mmCIF in Structural Bioinformatics

John Westbrook

Rutgers, The State University of New Jersey

25–29 August 2013

ECM 28

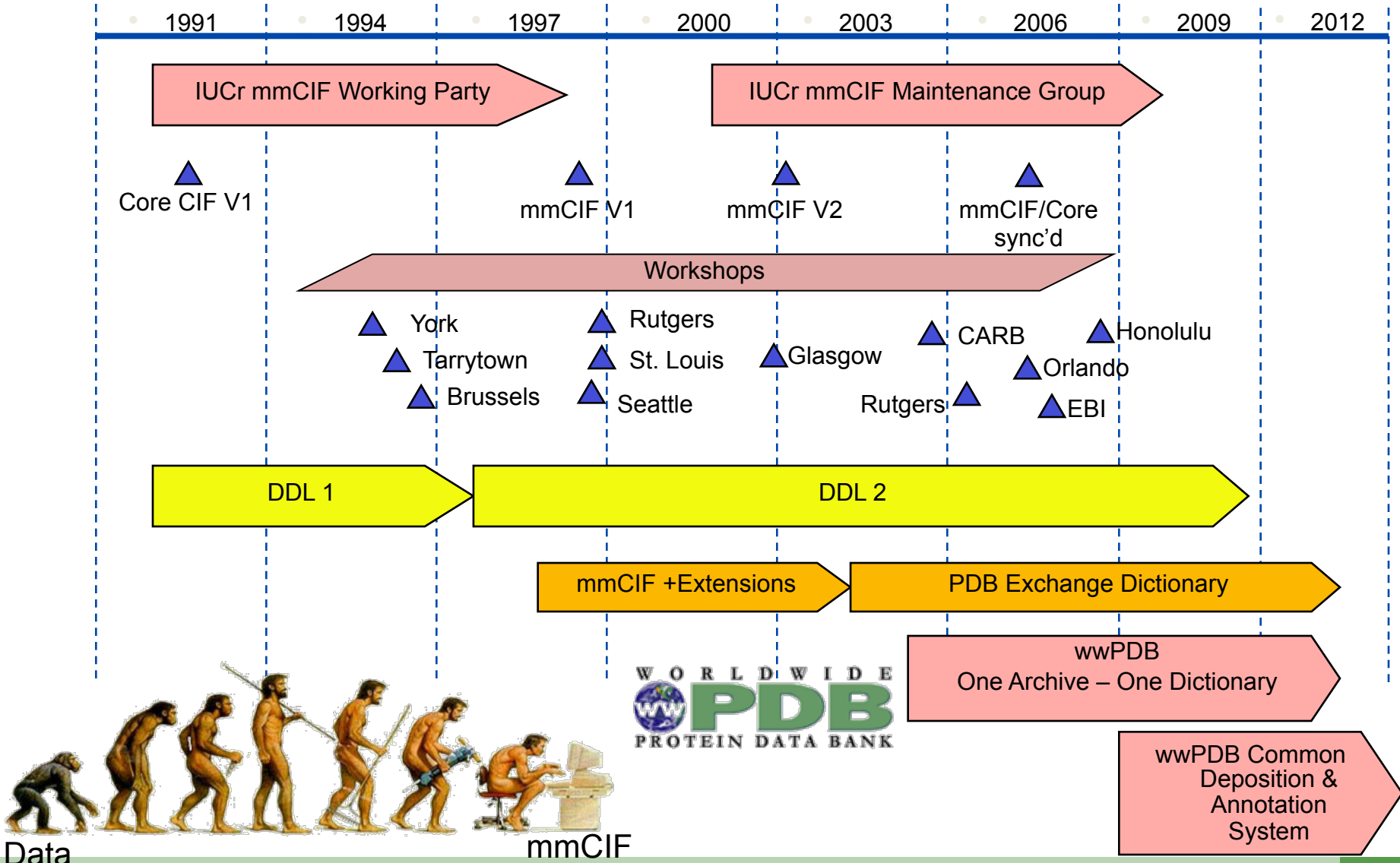
WORLDWIDE PDB
PROTEIN DATA BANK

www.wwpdb.org

Overview

- Brief history of mmCIF development & implementation
- How the wwPDB archives structure, experimental, and reference data
- How mmCIF is helping to address current challenges in data archiving
- Recent developments in wwPDB data deposition and delivery

PDBx/mmCIF Development Timeline



PDB Exchange Dictionary Scientific Content

- Coordinate and supporting primary data
- Experimental descriptions for: X-ray/Neutron diffraction, NMR, Electron Microscopy, SAS & hybrid methods
- Protein production
- Molecular and chemical representation
- Biological and functional annotation
- Additional derivative data –
  - Functional assemblies, validation details, coordinate frame transformations, secondary & tertiary structural features, nucleic acid structural features, …

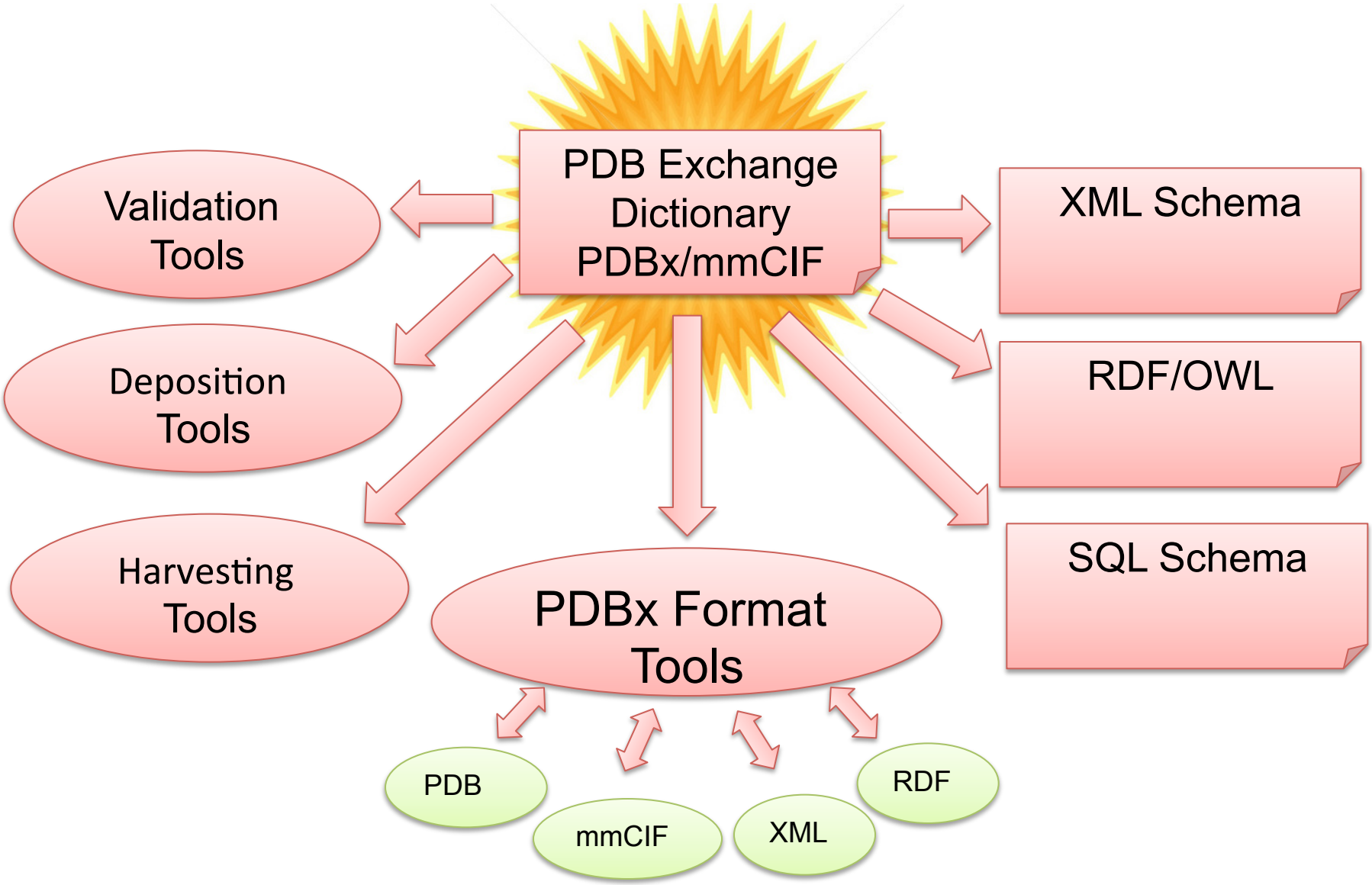
http://mmcif.pdb.org/

PDB Exchange Dictionary Metadata Content

- Features of Data Items
  - Definitions and examples
  - Data types (primitives & regular expression patterns)
  - Boundary values
  - Controlled vocabularies
- Simple organization
  - Tables and columns (categories)
  - Related data item sets (subcategories)
  - Chapters (category groups)
- Associations
  - Referential integrity - parent-child relationships
  - Interdependencies/exclusivity
  - Methods

http://mmcif.pdb.org/

The Central Role of the Data Dictionary



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.

PDB Format Example

```
REMARK   3  DATA USED IN REFINEMENT.
REMARK   3   RESOLUTION RANGE HIGH (ANGSTROMS) : 1.57
REMARK   3   RESOLUTION RANGE LOW  (ANGSTROMS) : 23.00
REMARK   3   DATA CUTOFF            (SIGMA(F)) : 0.000
REMARK   3   COMPLETENESS FOR RANGE        (%) : NULL
REMARK   3   NUMBER OF REFLECTIONS             : 43316
REMARK   3
REMARK   3  FIT TO DATA USED IN REFINEMENT.
REMARK   3   CROSS-VALIDATION METHOD           : NULL
REMARK   3   FREE
REMARK   3   R VA
REMARK   3   R VA
REMARK   3   FREE
REMARK   3   FREE
REMARK   3   FREE
REMARK   3
```

- Record-oriented with fixed column format
- Metadata in semi-structured remarks
- Documentation by example
- Most widely used and supported archival format

```
ATOM      1  N   VAL
ATOM      2  CA  VAL A 363      21.557  -0.831  11.024  1.00 32.13           C
ATOM      3  C   VAL A 363      20.954  -1.757   9.943  1.00 31.73           C
ATOM      4  O   VAL A 363      19.737  -1.906   9.845  1.00 30.94           O
ATOM      5  CB  VAL A 363      21.883   0.552  10.391  1.00 33.45           C
```

PDBx/mmCIF Format Example

- Name – value pairs

```
_exptl.entry_id          1XBB
_exptl.method            'X-RAY DIFFRACTION'
_exptl.crystals_number   1
```

- Table

```
 loop
_data
_data
_data
_data
_database_PDB_rev.replaces
_database_PDB_rev.status
1 2004-11-02 2004-08-30 0 1XBB ?
2 2005-03-22 ?          1 1XBB ?
3 2009-02-24 ?          1 1XBB ?
```

- Simple syntax
- Named data items
- Data semantics defined in the PDBx data dictionary
- Software support in most popular languages

PDBML Example

```
<PDBx:entity_polyCategory>
     <PDBx:entity_poly entity_id="1">
        <PDBx:type>polypeptide(L)</PDBx:type>
        <PDBx:nstd_linkage>no</PDBx:nstd_linkage>
        <PDBx:nstd_monomer>no</PDBx:nstd_monomer>
        <PDBx:pdbx_seq_one_letter_code>
         DIVLTQSPASLSASVGETVTITCRASGNIHNYLAWYQQKQGKSPQLLVYYTTTLADG
     </
</PDB
```

- Three flavors of XML files:
  - fully marked-up files
  - files without atom records
  - files with a more space efficient encoding of atom records
- Follows naming and semantics of the PDBx data dictionary

PDBML

RDF Example

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xml" href="http://pdbj.org/rdf-supp/pdbj-rdf.xsl" ?>
<rdf:RDF xmlns:PDBo="http://pdbj.org/schema/pdbx-v40.owl#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
         xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <PDBo:PDBID rdf:about="http://pdbj.org/pdb/1GOF">
    <rdfs:label>1GOF</rdfs:label>
  </PDBo:PDBID>
  <rdf:Des
    <PDBo
    <PDBo
    <PDBo
    <PDBo
    <PDBo
    <PDBo
    <PDBo
    <PDBo
    <PDBo
    <PDBo
    <PDBo
    <PDBo
    <PDBo
    <PDBo:r
    <rdf:type rdf:resource="http://pdbj.org/schema/pdbx-v40.owl#entity"/>
  </rdf:Description>
</rdf:RDF>
```

- Entry point for semantic web and reasoning systems
- Translates data items in PDBx/mmCIF schema into triples with URL identifiers
- Follows naming and semantics of the PDBx data dictionary
- http://pdbj.org/rdf/<pdbID>/<categoryName>/<pkey1>,…
- For example, http://pdbj.org/rdf/1GOF/entity/1

Chemical Reference Data
Chemical Component Dictionary

- Library of all polymer and non-polymer chemical components in PDB
  - ~18,000 chemical component definitions
  - 400 additional definitions of amino acid protonation variants
- ~700 new components released this year
- ~1700 component definitions updated this year
- Complimentary to the CCP4 monomer library

Chemical Reference Data Example

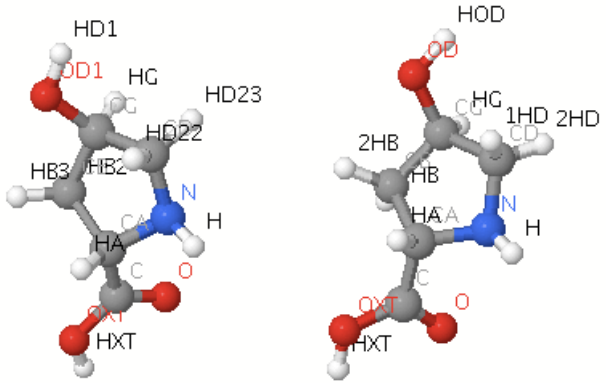
```
loop_
_chem_comp_atom.comp_id
_chem_comp_atom.atom_id
_chem_comp_atom.alt_atom_id
_chem_comp_atom.type_symbol
_chem_comp_atom.charge
_chem_comp_atom.pdbx_align
_chem_comp_atom.pdbx_aromatic_flag
_chem_comp_atom.pdbx_leaving_atom_flag
_chem_comp_atom.pdbx_stereo_config
_chem_comp_atom.model_Cartn_x
_chem_comp_atom.model_Cartn_y
_chem_comp_atom.model_Cartn_z
_chem_comp_atom.pdbx_model_Cartn_x_ideal
_chem_comp_atom.pdbx_model_Cartn_y_ideal
_chem_comp_atom.pdbx_model_Cartn_z_ideal
_chem_comp_atom.pdbx_ordinal
HYP  N    N    N  0  1  N  N  N  -3.366  16.585  44.188   0.168   1.360  -0.282  1
HYP  CA   CA   C  0  1  N  N  S  -2.955  15.768  43.044  -0.384  -0.003  -0.493  2
HYP  C    C    C  0  1  N  N  N  -1.447  15.609  43.030  -1.811  -0.072  -0.013  3
HYP  O    O    O  0  1  N  N  N  -0.722  16.484  43.503  -2.233   0.764   0.750  4
HYP  CB   CB   C  0  1  N  N  N  -3.408  16.578  41.829   0.515  -0.924   0.359  5
HYP  CG   CG   C  0  1  N  N  R  -4.437  17.482  42.330   1.847  -0.159   0.505  6
HYP  CD   CD   C  0  1  N  N  N  -4.068  17.803  43.753   1.640   1.159  -0.271  7
HYP  OD1  OD   O  0  1  N  N  N  -5.693  16.815  42.294   2.917  -0.911  -0.071  8
HYP  OXT  OXT  O  0  1  N  Y  N  -0.976  14.502  42.469  -2.614  -1.063  -0.433  9
HYP  H    H    H  0  1  N  Y  N  -3.980  16.047  44.765  -0.107   1.981  -1.028  10
HYP  HA   HA   H  0  1  N  N  N  -3.385  14.756  43.068  -0.325  -0.278  -1.546  11
HYP  HB2  1HB  H  0  1  N  N  N  -2.567  17.141  41.398   0.066  -1.092   1.337  12
HYP  HB3  2HB  H  0  1  N  N  N  -3.790  15.930  41.026   0.678  -1.873  -0.153  13
HYP  HG   HG   H  0  1  N  N  N  -4.508  18.399  41.726   2.052   0.048   1.555  14
HYP  HD22 1HD  H  0  0  N  N  N  -4.956  18.005  44.370   2.018   1.065  -1.289  15
HYP  HD23 2HD  H  0  0  N  N  N  -3.457  18.713  43.848   2.132   1.985   0.243  16
HYP  HD1  HOD  H  0  1  N  N  N  -5.999  16.666  43.181   3.780  -0.479  -0.009  17
HYP  HXT  HXT  H  0  1  N  N  N  -0.027  14.511  42.499  -3.520  -1.066  -0.098  18
#
```

Atom names

Stereochemistry & aromaticity

Model coordinates

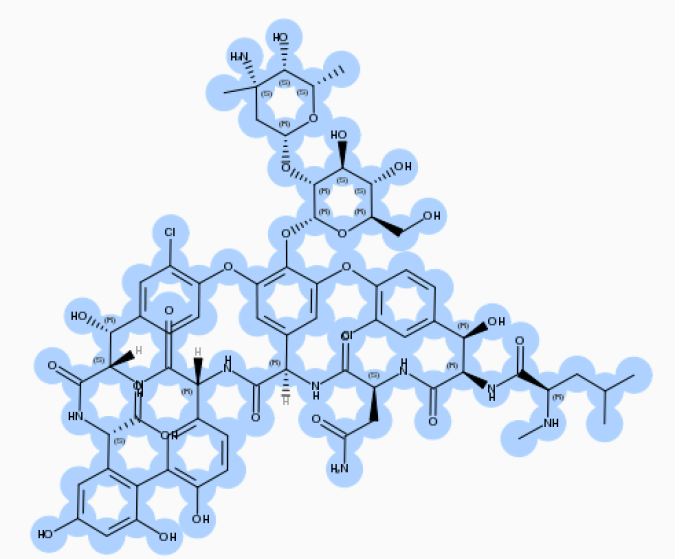
Ideal coordinates



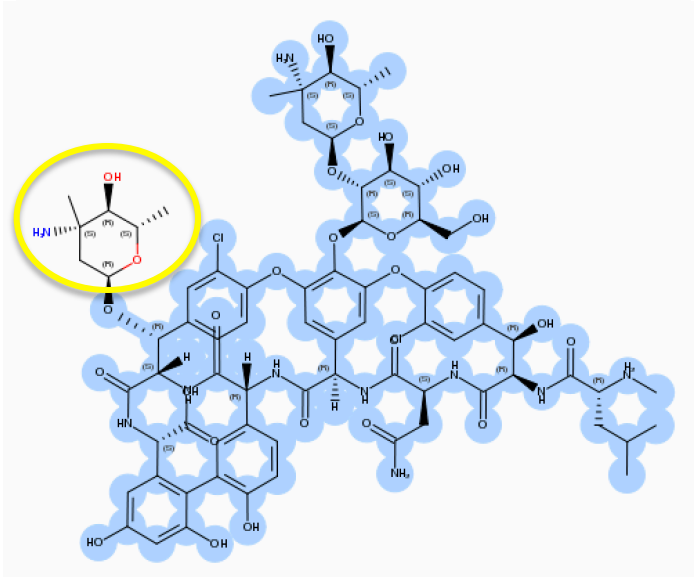
Biologically Interesting Reference Molecule Dictionary (BIRD)

- Contains ~630 chemical definitions for peptide inhibitors and antibiotics
- Unifies the representation of small polymers and single molecules with substantially polymeric chemical structure
- Provides structural and functional annotations
- Designed to facilitate both sequence and detailed chemical structure searches

Target

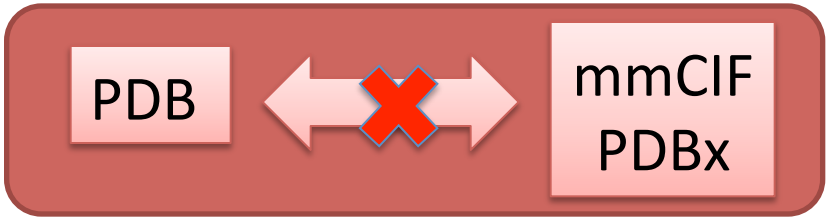


Hit

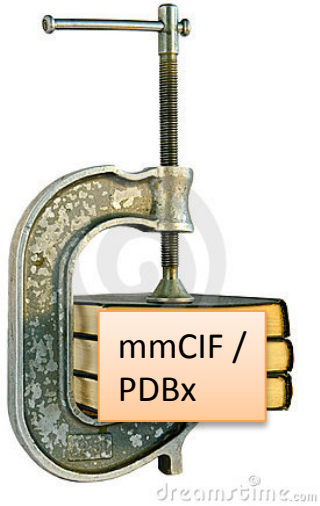


Keeping Pace with Structural Biology

- The most enduring and widely used archival PDB format is not keeping pace with new science and technology.
- Efforts to work around PDB format limitations are increasingly problematic.

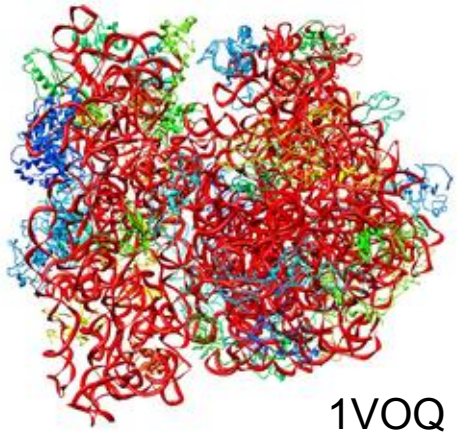


Reversible translation no longer practical.



Challenges of Molecular Size

- PDB column format limitations
  - 1-character for polymer chain labels
  - 5-characters for atom serial numbers
  - 3-characters for monomers and ligand identifiers
  - 5-characters for atom names
  - F8.3 for model coordinates
- Implications –
  - Maximum of 62 chains (upper and lower case!)
  - Maximum of 99,999 atoms
    - Requires splitting structures across multiple entries (5 ribosomes in ASU stored in 10 PDB entries!)
    - Map and experimental validation are difficult for split entries
  - Cannot use standard monomer & ligand nomenclatures (e.g. carbohydrates & protonation variants)
  - Cannot use conventional atom names in large ligands
  - Limits molecular dimension (< 9999.999 Angstroms)



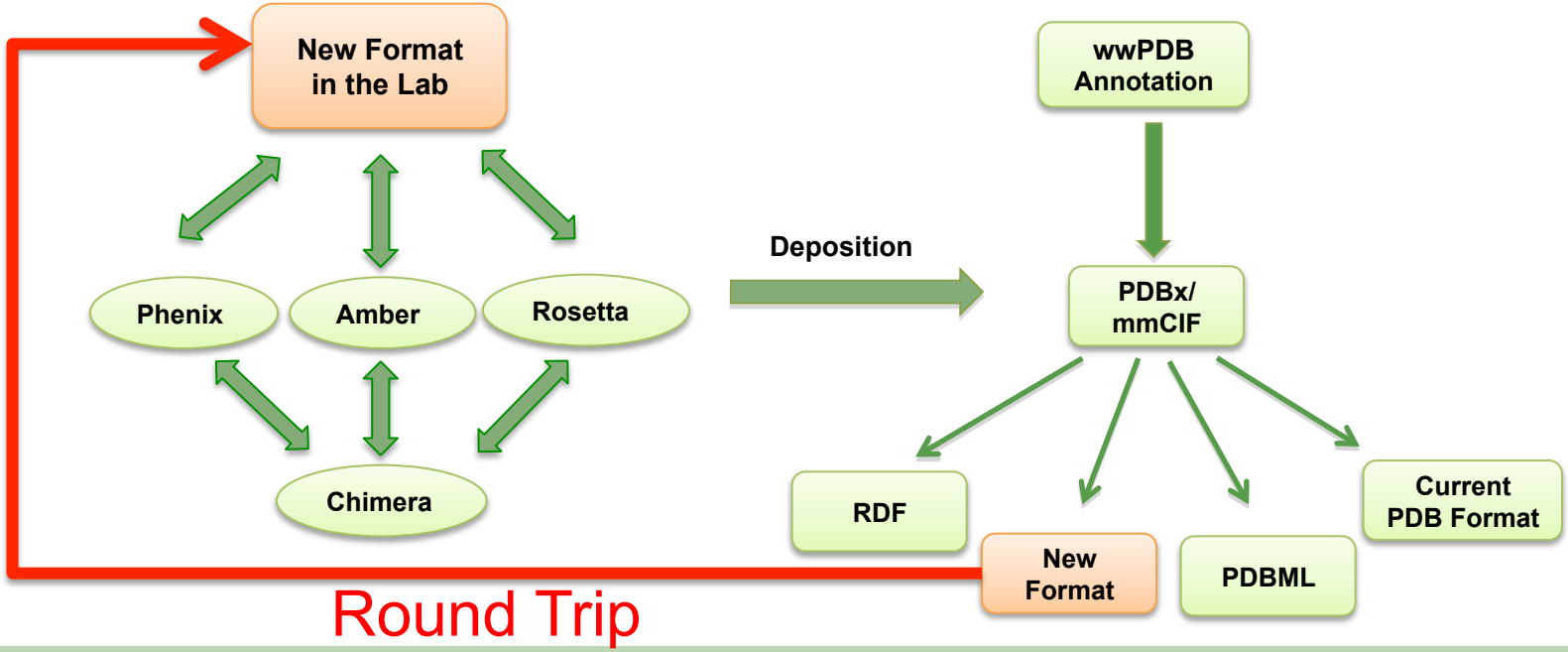
1VOQ

Representing Evolving Content

- PDB Record format limitations
  - The small number of named records are not extensible (e.g. ATOM, CONECT, SEQRES,…).
  - Text REMARKS with *ad hoc* formats hold all other meta-data.
- Implications
  - No bond orders are specified for ligands
  - ATOM records tailored for traditional X-ray methods and inflexible for newer methods (e.g. TLS groups)
  - Sequence and coordinate residue correspondence is ambiguous and there is no support for heterogeneity.
  - Growing diversity and complexity of REMARK records has become unmanageable for both **deposition** and **archiving**
  - No standard and extensible way to represent and document meta-data

Key Format Goals for PDB

- Represent all PDB model structure, supporting experimental and metadata
- Provide a working format for data exchange between the laboratory and the archive
- Support the entire structure biology pipeline: model-building, refinement, visualization, validation, analysis, simulation, prediction, …



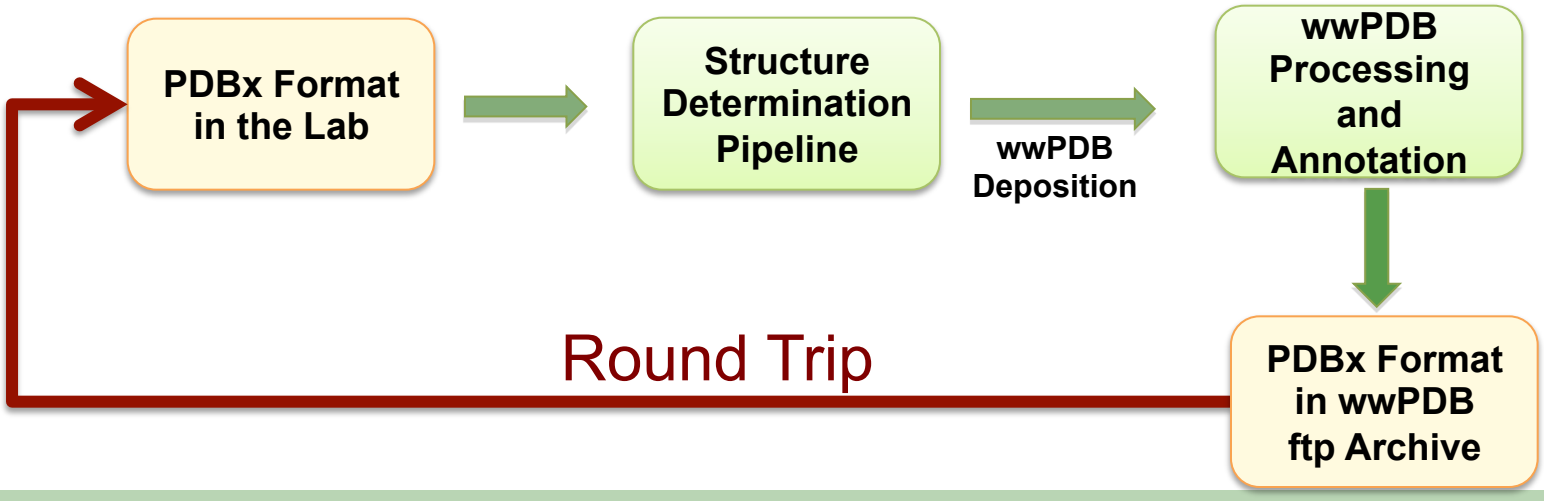
Finding a Simple Format Alternative

- 2010 – started process of defining new format, consulting many software developers
- 2011 – Developers Workshop - agreement to adopt PDBx (mmCIF) as the new format and to phase out the old PDB format
  - Commitments from CCP4, Phenix and Global Phasing (*i.e.*, ~85% of all PDB depositions)
  - Agreement on managing development between these software providers and wwPDB
  - Established PDBx Deposition Working Group
- 2013 - Working Group recommendations and implementations in CCP4 and Phenix.

PDBx/mmCIF Deposition Working Group

PDBx Deposition Working Group
Refinement Developers Workshop 2011 - EBI

- In 2011, charged with finding a “round trip” single format that can handle complex data not supported by the PDB file format
- Consensus reached on using dictionary-driven PDBx format
- Implementations delivered in January 2013



Working Group Recommendations

Announced 22-May-2013

Format extensions for large structures:

- Atom serial numbers (1 to the number atoms)
- Chain identifiers up to 4 characters
- Cartesian coordinates with field widths as required and 3 decimal places
- B-factors and occupancies with 3 decimal places precision.
- Implement extensions as required to as maintain backward compatibility

Transitional Home for Large Structures

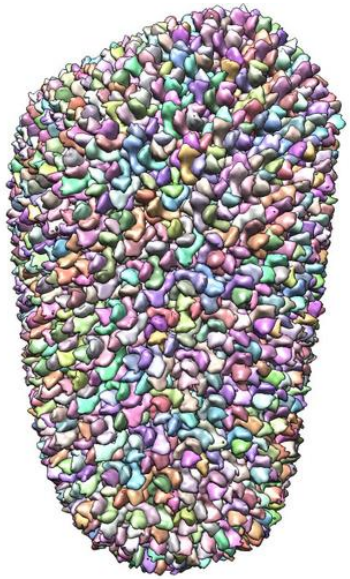
Large single entries are now stored separately on the wwPDB ftp site, and PDB internally produces *divided/split* PDB format files.

ftp://ftp.wwpdb.org/pub/pdb/data/large_structures/mmCIF/
ftp://ftp.wwpdb.org/pub/pdb/data/large_structures/XML/

HIV-1 Capsid 3J3Q –

- 1356 chains
- >2M atoms
- 25 – PDB format entries

3J3Y



3J3Q

Providing Format Compatibility

- Adopt a *PDB friendly* mmCIF/PDBx style -
  - All records on a single text line
  - Columns presented in standard column order.
  - Tabular presentation with leading record names
    (*e.g.* ATOM, CELL, REFINE)
  - Method independent features in left-most columns
    (*e.g.* identifiers & coordinates)
  - Method specific features in the right-most columns
    (*e.g.* ADPs, NMR order/disorder parameters)
  - Continue to support PDB nomenclature semantics
    (*e.g*. PDB style chains, residue numbering, and insertion codes)
- Large entries will be internally converted to divided/split PDB format files.

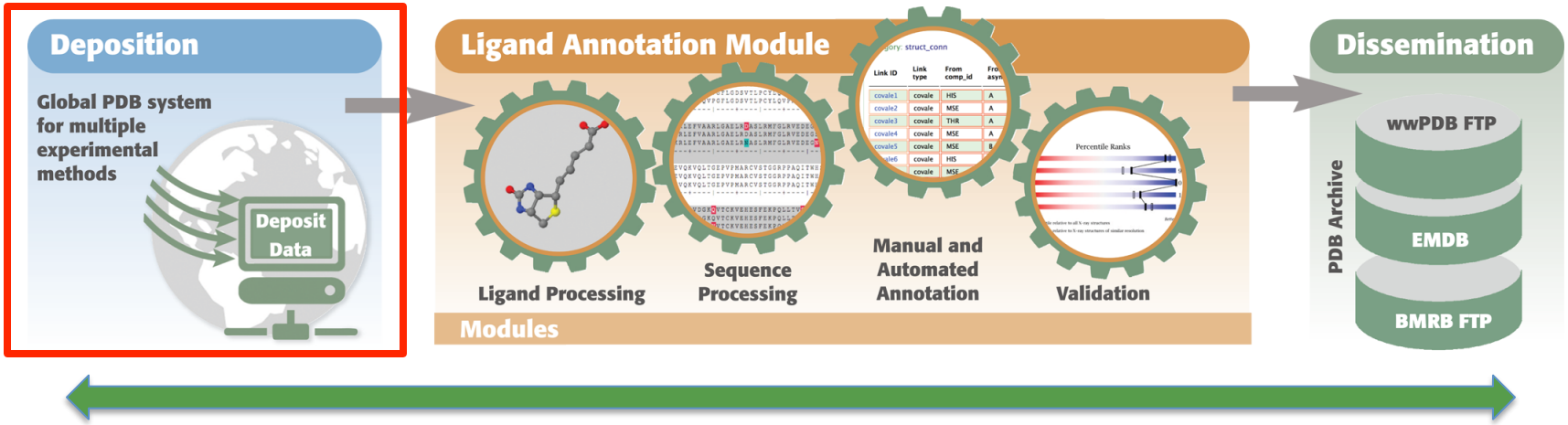
PDB

```
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
```

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
```

New wwPDB Deposition & Annotation System



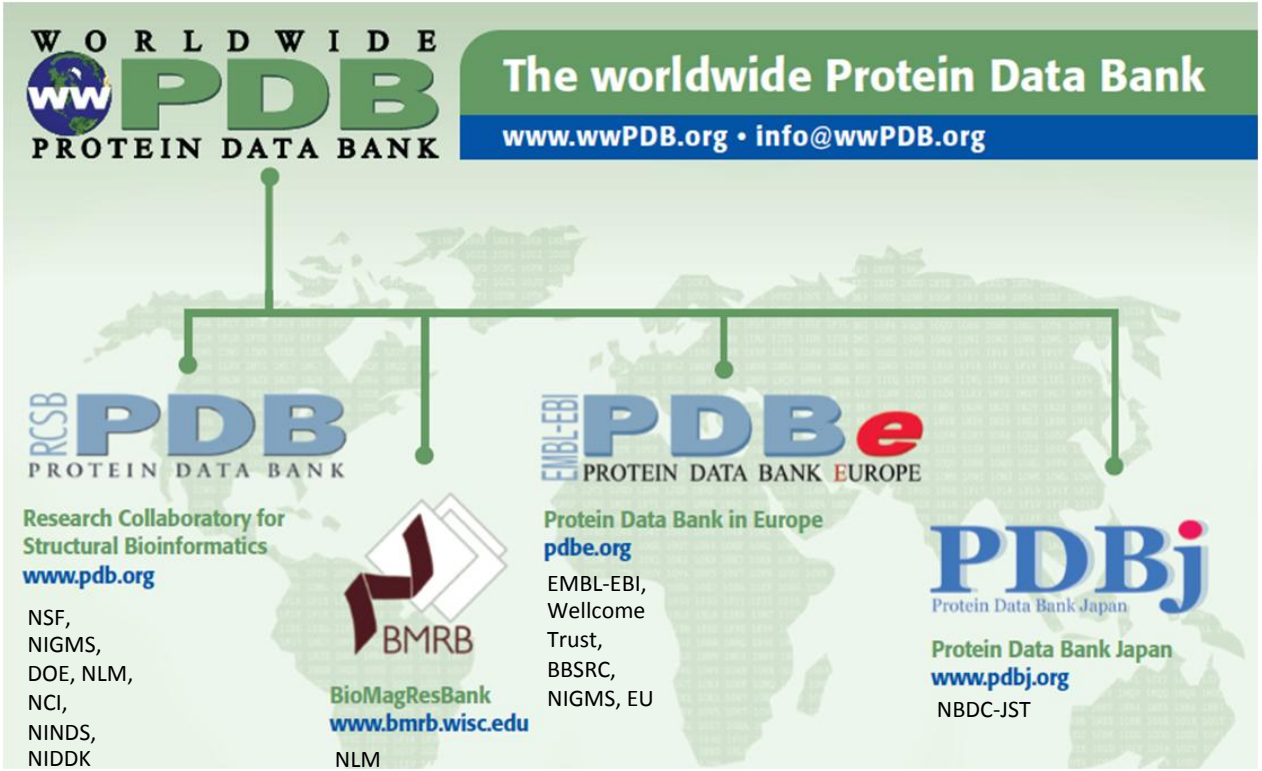
mmCIF/PDBx

End-to-end support for PDBx/mmCIF

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** - **Chimera**, **Jmol**, **OpenRasMol**

PDB actively working with community developers to help fill in missing functionalities. Two workshops scheduled in Fall 2013 …


WORLDWIDE PDB
PROTEIN DATA BANK
The worldwide Protein Data Bank
www.wwPDB.org • info@wwPDB.org
RCSB PDB
PROTEIN DATA BANK
Research Collaboratory for Structural Bioinformatics
www.pdb.org
NSF, NIGMS, DOE, NLM, NCI, NINDS, NIDDK
BMRB
BioMagResBank
www.bmrb.wisc.edu
NLM
EMBL-EBI PDBe
PROTEIN DATA BANK EUROPE
Protein Data Bank in Europe
pdbe.org
EMBL-EBI, Wellcome Trust, BBSRC, NIGMS, EU
PDBj
Protein Data Bank Japan
Protein Data Bank Japan
www.pdbj.org
NBDC-JST

wwPDB