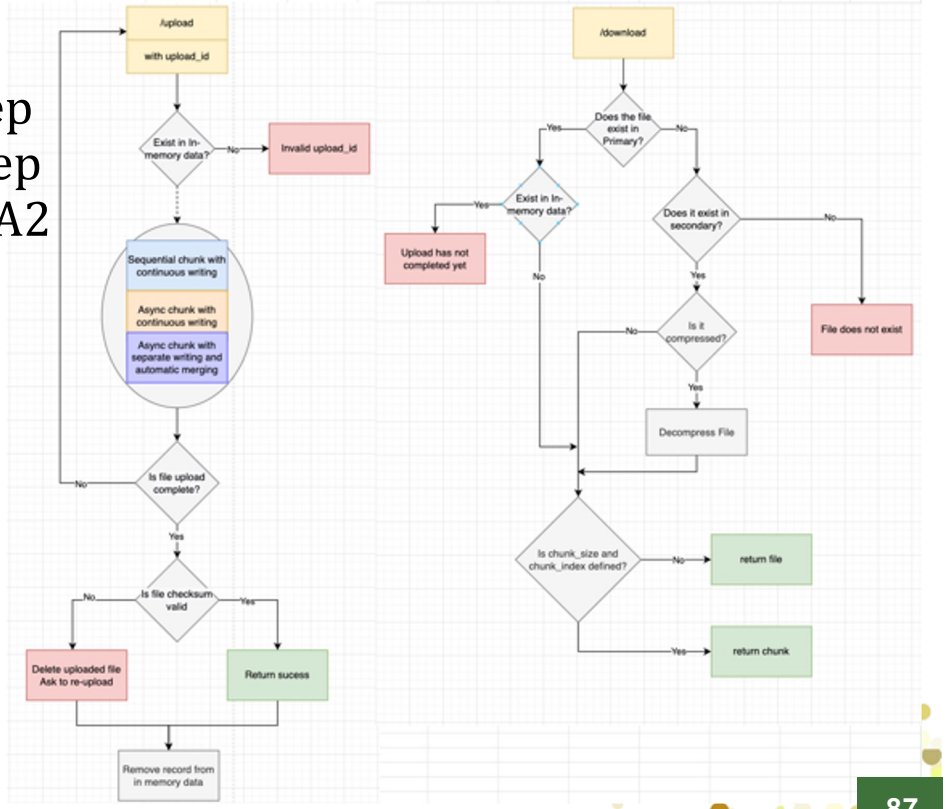
Project 1: Progress in 2021-22 (SA3)

- Installation procedure of wwPDB OneDep documented and simplified
  - from 133 manual steps to single installation script for PDBe environment
  - Continue improving installation script to support new partner with common steps and site specific setting
- 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: 2021-23 (SA3)

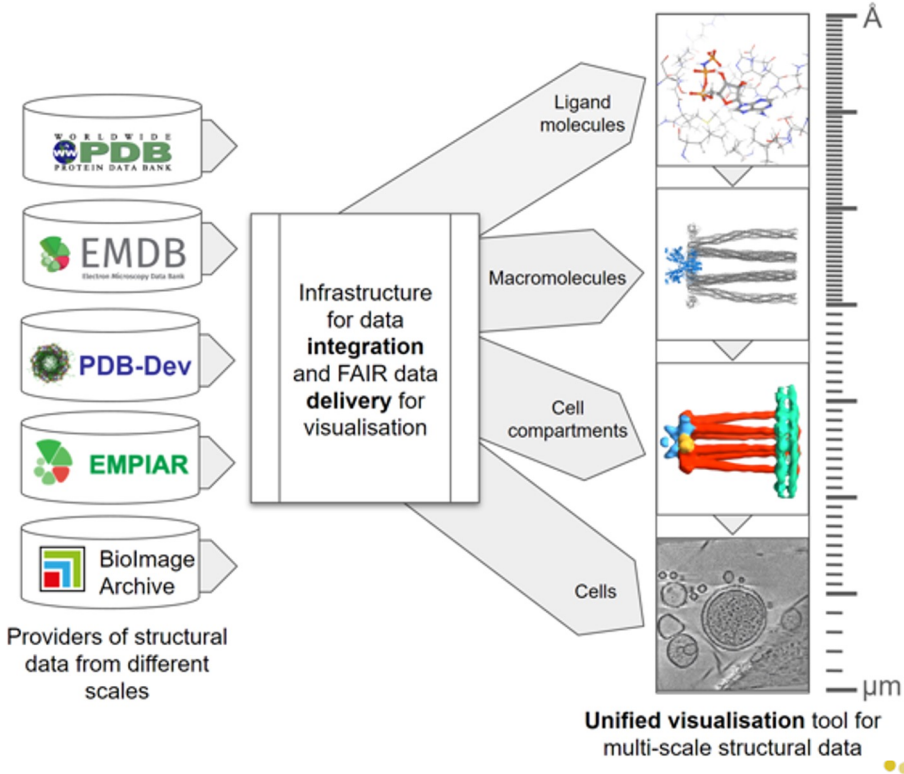
- File handling API to provide
  - Single access point for data in OneDep
  - Data exchange mechanism for OneDep
  - Groundwork for data exchange for SA2
  - Resumable upload functionality
  - Supporting asynchronous transfers
  - Handle file compression

- Would be rolled out initially on the deposition side.



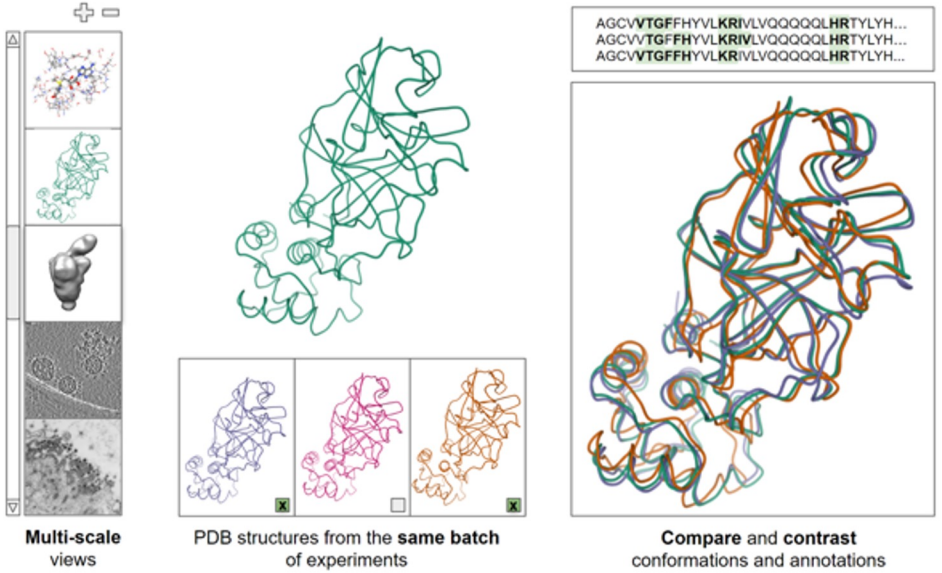
Project 2: Visualization Infrastructure

- 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; and
    c. Enabling comparisons and analyses of multiscale structures.



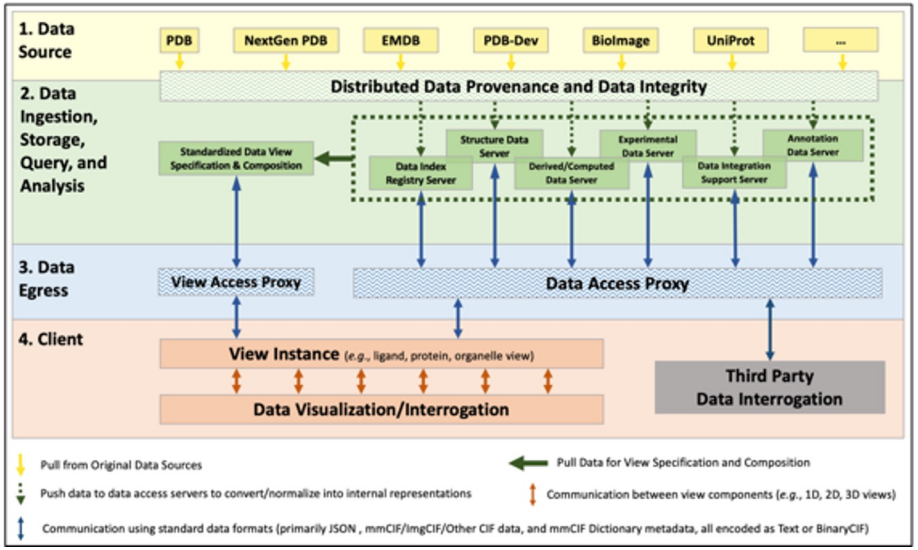
Project 2: Visualization Infrastructure

- 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.



Project 2: Visualization Infrastructure

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



Project 2: Progress in 2022-2023

- SA1 - Augment Mol-star for seamless operation across length scales ranging from atoms to cells with multiscale structure capability
    a. Data standards to be developed
        - Current Mol* uses BinaryCIF to transmit structure data, volume/density data is compressed using the same technology. Coarse-grained representations of multiscale structures are supported throughout Mol*.
    b. API to support multiscale data and multi-structure alignment
- SA2 - Extend 3D visualization beyond structure to include display of associated biological and functional annotations
    a. API to facilitate efficient mapping and data integration for chemical, polymer sequence, and 3D structure comparison.
    b. 2D topology viewer for proteins and RNA is in place at PDBe. Investigating its inclusion into the project.
- 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.
    a. Views specification and infrastructure design are being developed