Protein Data Bank

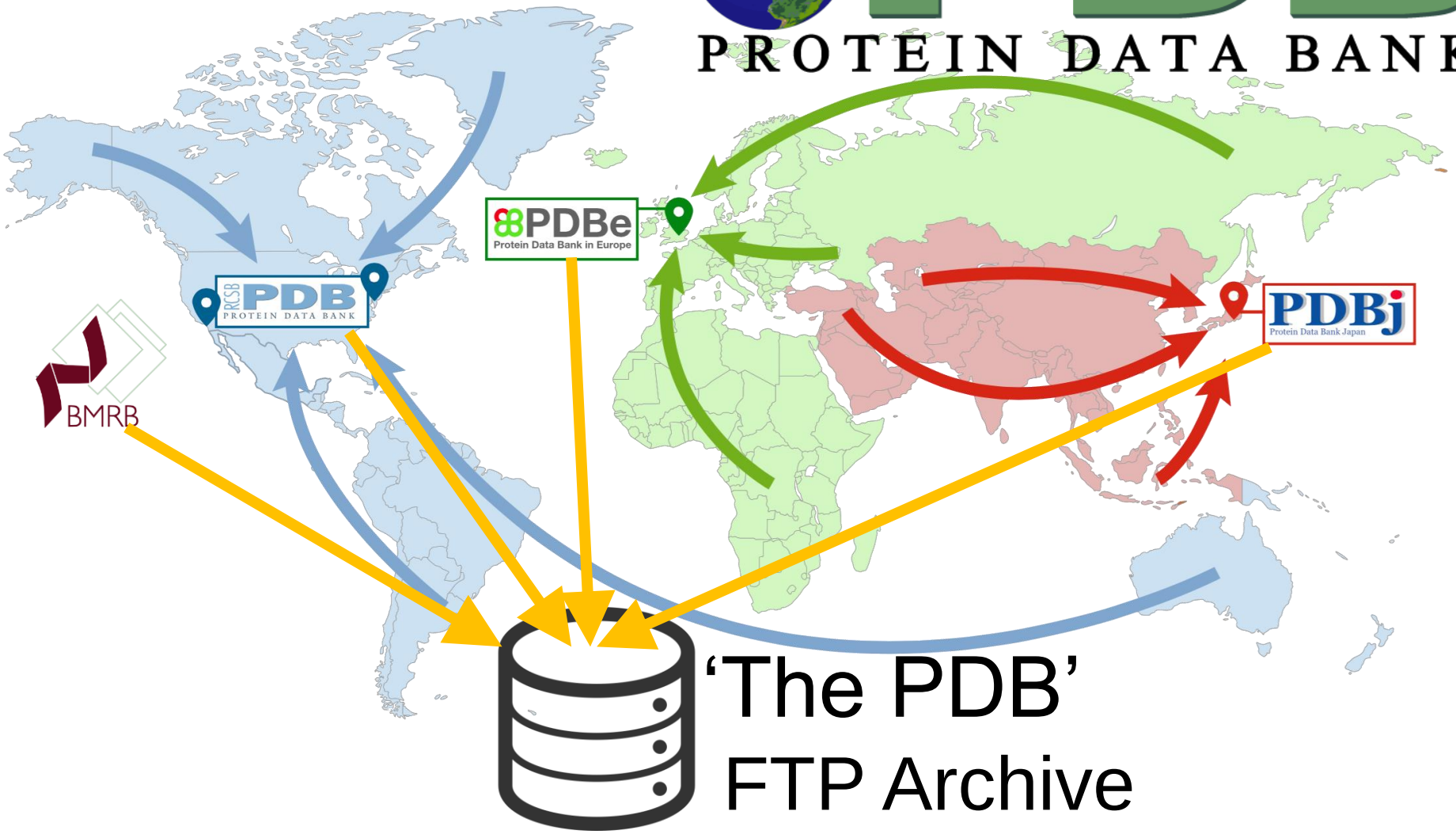
Introduction to the mmCIF dictionary

Introduction to the mmCIF dictionary

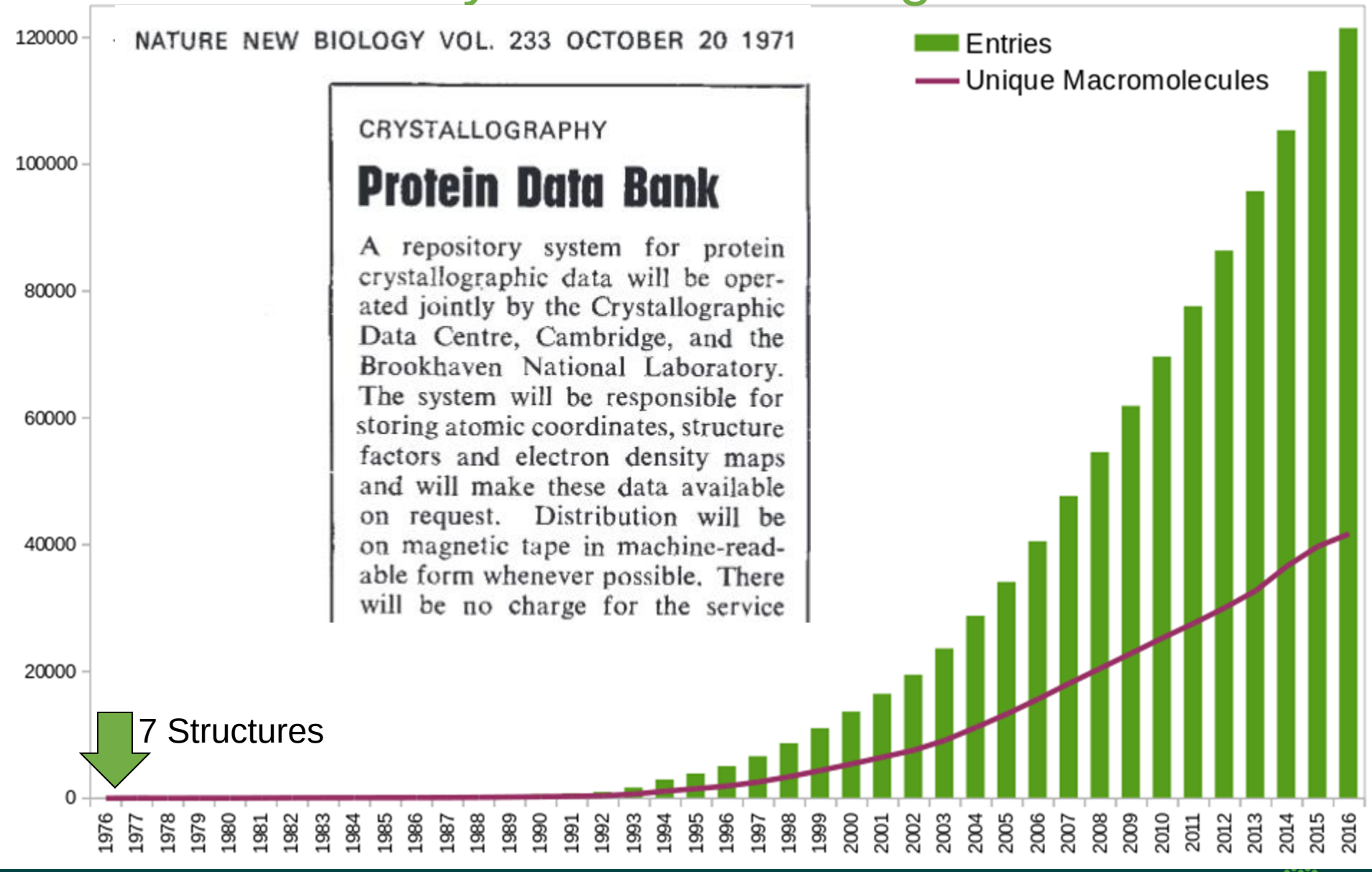
- Who are the wwPDB?
- What do we use the mmCIF dictionary for?
- Core mmCIF dictionary
- mmCIF dictionary extensions

Who are the PDB?

WORLDWIDE PDB
PROTEIN DATA BANK



Oldest freely available biological database

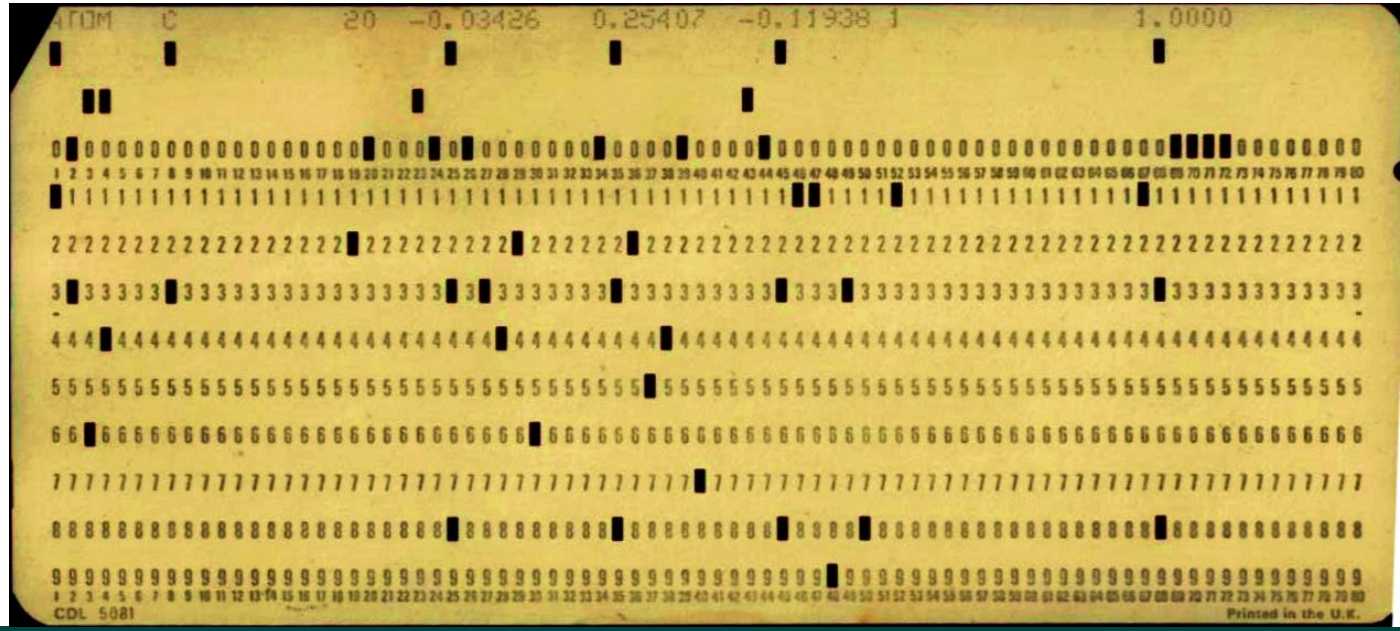


Move to mmCIF

- 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



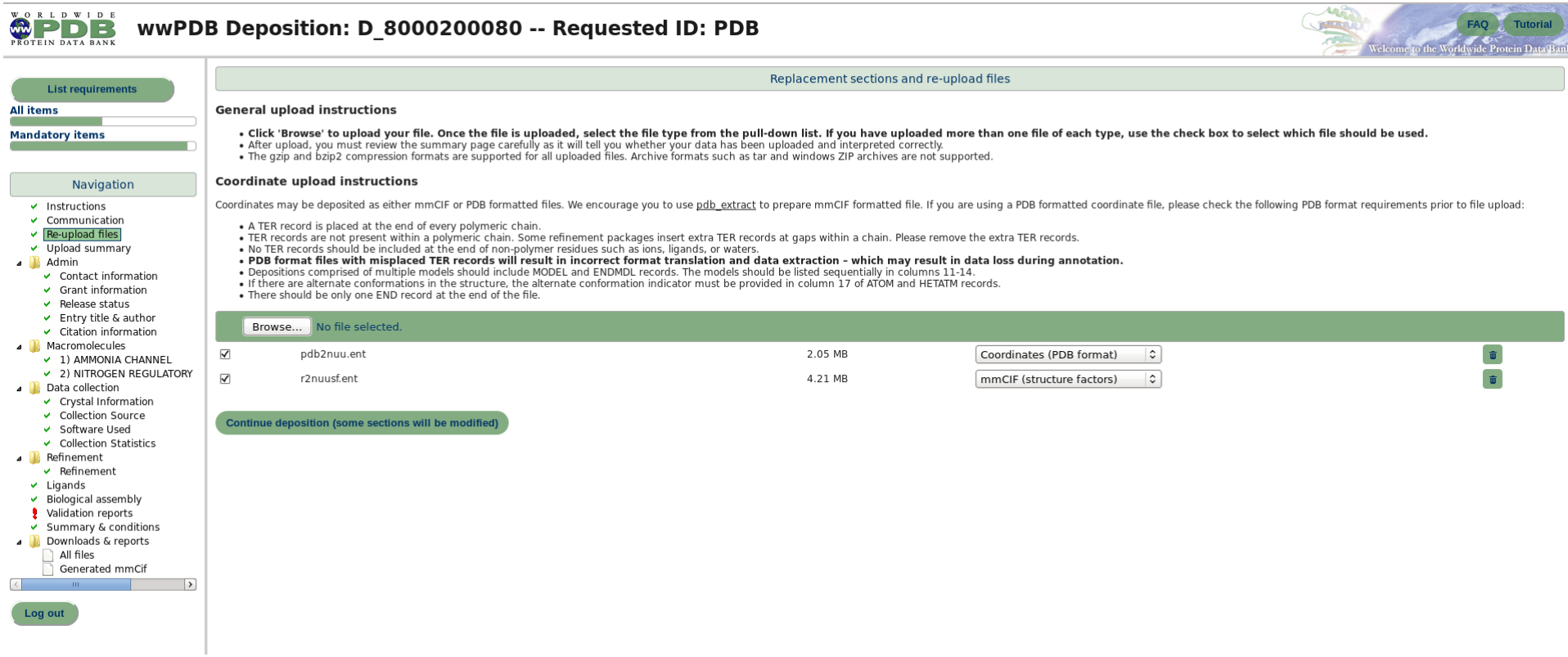
mmCIF dictionary resources

- 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



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

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-Za-z0-9*|+-]*

    - Enumeration

Allowed Value
ABS
ABSCALE
ABSCOR
ACORN
ADDREF
ADSC
AMBER
AMPLE
AMoRE

    - Boundary conditions

Allowed Boundary Conditions

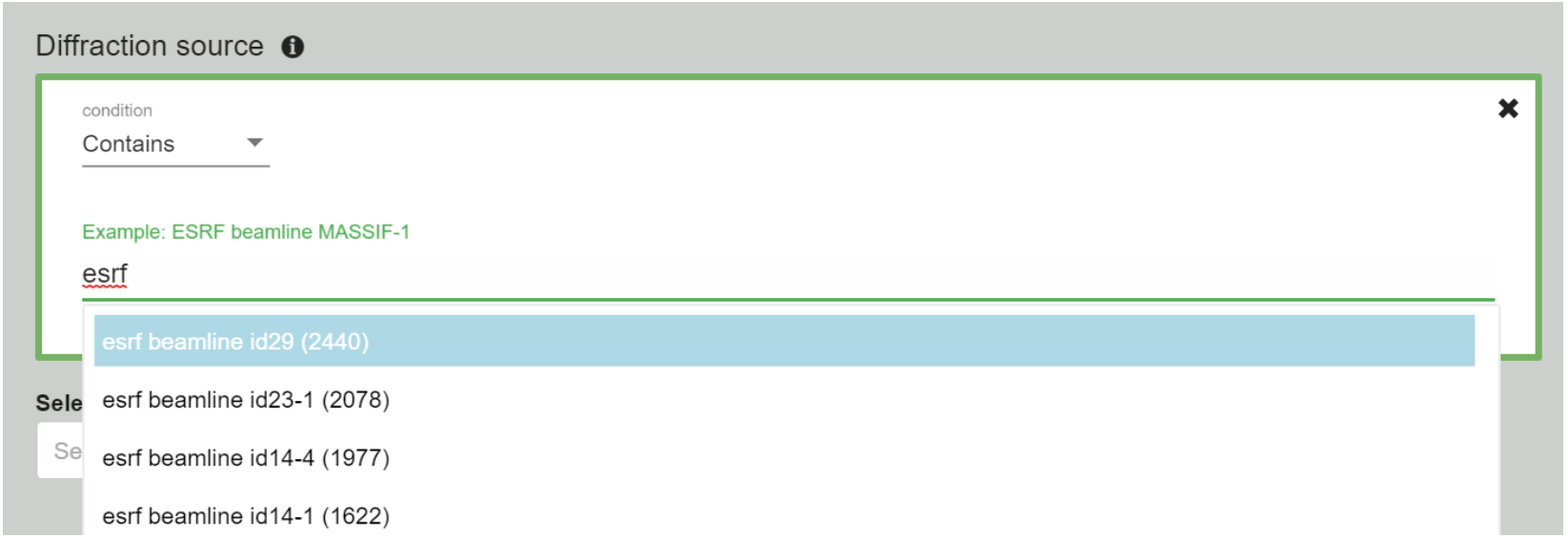
Minimum Value	Maximum Value
0.0	14.0

Advisory Boundary Conditions

Minimum Value	Maximum Value
3.5	10

Data validation

- Why have enumerations?
  - So users can find data



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 ! Warning: Expected a value between 80.0 and 300.0

Data collection temperature details:

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 ⚠ Value not in the enumeration list for this item.

Detector type*: ⚠ No value present for this mandatory item.

Date of collection*:

Optics:

DECTRIS EIGER X 16M
DECTRIS EIGER X 500K
DECTRIS EIGER R 1M
DECTRIS EIGER R 4M
DECTRIS EIGER X 1M
DECTRIS EIGER X 9M
DECTRIS EIGER X 4M

Continue to next section

Sanity Check

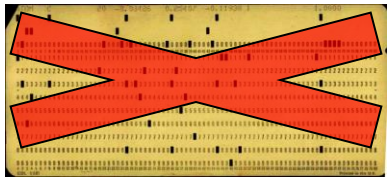
Controlled vocabulary

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 transistion

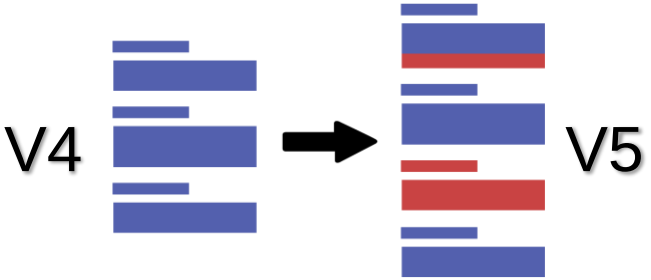
PDBx/mmCIF V5



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.bmrb.wisc.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

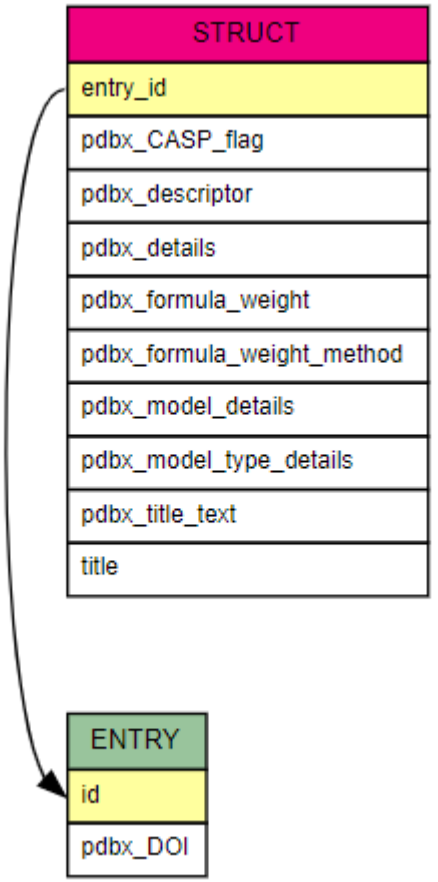
- mmCIF is effectively a relational database with primary and foreign keys relating tables (categories) within the file

Hierarchy within an mmCIF file

- Top level is
  - entry.id
- This has lots of children
- http://mmcif.wwpdb.org/dictionaries/mmcif_pdbx_v50.dic/Items/_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/Items/_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