BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 7 2005, pages 988–992
doi:10.1093/bioinformatics/bti082

Structural bioinformatics

PDBML: the representation of archival macromolecular structure data in XML

John Westbrook[1,*], Nobutoshi Ito[2], Haruki Nakamura[3], Kim Henrick[4] and Helen M. Berman[1]

[1]Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, 610 Taylor Road, Piscataway, NJ 08854, USA, [2]Protein Data Bank Japan (PDBj), School of Medical Science, Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan, [3]Protein Data Bank Japan (PDBj), Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan and [4]EMBL Outstation, The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received on April 14, 2004; revised on May 10, 2004; accepted on October 8, 2004
Advance Access publication October 27, 2004

ABSTRACT

Summary: The Protein Data Bank (PDB) has recently released versions of the PDB Exchange dictionary and the PDB archival data files in XML format collectively named PDBML. The automated generation of these XML files is driven by the data dictionary infrastructure in use at the PDB. The correspondences between the PDB dictionary and the XML schema metadata are described as well as the XML representations of PDB dictionaries and data files.

Availability: The current software translated XML schema file is located at http://deposit.pdb.org/pdbML/pdbx-v1.000.xsd, and on the PDB mmCIF resource page at http://deposit.pdb.org/mmcif/. PDBML files are stored on the PDB beta ftp site at ftp://beta.rcsb.org/pub/pdb/uniformity/data/XML

Contact: jwest@rcsb.rutgers.edu

The Protein Data Bank (PDB) (Bernstein *et al.*, 1977; Berman *et al.*, 2000) is the single worldwide repository for macromolecular structure data. For more than 30 years (Bernstein *et al.*, 1977), the PDB has used a column-oriented data format to store archival entries. This format resembles many other data formats constrained by the limitations of paper punch card technology. Examples of the data format are shown in Figure 1.

The representation of coordinate, sequence, secondary structure and citation data in the PDB has remained remarkably stable since the original format definition in 1972. The data records in the PDB format are prefixed with a record tag (e.g. CRYST1, ATOM) followed by individual items of data (Figure 1a). The specifications for the records in this data format are described informally in the PDB Content Guide: Atomic Coordinate Entry Format Description (Callaway *et al.*, 1996). The description of the experimental details of structure determination has been encoded largely in the form of remark records. Although these records have some internal structure, the organization of these records has changed over time. For example, in Figure 1b the details of refinement are presented as unstructured text, and in Figure 1c these details are presented as semi-structured remark records.

The growing interest in database development and electronic publication in the late 1980s created a need for a more structured representation of PDB data. In 1990, the International Union of Crystallography (IUCr) commissioned a working group (see http://ndbserver.rutgers.edu/mmcif/background/index.html; Fitzgerald *et al.*, 2004) to develop macromolecular extensions to the data representation used to describe small molecule structures and crystallographic structure determination, called the Crystallographic Information File (CIF) (Hall *et al.*, 1991). The CIF representation had been designed and deployed by the IUCr to support electronic publication of small molecule crystal structures. The efforts of the working group and many community experts lead to the development of the macromolecular Crystallographic Information Framework (mmCIF) dictionary. The first version of this data dictionary released in 1996 contained 1700 data definitions (Bourne *et al.*, 1997). The content of the mmCIF dictionary, a superset of PDB crystallographic content, included detailed definitions describing macromolecular structure and the current state of the macromolecular crystallographic experiment.

In 1998, the Research Collaboratory for Structure Bioinformatics (RCSB) assumed the management of the PDB. RCSB adopted the mmCIF data dictionary as the foundation of their data processing and data management infrastructure. The members of the worldwide PDB (wwPDB) (Berman *et al.*, 2003), that includes the RCSB, the Macromolecular Structure Database (MSD) at the European Bioinformatics Institute (EBI) and the PDBj at Osaka University, have collaborated to extend the mmCIF dictionary to include all of the data managed and distributed by the PDB. These extensions include data definitions describing internal bookkeeping, non-crystallographic structure determination methods (e.g. NMR and cryo-electron microscopy), greater detail in experimental crystallography and the details of protein production. These extensions are collected into the PDB Exchange data dictionary (Westbrook *et al.*, 2004a). This data dictionary provides the foundation for the generation of XML schema

[*]To whom correspondence should be addressed.

```
(a) Structured PDB records

CRYST1   63.150   83.590   53.800  90.00  99.34  90.00 P 21          4
ORIGX1      .963457   .136613   .230424       16.61000
ORIGX2     -.158977   .983924   .081383       13.72000
ORIGX3     -.215598  -.115048   .969683       37.65000
SCALE1      .015462   .002192   .003698         .26656
SCALE2     -.001902   .011771   .000974         .16413
SCALE3     -.001062  -.001721   .018728         .75059
ATOM      1  N   VAL A   1       6.130  16.559   4.905  7.00 41.29
ATOM      2  CA  VAL A   1       6.870  17.784   4.702  6.00 41.33
ATOM      3  C   VAL A   1       8.377  17.548   4.913  6.00 31.64
ATOM      4  O   VAL A   1       8.820  16.980   5.922  8.00 38.31
ATOM      5  CB  VAL A   1       6.345  18.763   5.731  6.00 52.26
ATOM      6  CG1 VAL A   1       6.745  20.188   5.356  6.00 52.75
ATOM      7  CG2 VAL A   1       4.826  18.612   5.847  6.00 58.75
```

```
(b) Unstructured PDB records

REMARK   3 REFINEMENT. BY THE METHOD OF JACK AND LEVITT.  THE SCALE
REMARK   3  FACTOR BETWEEN ENERGY AND X-RAY FORMS WAS VARIED BETWEEN
REMARK   3  .00025 AND .0005 TO MAINTAIN THE RMS VARIATION OF THE C-C
REMARK   3  SINGLE BONDS BETWEEN 0.02 AND 0.03 ANGSTROMS.  THE IRON
REMARK   3  ATOMS WERE UNRESTRAINED IN ORDER TO AVOID ANY POSSIBLE
REMARK   3  BIAS IN THEIR POSITIONS.  THE FINAL R VALUE IS 0.16.
```

```
(c) Semi-structured PDB records

REMARK   3  FIT TO DATA USED IN REFINEMENT.
REMARK   3   CROSS-VALIDATION METHOD          : THROUGHOUT
REMARK   3   FREE R VALUE TEST SET SELECTION  : RANDOM
REMARK   3   R VALUE            (WORKING SET) : 0.213
REMARK   3   FREE R VALUE                     : 0.257
REMARK   3   FREE R VALUE TEST SET SIZE   (%) : 7.500
REMARK   3   FREE R VALUE TEST SET COUNT      : 2532
REMARK   3   ESTIMATED ERROR OF FREE R VALUE  : 0.005
```

Fig. 1. Excerpts of records from a PDB data files. (**a**) Structured PDB records describing crystallographic cell constants (CRYST1), transformation matrices between orthogonal and fractional coordinates (ORIGX and SCALE) and the atomic coordinates (ATOM). (**b**) Unstructured PDB records describing the details of crystallographic refinement used in PDB data files before 1996. (**c**) Semi-structured PDB records describing crystallographic refinement used in PDB data files after 1996.

(World Wide Web Consortium, 2001a,b,c) and XML data files described in the remainder of this article.

XML SCHEMA FOR PDB DATA, PDBML

The representation of PDB data in XML builds from the content of the PDB Exchange dictionary, both for assignment of data item names and for defining data organization. Although presented in very different syntaxes, the PDB Exchange and XML representations use the same logical data organization. A side effect of maintaining a logical correspondence with the PDB Exchange representation is that the PDBML files lack the hierarchical structure characteristic of many XML data applications. However, preserving the logical data model of the PDB Exchange dictionary has three important advantages. First, the semantics of PDB data are completely preserved across the two formats. Second, the translation of the PDB Exchange dictionary and PDB Exchange data files to XML is greatly simplified. Third, the straightforward mapping of PDB data to relational database systems is retained.

The correspondences between the metadata attributes used in the PDB Exchange dictionary (Westbrook and Bourne, 2000; Westbrook *et al*., 2004b) and those of XML schema are summarized in Table 1. The top level of scope in the PDB Exchange dictionary or data file is the data block. The data block encloses complete data dictionaries or data entries. The dictionary data block is mapped to

Table 1. Summary of the correspondences between PDB Exchange data dictionary and XML schema metadata

PDB Exchange data dictionary attributes	XML schema mapping
Data block	Root level *schema element*
Category groups	
Categories	*complexTypes*
Definition	*annotation* and *documentation* elements
Examples	*annotation* and *documentation* elements
Primary keys	*attributes* of the data category
Items	*elements* of the data category
Definition	*annotation* and *documentation* elements
Examples	annotation and documentation elements
Data types	*simpleTypes*
Range restrictions and allowed values	*restrictions* within *simpleTypes* or *unions* of *simpleTypes*
Mandatory data code	Element attributes *minOccurs* and *nillable*
Parent-child relationships	*key/keyref* elements
Interdependency/exclusivity	
Units of measurement	Additional *fixed attributes*
Subcategory membership	

```
(a) Abbreviated example PDB exchange data file

data_EXAMPLE
_entity_poly.entity_id    1
_entity_poly.type         polypeptide(L)
_entity_poly.nstd_linkage no
_entity_poly.nstd_monomer no
_entity_poly.pdbx_seq_one_letter_code
 ;DIVLTQSPASLSASVGETVTITCRASGNIHNYLAWYQQKQGKSPQLLVYYTTTLADG
VPSRFSGSGSGTQYSLKINSLQPEDFGSYYCQHFWSTPRTFGGGTKLEIK
;
_entity_poly.pdbx_seq_one_letter_code_can
;DIVLTQSPASLSASVGETVTITCRASGNIHNYLAWYQQKQGKSPQLLVYYTTTLADG
VPSRFSGSGSGTQYSLKINSLQPEDFGSYYCQHFWSTPRTFGGGTKLEIK
;
```

```
(b) Abbreviated example XML data file

<?xml version="1.0" encoding="UTF-8" ?>
<PDBx:datablock  datablockName="EXAMPLE"
   xmlns:PDBx="http://deposit.pdb.org/pdbML/pdbx-v1.000.xsd"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xsi:schemaLocation="http://deposit.pdb.org/pdbML/pdbx-v1.000.xsd
            pdbx-v1.000.xsd">
   <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
          VPSRFSGSGSGTQYSLKINSLQPEDFGSYYCQHFWSTPRTFGGGTKLEIK
         </PDBx:pdbx_seq_one_letter_code>
         <PDBx:pdbx_seq_one_letter_code_can>
          DIVLTQSPASLSASVGETVTITCRASGNIHNYLAWYQQKQGKSPQLLVYYTTTLADG
          VPSRFSGSGSGTQYSLKINSLQPEDFGSYYCQHFWSTPRTFGGGTKLEIK
         </PDBx:pdbx_seq_one_letter_code_can>
      </PDBx:entity_poly>
   </PDBx:entity_polyCategory>
</PDBx:datablock>
```

Fig. 2. Examples of PDB Exchange data and PDBML data representations. (**a**) PDB Exchange data file example with a single category describing some of the features of polymer molecule, (**b**) the corresponding example of polymer description in a PDBML data file.

the standard top-level XML *schema* element, and the data file data block is mapped to a *datablock* element. The *schema* and *datablock* elements provide namespace definitions, linkages to the supporting XML schema definition documents and linkages to the location of the current supporting schema.

Category or table definitions in the Exchange dictionary are described as XML *complexTypes*. The category definition and examples are mapped to XML *annotation* and *documentation* elements. The data items within the category are defined as an unordered sequence of XML elements named according to the attribute portion of their dictionary equivalents. The special data items that form the primary key for the category are defined as XML attributes.

Individual data items have a definition and an optional set of examples. The item-level definition and examples are mapped to XML *annotation* and *documentation* elements. Parent–child relationships between data items in the Exchange dictionary are represented as XML *key* and *keyref* elements. All parent data items are identified as named XML *keys*, and their associated children are identified as named XML *keyrefs*. Primitive data types in the Exchange dictionary are described as XML *simpleTypes*. Allowed ranges are represented as *restriction* elements within *simpleTypes*. Complicated boundary conditions are represented as unions of *simpleTypes* containing *restriction* elements. Controlled vocabularies and allowed values are represented as *simpleTypes* with *restrictions* including *enumeration* elements. Where physical units of measurement are included in a definition in the absence of any other range restrictions, this information is mapped to a XML *simpleContent* element containing a *fixed attribute* element representing the measurement units. There are currently no mappings for the item-level dictionary attributes describing item-level interdependency, exclusivity or subcategory membership.

The correspondences between the PDB dictionary and XML schema metadata described in this section make the automatic translation of the PDB dictionary to XML schema possible. The current software translated XML schema file is located at http://deposit.pdb.org/pdbML/pdbx-v1.000.xsd, and on the PDB mmCIF resource page at http://deposit.pdb.org/mmcif/.

PDBML DATA FILES

The PDBML data files follow the same logical organization as their PDB Exchange data file counterparts. Figure 2 provides an abbreviated example comparing the presentation of a category describing polymer features in the two syntaxes. In Figure 2a, a single row of the *entity_poly* data category is illustrated within a data block named *EXAMPLE*. The corresponding XML representation of this information is shown in Figure 2b. Here the root-level enclosing XML *datablock* element identifies the namespace and the associated schema files. The *entity_poly* data category is enclosed by an XML *entity_polyCategory* element. Each row of the category is defined within an XML *entity_poly* element where the category key, *entity_id*, is included as an XML attribute.

```
(a) Example of a fully marked-up PDBML atom record

<PDBx:atom_siteCategory>
      <PDBx:atom_site id="1">
         <PDBx:group_PDB>ATOM</PDBx:group_PDB>
         <PDBx:type_symbol>N</PDBx:type_symbol>
         <PDBx:label_atom_id>N</PDBx:label_atom_id>
         <PDBx:label_comp_id>ASP</PDBx:label_comp_id>
         <PDBx:label_asym_id>A</PDBx:label_asym_id>
         <PDBx:label_entity_id>1</PDBx:label_entity_id>
         <PDBx:label_seq_id>1</PDBx:label_seq_id>
         <PDBx:Cartn_x>23.482</PDBx:Cartn_x>
         <PDBx:Cartn_y>-0.621</PDBx:Cartn_y>
         <PDBx:Cartn_z>-1.419</PDBx:Cartn_z>
         <PDBx:occupancy>1.00</PDBx:occupancy>
         <PDBx:B_iso_or_equiv>35.27</PDBx:B_iso_or_equiv>
         <PDBx:auth_seq_id>1</PDBx:auth_seq_id>
         <PDBx:auth_comp_id>ASP</PDBx:auth_comp_id>
         <PDBx:auth_asym_id>A</PDBx:auth_asym_id>
         <PDBx:auth_atom_id>N</PDBx:auth_atom_id>
         <PDBx:pdbx_PDB_model_num>1</PDBx:pdbx_PDB_model_num>
      </PDBx:atom_site>
</PDBx:atom_siteCategory>

(b) Example of the alternative simplified XML PDB atom record

<category_atom_record>
  <atom_record id="1">
   ATOM 1 A A 1 1 ? . ASP ASP N N N 23.482 -0.621 -1.419 1.00 35.27
  </atom_record>
</category_atom_record>
```

Fig. 3. Examples of PDBML atom records. (**a**) Example of a fully marked-up PDBML atom record. The content of this record is equivalent to the content of the PDB Exchange data file. Empty data records are not translated to the XML data file. (**b**) Example of a simplified PDBML atom record. The information in this record is also the equivalent to the PDB Exchange data file; however, it is formatted as a white-space delimited string.

The remaining data items in the row are represented as XML elements.

The XML organization illustrated in Figure 2b is repeated for each data category in the data file. Because of its size the *atom_site* category is also represented in an alternative form. Examples of the fully marked-up atom record and the simplified alternative are shown in Figure 3. The alternative representation of the *atom_site* category in Figure 3b simplifies the fully marked-up style in Figure 3a by presenting the data items within the *atom_site* category in a white-space delimited string. The current schema fragment describing the alternative *atom_site* representation is located at http://deposit.pdb.org/pdbML/pdbx-v1.000-alt.xsd.

PDBML files are stored on the PDB beta ftp site at ftp://beta.rcsb.org/pub/pdb/uniformity/data/XML. The files are updated during each weekly PDB update. These files are currently under beta test. Comments and data issues related to these files may be reported at http://pdb-forum.rutgers.edu/. Three XML data files are produced from each PDB Exchange data file. One XML file contains the fully marked-up translation of the PDB Exchange data file. A second XML file contains the full PDB Exchange data file content omitting coordinate data. A third XML file contains only the simplified representation of the coordinate data in which each atom record is marked up as a single XML string.

SUPPORTING SOFTWARE TOOLS

The XML schema and data files described in this article are produced by software translation of the PDB Exchange dictionary and data files, respectively. The software tools that RCSB has developed to automate the production of XML schema and dictionaries can be downloaded from the website http://deposit.pdb.org/mmcif/MMCIF-XML-UTIL/. The molecular graphics viewer, PDBjViewer (Kinoshita and Nakamura, in press) that PDBj has developed can parse the current PDBML files to display macromolecular structures http://www.pdbj.org/PDBjViewer/. These tools are available in full source under an Open Source software license. The XML-based Protein Structure Search Service (xPSSS) is a browser with the XPath-SOAP service, based on the PDBML files using a native XML-DB at PDBj http://www.pdbj.org/xpsss/.

ACKNOWLEDGEMENTS

The authors thank Ms Kaori Kobayashi, Mr Hisashi Sakamoto, Ms Reiko Yamashita, Dr Daron Standley and Dr Arno Paehler at PDBj for their help to develop the schema and validation system for the PDBML. The RCSB PDB is operated by Rutgers, The State University of New Jersey; The San Diego Supercomputer Center at the University of California, San Diego; and the Center for Advanced Research in Biotechnology of the National Institute of Standards and Technology—three members of the Research Collaboratory for Structural Bioinformatics (RCSB). The work reported in this paper has been supported by grants from the NSF, NIGMS, DOE, NLM, NCI, NCRR, NIBIB and NINDS. The MSD-EBI is supported by funds from the Wellcome Trust, the European Union (TEMBLOR, NMRQUAL, SPINE, AUTOSTRUCT, and IIMS awards), CCP4, the Biotechnology and Biological Sciences Research Council (UK), the Medical Research Council (UK), and the European Molecular

Biology Laboratory. The PDBj is supported by grant-in-aid from the Institute for Bioinformatics Research and Development of the Japan Science and Technology Agency (BIRD-JST), and the Ministry of Education, Culture, Sports, Science and Technology (MEXT).

REFERENCES

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig, H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Berman,H.M., Henrick,K. and Nakamura,H. (2003) Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*, **10**, 980.

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F.Jr., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.

Bourne,P.E., Berman,H.M., Watenpaugh,K., Westbrook,J.D. and Fitzgerald,P.M.D. (1997) The macromolecular Crystallographic Information File (mmCIF). *Meth. Enzymol.*, **277**, 571–590.

Callaway,J., Cummings,M., Deroski,B., Esposito,P., Forman,A., Langdon,P., Libeson,M., McCarthy,J., Sikora,J., Xue,D., *et al*. 1996. *Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description*. Brookhaven National Laboratory.

Fitzgerald,P.M.D., Westbrook,J.D., Bourne,P.E., McMahon,B., Watenpaugh,K.D. and Berman,H.M. (2004) Classification and use of macromolecular data. In Hall,S.R. and McMahon,B. (eds), *International Tables for Crystallography vol G*. Kluwer Academic Publishers, Dordrecht (in press).

Hall,S.R., Allen,A.H. and Brown,I.D. (1991) The Crystallographic Information File (CIF): a new standard archive file for crystallography. *Acta Crystallogr A*, **47**, 655–685.

Kinoshita,K. and Nakamura,H. (2004) eF-site and PDBjViewer: database and viewer for protein functional sites. *Bioinformatics*, **20**, 1329–1330.

Westbrook,J. and Bourne,P.E. (2000) STAR/mmCIF: an extensive ontology for macromolecular structure and beyond. *Bioinformatics*, **16**, 159–168.

Westbrook,J., Henrick,K., Ulrich,E.L. and Berman,H.M. (2004a) The Protein Data Bank exchange dictionary. In Hall,S.R. and McMahon,B. (eds), *International Tables for Crystallography*. Kluwer Academic Publishers, Dordrecht (in press).

Westbrook,J.D., Berman,H.M. and Hall,S.R. (2004b) Specification of a relational Dictionary Definition Language (DDL2). In Hall,S.R. and McMahon,B. (eds), *International Tables for Crystallography*. Kluwer Academic Publishers, Dordrecht (in press).

World Wide Web Consortium (2001a) In Fallside,D.C. (eds), *XML Schema Part 0: Primer, W3C Recommendation*. http://www.w3.org/TR/2001/REC-xmlschema-2000-20010502/.

World Wide Web Consortium (2001b) In Thompson,H.S., Beech,D., Maloney,M. and Mendelsohn,N. (eds), *XML Schema Part 1: Structures W3C Recommendation*. W3C. http://www.w3.org/TR/2001/REC-xmlschema-2001–20010502/.

World Wide Web Consortium (2001c) In Biron,P.V. and Malhotra,A. (eds), *XML Schema Part 2: Datatypes W3C Recommendation*. W3C pp. http://www.w3.org/TR/xmlschema-2/.