

Protein Data Bank

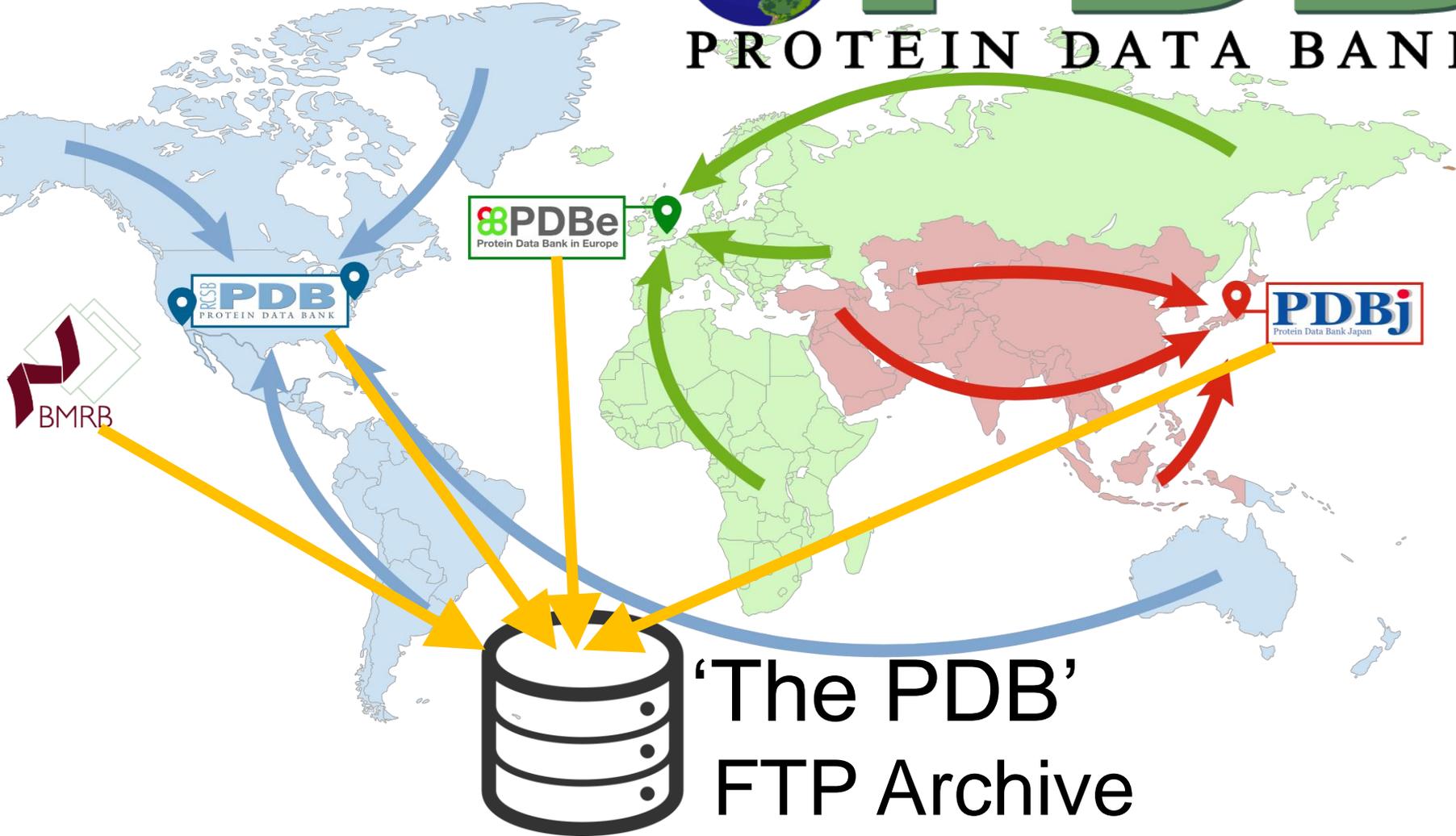
Introduction to the mmCIF dictionary

Introduction to the mmCIF dictionary

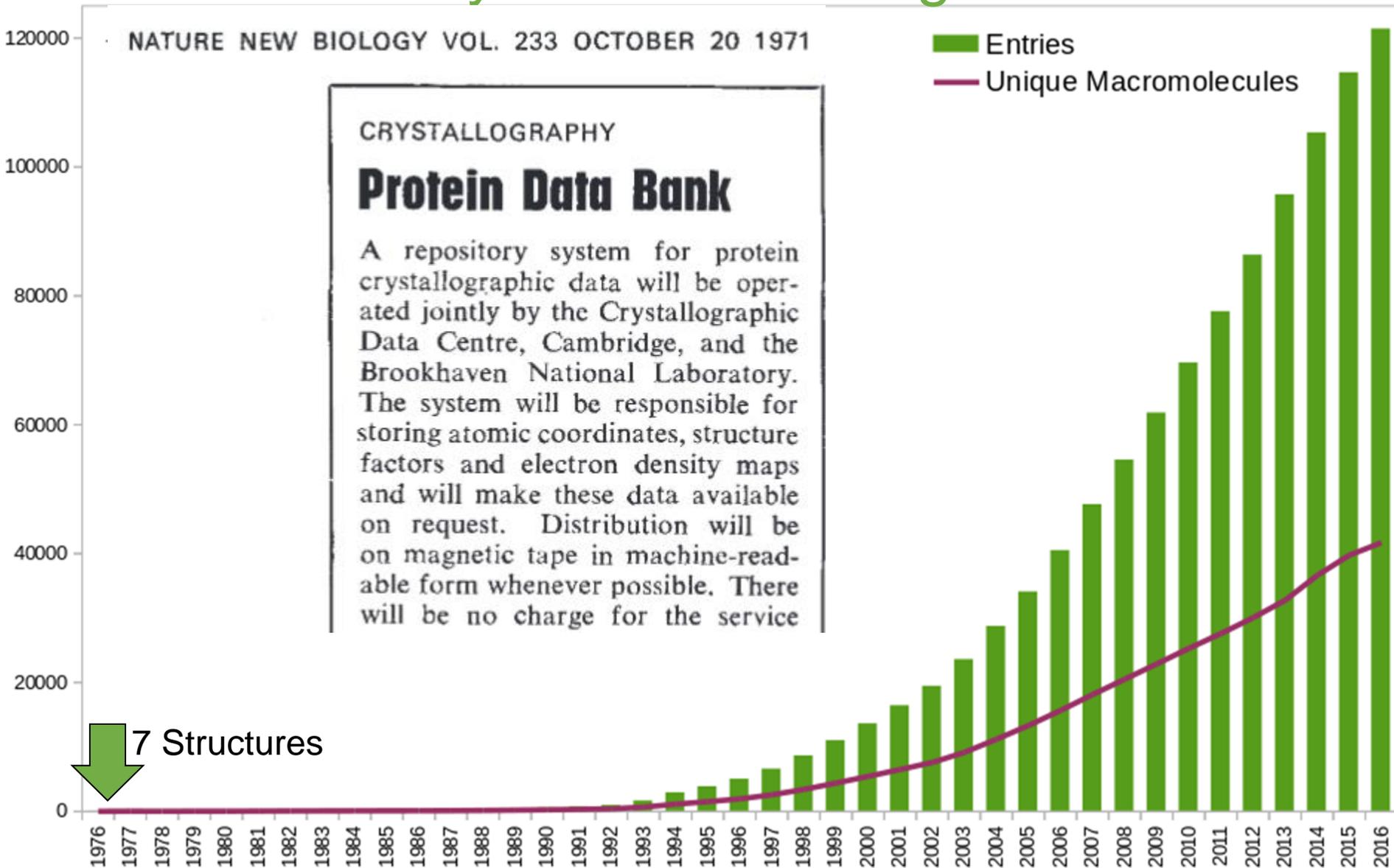
- Who are the wwPDB?
- What do we use the mmCIF dictionary for?
- Core mmCIF dictionary
- mmCIF dictionary extensions

Who are the PDB?

W O R L D W I D E
wwPDB
P R O T E I N D A T A B A N K



Oldest freely available biological database

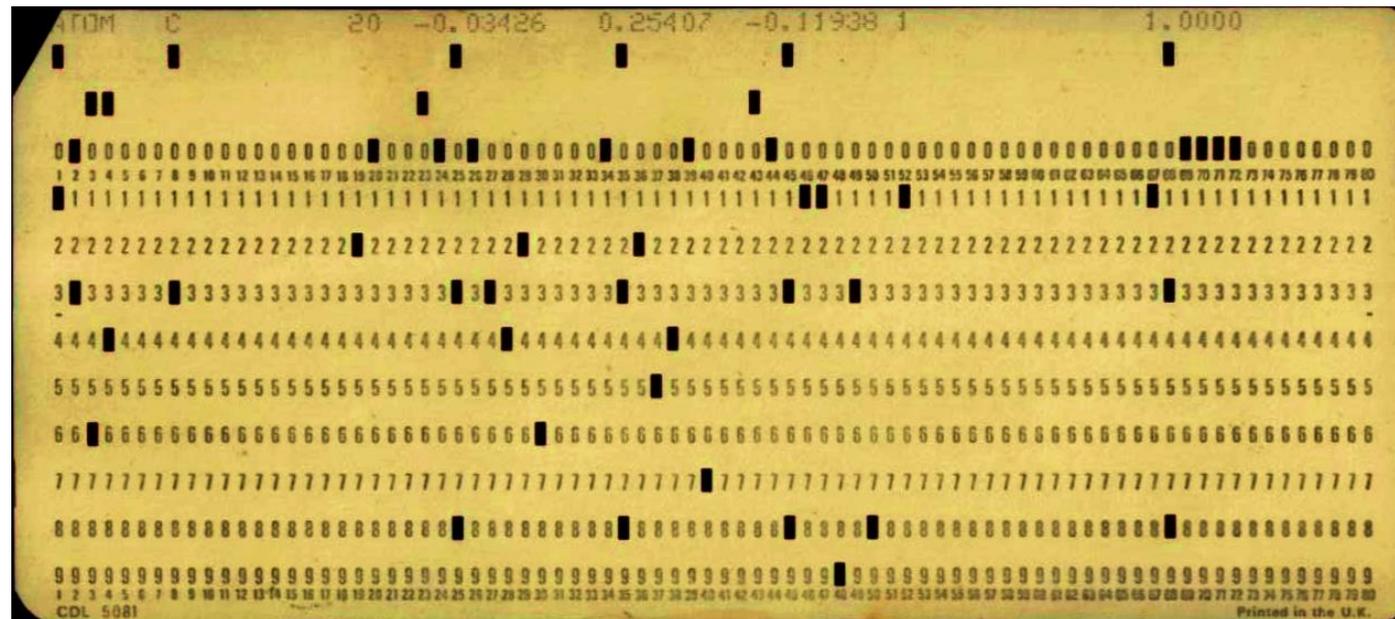


Move to mmCIF

- <http://mmcif.wwpdb.org/docs/tutorials/glossary/early-history.html>
- CIF is a subset of STAR
 - S.R. Hall (1991) The STAR File: A new format for electronic data transfer and archiving. *J. Chem. Inf. Comp. Sci.*, **31**, 326-333.
- 1990: CIF (Crystallographic Information File) introduced
- 1990: Working group setup to create mmCIF (macromolecular CIF) from CIF
- 1997: mmCIF introduced – based on DDL2.2.1
 - Took core data items from CIF dictionary
- 2014: mmCIF became the wwPDB core format – replacing PDB
- 2019: mmCIF mandatory for diffraction entries

Move to mmCIF

- PDB format based on punch cards
- Last updated in 2012
- Column limited format
 - 99,999 atoms
 - 62 chains



mmCIF dictionary resources

- <http://mmcif.wwpdb.org/>
- Includes the history of mmCIF
- File syntax -
<http://mmcif.wwpdb.org/docs/tutorials/mechanics/pdbx-mmcif-syntax.html>
- Current version V5.0
- Documentation provides details of
 - mandatory items
 - enumerations
- Contains links to parsers for various languages

What do we use the mmCIF dictionary for?

- Validation of data in the PDB archive
 - Numerical limits on values
 - Enumeration lists
- OneDep deposition system

OneDep deposition system



List requirements

All items

Mandatory items

Navigation

- ✓ Instructions
- ✓ Communication
- ✓ **Re-upload files**
- ✓ Upload summary
- Admin
 - ✓ Contact information
 - ✓ Grant information
 - ✓ Release status
 - ✓ Entry title & author
 - ✓ Citation information
- Macromolecules
 - ✓ 1) AMMONIA CHANNEL
 - ✓ 2) NITROGEN REGULATORY
- Data collection
 - ✓ Crystal Information
 - ✓ Collection Source
 - ✓ Software Used
 - ✓ Collection Statistics
- Refinement
 - ✓ Refinement
 - ✓ Ligands
 - ✓ Biological assembly
 - Validation reports
 - Summary & conditions
- Downloads & reports
 - All files
 - Generated mmCif

Log out

Replacement sections and re-upload files

General upload instructions

- Click 'Browse' to upload your file. Once the file is uploaded, select the file type from the pull-down list. If you have uploaded more than one file of each type, use the check box to select which file should be used.
- After upload, you must review the summary page carefully as it will tell you whether your data has been uploaded and interpreted correctly.
- The gzip and bzip2 compression formats are supported for all uploaded files. Archive formats such as tar and windows ZIP archives are not supported.

Coordinate upload instructions

Coordinates may be deposited as either mmCIF or PDB formatted files. We encourage you to use [pdb_extract](#) to prepare mmCIF formatted file. If you are using a PDB formatted coordinate file, please check the following PDB format requirements prior to file upload:

- A TER record is placed at the end of every polymeric chain.
- TER records are not present within a polymeric chain. Some refinement packages insert extra TER records at gaps within a chain. Please remove the extra TER records.
- No TER records should be included at the end of non-polymer residues such as ions, ligands, or waters.
- **PDB format files with misplaced TER records will result in incorrect format translation and data extraction - which may result in data loss during annotation.**
- Depositions comprised of multiple models should include MODEL and ENDMDL records. The models should be listed sequentially in columns 11-14.
- If there are alternate conformations in the structure, the alternate conformation indicator must be provided in column 17 of ATOM and HETATM records.
- There should be only one END record at the end of the file.

Browse... No file selected.

<input checked="" type="checkbox"/>	pdb2nuu.ent	2.05 MB	Coordinates (PDB format)	
<input checked="" type="checkbox"/>	r2nuusf.ent	4.21 MB	mmCIF (structure factors)	

Continue deposition (some sections will be modified)

mmCIF is the PDB main distribution format

- Coordinate file
- Structure factor file
- Map coefficient file
- NMR chemical shift file
- Chemical component (ligands) definition

Organisation of mmCIF

- mmCIF is organised into
 - Data blocks
 - Groups of categories
 - Categories
 - Items
- <http://mmcif.wwpdb.org/docs/tutorials/mechanics/pdbx-mmcif-dict-struct.html>

Organisation of mmCIF

- Data blocks
 - Top level organisation of the mmCIF
 - Denoted as
 - data_
 - Files can contain multiple data blocks – each with a unique name
 - data_first

Organisation of mmCIF

- Groups
 - http://mmcif.wwpdb.org/dictionaries/mmcif_pdbx_v50.dic/Groups/index.html
 - 41 groups
 - Used for organisation of the dictionary, not exported to files
 - Include extensions incorporated into the main dictionary
 - Chemical component (CCD)
 - validate
 - XFEL
 - EM
 - pdbx

Organisation of mmCIF

- Categories
 - http://mmcif.wwpdb.org/dictionaries/mmcif_pdbx_v50.dic/Categories/index.html
 - 553 categories
 - Prefixed with “_” in mmCIF files
 - E.g. _software
 - Coloured in the web page if used in current PDB or CCD entries

cell

cell_measurement

cell_measurement_refln

chem_comp

chem_comp_angle

chem_comp_atom

chem_comp_bond



Organisation of mmCIF

- Items
- found after a “.” following a category
 - E.g. software.name
- Must be unique per category

Organisation of mmCIF

- Values

- Values are controlled by

- Regular expressions

boolean	char	YES NO
citation_doi	char	10\..*
code	char	[] [_ , . ; : " & < > () / \ { } ' ` ~ ! @ # \$ % A - Z a - z 0 - 9 * + -] *

- Enumeration

Allowed Value
ABS
ABSCALE
ABSCOR
ACORN
ADDRF
ADSC
AMBER
AMPLE
AMoRE

- Boundary conditions

Allowed Boundary Conditions	
Minimum Value	Maximum Value
0.0	14.0

Advisory Boundary Conditions	
Minimum Value	Maximum Value
3.5	10

Data validation

- Why have enumerations?
 - So users can find data

Diffraction source ⓘ

condition

Contains ▾

Example: ESRF beamline MASSIF-1

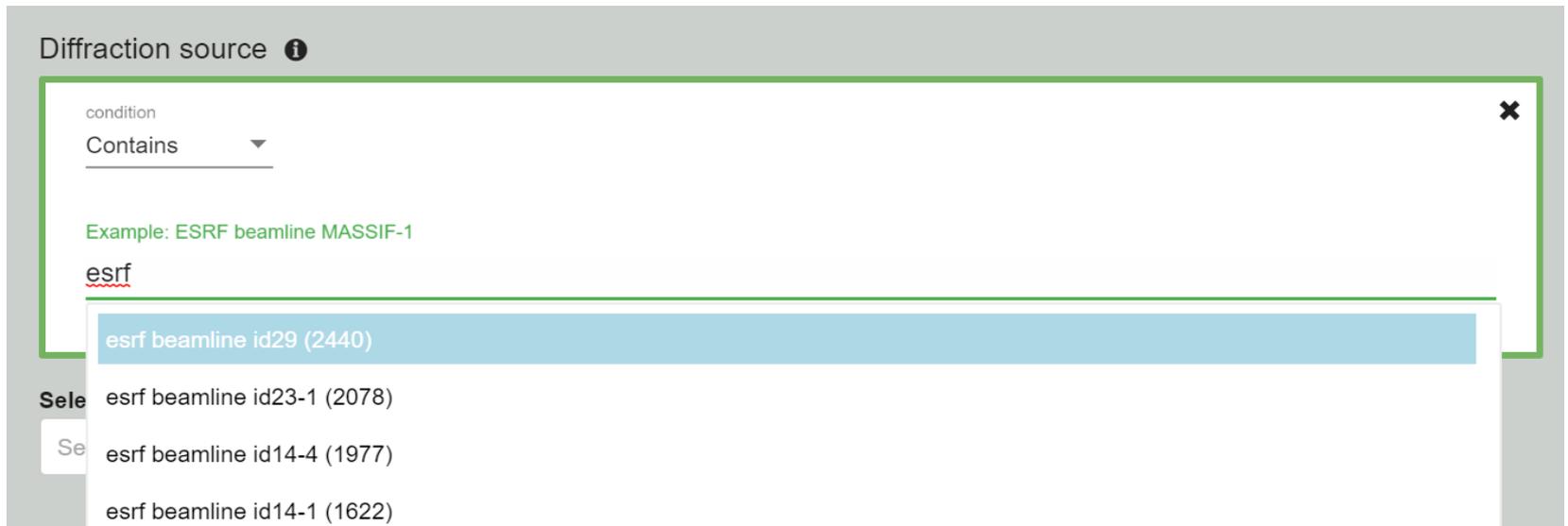
esrf

esrf beamline id29 (2440)

Sele esrf beamline id23-1 (2078)

Se esrf beamline id14-4 (1977)

esrf beamline id14-1 (1622)



Additional checks using mmCIF

- http://mmcif.wwpdb.org/dictionaries/mmcif_pdbx_v50.dic/Items/_diffrn_detector.type.html
- Details column used to link enumerated values to
- http://mmcif.wwpdb.org/dictionaries/mmcif_pdbx_v50.dic/Items/_diffrn_detector.detector.html

Controlled Vocabulary at Deposition

[View/Hide Table](#)

Allowed Value

Details

ADSC HF-4M

PIXEL

- `_diffrn_detector.detector` is filled automatically when depositors select a value for `_diffrn_detector.type`
- This is used by the OneDep deposition system to ensure data is consistent

Additional checks using mmCIF

- Similar checks added for other categories
- `_software.name` vs `_software.classification`
- `_diffrn_source.source` vs
`_diffrn_radiation.pdbx_scattering_type`
- `_diffrn_source.type` vs `_diffrn_source.source`

Diffraction ID*: 1
Data collection temperature (K)*: 10
Data collection temperature details:

! Warning: Expected a value between 80.0 and 300.0

Sanity Check

Diffraction Radiation

Protocol*: LAUE MAD SINGLE WAVELENGTH
Monochromator:
Scattering type*: electron neutron x-ray

Diffraction Source

Source type*: SYNCHROTRON
Source details*: DIAMOND BEAMLINE I02
Wavelength or list of wavelengths used for this experiment (Å)*: 15

! Warning: Expected a value between 0.5 and 2.0

Diffraction Detector

Detector*: EIG
Detector type*: DECTRIS EIGER X 16M
Date of collection*:
Optics:
DECTRIS EIGER X 500K
DECTRIS EIGER R 1M
DECTRIS EIGER R 4M
DECTRIS EIGER X 1M
DECTRIS EIGER X 9M
DECTRIS EIGER X 4M

! Value not in the enumeration list for this item.

! No value present for this mandatory item.

Controlled vocabulary

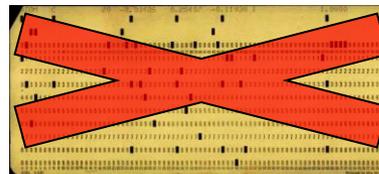
Continue to next section

Additional data we collect in mmCIF

- Admin information
 - Who are you?
 - When do you want your entry released?
- Sample description
 - Including sequence
- Data collection statistics
 - Where did you collect your data?
- Refinement statistics

V4 to V5 transition

PDBx/mmCIF V5

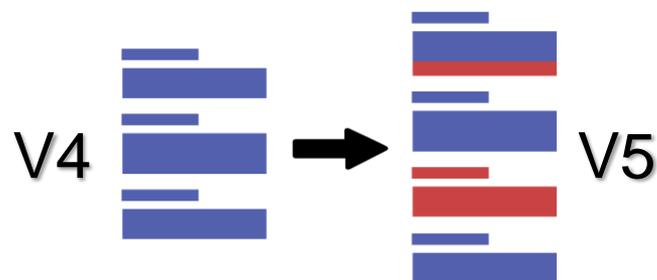


More metadata for EM derived models

Better audit information

Some data standardisation

work still on-going to clean up data



Experimental data

- Structure factors are provided in mmCIF format
- https://ftp.wwpdb.org/pub/pdb/data/structures/divided/structure_factors/cb/r1cbssf.ent.gz
- <http://www.ebi.ac.uk/pdbe/entry-files/download/r1cbssf.ent>

Electron Microscopy

- Originally an extension to the mmCIF dictionary – now incorporated in V5.0
- Categories prefixed with `_em_`

NMR STAR and NEF

- NMR STAR is used by BMRB to describe NMR experimental data
 - <http://www.bmrwisc.edu/formats.shtml>
- This is a STAR format file which can be interconverted to mmCIF format – this interconversion happens in OneDep
- NMR restraints are currently accepted in any format – this is not useful for archive keeping.
- This is being addressed with the introduction of NMR Exchange Format (NEF)
 - <https://www.nature.com/articles/nsmb.3041>
- NEF is a STAR format file which contains a subset of NMR STAR
 - <https://github.com/NMRExchangeFormat/NEF/>

Hierarchy within an mmCIF file

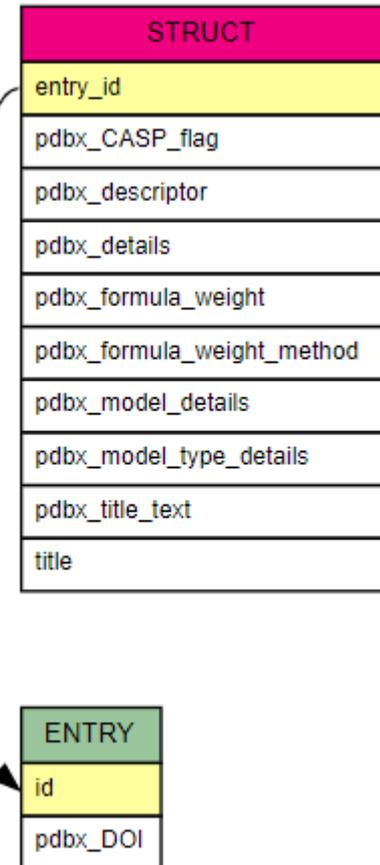
- mmCIF is effectively a relational database with primary and foreign keys relating tables (categories) within the file

Hierarchy within an mmCIF file

- Top level is
 - entry.id
- This has lots of children
- http://mmcif.wwpdb.org/dictionaries/mmcif_pdbx_v50.dic/terms/_entry.id.html
- Each child item has the item name “.entry_id” to denote its parent as “entry.id”
- For example “struct.entry_id”
- http://mmcif.wwpdb.org/dictionaries/mmcif_pdbx_v50.dic/terms/_struct.entry_id.html

Hierarchy within an mmCIF file

- This is displayed graphically on the wwPDB pages
- The keys (as per a relational database) in each category are shown in yellow
- Some categories are a bit more complicated
- http://mmcif.wwpdb.org/dictionaries/mmcif_pdbx_v50.dic/Categories/refine.html



Using an mmCIF file

- mmCIF files are human readable, but designed to be parsable by a computer
- Parsers are available for various languages
 - <http://mmcif.wwpdb.org/docs/software-resources.html>
- Another good resource is the documentation for Gemmi CIF parser
 - <https://gemmi.readthedocs.io/en/latest/cif-parser.html#>

Using an mmCIF file

- As mmCIF is like a mini relational database at PDB sites load the PDB archive into databases
- This allows archive wide queries to be made and allows searching across the whole PDB archive
- This will be covered in the tutorial