PDBj Luncheon

Making full use of the wwPDB validation reports

wwPDB検証レポートの使用法

The 56th Ann. Meeting of BSJ, Okayama, 16 Sep. 2018

Masashi Yokochi
横地政志

PDBj-BMRB, Institute for Protein Research, Osaka University

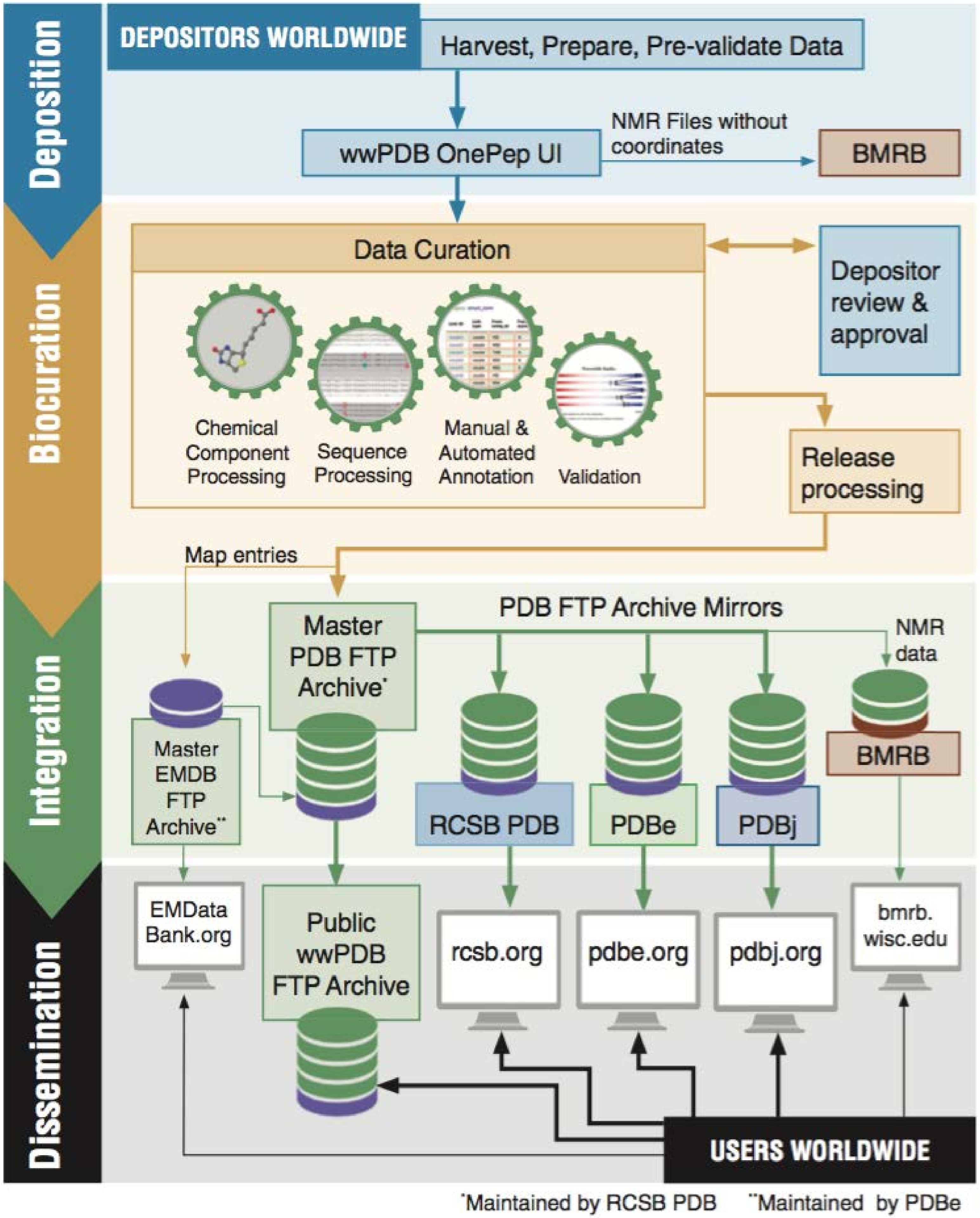
OneDep

“OneDep” is current PDB deposition and annotation system used since January 2016.

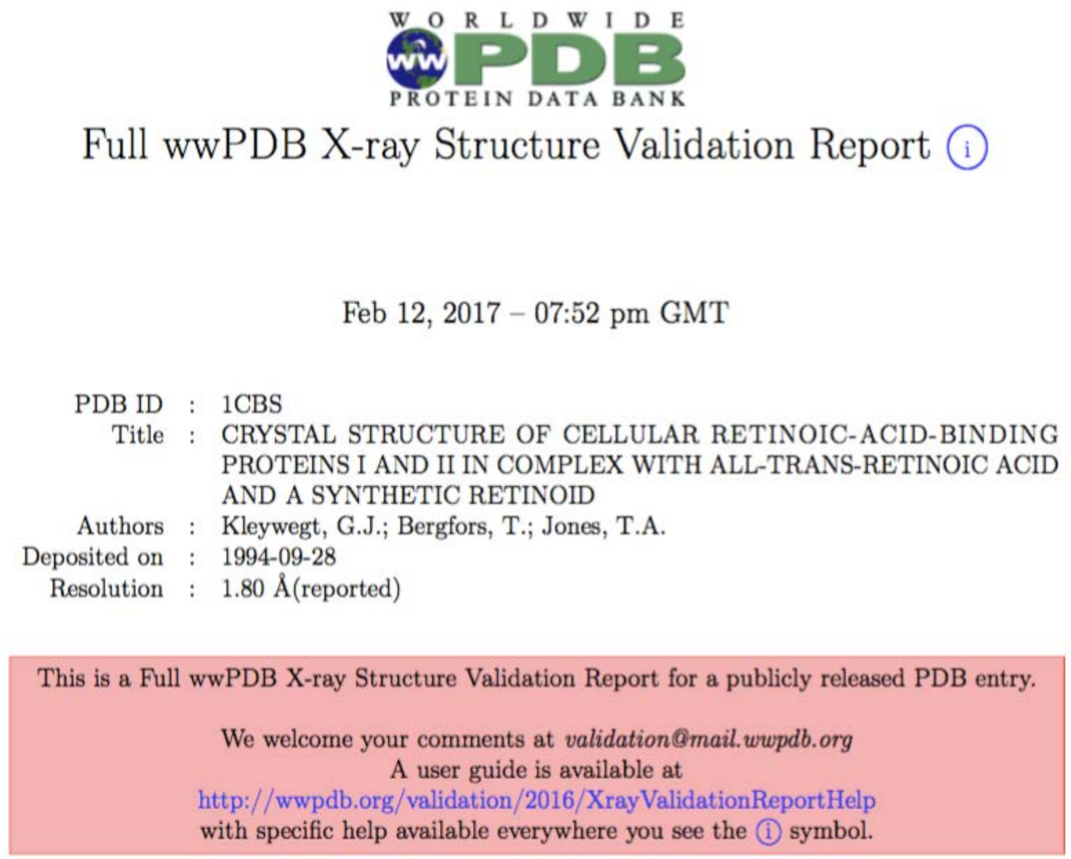
Features:

- A common, web-based deposition interface
- Minimization of manual entry
- Allows submission based on existing depositions
- Enables replacement of coordinate and experimental file prior to submission and after processing
- Preview and download PDB files after submission
- Supports hybrid methods for structure determination
- mmCIF is the master file format, instead of legacy PDB format
- Improved checking for ligand chemistry and polymer sequence consistency
- Communication with PDB annotation staff using web-based interface
- Validation based on recommendations from community Task Forces.

Young, JY et al., Structure, 25(3), 536-545, 2017



Overview of OneDep system



WORLDWIDE PDB PROTEIN DATA BANK

Full wwPDB X-ray Structure Validation Report ⓘ

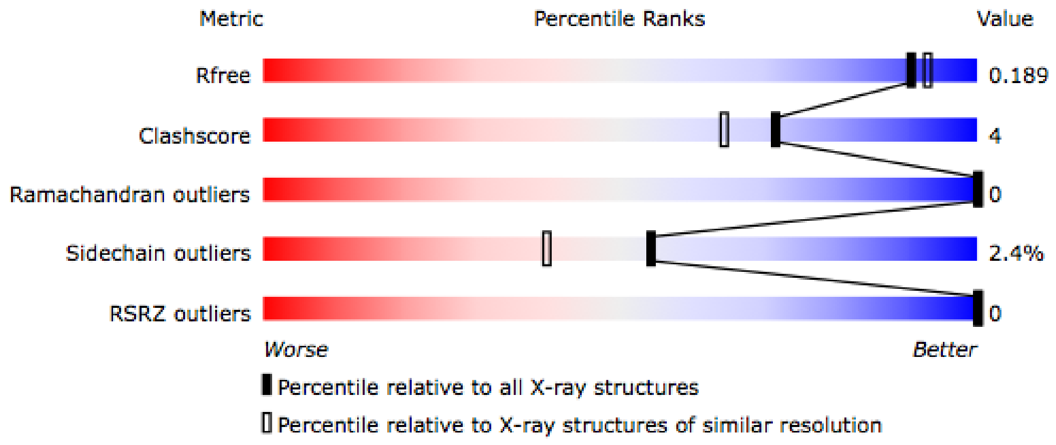
Feb 12, 2017 – 07:52 pm GMT

PDB ID : 1CBS
Title : CRYSTAL STRUCTURE OF CELLULAR RETINOIC-ACID-BINDING PROTEINS I AND II IN COMPLEX WITH ALL-TRANS-RETINOIC ACID AND A SYNTHETIC RETINOID
Authors : Kleywegt, G.J.; Bergfors, T.; Jones, T.A.
Deposited on : 1994-09-28
Resolution : 1.80 Å(reported)

This is a Full wwPDB X-ray Structure Validation 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/validation/2016/XrayValidationReportHelp
with specific help available everywhere you see the ⓘ symbol.

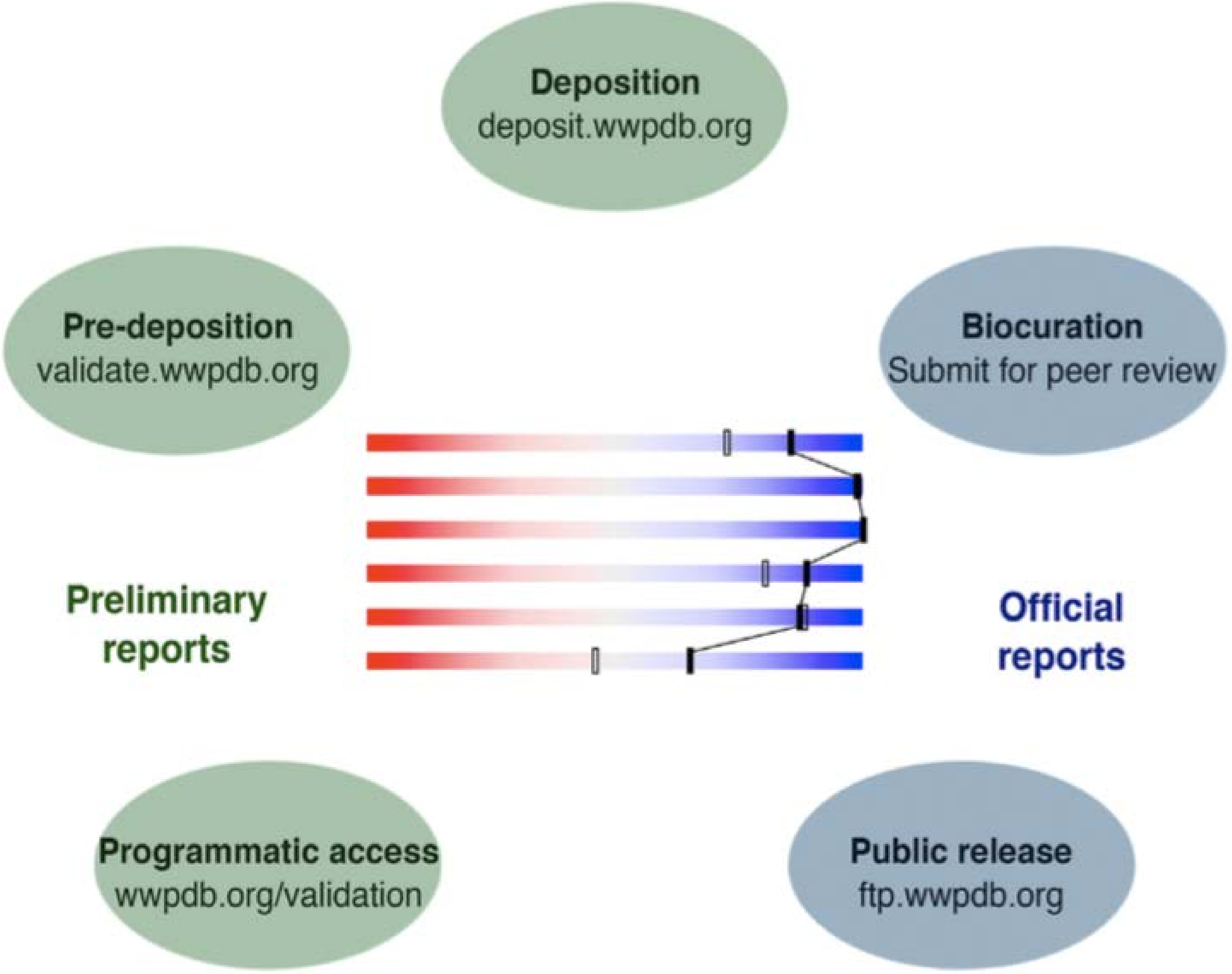
wwPDB validation report (PDF)



Percentile view of validation report

Young, JY et al., Structure, 25(3), 536–545, 2017

wwPDB validation reports



Deposition
deposit.wwpdb.org

Pre-deposition
validate.wwpdb.org

Biocuration
Submit for peer review

Preliminary reports

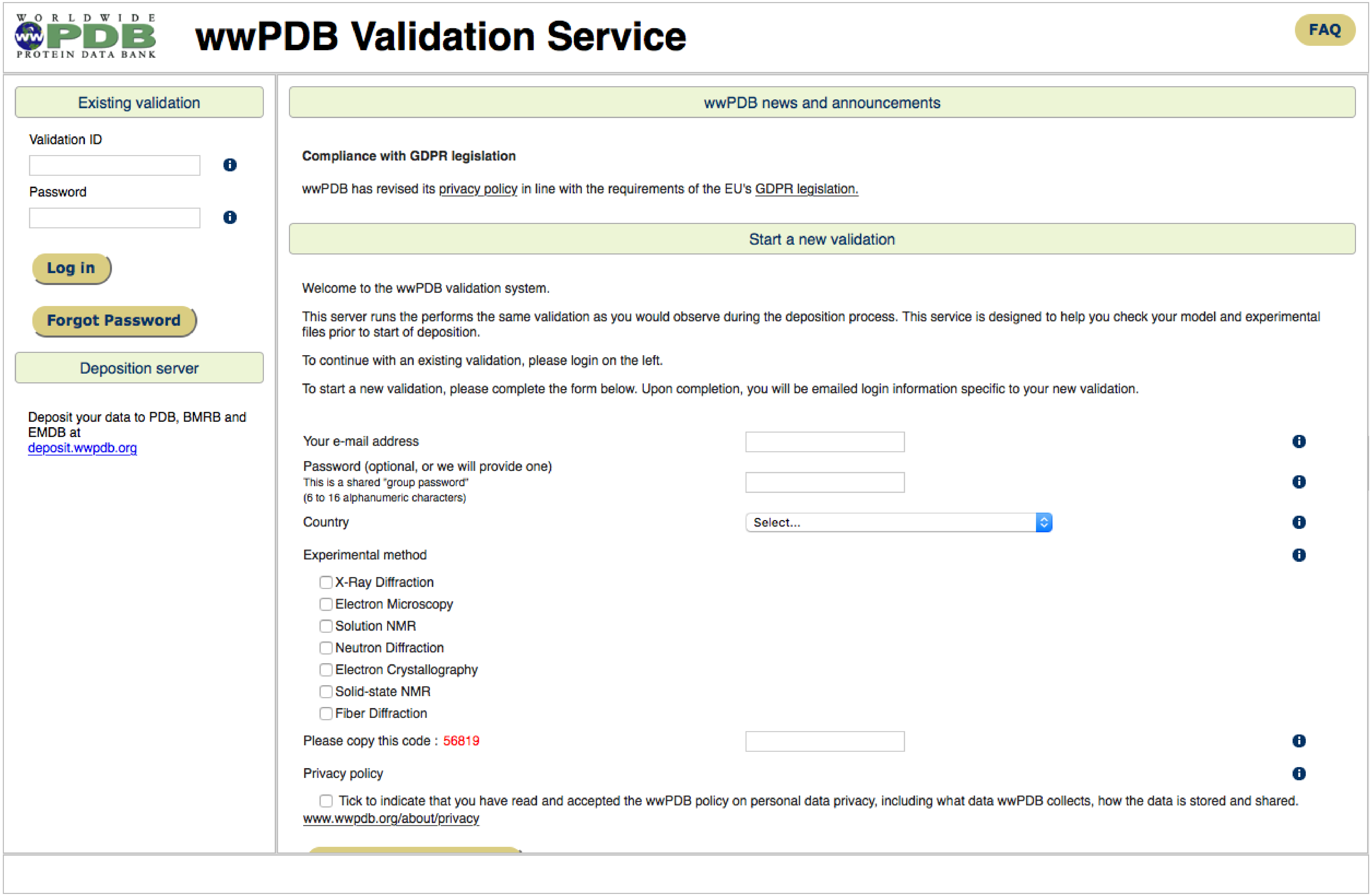
Official reports

Programmatic access
wwpdb.org/validation

Public release
ftp.wwpdb.org

Structure 25, 1916–1927, 2017

Stand-alone wwPDB validation server



wwPDB Validation Service

Validation ID

Password

Deposit your data to PDB, BMRB and EMDB at deposit.wwpdb.org

Compliance with GDPR legislation

wwPDB has revised its privacy policy in line with the requirements of the EU's GDPR legislation.

Welcome to the wwPDB validation system.

This server runs the performs the same validation as you would observe during the deposition process. This service is designed to help you check your model and experimental files prior to start of deposition.

To continue with an existing validation, please login on the left.

To start a new validation, please complete the form below. Upon completion, you will be emailed login information specific to your new validation.

Your e-mail address

Password (optional, or we will provide one)
This is a shared "group password"
(6 to 16 alphanumeric characters)

Country: Select...

Experimental method

- □ X-Ray Diffraction
- □ Electron Microscopy
- □ Solution NMR
- □ Neutron Diffraction
- □ Electron Crystallography
- □ Solid-state NMR
- □ Fiber Diffraction

Please copy this code : 56819

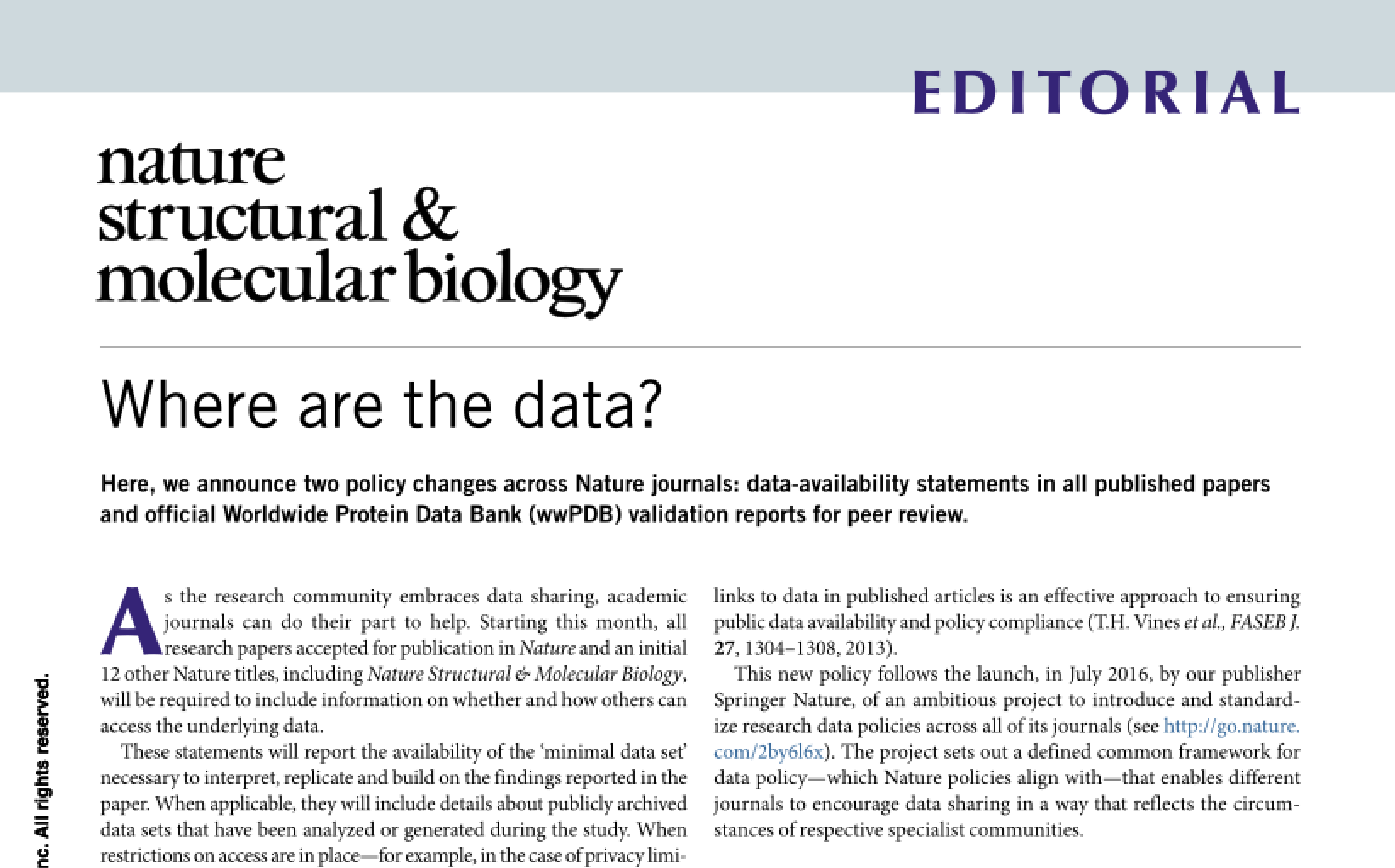
Privacy policy

□ Tick to indicate that you have read and accepted the wwPDB policy on personal data privacy, including what data wwPDB collects, how the data is stored and shared. www.wwpdb.org/about/privacy

https://validate.wwpdb.org

wwPDB validation report is now required for publication

Nat. Struct. Mol. Biol., 10(23), 871, 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.

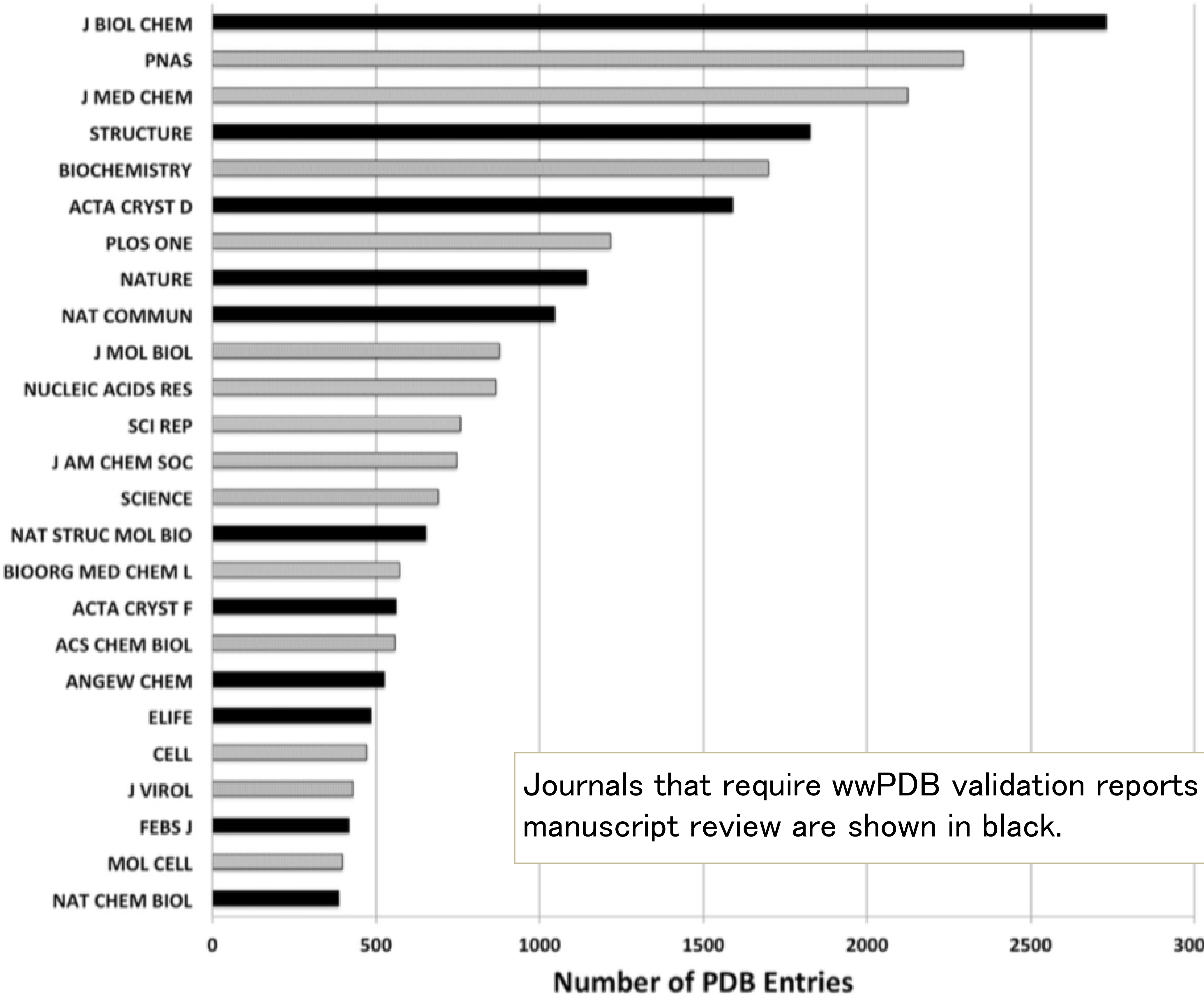
As the research community embraces data sharing, academic journals can do their part to help. Starting this month, all research papers accepted for publication in *Nature* and an initial 12 other Nature titles, including *Nature Structural & Molecular Biology*, will be required to include information on whether and how others can access the underlying data.

These statements will report the availability of the 'minimal data set' necessary to interpret, replicate and build on the findings reported in the paper. When applicable, they will include details about publicly archived data sets that have been analyzed or generated during the study. When restrictions on access are in place—for example, in the case of privacy limi-

links to data in published articles is an effective approach to ensuring public data availability and policy compliance (T.H. Vines *et al.*, *FASEB J.* **27**, 1304–1308, 2013).

This new policy follows the launch, in July 2016, by our publisher Springer Nature, of an ambitious project to introduce and standardize research data policies across all of its journals (see http://go.nature.com/2by6l6x). The project sets out a defined common framework for data policy—which Nature policies align with—that enables different journals to encourage data sharing in a way that reflects the circumstances of respective specialist communities.

Papers describing PDB structure from 2012 to 2016



Journals that require wwPDB validation reports for manuscript review are shown in black.

Structure 25, 1916–1927, 2017

Validation software utilized for generation of wwPDB validation report (2018)

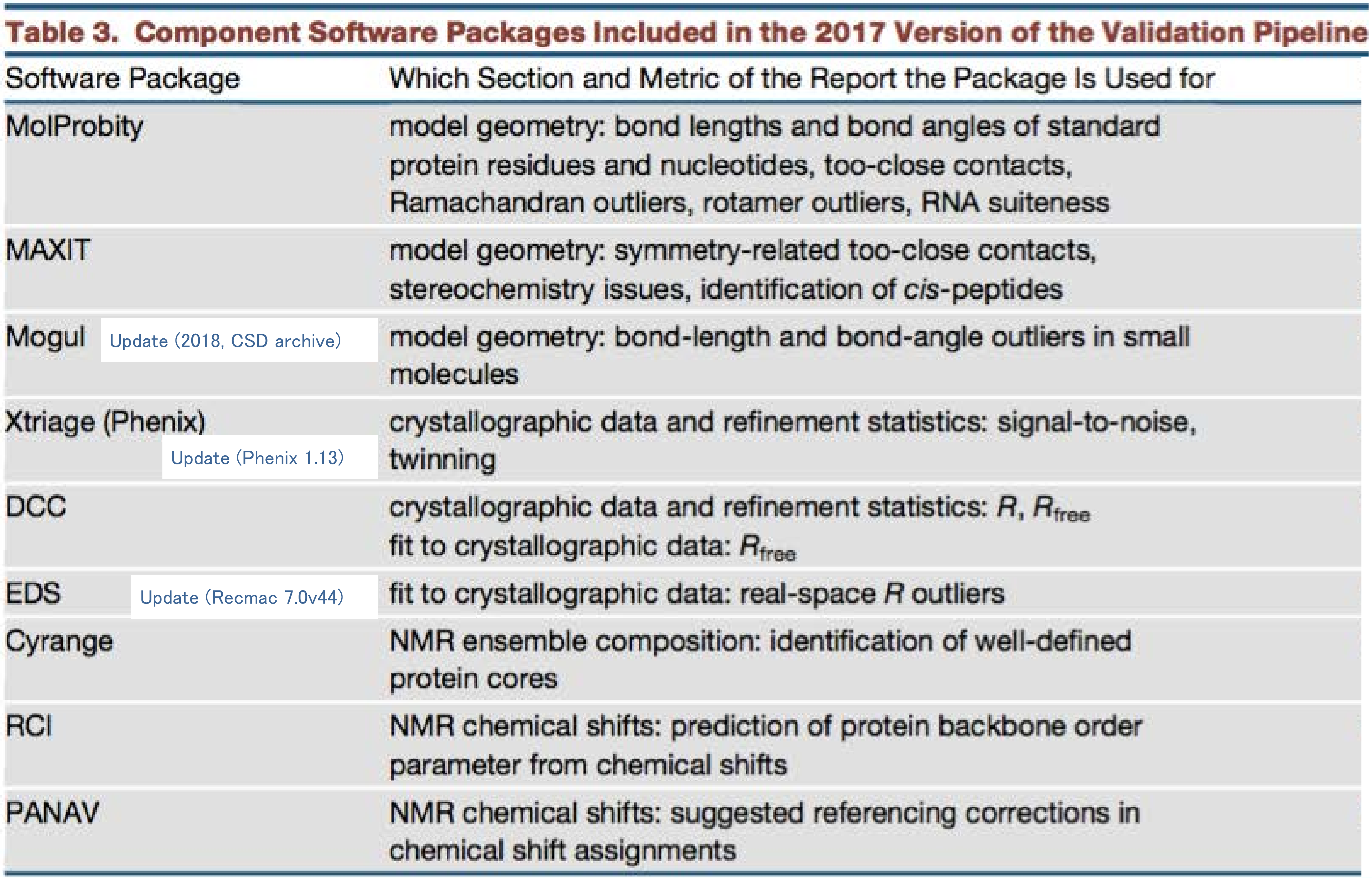


Table 3. Component Software Packages Included in the 2017 Version of the Validation Pipeline

Software Package	Which Section and Metric of the Report the Package Is Used for
MolProbity	model geometry: bond lengths and bond angles of standard protein residues and nucleotides, too-close contacts, Ramachandran outliers, rotamer outliers, RNA suiteness
MAXIT	model geometry: symmetry-related too-close contacts, stereochemistry issues, identification of *cis*-peptides
Mogul Update (2018, CSD archive)	model geometry: bond-length and bond-angle outliers in small molecules
Xtriage (Phenix) Update (Phenix 1.13)	crystallographic data and refinement statistics: signal-to-noise, twinning
DCC	crystallographic data and refinement statistics: R, R_{free} fit to crystallographic data: R_{free}
EDS Update (Recmac 7.0v44)	fit to crystallographic data: real-space R outliers
Cyrange	NMR ensemble composition: identification of well-defined protein cores
RCI	NMR chemical shifts: prediction of protein backbone order parameter from chemical shifts
PANAV	NMR chemical shifts: suggested referencing corrections in chemical shift assignments

Percentile statistics reflecting the state of the archive on December 31st 2017.

Structure 25, 1916–1927, 2017

Validation metrics in wwPDB validation reports

X-ray/EM/NMR

- Geometric & conformational
  - bond, angle, planarity
  - protein backbone conformation
  - protein side-chain conformation
- Atomic & molecular interaction
  - all-atom contacts
  - under packing
  - hydrogen bond quality
- Non-protein
  - nucleic acids (RNA pucker, suite)
  - carbohydrates (N-glycan core)
  - ligands (CSD)
  - ions & other solvent
- Incomplete model (e.g. CA_ONLY)

X-ray

- Structure factor & electron density
  - Wilson plot, outliers, tNCS
  - wrong symmetry
  - twinning
  - agreement (R_{free}, RSR, RSCC)

NMR

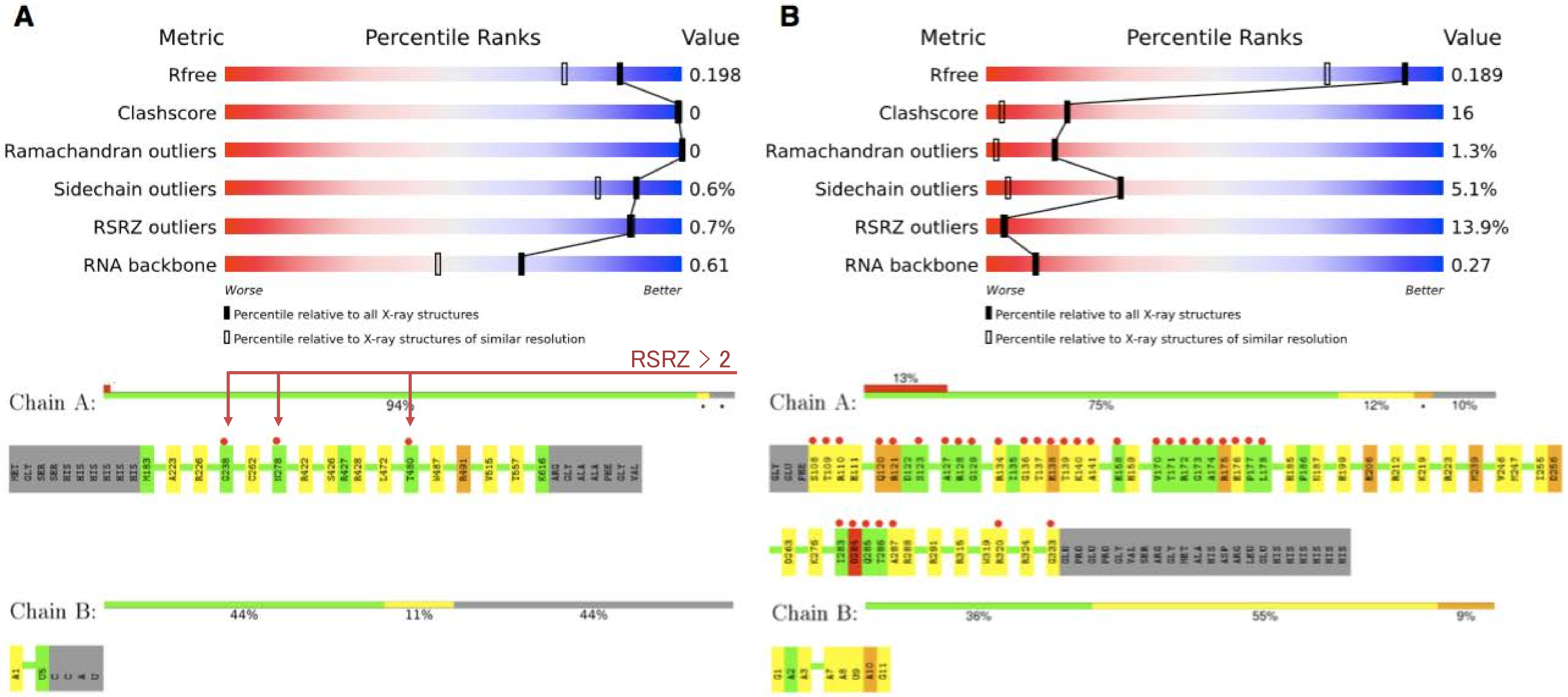
- Chemical shifts
  - completeness
  - outliers
  - estimated reference error
  - random coil index
- Structure ensembles
  - representative model (medoid)
  - domain detection

Caveat:

LLDF (Local Ligand Density Fit) has been replaced by a combination of RSR (Real-space R factor) > 0.4 and RSCC (Real-space correlations coefficient) < 0.8 since this March.

wwPDB validation report PDFs

Summary quality metrics in wwPDB validation reports



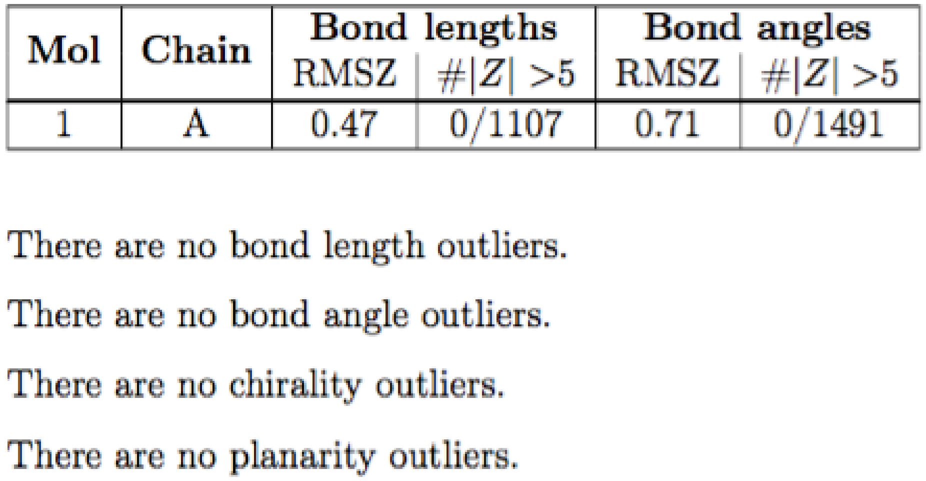
no geometric outliers 1 type of outlier 2 types of outliers 3 types of outliers no coordinates

wwPDB validation PDFs are easily reviewed and shared an assessment of structure quality.

Structure 25, 1916–1927, 2017

wwPDB validation report PDFs

- **Standard geometry**



Mol	Chain	Bond lengths		Bond angles	
		RMSZ	#|Z| >5	RMSZ	#|Z| >5
1	A	0.47	0/1107	0.71	0/1491

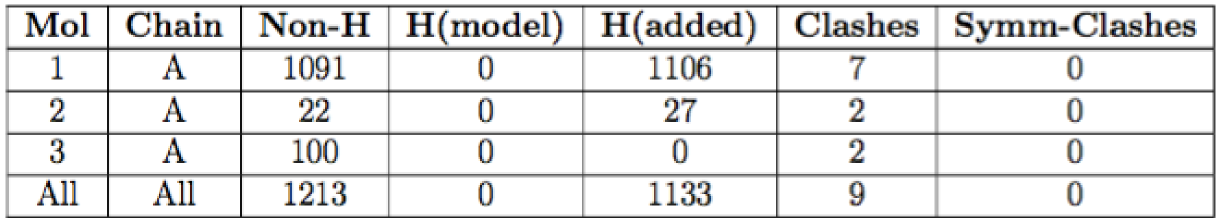
There are no bond length outliers.

There are no bond angle outliers.

There are no chirality outliers.

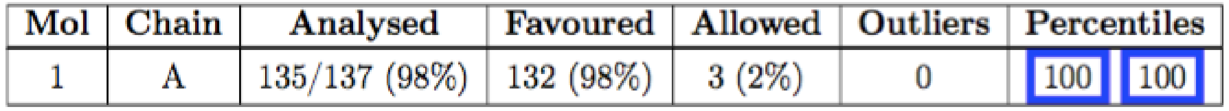
There are no planarity outliers.

- **Too close contacts**



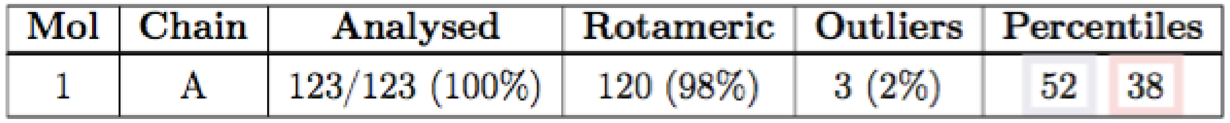
Mol	Chain	Non-H	H(model)	H(added)	Clashes	Symm-Clashes
1	A	1091	0	1106	7	0
2	A	22	0	27	2	0
3	A	100	0	0	2	0
All	All	1213	0	1133	9	0

- **Protein backbones**



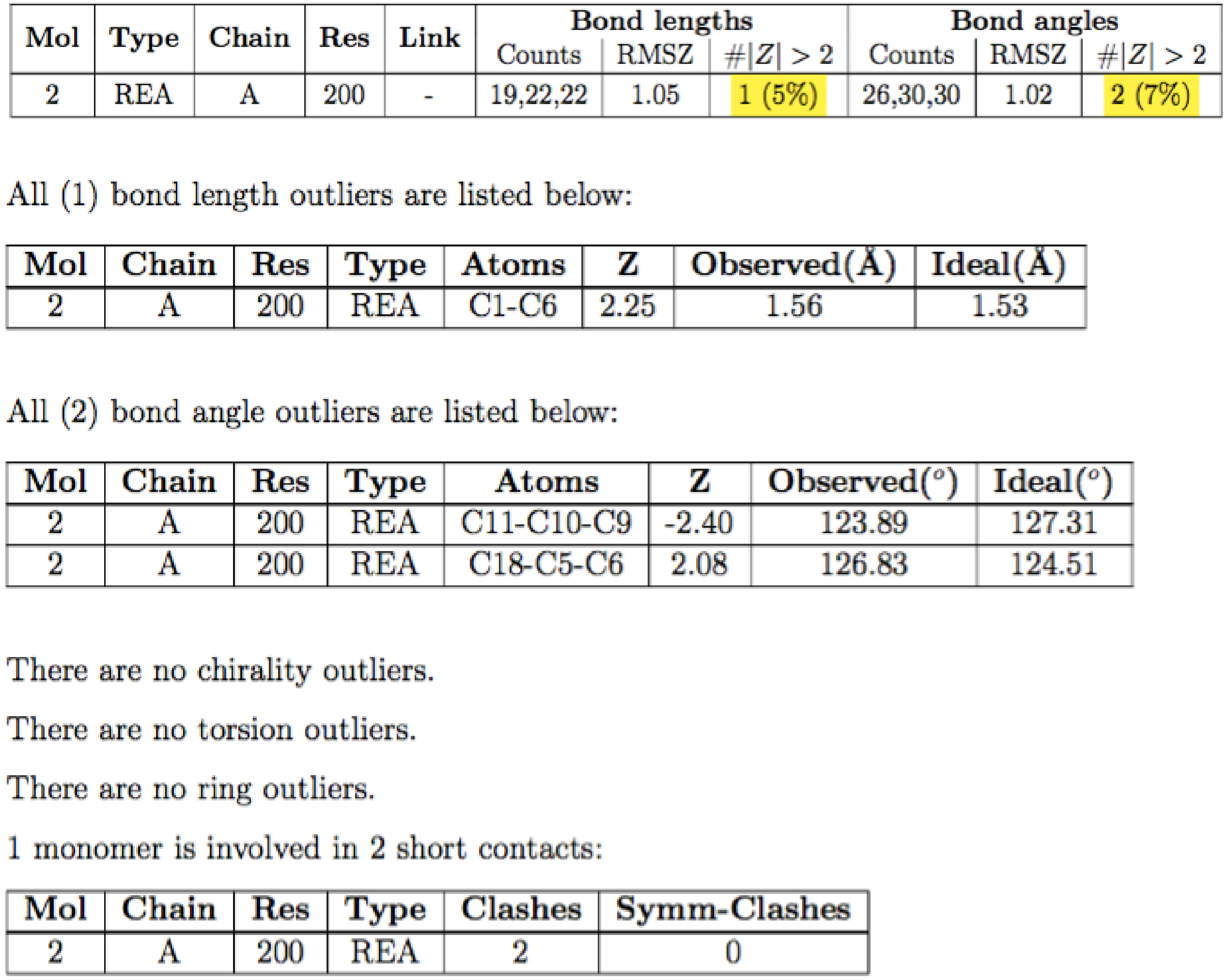
Mol	Chain	Analysed	Favoured	Allowed	Outliers	Percentiles	
1	A	135/137 (98%)	132 (98%)	3 (2%)	0	100	100

- **Protein sidechains**



Mol	Chain	Analysed	Rotameric	Outliers	Percentiles	
1	A	123/123 (100%)	120 (98%)	3 (2%)	52	38

- **Ligand geometry**



Mol	Type	Chain	Res	Link	Bond lengths			Bond angles		
					Counts	RMSZ	#|Z| > 2	Counts	RMSZ	#|Z| > 2
2	REA	A	200	-	19,22,22	1.05	1 (5%)	26,30,30	1.02	2 (7%)

All (1) bond length outliers are listed below:

Mol	Chain	Res	Type	Atoms	Z	Observed(Å)	Ideal(Å)
2	A	200	REA	C1-C6	2.25	1.56	1.53

All (2) bond angle outliers are listed below:

Mol	Chain	Res	Type	Atoms	Z	Observed(°)	Ideal(°)
2	A	200	REA	C11-C10-C9	-2.40	123.89	127.31
2	A	200	REA	C18-C5-C6	2.08	126.83	124.51

There are no chirality outliers.

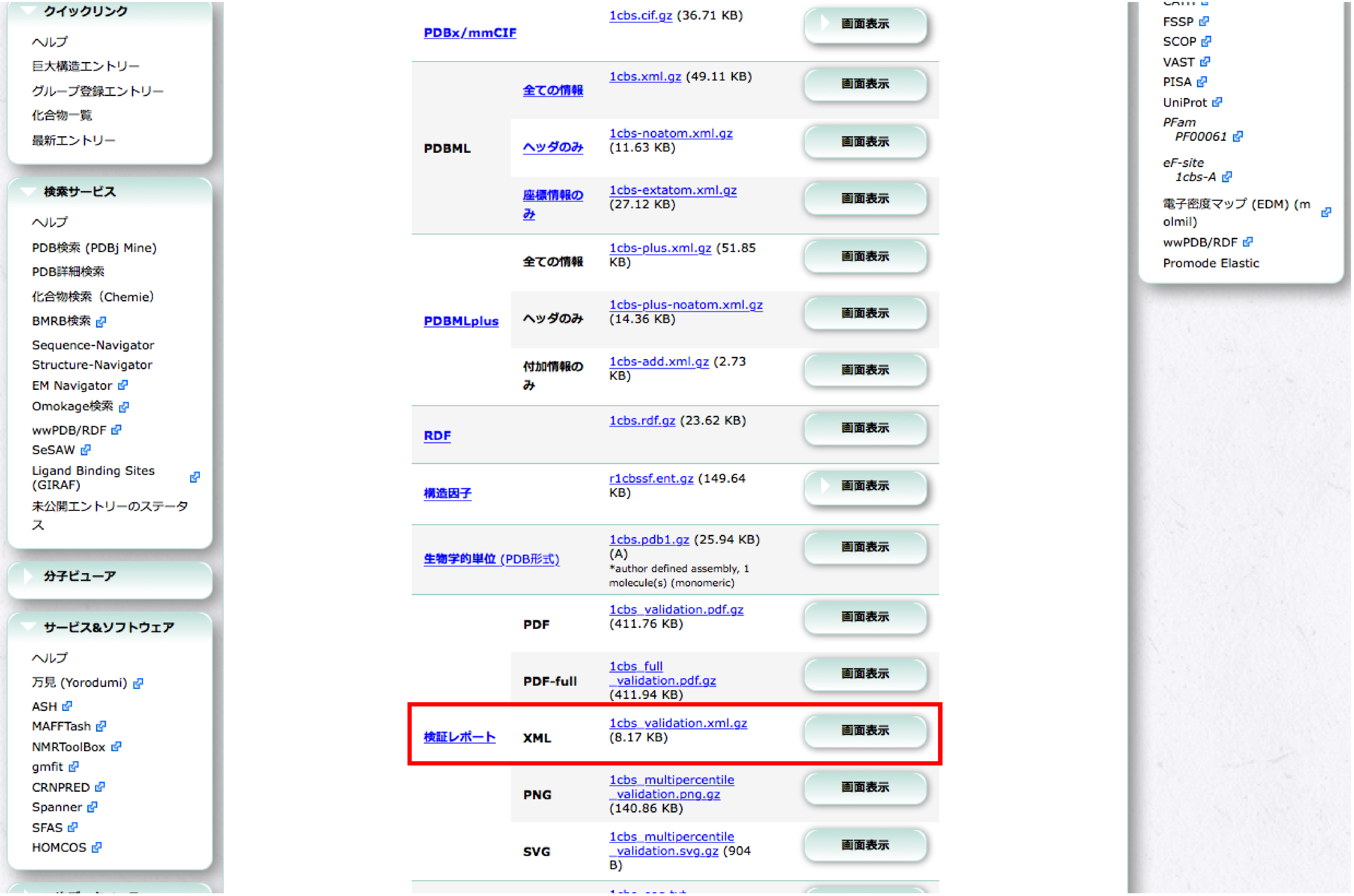
There are no torsion outliers.

There are no ring outliers.

1 monomer is involved in 2 short contacts:

Mol	Chain	Res	Type	Clashes	Symm-Clashes
2	A	200	REA	2	0

Where is the detailed validation data?



クイックリンク

- ヘルプ
- 巨大構造エントリー
- グループ登録エントリー
- 化合物一覧
- 最新エントリー

検索サービス

- ヘルプ
- PDB検索 (PDBj Mine)
- PDB詳細検索
- 化合物検索（Chemie)
- BMRB検索
- Sequence-Navigator
- Structure-Navigator
- EM Navigator
- Omokage検索
- wwPDB/RDF
- SeSAW
- Ligand Binding Sites (GIRAF)
- 未公開エントリーのステータス

分子ビューア

サービス&ソフトウェア

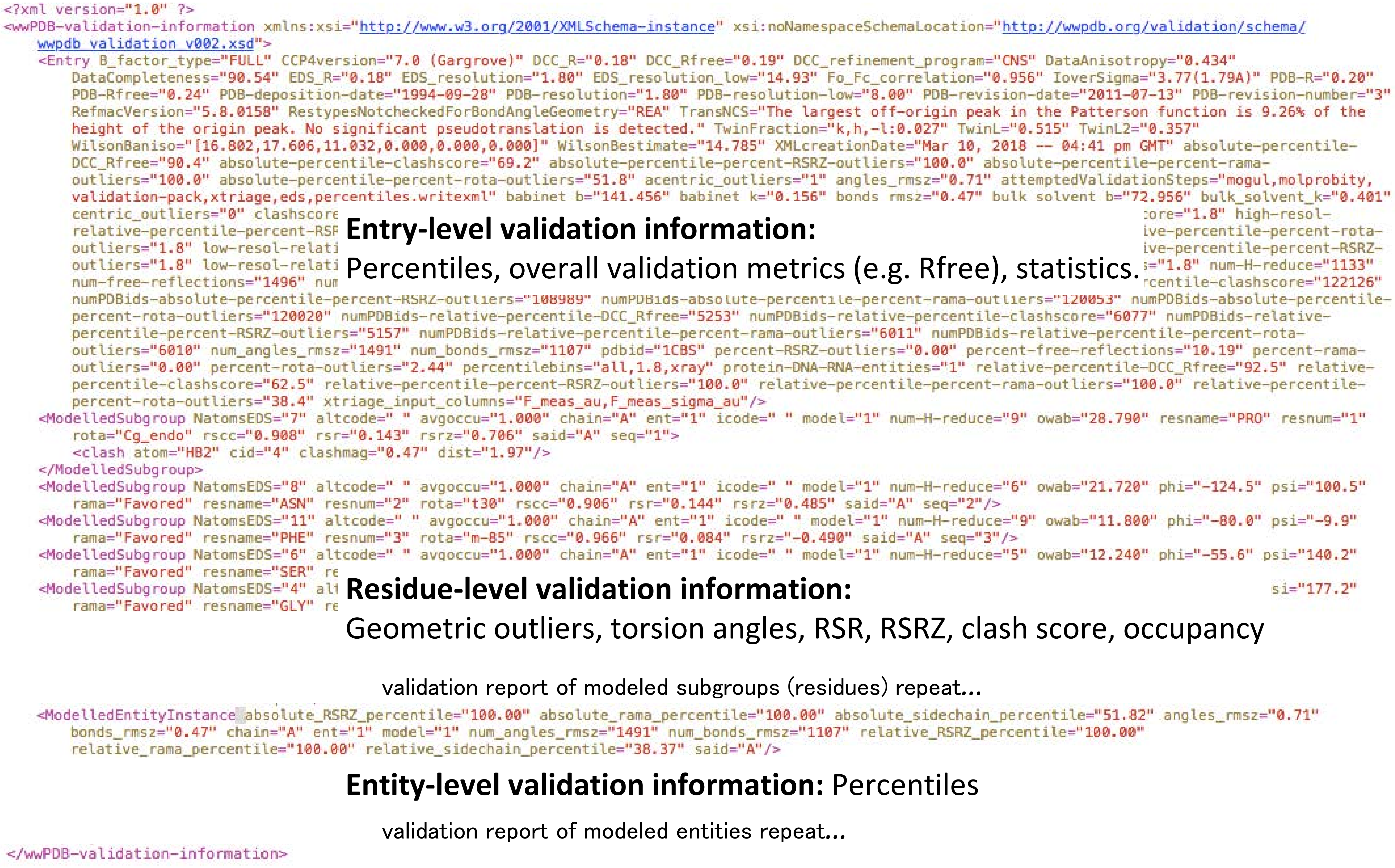
- ヘルプ
- 万見 (Yorodumi)
- ASH
- MAFFTash
- NMRToolBox
- gmfit
- CRNPRED
- Spanner
- SFAS
- HOMCOS

PDBx/mmCIF		1cbs.cif.gz (36.71 KB)	画面表示
PDBML	全ての情報	1cbs.xml.gz (49.11 KB)	画面表示
	ヘッダのみ	1cbs-noatom.xml.gz (11.63 KB)	画面表示
	座標情報のみ	1cbs-extatom.xml.gz (27.12 KB)	画面表示
PDBMLplus	全ての情報	1cbs-plus.xml.gz (51.85 KB)	画面表示
	ヘッダのみ	1cbs-plus-noatom.xml.gz (14.36 KB)	画面表示
	付加情報のみ	1cbs-add.xml.gz (2.73 KB)	画面表示
RDF		1cbs.rdf.gz (23.62 KB)	画面表示
構造因子		r1cbssf.ent.gz (149.64 KB)	画面表示
生物学的単位 (PDB形式)		1cbs.pdb1.gz (25.94 KB) (A) *author defined assembly, 1 molecule(s) (monomeric)	画面表示
検証レポート	PDF	1cbs_validation.pdf.gz (411.76 KB)	画面表示
	PDF-full	1cbs_full_validation.pdf.gz (411.94 KB)	画面表示
	XML	1cbs_validation.xml.gz (8.17 KB)	画面表示
	PNG	1cbs_multipercentile_validation.png.gz (140.86 KB)	画面表示
	SVG	1cbs_multipercentile_validation.svg.gz (904 B)	画面表示

- CATH
- FSSP
- SCOP
- VAST
- PISA
- UniProt
- *PFam*
  PF00061
- *eF-site*
  1cbs-A
- 電子密度マップ (EDM) (molmil)
- wwPDB/RDF
- Promode Elastic

https://pdbj.org

Detailed wwPDB validation reports (XML)



```
<?xml version="1.0" ?>
<wwPDB-validation-information xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:noNamespaceSchemaLocation="http://wwpdb.org/validation/schema/
    wwpdb_validation_v002.xsd">
    <Entry B_factor_type="FULL" CCP4version="7.0 (Gargrove)" DCC_R="0.18" DCC_Rfree="0.19" DCC_refinement_program="CNS" DataAnisotropy="0.434"
        DataCompleteness="90.54" EDS_R="0.18" EDS_resolution="1.80" EDS_resolution_low="14.93" Fo_Fc_correlation="0.956" IoverSigma="3.77(1.79A)" PDB-R="0.20"
        PDB-Rfree="0.24" PDB-deposition-date="1994-09-28" PDB-resolution="1.80" PDB-resolution-low="8.00" PDB-revision-date="2011-07-13" PDB-revision-number="3"
        RefmacVersion="5.8.0158" RestypesNotcheckedForBondAngleGeometry="REA" TransNCS="The largest off-origin peak in the Patterson function is 9.26% of the
        height of the origin peak. No significant pseudotranslation is detected." TwinFraction="k,h,-l:0.027" TwinL="0.515" TwinL2="0.357"
        WilsonBaniso="[16.802,17.606,11.032,0.000,0.000,0.000]" WilsonBestimate="14.785" XMLcreationDate="Mar 10, 2018 -- 04:41 pm GMT" absolute-percentile-
        DCC_Rfree="90.4" absolute-percentile-clashscore="69.2" absolute-percentile-percent-RSRZ-outliers="100.0" absolute-percentile-percent-rama-
        outliers="100.0" absolute-percentile-percent-rota-outliers="51.8" acentric_outliers="1" angles_rmsz="0.71" attemptedValidationSteps="mogul,molprobity,
        validation-pack,xtriage,eds,percentiles.writexml" babinet b="141.456" babinet k="0.156" bonds rmsz="0.47" bulk solvent b="72.956" bulk_solvent_k="0.401"
        centric_outliers="0" clashscore                                                                                   core="1.8" high-resol-
        relative-percentile-percent-RSR                                                                                   ive-percentile-percent-rota-
        outliers="1.8" low-resol-relati                                                                                   ive-percentile-percent-RSRZ-
        outliers="1.8" low-resol-relati                                                                                   ="1.8" num-H-reduce="1133"
        num-free-reflections="1496" num                                                                                   rcentile-clashscore="122126"
        numPDBids-absolute-percentile-percent-RSRZ-outliers="108989" numPDBids-absolute-percentile-percent-rama-outliers="120053" numPDBids-absolute-percentile-
        percent-rota-outliers="120020" numPDBids-relative-percentile-DCC_Rfree="5253" numPDBids-relative-percentile-clashscore="6077" numPDBids-relative-
        percentile-percent-RSRZ-outliers="5157" numPDBids-relative-percentile-percent-rama-outliers="6011" numPDBids-relative-percentile-percent-rota-
        outliers="6010" num_angles_rmsz="1491" num_bonds_rmsz="1107" pdbid="1CBS" percent-RSRZ-outliers="0.00" percent-free-reflections="10.19" percent-rama-
        outliers="0.00" percent-rota-outliers="2.44" percentilebins="all,1.8,xray" protein-DNA-RNA-entities="1" relative-percentile-DCC_Rfree="92.5" relative-
        percentile-clashscore="62.5" relative-percentile-percent-RSRZ-outliers="100.0" relative-percentile-percent-rama-outliers="100.0" relative-percentile-
        percent-rota-outliers="38.4" xtriage_input_columns="F_meas_au,F_meas_sigma_au"/>
```

Entry-level validation information:

Percentiles, overall validation metrics (e.g. Rfree), statistics.

```
    <ModelledSubgroup NatomsEDS="7" altcode=" " avgoccu="1.000" chain="A" ent="1" icode=" " model="1" num-H-reduce="9" owab="28.790" resname="PRO" resnum="1"
        rota="Cg_endo" rscc="0.908" rsr="0.143" rsrz="0.706" said="A" seq="1">
        <clash atom="HB2" cid="4" clashmag="0.47" dist="1.97"/>
    </ModelledSubgroup>
    <ModelledSubgroup NatomsEDS="8" altcode=" " avgoccu="1.000" chain="A" ent="1" icode=" " model="1" num-H-reduce="6" owab="21.720" phi="-124.5" psi="100.5"
        rama="Favored" resname="ASN" resnum="2" rota="t30" rscc="0.906" rsr="0.144" rsrz="0.485" said="A" seq="2"/>
    <ModelledSubgroup NatomsEDS="11" altcode=" " avgoccu="1.000" chain="A" ent="1" icode=" " model="1" num-H-reduce="9" owab="11.800" phi="-80.0" psi="-9.9"
        rama="Favored" resname="PHE" resnum="3" rota="m-85" rscc="0.966" rsr="0.084" rsrz="-0.490" said="A" seq="3"/>
    <ModelledSubgroup NatomsEDS="6" altcode=" " avgoccu="1.000" chain="A" ent="1" icode=" " model="1" num-H-reduce="5" owab="12.240" phi="-55.6" psi="140.2"
        rama="Favored" resname="SER" re
    <ModelledSubgroup NatomsEDS="4" alt                                                                                 si="177.2"
        rama="Favored" resname="GLY" re
```

Residue-level validation information:

Geometric outliers, torsion angles, RSR, RSRZ, clash score, occupancy

validation report of modeled subgroups (residues) repeat...

```
    <ModelledEntityInstance absolute_RSRZ_percentile="100.00" absolute_rama_percentile="100.00" absolute_sidechain_percentile="51.82" angles_rmsz="0.71"
        bonds_rmsz="0.47" chain="A" ent="1" model="1" num_angles_rmsz="1491" num_bonds_rmsz="1107" relative_RSRZ_percentile="100.00"
        relative_rama_percentile="100.00" relative_sidechain_percentile="38.37" said="A"/>
```

Entity-level validation information: Percentiles

validation report of modeled entities repeat...

```
</wwPDB-validation-information>
```

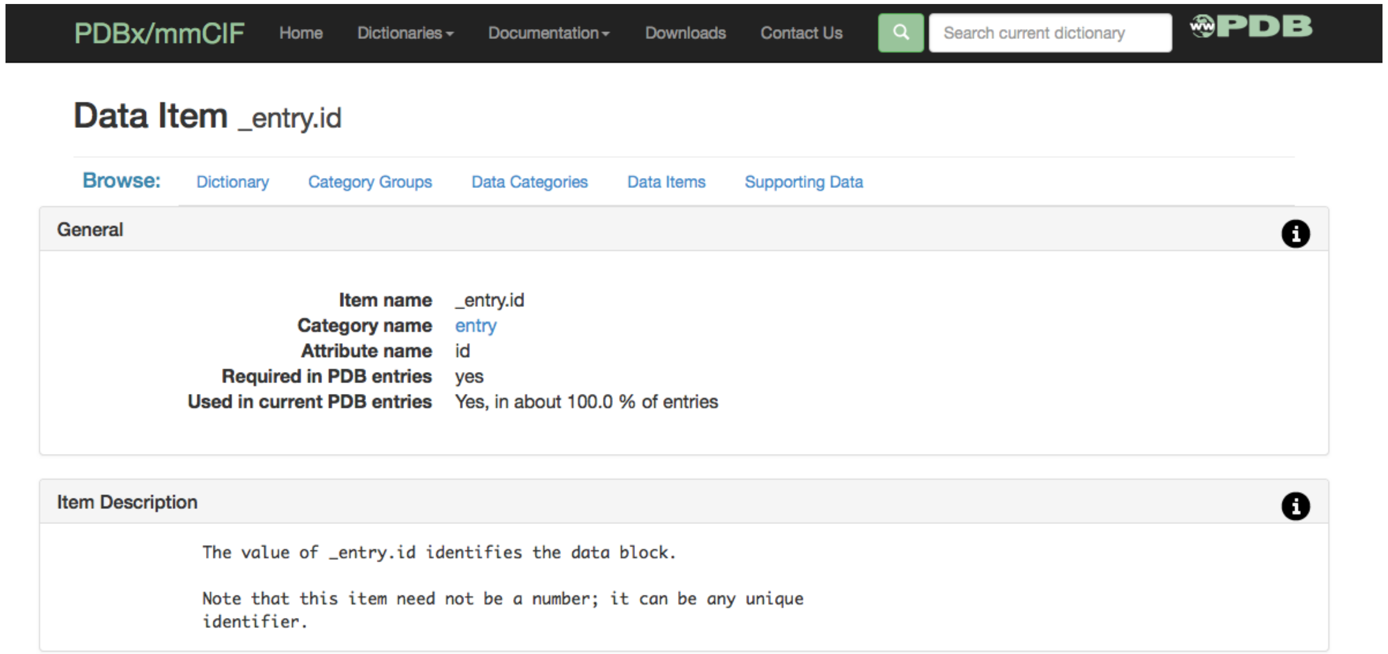
Known issues on wwPDB validation reports (XML)

- Incompatible naming with the PDB’s manner (PDBx/mmCIF).
  - said -> _atom_site.auth_asym_id
  - cid -> _pdbx_validate_close_contact.id
- Fat attributes
  - It leads to an inefficient search, where program always travels all instances, regardless of whether data exists or not.
- No categories
  - All metrics are tied to entry, entity, and monomer. It is simple, but …
  - It needs tweak for description for
    - angles between adjacent monomers
    - steric collision between entities, or different asymmetric units
    - NMR ensemble structure model
- Name collision
  - For example, ‘value’ data item indicates one of either chemical shift outlier, random coil index, referencing offset, unparsed chemical shift or unmapped chemical shift.

Canonical representations of PDB archives

	PDBx/mmCIF	PDBML	wwPDB/RDF
Coordinate	yes	yes or no	–
Metadata (entity, citation, …)	yes	yes	yes
Human-readability	yes	–	–
Searchable	no standard (visit PDBj sites)	yes (XQuery)	yes (SPARQL)
URI	–	–	supported
Purpose	data processing	data exchange	knowledge sharing
Example	_entry.id 1CBS	<PDBx:entry id=“1CBS"/>	<PDBo:entry.id>1CBS</PDBo:entry.id>

All PDB archives use the same categories and items defined in the PDBx/mmCIF Dictionary.



Data Item _entry.id

General

Item name	_entry.id
Category name	entry
Attribute name	id
Required in PDB entries	yes
Used in current PDB entries	Yes, in about 100.0 % of entries

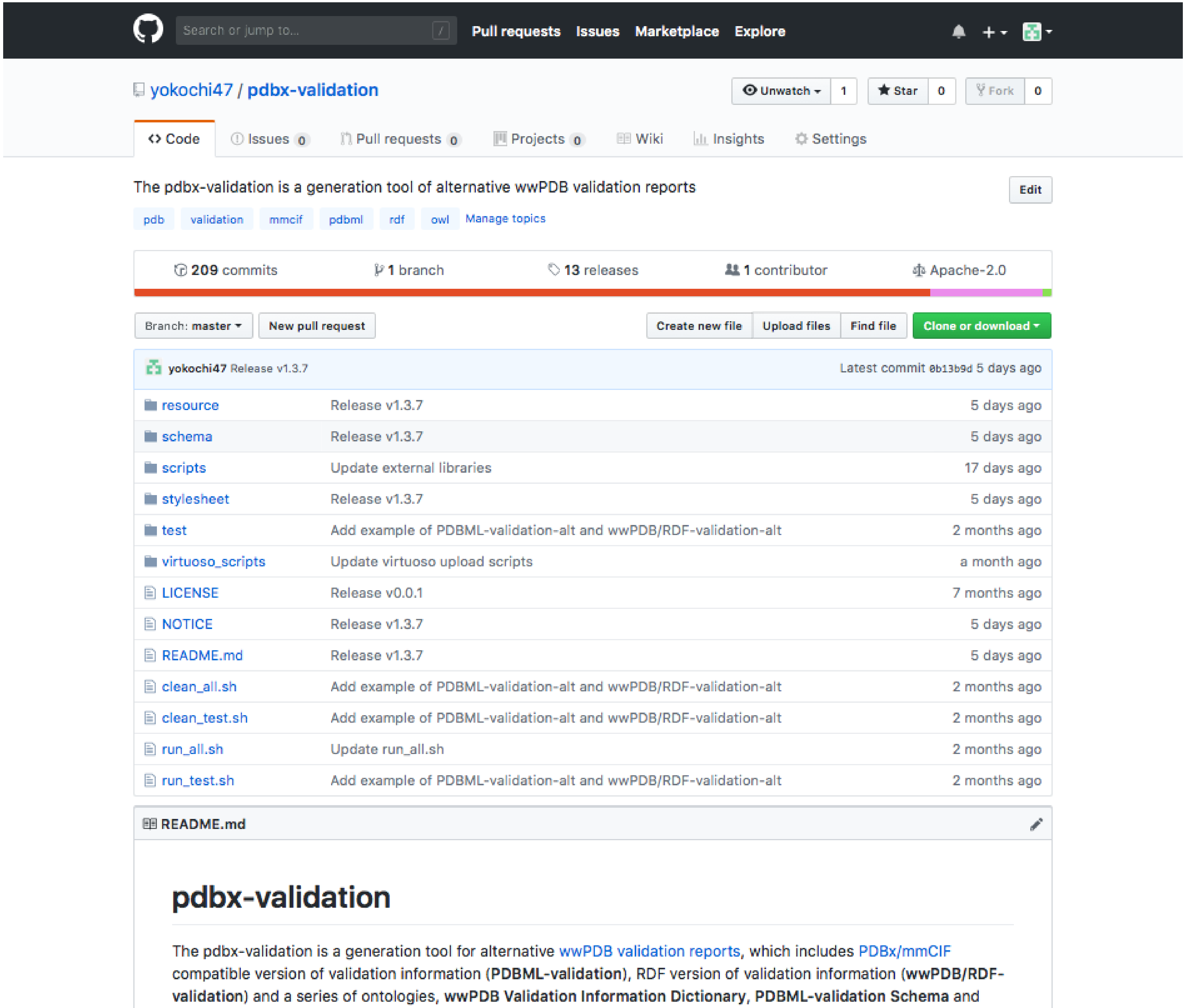
Item Description

```
The value of _entry.id identifies the data block.

Note that this item need not be a number; it can be any unique
identifier.
```

http://mmcif.wwpdb.org

Reorganization of wwPDB validation reports in manner of the PDBx/mmCIF



yokochi47 / pdbx-validation

The pdbx-validation is a generation tool of alternative wwPDB validation reports

209 commits | 1 branch | 13 releases | 1 contributor | Apache-2.0

yokochi47 Release v1.3.7 — Latest commit 0b13b9d 5 days ago

resource	Release v1.3.7	5 days ago
schema	Release v1.3.7	5 days ago
scripts	Update external libraries	17 days ago
stylesheet	Release v1.3.7	5 days ago
test	Add example of PDBML-validation-alt and wwPDB/RDF-validation-alt	2 months ago
virtuoso_scripts	Update virtuoso upload scripts	a month ago
LICENSE	Release v0.0.1	7 months ago
NOTICE	Release v1.3.7	5 days ago
README.md	Release v1.3.7	5 days ago
clean_all.sh	Add example of PDBML-validation-alt and wwPDB/RDF-validation-alt	2 months ago
clean_test.sh	Add example of PDBML-validation-alt and wwPDB/RDF-validation-alt	2 months ago
run_all.sh	Update run_all.sh	2 months ago
run_test.sh	Add example of PDBML-validation-alt and wwPDB/RDF-validation-alt	2 months ago

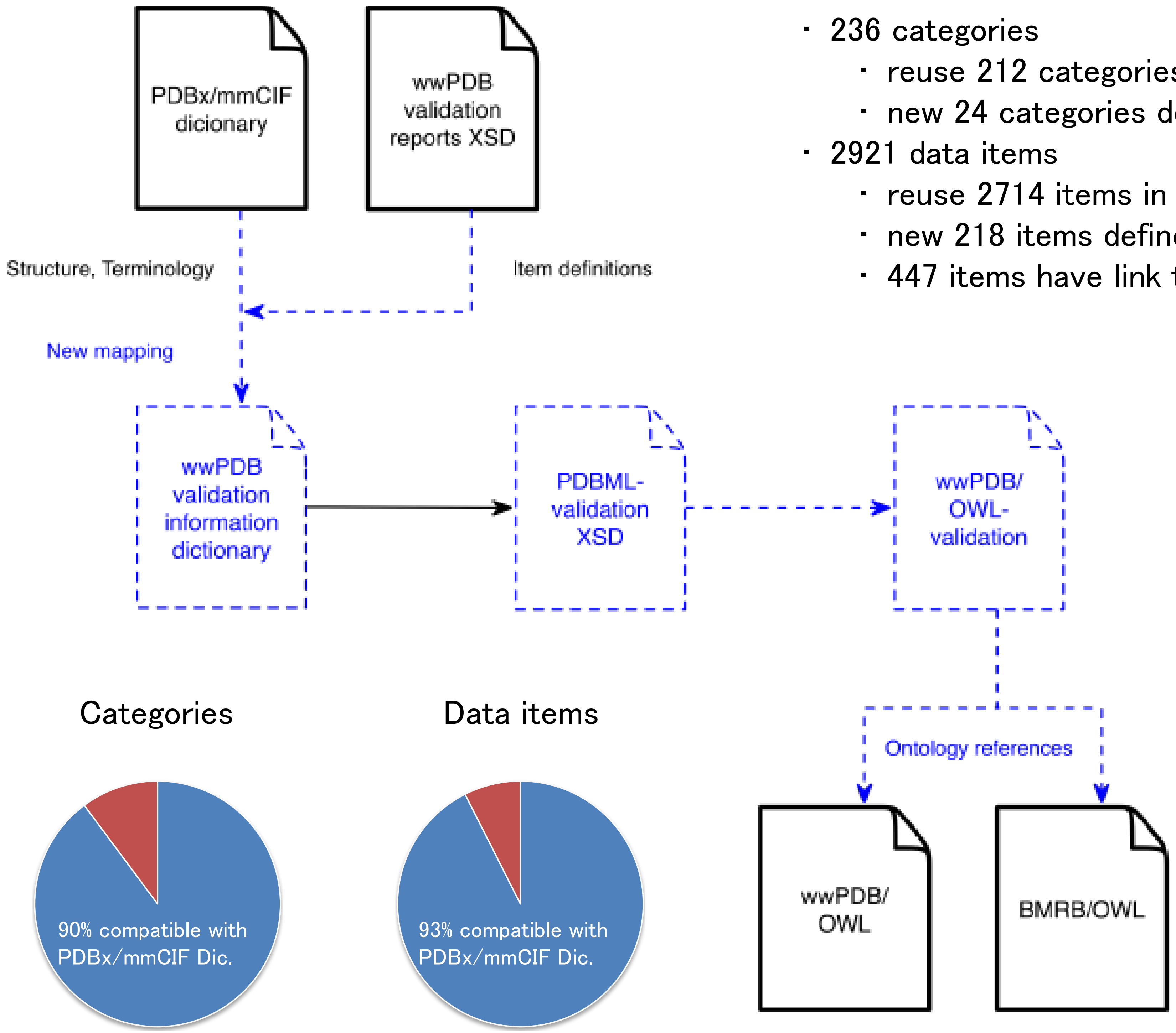
README.md

pdbx-validation

The pdbx-validation is a generation tool for alternative wwPDB validation reports, which includes PDBx/mmCIF compatible version of validation information (**PDBML-validation**), RDF version of validation information (**wwPDB/RDF-validation**) and a series of ontologies, **wwPDB Validation Information Dictionary**, **PDBML-validation Schema** and

https://github.com/yokochi47/pdbx-validation

Reorganization of wwPDB validation reports in manner of the PDBx/mmCIF



wwPDB Validation Information Dictionary, PDBML-validation Schema (XSD)

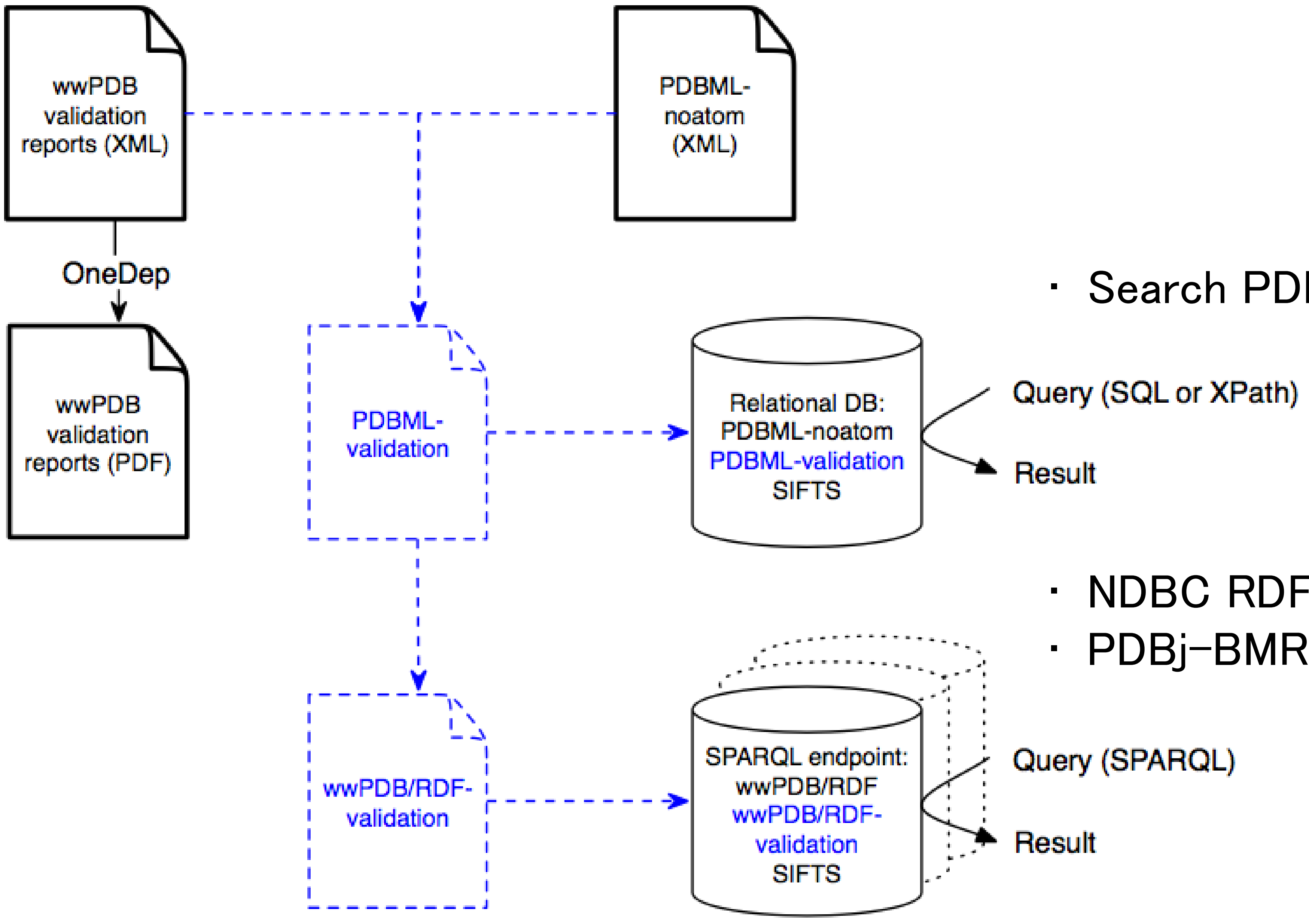
- 236 categories
  - reuse 212 categories in PDBx/mmCIF Dic.
  - new 24 categories defined
- 2921 data items
  - reuse 2714 items in PDBx/mmCIF Dic.
  - new 218 items defined
  - 447 items have link to validation reports' XSD

wwPDB/OWL-validation

- As for relation with wwPDB/OWL
  - 25 same classes
  - 427 equivalent classes
  - 3795 equivalent properties
- As for relation with BMRB/OWL
  - 134 equivalent properties

https://github.com/yokochi47/pdbx-validation

Semantic extension of the wwPDB validation reports and planned Web applications



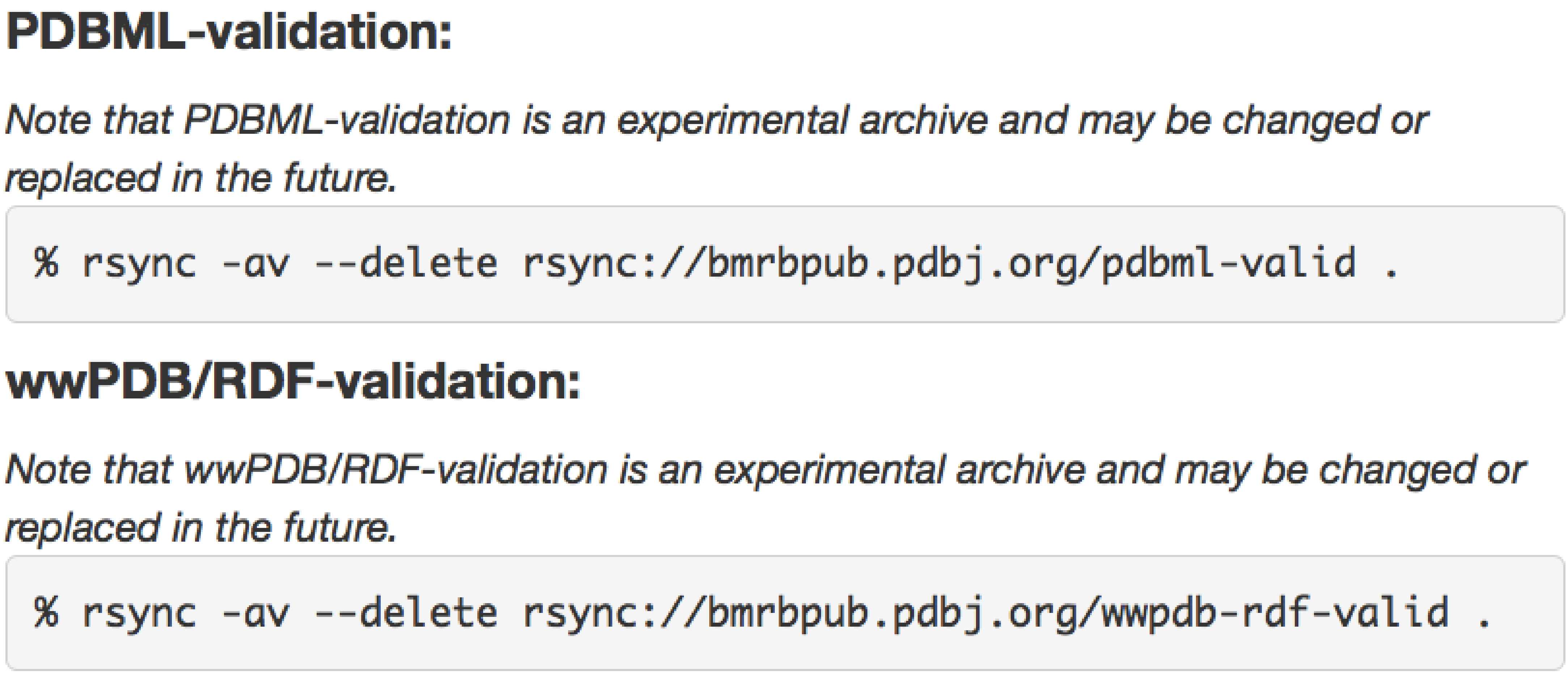
wwPDB validation reports (XML)
PDBML-noatom (XML)
OneDep
wwPDB validation reports (PDF)
PDBML-validation
Relational DB: PDBML-noatom PDBML-validation SIFTS
Query (SQL or XPath)
Result
wwPDB/RDF-validation
SPARQL endpoint: wwPDB/RDF wwPDB/RDF-validation SIFTS
Query (SPARQL)
Result

- Search PDB (PDBj Mine)

- NDBC RDF-portal
- PDBj-BMRB endpoint

Search PDB (PDBj-Mine)	https://pdbj.org
NDBC RDF-portal	https://integbio.jp/rdf/
PDBj-BMRB endpoint	https://bmrbpub.pdbj.org

PDBML-validation and wwPDB/RDF-validation



PDBML-validation:

Note that PDBML-validation is an experimental archive and may be changed or replaced in the future.

```
% rsync -av --delete rsync://bmrbpub.pdbj.org/pdbml-valid .
```

wwPDB/RDF-validation:

Note that wwPDB/RDF-validation is an experimental archive and may be changed or replaced in the future.

```
% rsync -av --delete rsync://bmrbpub.pdbj.org/wwpdb-rdf-valid .
```

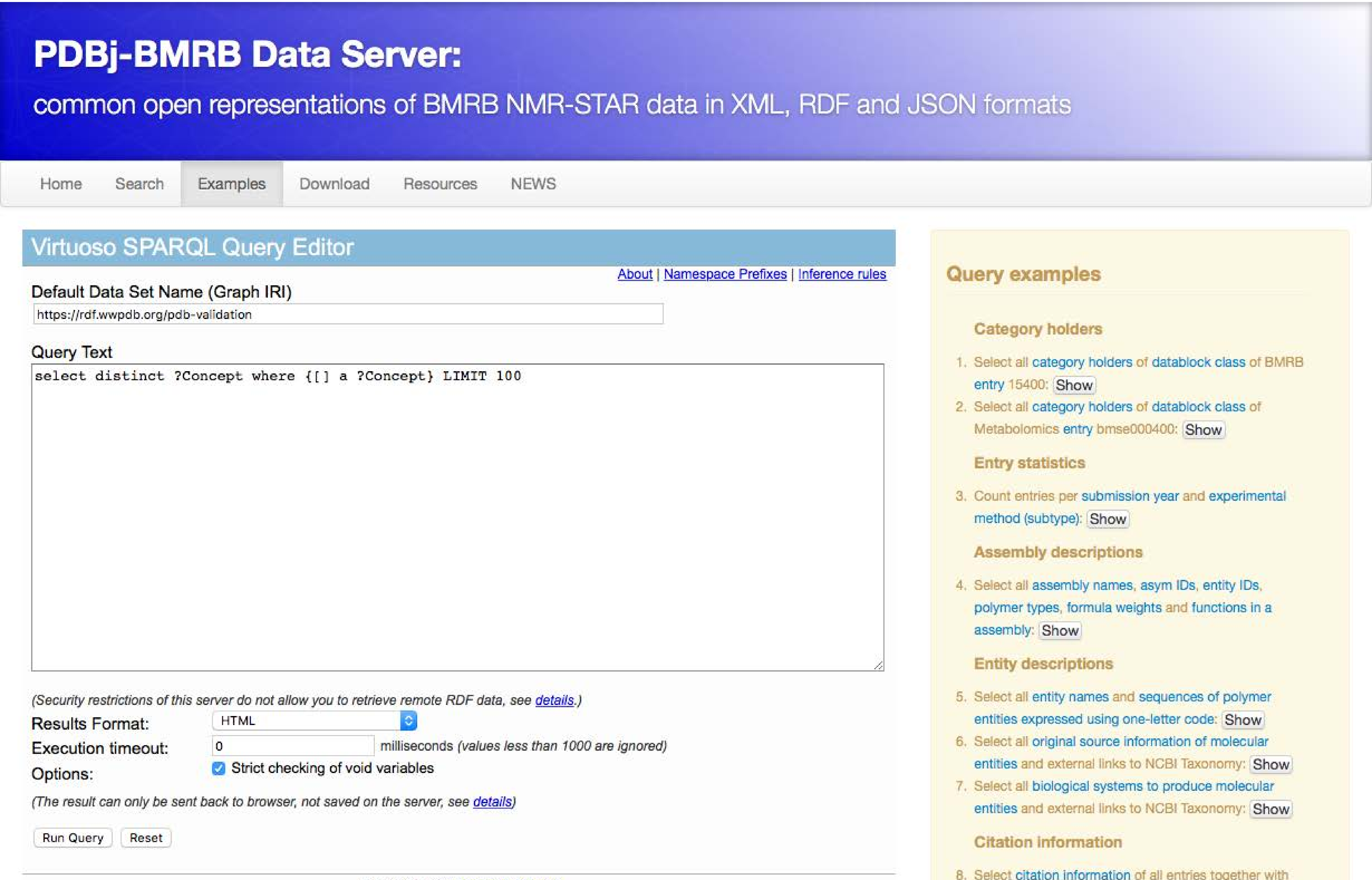
PostgreSQL dump image:

category	description	size (GB)
pdbx_dcc_map	output of MAPMAN used by DCC (RSR, RSCC, LLDF)	10
pdbx_poly_seq_scheme	residue nomenclature mapping for polymer entities	8
struct_mon_prot	structure properties of a protein	6
entity_poly_seq	sequence of monomers in a polymer	2
pdbx_validate_close_contact	close contact with regard to the distance expected	2

https://bmrbpub.pdbj.org

SPARQL endpoint contains wwPDB/RDF-validation graph

https://bmrbpub.pdbj.org



PDBj-BMRB Data Server:

common open representations of BMRB NMR-STAR data in XML, RDF and JSON formats

Virtuoso SPARQL Query Editor

About | Namespace Prefixes | Inference rules

Default Data Set Name (Graph IRI)

https://rdf.wwpdb.org/pdb-validation

Query Text

```
select distinct ?Concept where {[] a ?Concept} LIMIT 100
```

(Security restrictions of this server do not allow you to retrieve remote RDF data, see details.)

Results Format: HTML

Execution timeout: 0 milliseconds *(values less than 1000 are ignored)*

Options: ☑ Strict checking of void variables

(The result can only be sent back to browser, not saved on the server, see details)

Query examples

Category holders

1. Select all category holders of datablock class of BMRB entry 15400: Show
2. Select all category holders of datablock class of Metabolomics entry bmse000400: Show

Entry statistics

3. Count entries per submission year and experimental method (subtype): Show

Assembly descriptions

4. Select all assembly names, asym IDs, entity IDs, polymer types, formula weights and functions in a assembly: Show

Entity descriptions

5. Select all entity names and sequences of polymer entities expressed using one-letter code: Show
6. Select all original source information of molecular entities and external links to NCBI Taxonomy: Show
7. Select all biological systems to produce molecular entities and external links to NCBI Taxonomy: Show

Citation information

8. Select citation information of all entries together with

Example #1: Search wwPDB/RDF-validation with SPARQL

Search all enzyme-ligand complexes of which real space R-factor (RSR) of ligand is less than 10%. (showing only essential part of about 30 line-SPARQL query)

```
PREFIX PDBov: <https://rdf.wwpdb.org/schema/pdbx-validation-v1.owl#>
SELECT ?PDB_ID ?enzyme ?ligand ?comp_id MIN(?RSR AS ?minRSR)
FROM <http://rdf.wwpdb.org/pdb-validation>
WHERE {
 ?entity PDBov:link_to_enzyme ?link_to_enzyme ;
        PDBov:entity.pdbx_description ?enzyme ;
        PDBov:of_datablock ?datablock .

 BIND (SUBSTR(STR(?datablock),38,4) AS ?PDB_ID)


…

 FILTER (?ligand != "water" && !STRENDS(?ligand, "ION"))

…

 ?dcc_map PDBov:pdbx_dcc_map.auth_asym_id ?asym_id ;
         PDBov:pdbx_dcc_map.auth_comp_id ?comp_id ;
         PDBov:pdbx_dcc_map.RSR ?RSR .

 FILTER (xsd:float(?RSR) < 0.1)

} GROUP BY ?PDB_ID ?enzyme ?ligand ?comp_id
```

✍ selection of enzyme

✍ ligand selection: non-polymer,
not water, not ion

✍ RSR ＜ 0.1

Example #1: Search wwPDB/RDF-validation with SPARQL

Found 15k pairs of enzyme-ligand complexes of which real space R-factor (RSR) of ligand is less than 10%.

PDB ID, enzyme name, ligand name, ligand (3-letters code), minimum RSR value of the ligand

```
"4CK1","INTEGRASE","(4-CARBOXY-1,3-BENZODIOXOL-5-YL)METHYL-[[2-[(4-
METHOXYPHENYL)METHYLCARBAMOYL]PHENYL]METHYL]AZANIUM","OM1","0.081"
"2IOD","Dihydroflavonol 4-reductase","NADP NICOTINAMIDE-ADENINE-DINUCLEOTIDE
PHOSPHATE","NAP","0.091"
"5BYR","Iron hydrogenase 1","FE2/S2 (INORGANIC) CLUSTER","FES","0.096"
"2PU0","Enolase","PHOSPHONOACETOHYDROXAMIC ACID","PAH","0.075"
"4LV2","Beta-lactamase","[1-(6-chloropyrimidin-4-yl)-1H-pyrazol-4-yl]boronic acid","N95","0.083"
"3OLE","Pancreatic alpha-amylase","ALPHA-D-GLUCOSE","GLC","0.084"
"4MOR","Pyranose 2-oxidase","DODECAETHYLENE GLYCOL","12P","0.093"
"4FKX","Nucleoside diphosphate kinase","CYTIDINE-5'-DIPHOSPHATE","CDP","0.073"
"4JPU","Cytochrome c peroxidase","PROTOPORPHYRIN IX CONTAINING FE","HEM","0.094"
"1GTV","THYMIDYLATE KINASE","THYMIDINE-5'-DIPHOSPHATE","TYD","0.090"
"5IA2","7-(5-hydroxy-2-methylphenyl)-8-(2-methoxyphenyl)-1-methyl-1H-imidazo[2,1-f]purine-
2,4(3H,8H)-dione","7-(5-hydroxy-2-methylphenyl)-8-(2-methoxyphenyl)-1-methyl-1H-imidazo[2,1-
f]purine-2,4(3H,8H)-dione","L66","0.077"
"1NFQ","Putative oxidoreductase Rv2002","1,4-DIHYDRONICOTINAMIDE ADENINE
DINUCLEOTIDE","NAI","0.090”
```

Example #2: Search wwPDB/RDF-validation with SPARQL

Search all enzyme–ligand complexes of which percentage of outlier in real space R–factor, defined by Z–score (RSRZ) is larger than 2, of enzyme is less than 1%. (showing only essential part of about 20 line–SPARQL query)

```
PREFIX PDBov: <https://rdf.wwpdb.org/schema/pdbx-validation-v1.owl#>
SELECT ?PDB_ID ?Enzyme (GROUP_CONCAT(?Ligand; SEPARATOR=",") AS ?Ligands)
?RSRZ_outliers_percent
FROM <http://rdf.wwpdb.org/pdb-validation>
WHERE {
 ?map_overall PDBov:pdbx_dcc_map_overall.entry_id ?PDB_ID ;
            PDBov:pdbx_dcc_map_overall.RSRZ_outliers_percent ?RSRZ_outliers_percent .

 FILTER (xsd:float(?RSRZ_outliers_percent) < 0.01)

 BIND (IRI(CONCAT("https://rdf.wwpdb.org/pdb-validation/", ?PDB_ID, "/entityCategory")) AS
?entity_category)

 ?entity_category PDBov:has_entity ?entity .

 ?entity PDBov:link_to_enzyme ?link_to_enzume ;
        PDBov:entity.pdbx_description ?Enzyme .

…

}
```

✍ % of outliers in RSRZ < 1%

✍selection of enzyme

✍ligand selection (omission)

Example #2: Search wwPDB/RDF-validation with SPARQL

Found 5k pairs of enzyme-ligand complexes of which percentage of outlier in real space R-factor (RSRZ) of enzyme is less than 1%, 1k pairs for 0%.

PDB ID, enzyme name, ligand name, percentage of outliers in RSR value of the enzyme

```
"1BUL","NMC-A BETA-LACTAMASE","2-(1-CARBOXY-2-HYDROXY-2-METHYL-PROPYL)-5,5-DIMETHYL-
THIAZOLIDINE-4-CARBOXYLIC ACID,2-(N-MORPHOLINO)-ETHANESULFONIC ACID","0.00"
"5A1G","S-ADENOSYLMETHIONINE SYNTHASE ISOFORM TYPE-
2","(DIPHOSPHONO)AMINOPHOSPHONIC ACID,[(3S)-3-amino-3-carboxypropyl]{[(2S,3S,4R,5R)-5-(6-
amino-9H-purin-9-yl)-3,4-dihydroxytetrahydrofuran-2-yl]methyl}ethylsulfonium,(4S)-2-METHYL-2,4-
PENTANEDIOL,IMIDAZOLE","0.00"
"2DRS","Xylanase Y","GLYCEROL","0.00"
"2AS1","Cytochrome c peroxidase, mitochondrial","PROTOPORPHYRIN IX CONTAINING
FE,THIOPHENE-3-CARBOXIMIDAMIDE","0.00"
"1H4W","TRYPSIN IVA","BENZAMIDINE","0.00"
"142L","T4 LYSOZYME","BETA-MERCAPTOETHANOL","0.00"
"4CIK","PLASMINOGEN","5-[(2R,4S)-2-(phenylmethyl)piperidin-4-yl]-1,2-oxazol-3-one","0.00"
"4L4O","Endo-1,4-beta-xylanase","TRIS-HYDROXYMETHYL-METHYL-AMMONIUM","0.00"
"4G5P","Epidermal growth factor receptor","N-{4-[(3-chloro-4-fluorophenyl)amino]-7-[(3S)-
tetrahydrofuran-3-yloxy]quinazolin-6-yl}-4-(dimethylamino)butanamide","0.00"
"3GA6","Exodeoxyribonuclease","GLYCEROL","0.00"
```

Semantic extension of the wwPDB validation reports

	PDF	XML	PDBML-validation	wwPDB/RDF-validation
Human-readability	yes	–	–	–
Validation information	summary	full	full	full
Searchable	no	yes (XQuery)	yes (SQL, XQuery)	yes (SPARQL)
PDBx/mmCIF	–	–	~90% compatible	~90% compatible
URI	–	–	–	supported
Purpose	peer review	data exchange	quick search, data exchange	knowledge sharing

⇕

wwPDB/RDF*, PDBj-SIFTS*

*Nucleic Acids Research, 2017, Vol. 45, D282-D288

SIFTS: Structure Integration with Function, Taxonomy and Sequences resource

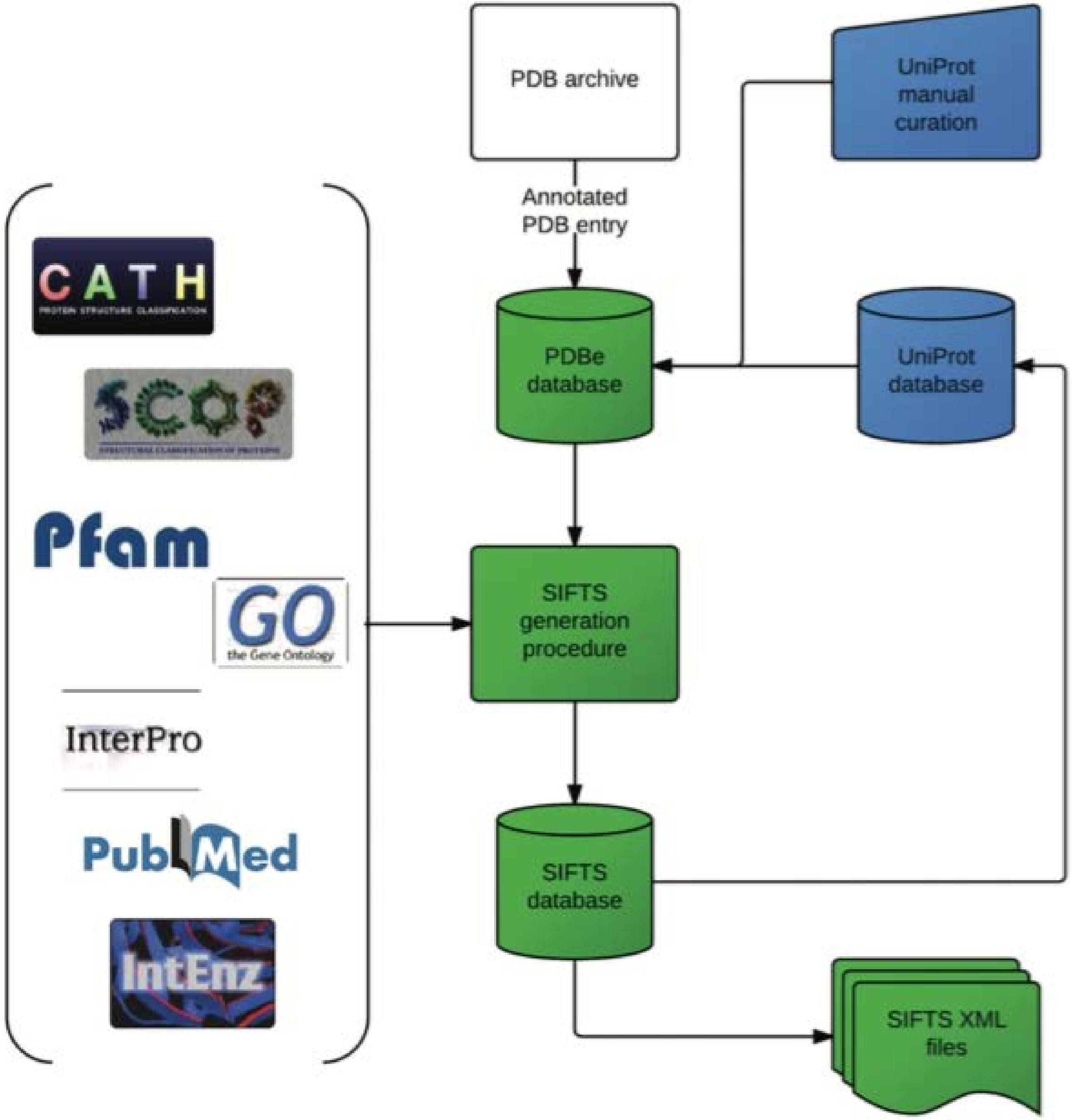


Figure 1. The SIFTS pipeline combines manual and automated processes to produce up-to-date residue-level mappings between proteins in the PDB and their corresponding UniProtKB entry. The pipeline also enriches the annotations of proteins in the PDB by adding data from other biological resources. The SIFTS data are distributed in XML format.

Nucleic Acids Research, 2013, Vol. 41, D483–D489
Nucleic Acids Research, 2017, Vol. 45, D282–D288

Summary

- Pointed out problems of the current wwPDB validation reports when used as bulk data.
- A proposal of semantically enhanced version of wwPDB validation reports, which is highly compatible with the PDB’s assets (PDBx/mmCIF manner).
- PDBML-validation, wwPDB/RDF-validation archives are available.
- Release of PostgreSQL database snapshot dedicated to the validation reports.

Coming soon…

- Official release from PDBj’s FTP and wwPDB/RDF servers
  https://pdbj.org https://rdf.wwpdb.org
- Integration of the validation reports into PDBj Mine, and PDBj-BMRB search service.
- Preparation of SPARQL queries interplaying with wwPDB/RDF, PDBj-SIFTS, and so on.

Download archives, Feedback, Development, …

https://github.com/yokochi47/pdbx-validation

Nucleic Acids Research, 2017, Vol. 45, D282-D288
Protein Science, 2018, Vol. 27, 95-120