The 54th Ann. Meeting of BSJ
Tsukuba, 25 November 2016

New Issues in the wwPDB and the PDBj

Haruki Nakamura
中村 春木

Institute for Protein Research
Osaka University
大阪大学蛋白質研究所

harukin@protein.osaka-u.ac.jp

http://pdbj.org/, http://wwpdb.org/

Kinjo et al. (2017) Nucl. Acids Res. doi: 10.1093/nar/gkw962.

Protein Data Bank Japan (PDBj): updated user interfaces, resource description framework, analysis tools for large structures

Akira R. Kinjo*, Gert-Jan Bekker, Hirofumi Suzuki, Yuko Tsuchiya, Takeshi Kawabata, Yasuyo Ikegawa and Haruki Nakamura

Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan

Received September 15, 2016; Revised October 06, 2016; Editorial Decision October 07, 2016; Accepted October 11, 2016

ABSTRACT

The Protein Data Bank Japan (PDBj, http://pdbj.org), a member of the worldwide Protein Data Bank (wwPDB), accepts and processes the deposited data of experimentally determined macromolecular structures. While maintaining the archive in collaboration with other wwPDB partners, PDBj also provides a wide range of services and tools for analyzing structures and functions of proteins. We herein outline the updated web user interfaces together with RESTful web services and the backend relational database that support the former. To enhance the interoperability of the PDB data, we have previously developed PDB/RDF, PDB data in the Resource Description Framework (RDF) format, which is now a wwPDB standard called wwPDB/RDF. We have enhanced the connectivity of the wwPDB/RDF data by incorporating various external data resources. Services for searching, comparing and analyzing the ever-increasing large structures determined by hybrid methods are also described.

major changes regarding user interfaces and analysis tools as well as additional data provided. The previously described Resource Description Framework (RDF) format, PDB/RDF, is now one of the wwPDB standard formats called wwPDB/RDF and is enhanced with supplementary information in order to connect PDB data with other biological data resources.

USER INTERFACES

User interfaces include interactive (and graphical) web interfaces for humans and RESTful web services for computer programs. We also expose our backend database in the forms of web services or dump files for enabling very complex queries. These are described in turn.

Web interface

The web interface of PDBj was updated to provide a uniform integrated interface for the available services as well as to provide a scalable interface for devices ranging from smartphones to workstations. This update incorporates several innovative/renovative features as described below.

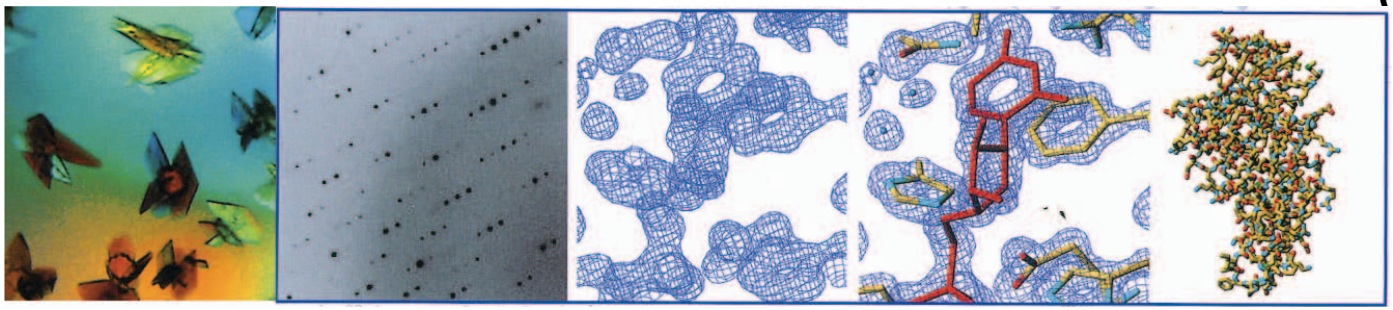
We have implemented various functionalities to ease

PDB: Protein Data Bank since 1971

Collects 3D structural information of biomolecules

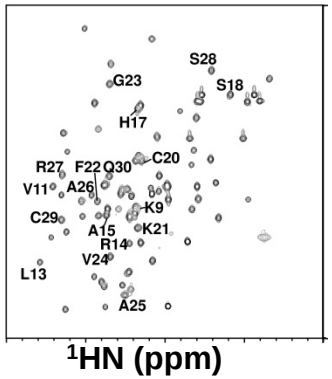
- **X-ray Crystallography**

XFEL & Synchrotron (SACLA & SPring-8)



Crystal ⇒ X-ray diffraction ⇒ Electron density map ⇒ Atomic model

- **Nuclear Magnetic Resonance (NMR)**



NMR spectrometer with super-conducting magnet (@IPR)

CryoEM (@IPR)

- **Cryo-Electron Microscopy**

Protein Data Bank: PDB

PDB History

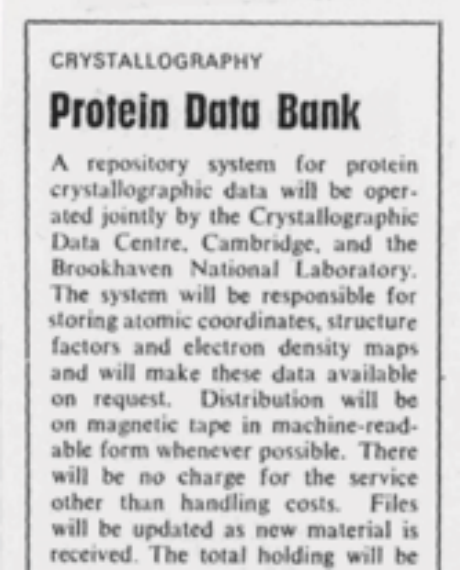
1960s: Protein crystallography started

- **Myoglobin, Hemoglobin, Lysozyme, etc**

1970s

- **PDB started with 7 data (October, 1971) at Brookhaven National Laboratory, USA**
- **PR-Osaka Univ. determined cytochrome *c* and deposited it to PDB**
- **Magnetic tapes were distributed**

Max Perutz & John Kendrew (Nobel Prize in Chemistry, 1962)



CRYSTALLOGRAPHY

Protein Data Bank

A repository system for protein crystallographic data will be operated jointly by the Crystallographic Data Centre, Cambridge, and the Brookhaven National Laboratory. The system will be responsible for storing atomic coordinates, structure factors and electron density maps and will make these data available on request. Distribution will be on magnetic tape in machine-readable form whenever possible. There will be no charge for the service other than handling costs. Files will be updated as new material is received. The total holding will be

PDBj Protein Data Bank Japan
WORLDWIDE PDB PROTEIN DATA BANK

Protein Data Bank: PDB

1980s : Rapid data increase (IUCr recommendation "Data deposition to PDB is mandatory for paper submission to journals")

1990s : RCSB-PDB started in USA

2000s : Foundation of wwPDB in 2003

Further many data by structural genomics

2010s : New methodologies are applied

More than 124,000 data

Towards "Big Data"

PDBj Protein Data Bank Japan

WORLDWIDE PDB PROTEIN DATA BANK

Protein Data Bank Japan

http://pdbj.org/

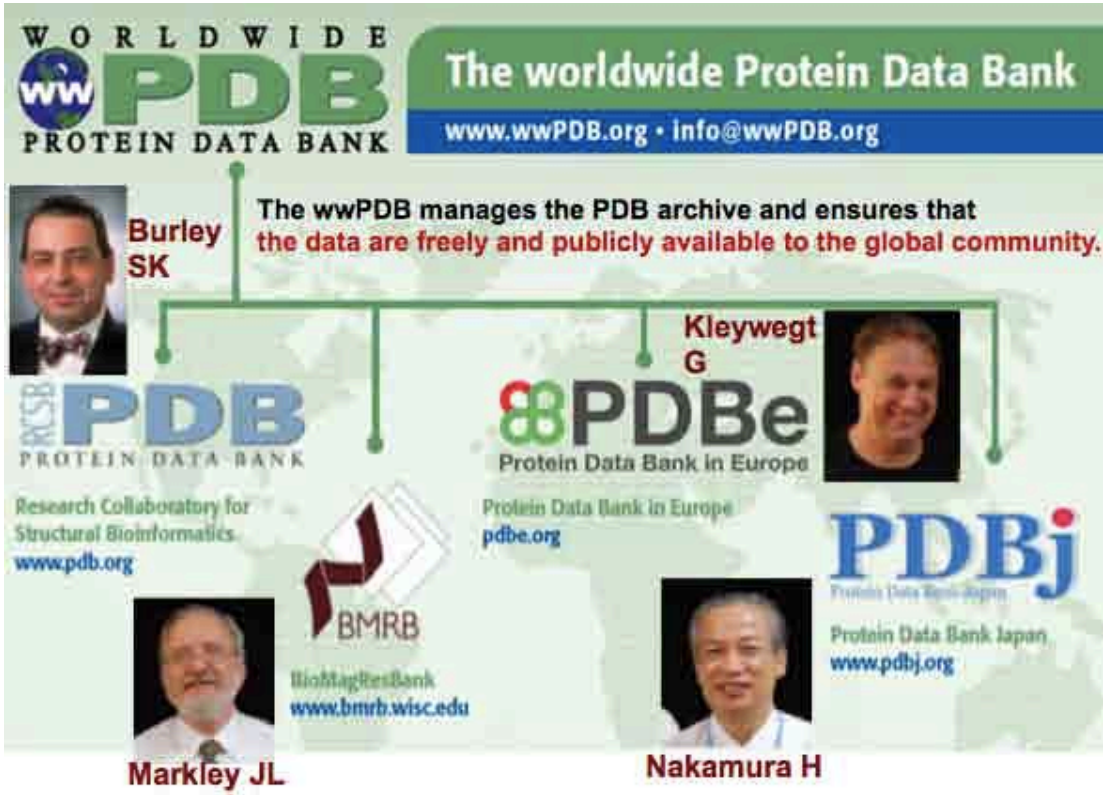
PDBj staffs (April 2016)

PDBj-BMRB staffs

Since 2001, PDBj has been managed at Institute for Protein Research, Osaka University as a member of the wwPDB, to curate and process the deposited data for an open and single archive.

The wwPDB (world-wide PDB)

wwPDB since 2003



wwpdb.org

12th Annual wwPDB Advisory Committee meeting held on 2 Oct 2015 at IPR, Osaka University

PDBj Protein Data Bank Japan
WORLDWIDE PDB PROTEIN DATA BANK

Activities/Services of each member of the wwPDB

- "Data-in" activity, common in all the wwPDB members with high quality control. For the purpose, data deposition, validation system, versioning system, and new format are developed
- "Data-out" services, common archive as the ftp site and the characteristic services by each wwPDB member

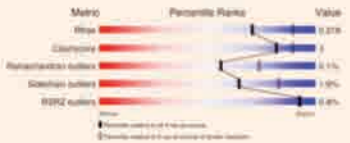
PDBj Protein Data Bank Japan

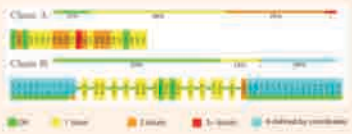
WORLDWIDE PDB PROTEIN DATA BANK

OneDep: Unified Deposition Portal for the Protein Data Bank

wwPDB Partners - RCSB PDB, PDBe, PDBj, and BMRB

OneDep





- Overall quality at a glance
- Residue-property plots
- X-ray fit of model and data
- NMR ensemble analysis
- NMR chemical shift validation

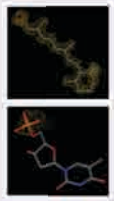
2. Pre-deposition Validation
validate.wwpdb.org

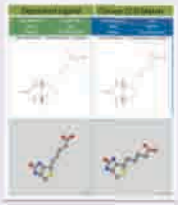
System supports
- X-ray, NMR, and 3DEM
- File re-upload
- Submissions based on existing depositions

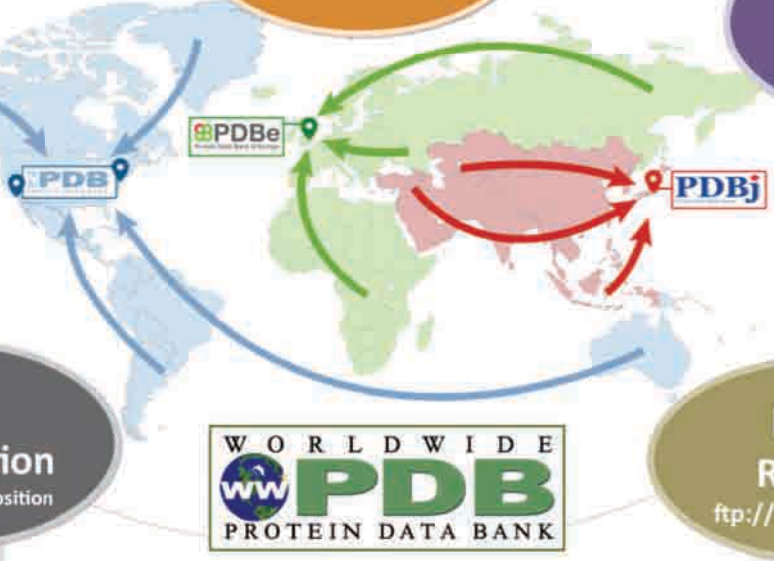
3. Deposition
deposit.wwpdb.org

- Sequence and taxonomy
- Electron density fit
- Stereochemistry
- Chemical nomenclature

4. Biocuration







- Generate atomic coordinate and experimental data files in PDBx/mmCIF format
- Assemble mandatory data items for deposition
- Complete sample polymer sequences including unobserved regions
- Source organism(s)
- Ligand and non-standard chemical details
- Macromolecular assembly and experimental support

1. File Preparation
wwpdb.org/deposition

5. Public Release
ftp://ftp.wwpdb.org

- Entries are released following Depositor instructions at deposition:
  - Hold until publication
  - Hold for a period of time (1 year max.)
  - Release immediately
- Journal policies governing PDB entry release take precedence over author-requested embargo periods

WORLDWIDE PDB PROTEIN DATA BANK

OneDep, a unified system for deposition, biocuration, and validation of experimentally determined X-ray, NMR, and 3DEM macromolecular structures, has been developed as a global project by the Worldwide PDB partners (wwPDB). The system unifies deposition, biocuration, and validation procedures across all wwPDB, EMDB, and BMRB deposition centers with a focus on simplifying and improving the process for the Depositor, improving data quality and completeness, and addressing challenges of the increasing number, size, and complexity of structures deposited to the Protein Data Bank archive.

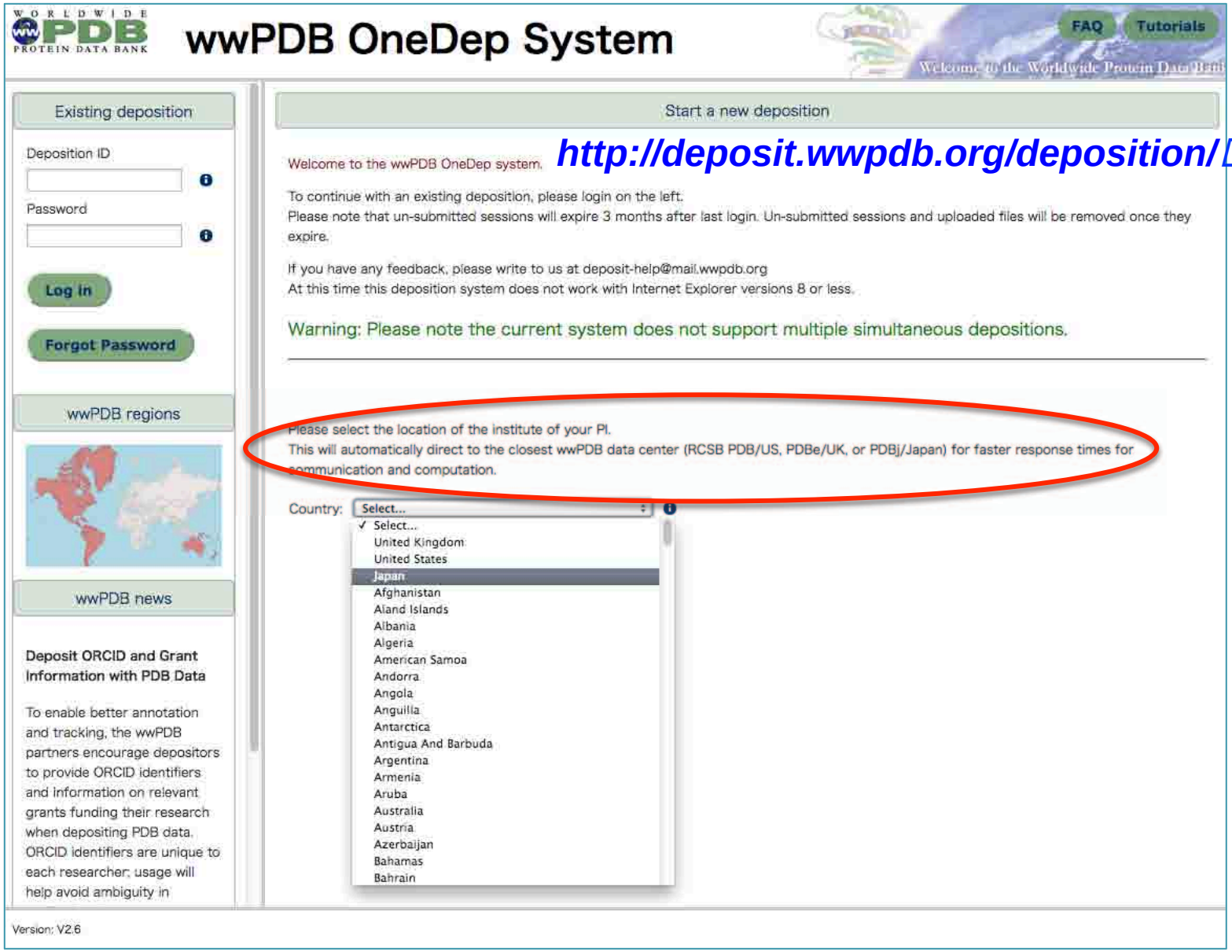
Acknowledgements

OneDep: Unified Deposition Portal for the wwPDB

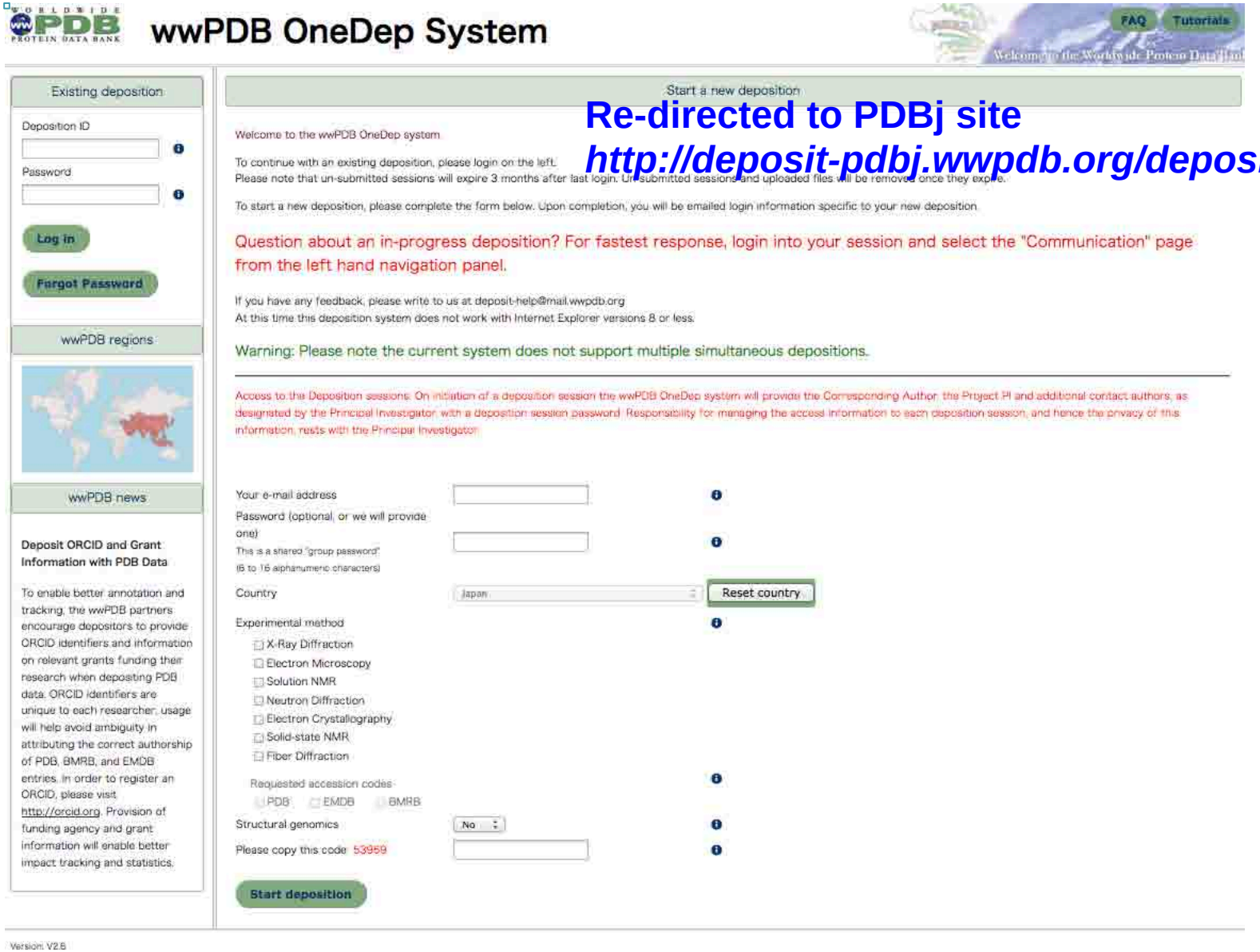
wwpdb.org

OneDep

OneDep: Unified Deposition Portal for the wwPDB



OneDep: Unified Deposition Portal for the wwPDB

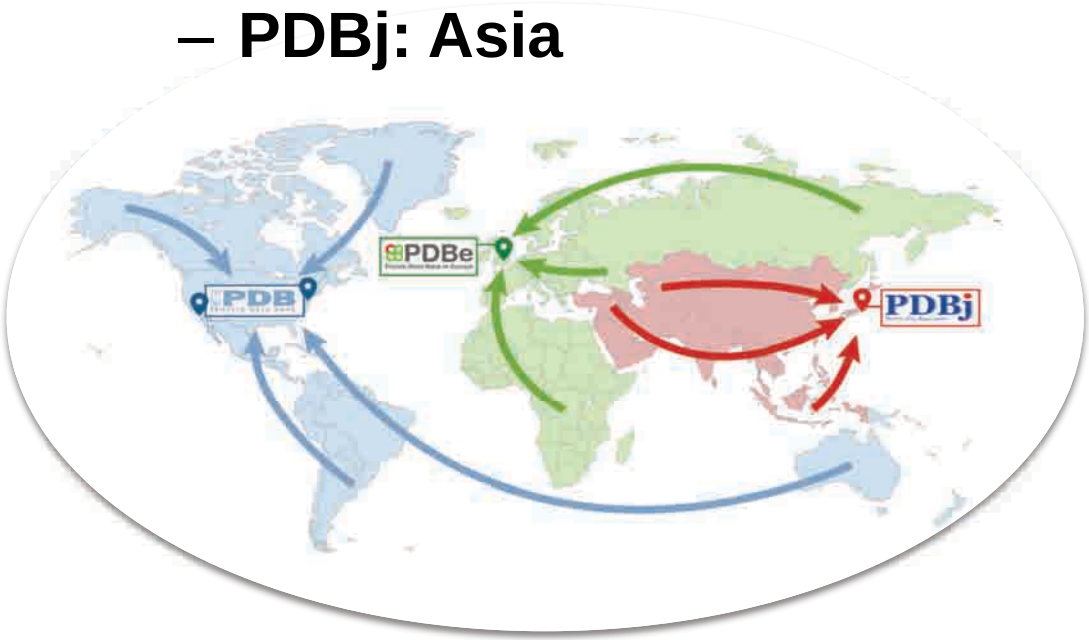


Re-directed to PDBj site

http://deposit-pdbj.wwpdb.org/deposition

PDB Depositions & Annotations

- **As of 2016 region-based processing of D&A-deposited entries:**
  - **RCSB PDB: Americas & Oceania**
  - **PDBe: Europe & Africa**
  - **PDBj: Asia**

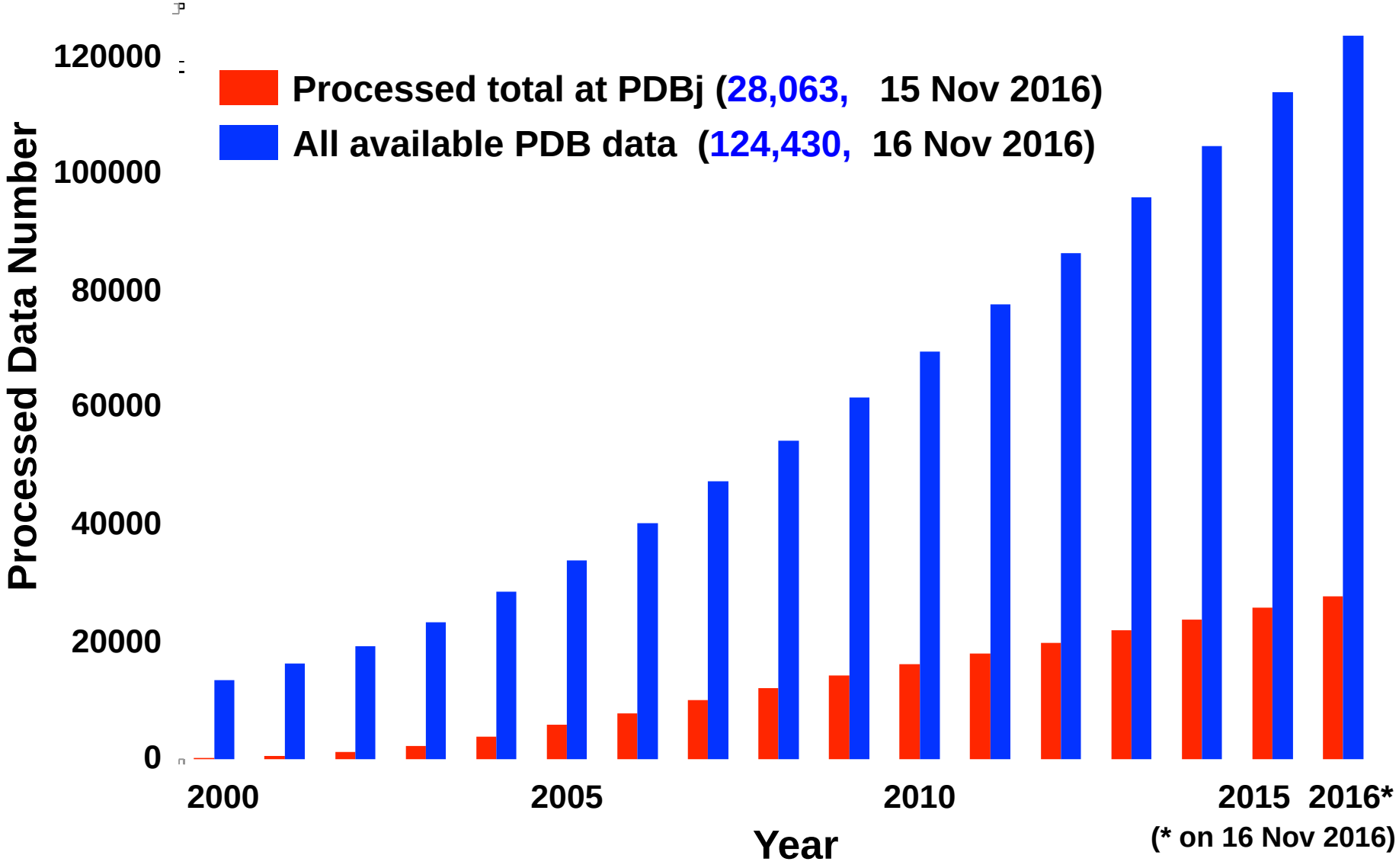


Year	Total Depositions	Processed By		
		RCSB PDB	PDBj	PDBe
2000	2983	2297	158	528
2001	3287	2408	383	496
2002	3565	2401	657	507
2003	4830	3135	1026	669
2004	5508	3082	1614	812
2005	6678	3563	2110	1005
2006	7282	4252	1945	1085
2007	8130	4703	2299	1128
2008	7073	4106	1994	973
2009	8300	5069	2173	1058
2010	8878	5464	2041	1373
2011	9250	5938	1816	1496
2012	9972	6408	1888	1676
2013	10566	6652	2128	1786
2014	10364	6038	1781	2545
2015	10954	4843	2100	4011
2016	8641	4034	1666	2941
TOTAL	126261	74393	27779	24089

125K **74K** **28K** **24K**

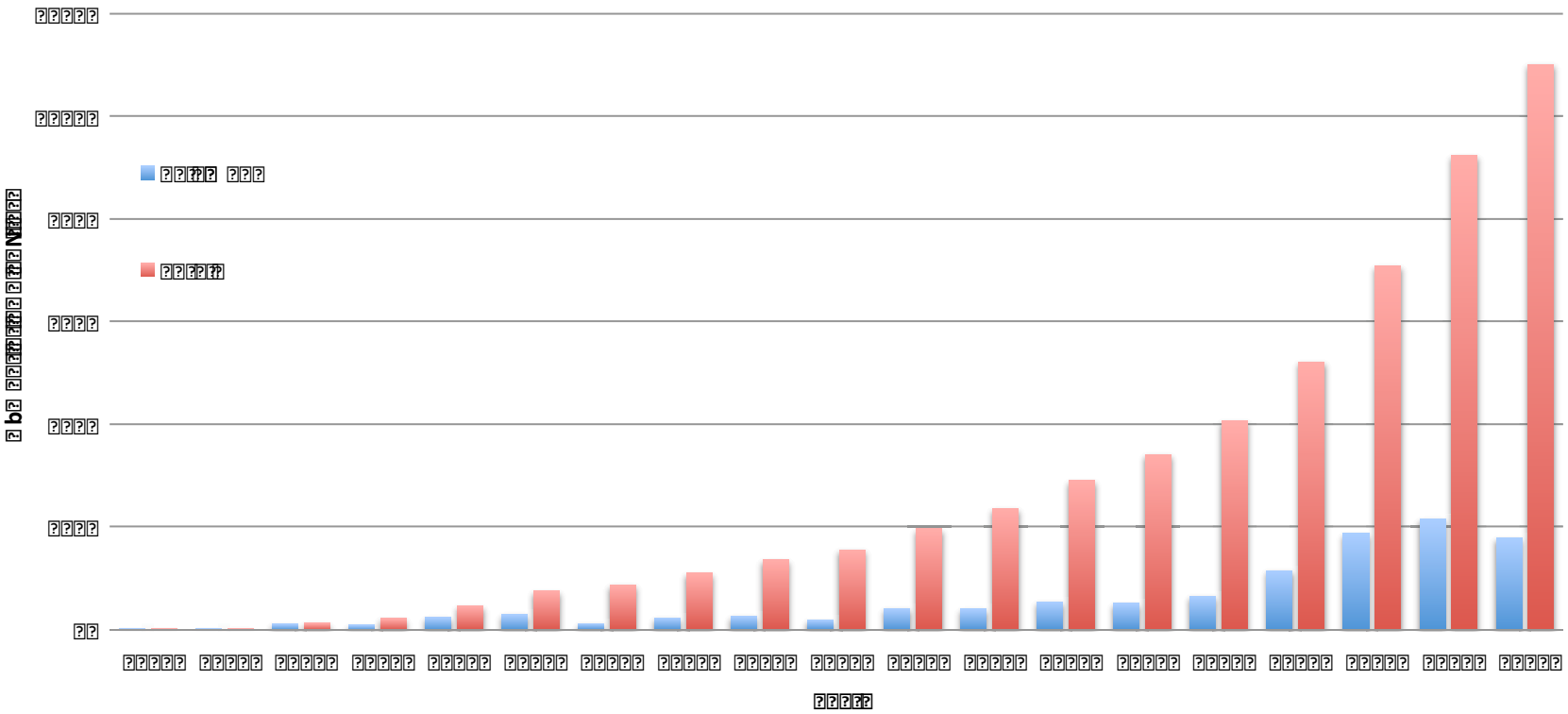
Last Updated: 15 September 2015

Data-in at PDBj and wwPDB



PDBj curates and processes about a Quarter of the deposited data, mainly from Asian and Oceania regions

Growth of PDB EM Entries



As of August 1, 2016, >1100 EM entries in the PDB archive

178 new entries released Jan 1 - Aug 1, 2016

Collection of ORCID IDs

Open Researcher and Contributor ID

http://orcid.org

Example: orcid.org/0000-0001-6690-5863

- **wwPDB Begins to use in April 2016**
- **Mandatory going forward in 2018**

Validation Report with experimental data

WORLDWIDE PDB
PROTEIN DATA BANK

wwPDB X-ray Structure Validation Summary Report (i)

Feb 28, 2014 – 04:13 AM GMT

PDB ID	:	3WDZ
Title	:	Crystal Structure of Keap1 in Complex with phosphorylated p62
Authors	:	Fukutomi, T.; Takagi, K.; Mizushima, T.; Tanaka, K.; Komatsu, M.; Yamamoto, M.
Deposited on	:	2013-06-26
Resolution	:	2.60 Å(reported)

This is a wwPDB validation summary report for a publicly released PDB entry.
We welcome your comments at validation@mail.wwpdb.org
A user guide is available at http://wwpdb.org/ValidationPDFNotes.html

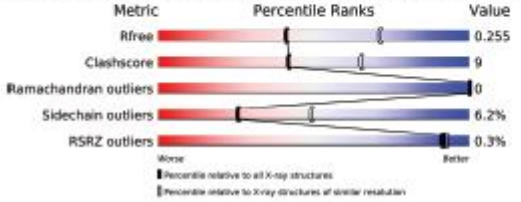
Top page of Validation report

Mandatory Experimental data
- **X-ray: Structure Factor**
- **NMR: Chemical Shifts and Distance restraints**
- **EM: 3DEM volume map**

1 Overall quality at a glance (i)

The reported resolution of this entry is 2.60 Å.

Percentile scores (ranging between 0-100) for global validation metrics of the entry are shown in the following graphic. The table shows the number of entries on which the scores are based.

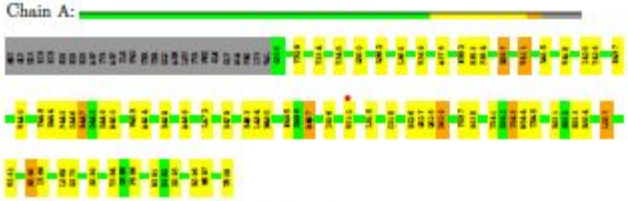


3 Residue-property plots (i)

These plots are drawn for all protein, RNA and DNA chains in the entry. The first graphic for a chain summarises the proportions of errors displayed in the second graphic. The second graphic shows the sequence view annotated by issues in geometry and electron density. Residues are color-coded according to the number of geometric quality criteria for which they contain at least one outlier: green = 0, yellow = 1, orange = 2 and red = 3 or more. A red dot above a residue indicates a poor fit to the electron density (RSRZ > 2). Stretches of 2 or more consecutive residues without any outlier are shown as a green connector. Residues present in the sample, but not in the model, are shown in grey.

- Molecule 1: Kelch-like ECH-associated protein 1

Chain A:



- Molecule 2: Peptide from Sequestosome-1

Chain B:

Validation report

Nature Structural & Molecular Biology 23, 871 (2016)
doi:10.1038/nsmb.3307, Published online 05 October 2016

EDITORIAL

nature structural & molecular biology

Where are the data?

Here, we announce two policy changes across Nature journals: data-availability statements in all published papers and official Worldwide Protein Data Bank (wwPDB) validation reports for peer review.

… We are now taking a further step and are requesting official wwPDB validation reports for peer review. These reports are made available by the wwPDB after data deposition (http://www.wwpdb.org/validation/validation-reports). Other Nature journals will soon follow suit.

Versioning of PDB entries

Current Issue

- Loss of connection between PDB ID and Publication

Introduction of a new PDB ID code format

- Allow more informative and transparent delivery of revised data files
- With PDB prefix and extension of 4 characters (e.g. from “1ABC” to “PDB_00001ABC”)
- Example: PDB_00001ABC_XYZ_V2-2.cif.gz

Implementation plan

- Create new versioned ftp tree containing the latest minor revision to each major version
- Continue current ftp tree with current file naming convention.

  Files in this branch will serve up the latest version of each data file

PDBj Protein Data Bank Japan
WORLDWIDE PDB PROTEIN DATA BANK

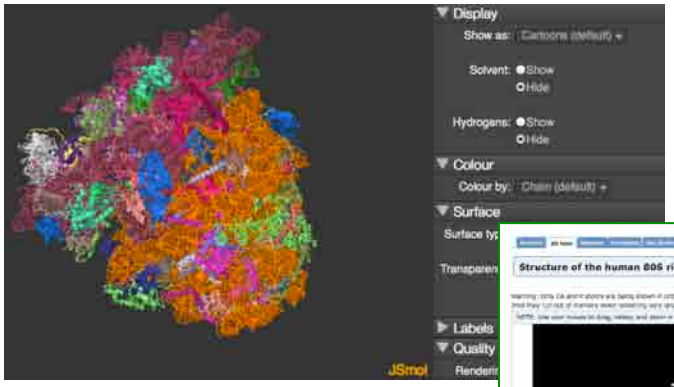
New format of PDB: PDBx/mmCIF

- **PDB format is almost 40 years old and does not support today's science.**
- **PDB Record format limitations**
  - **Max. 62 chains**
  - **Max. 99,999 atoms**
  - **No bond orders or chirality specified for ligands**
  - **No support for NMR, EM, hybrid methods, …**
  - **Meta-data specification cumbersome and inflexible**

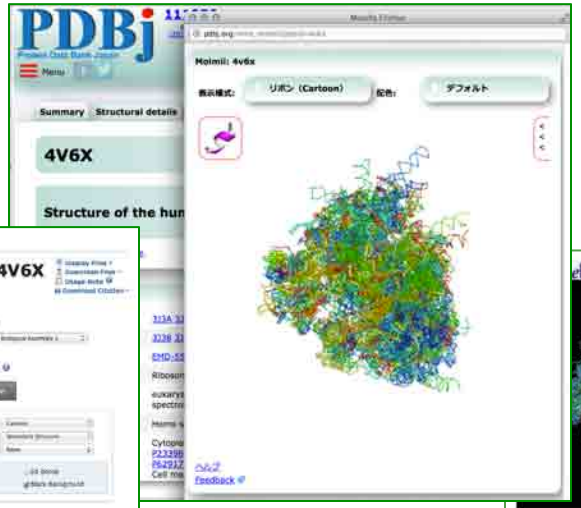
- **Preserve backward compatibility where possible**
- **web service to create PDB format data files**
- **Start in 2014**

Large Structures (476) in PDB

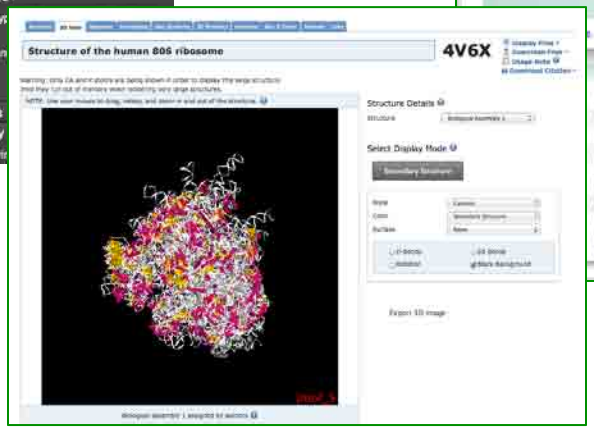
- As of December 2014, large structures are no longer split into multiple PDB entries but are **released as single PDBx/mmCIF and PDBML files**
  - Existing split entries were consolidated into single entries
  - A bundle of best-effort PDB files (PDB bundle) is available for every large structure through the wwPDB FTP sites
- Each wwPDB member provides **tools to view large structures**

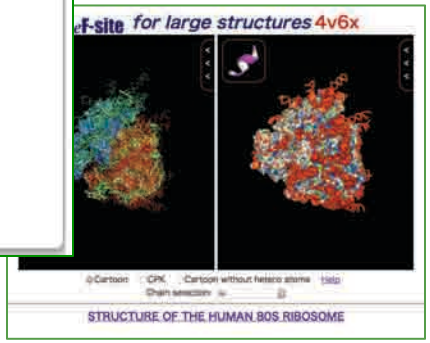


PDBe



PDBj





4v6x: structure of human 80S ribosome, 89 chains, 4 MDa

RCSB-PDB

Current Supported Archival Formats
protein structure format universe

PDB (*ca.* 1974)

PDBx/mmCIF (*ca.* 1997)

PDBML (*ca.* 2005)

RDF (*ca.* 2011)

In managing the formats, PDBx is the master format.

```
ATOM      1  N   GLN A  39      24.690 -27.754  24.275  1.00 60.76           N
ATOM      2  CA  GLN A  39      23.581 -26.768  24.416  1.00 60.98           C
ATOM      3  C   GLN A  39      23.990 -25.379  23.905  1.00 59.98           C
ATOM      4  O   GLN A  39      25.070 -25.209  23.330  1.00 60.25           O
ATOM      5  CB  GLN A  39      23.136 -26.685  25.878  1.00 60.69           C
ATOM      6  N   VAL A  40      23.115 -24.395  24.122  1.00 59.58           N
ATOM      7  CA  VAL A  40      23.342 -23.010  23.690  1.00 57.26           C
ATOM      8  C   VAL A  40      24.000 -22.152  24.778  1.00 56.00           C
ATOM      9  O   VAL A  40      23.992 -20.920  24.692  1.00 55.53           O
ATOM     10  CB  VAL A  40      22.015 -22.337  23.275  1.00 57.32           C
```

PDB

PDBx/mmCIF

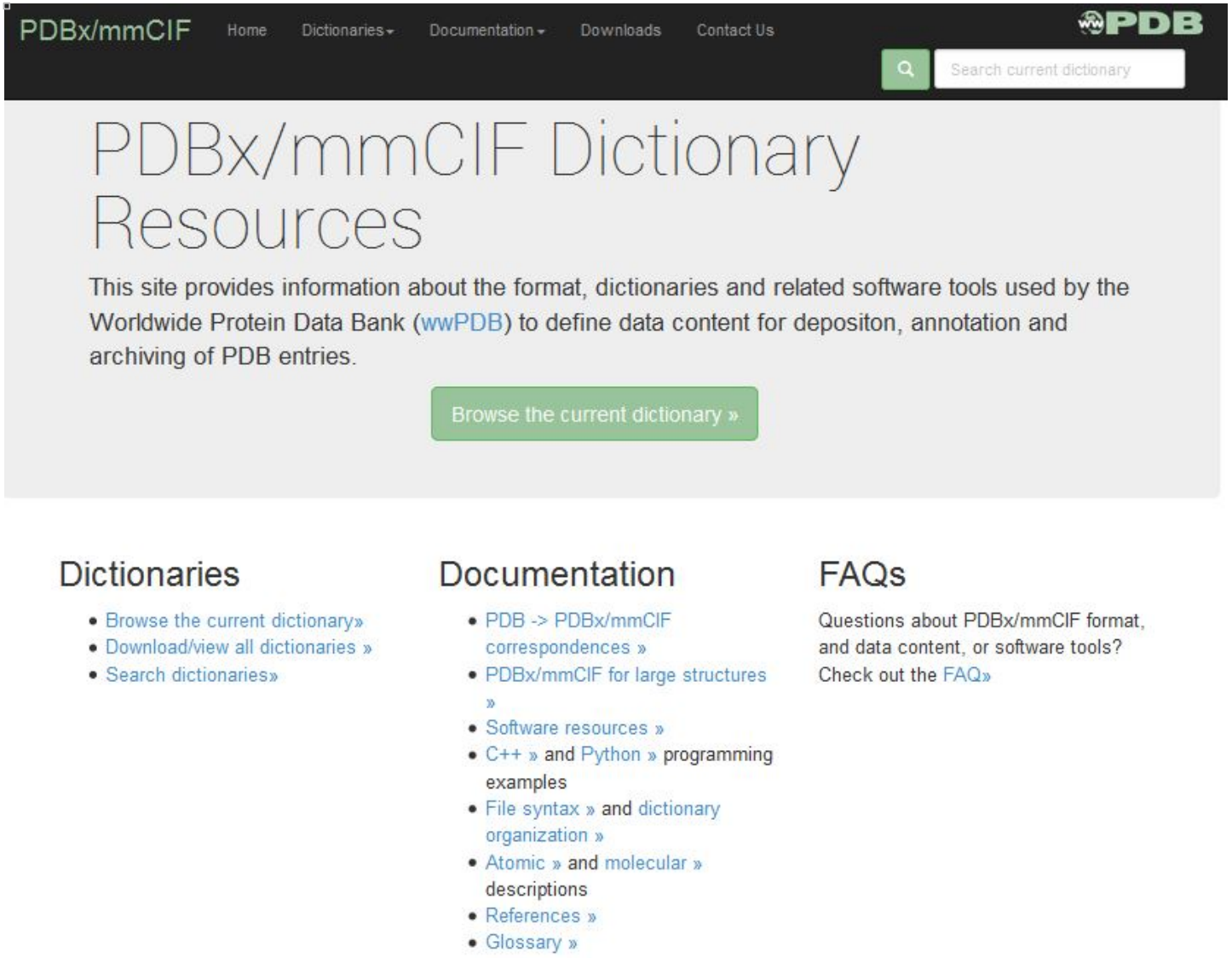
```
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.auth_atom_id
_atom_site.type_symbol
_atom_site.auth_comp_id
_atom_site.auth_asym_id
_atom_site.auth_seq_id
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.pdbx_PDB_model_num
_atom_site.occupancy
_atom_site.pdbx_auth_alt_id
_atom_site.B_iso_or_equiv
ATOM       1  N    N  GLN  A   39   24.690  -27.754   24.275  1  1.000  .  60.760
ATOM       2  CA   C  GLN  A   39   23.581  -26.768   24.416  1  1.000  .  60.980
ATOM       3  C    C  GLN  A   39   23.990  -25.379   23.905  1  1.000  .  59.980
ATOM       4  O    O  GLN  A   39   25.070  -25.209   23.330  1  1.000  .  60.250
ATOM       5  CB   C  GLN  A   39   23.136  -26.685   25.878  1  1.000  .  60.690
ATOM       6  N    N  VAL  A   40   23.115  -24.395   24.122  1  1.000  .  59.580
ATOM       7  CA   C  VAL  A   40   23.342  -23.010   23.690  1  1.000  .  57.260
ATOM       8  C    C  VAL  A   40   24.000  -22.152   24.778  1  1.000  .  56.000
ATOM       9  O    O  VAL  A   40   23.992  -20.920   24.692  1  1.000  .  55.530
ATOM      10  CB   C  VAL  A   40   22.015  -22.337   23.275  1  1.000  .  57.320
ATOM      11  N    N  ALA  A   41   24.560  -22.804   25.797  1  1.000  .  54.570
```

PDBx/mmCIF Software Support

- **Phenix and Refmac – produce native PDBx files for deposition**
- **MMDB** - macromolecular object library in CCP4
- **iotbx.cif/ucif** - CCTBx C++/Python IO library with dictionary validation
- **CCIF** – CCP4 C++ library with FORTRAN support and dictionary validation
- **CBFLib** - ANSI-C library for CIF & imgCIF files
- **mmLIB** - Python toolkit supporting CIF & mmCIF
- **BioPython** - Python toolkit for computational biology
- **PyCifRW** - Python CIF/mmCIF parsing tools
- **BioJava** - Java mmCIF IO package
- **STAR::Parser** – Perl mmCIF parser and molecular object library
- **RCSBTools** - C++/Python parsing and dictionary validation tools plus many other supporting format conversion and data management applications
- **Visualization** - **UCSF Chimera**, **Jmol**, **OpenRasMol, Coot, CCP4mg, jV, Molmil**

wwPDB Service site for a new format

http://mmcif.wwpdb.org/ or *http://mmcif.pdbj.org/*

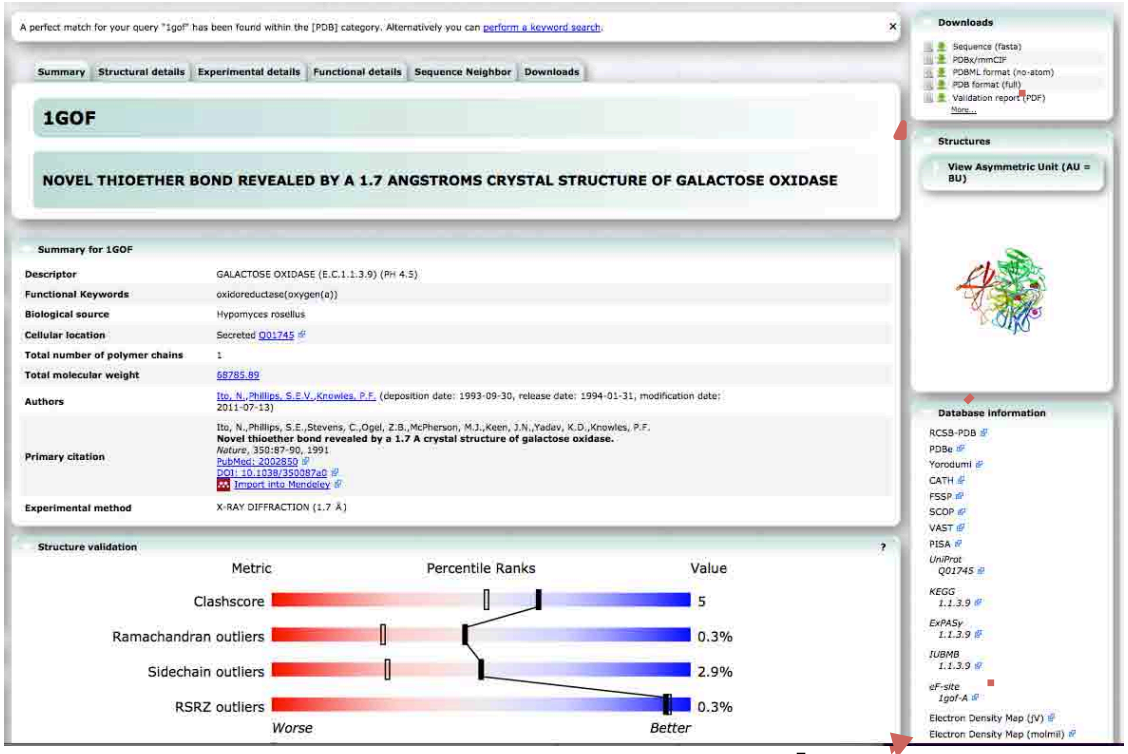


wwPDB Service site for a new format

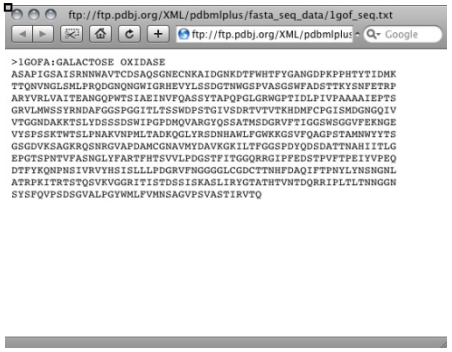
http://mmcif.pdbj.org/converter/index.php?l=en

Data–out from PDBj

http://pdbj.org/

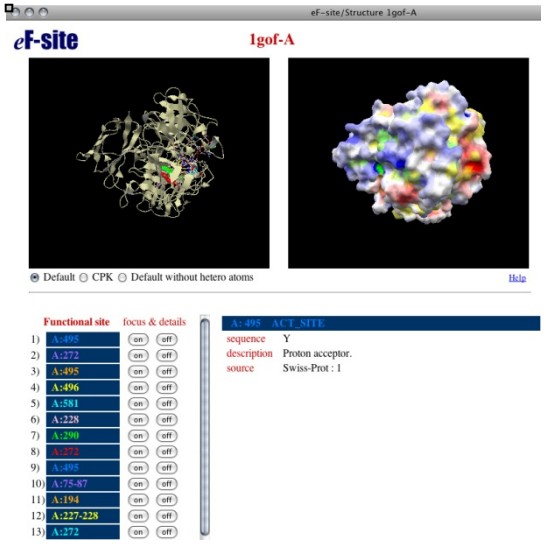


Amino acid sequence (FASTA)



Molecular surface DB: *eF-site*

http://ef-site.hgc.jp/eF-site/



Data viewer at PDBj

Graphic viewer: *jV* and *Molmil*

http://pdbj.org/jV/

Kinjo et al. NAR 40, D453 (2012)

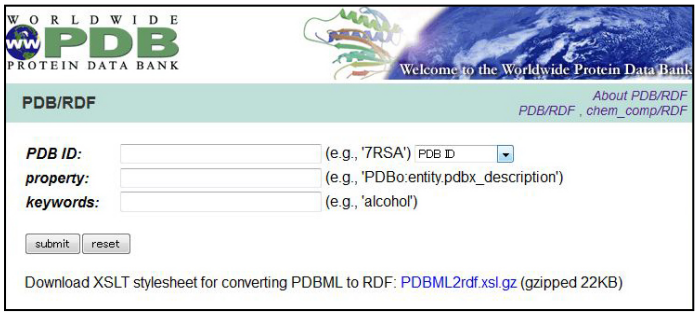
wwPDB/RDF format for Semantic Web

Service from wwPDB

Kinjo et al. (2017) Nucl. Acids Res. doi: 10.1093/nar/gkw962.
Yokochi et al. (2016) J. Biomed. Semantics, 7:16.

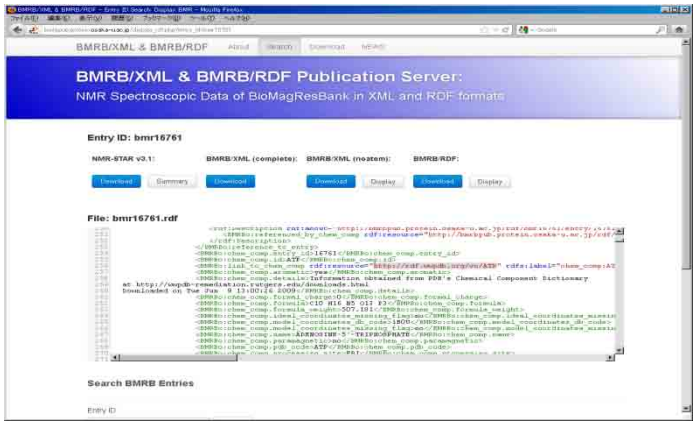
wwPDB/RDF

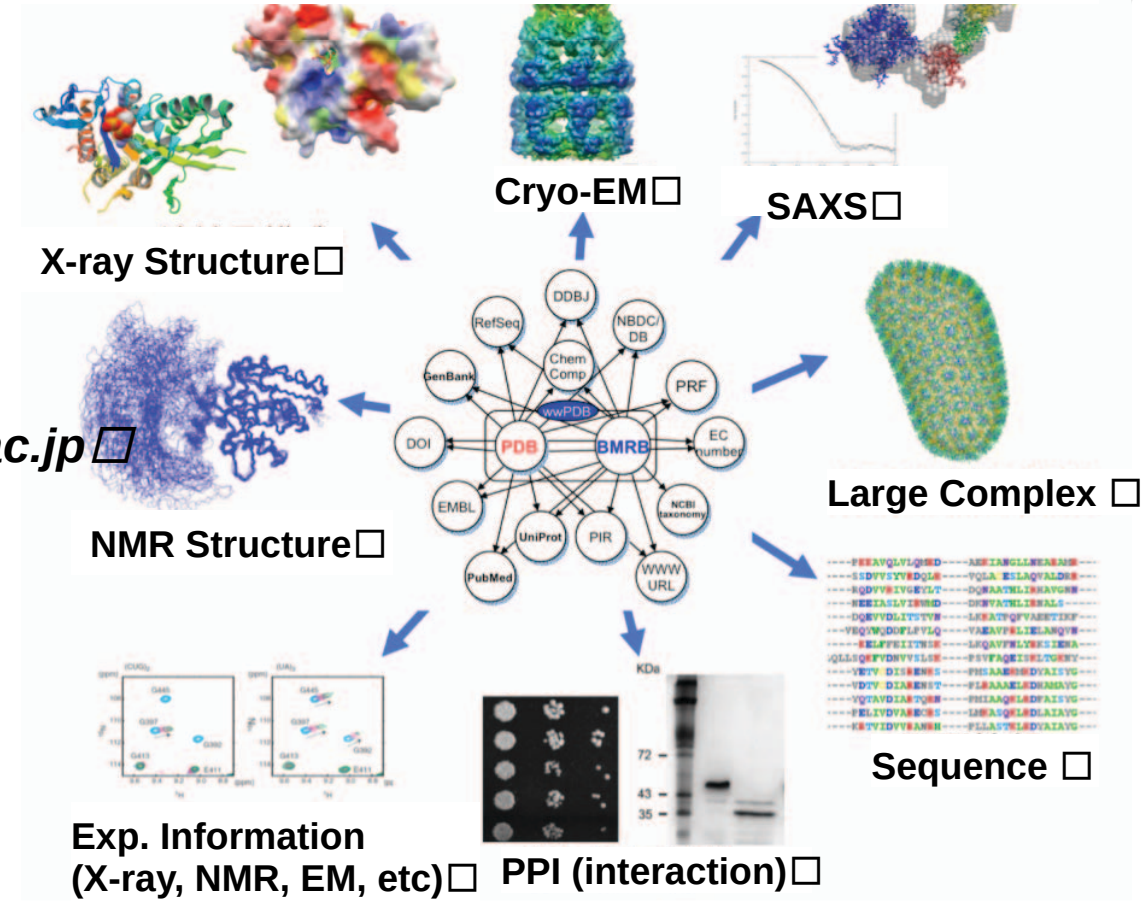
http://rdf.wwpdb.org/



BMRB/RDF

http://bmrbpub.protein.osaka-u.ac.jp



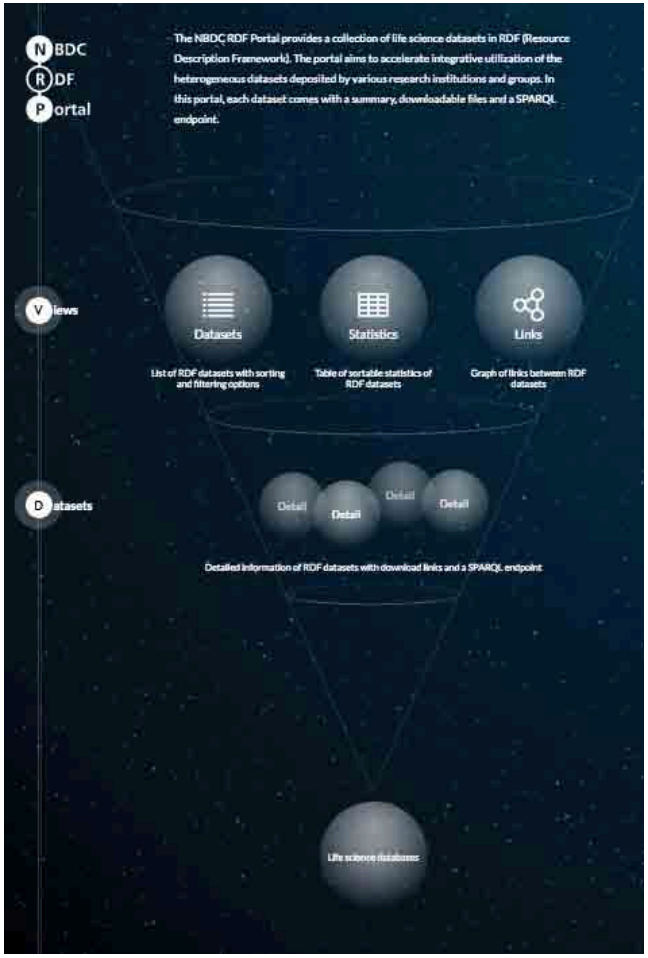


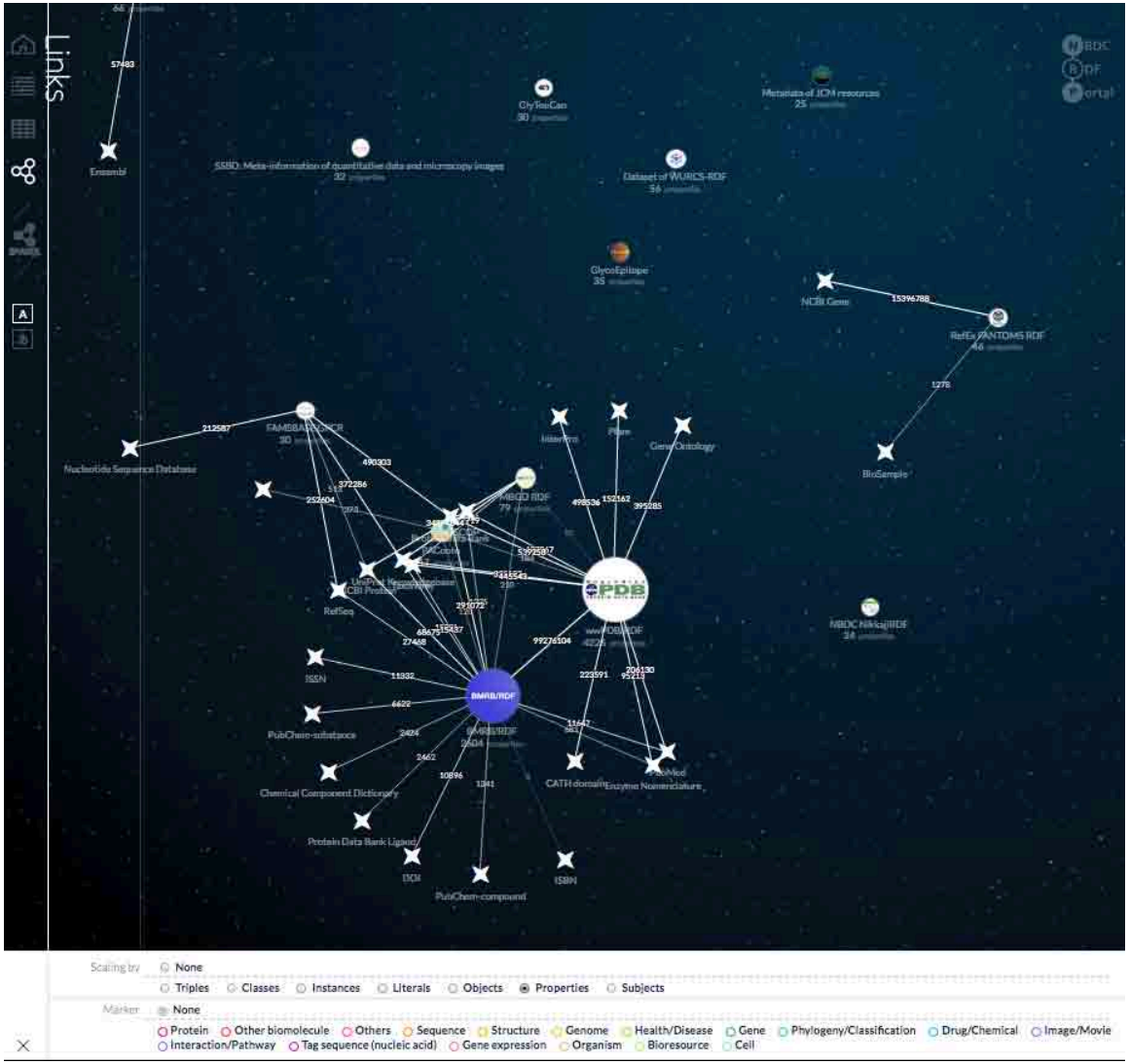

```
<rdf:Description rdf:about="http://rdf.wwpdb.org/pdb/1BY4">
  <rdf:type rdf:resource="http://purl.uniprot.org/core/Structure_Resource"/>
  <database rdf:resource="http://purl.uniprot.org/database/PDB"/>
  <method rdf:resource="http://purl.uniprot.org/core/X-Ray_Crystallography"/>
  <resolution rdf:datatype="http://www.w3.org/2001/XMLSchema#float">2.10</resolution>
</rdf:Description>
```

PDBj Protein Data Bank Japan

WORLDWIDE PDB PROTEIN DATA BANK

Integration to NBDC RDF-portal





NBDC RDF-Portal

https://integbio.jp/rdf/

PDBj Protein Data Bank Japan

WORLDWIDE PDB PROTEIN DATA BANK

PDBj staffs and collaborators

- **Head**
  - Nakamura, Haruki, Ph. D. (Prof., IPR, Osaka Univ.)
- **Group for PDB Database Curation**
  - Nakagawa, Atsushi, Ph. D. (Group Leader, Prof., IPR, Osaka Univ.)
  - Igarashi, Reiko (IPR, Osaka Univ.)
  - Kengaku, Yumiko (IPR, Osaka Univ.)
  - Cho, Hasumi, Ph. D. (IPR, Osaka Univ.)
  - Ikegawa, Yasuyo (IPR, Osaka Univ.)
  - Sato, Junko (IPR, Osaka Univ.)
- **Group for Development of new tools and services**
  - Kinjo, Akira R., Ph. D. (Assoc. Prof., IPR, Osaka Univ.)
  - Iwasaki, Kenji, Ph. D. (Assoc. Prof., IPR, Osaka Univ.)
  - Suzuki, Hirofumi, Ph. D. (IPR, Osaka Univ.)
  - Yamashita, Reiko (IPR, Osaka Univ.)
  - Kudou, Takahiro (IPR, Osaka Univ.)
  - Bekker, Gert-Jan (IPR, Osaka Univ.)
- **Group for BMRB**
  - Fujiwara, Toshimichi, Ph. D. (Group Leader, Prof. Osaka Univ.)
  - Kojima, Chojiro, Ph. D. (Assoc. Prof., IPR, Osaka Univ.)
  - Kobayashi, Naohiro, Ph. D. (IPR, Osaka Univ.)
  - Iwata, Takeshi (IPR, Osaka Univ.)
  - Yokochi, Masashi (IPR, Osaka Univ.)
- **Collaboratory Researchers**
  - Wako, Hiroshi, Ph. D. (Prof., Waseda Univ.) for ProMode
  - Endo, Shigeru, Ph. D. (Assoc. Prof., Kitasato Univ.) for ProMode
  - Ito, Nobutoshi, Ph. D. (Prof., Tokyo Medical and Dental Univ.)
  - Kinoshita, Kengo, Ph.D. (Prof., Tohoku Univ.) for *e*F-site
  - Standley, Daron, Ph. D. (Prof., IVR, Kyoto Univ.) for SeqNavi, StructNavi, SeSAW, and ASH
  - Katoh, Kazutaka, Ph. D. (Assoc. Prof., IFReC, Osaka Univ.) for MAFFTash
- **Secretary**
  - Haruki, Nahoko (IPR, Osaka Univ.)