PDBによってリリースされたXML
ＰＤＢＭＬ(仮称)の紹介

坂本　久※2、小林　香織※2、Daron Standley※2、伊藤　暢聡※3
中村　春木※1

※1 大阪大学蛋白質研究所
※2 科学技術振興事業団
※3 東京医科歯科大学

PDBj
Protein Data Bank Japan

紹介内容

- ＰＤＢＭＬの特徴
- ＰＤＢＭＬの構造について
- ＰＤＢＭＬ応用サービス
- ＸＭＬ　ＮａｔｉｖｅＤＢ＋ＸＱｕｅｒｙ検索を用いたＰＤＢＭＬ検索サービス

PDBMLとは

- Protein Data Bank
  Markup Language
- ベースはmmCIF
  カテゴリやアイテムの構造をタグ/属性に
- XMLバリデーションによる
  不正データ検出を強化

PDBj

PDBMLの特徴

- mmCIFのカテゴリ/アイテムの構造/名前との互換性を保持
- データの大半を占める座標データは
  外部ファイル化が可能
- XML Schema形式による構造の定義

URL: http://deposit.pdb.org/pdbML/pdbx-v0.904.xsd

外部ファイル版URL: http://deposit.pdb.org/pdbML/pdbx-v0.904-alt.xsd

PDBj

mmCIF とは

- 国際結晶学会(IUCr)主導で開発
- 低分子で広く用いられている CIF (Crystallographic Information Format)を生体高分子用に拡張したもの
- *name* と *value* の対で構成されており XML（*tag* と *content* の対）への変更が容易

```
_name value
```

↓

```
<tag> content </tag>
```

PDBj

mmCIFファイル例

```
data_1CRN
#
loop_
_audit_author.name
'Hendrickson, W.A.'
'Teeter, M.M.'
#
_cell.entry_id              1CRN
_cell.length_a              40.960
_cell.length_b              18.650
_cell.length_c              22.520
_cell.angle_alpha           90.00
_cell.angle_beta            90.77
_cell.angle_gamma           90.00
_cell.Z_PDB                 2
_cell.pdbx_unique_axis      ?
#
```

PDBj

```
loop_
_pdbx_poly_seq_scheme.asym_id
_pdbx_poly_seq_scheme.entity_id
_pdbx_poly_seq_scheme.seq_id
_pdbx_poly_seq_scheme.mon_id
_pdbx_poly_seq_scheme.ndb_seq_num
_pdbx_poly_seq_scheme.pdb_seq_num
_pdbx_poly_seq_scheme.auth_seq_num
_pdbx_poly_seq_scheme.pdb_mon_id
_pdbx_poly_seq_scheme.auth_mon_id
_pdbx_poly_seq_scheme.pdb_strand_id
A 1 1  THR  1  1  1  THR THR  A
A 1 2  THR  2  2  2  THR THR  A
A 1 3  CYS  3  3  3  CYS CYS  A
A 1 4  CYS  4  4  4  CYS CYS  A
A 1 5  PRO  5  5  5  PRO PRO  A
A 1 6  SER  6  6  6  SER SER  A
A 1 7  ILE  7  7  7  ILE ILE  A
A 1 8  VAL  8  8  8  VAL VAL  A
A 1 9  ALA  9  9  9  ALA ALA  A
A 1 10 ARG 10 10 10  ARG ARG  A
```

PDBj

PDBMLの構造(1)～Rootタグ～

- 全てのタグ/属性は名前空間"PDBx"に所属
- Rootタグ：<datablock>
  datablockName属性 → PDB IDを保持

```
<?xml version="1.0" encoding="UTF-8" ?>
<PDBx:datablock  datablockName="1CRN"
  xmlns:PDBx="http://deposit.pdb.org/pdbML/pdbx-0.904.xsd"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="……/pdbx-v0.904.xsd">
```

PDBj

PDBMLの構造(2) ～カテゴリのTag化～

- Rootタグ直下にmmCIFカテゴリがタグ化
  例えば…_entity_poly_seq カテゴリ
  →<PDBx:entity_poly_seqCategory>

```
<PDBx:datablock  datablockName="1CRN">
        :
    <PDBx:entity_poly_seqCategory>
        :
    </PDBx:entity_poly_seqCategory>
        :
</PDBx:datablock>
```

PDBj

PDBMLの構造(3) ～アイテムのTag化～

- <xxxCategory>タグ直下に<xxx>タグを配置(xxxはmmCIFカテゴリ名)
- アイテムはタグ化/属性化され<xxx>タグ以下に配置
- mmCIFで複数アイテムが_loopキーワードで繰り返し記述されている場合
  → <xxx>タグが繰り返し存在する

PDBj

例1: _entity カテゴリ

```
_entity.id                          1
_entity.type                        polymer
_entity.src_method                  man
_entity.pdbx_description            CRAMBIN
_entity.formula_weight              4743.503
_entity.pdbxnumber_of_molecules     1
```

```
<PDBx:entityCategory>
  <PDBx:entity id="1">
    <PDBx:type>polymer</PDBx:type>
    <PDBx:src_method>man</PDBx:src_method>
    <PDBx:pdbx_description>CRAMBIN</PDBx:pdbx_description>
    <PDBx:formula_weight>4743.503</PDBx:formula_weight>
    <PDBx:pdbx_number_of_molecules>1</PDBx:pdbx_number_of_molecules>
  </PDBx:entity>
</PDBx:entityCategory>
```

PDBj

例2: _entity_poly_seq カテゴリ

```
loop_
_entity_poly_seq.entity_id
_entity_poly_seq.num
_entity_poly_seq.mon_id
1 1  THR
1 2  THR
1 3  CYS
1 4  CYS
```

```
<PDBx:entity_poly_seqCategory>
   <PDBx:entity_poly_seq entity_id="1" num="1" mon_id="THR"/>
   <PDBx:entity_poly_seq entity_id="1" num="2" mon_id="THR"/>
   <PDBx:entity_poly_seq entity_id="1" num="3" mon_id="CYS"/>
   <PDBx:entity_poly_seq entity_id="1" num="4" mon_id="CYS"/>
        :
</PDBx:entity_poly_seqCategory>
```

PDBj

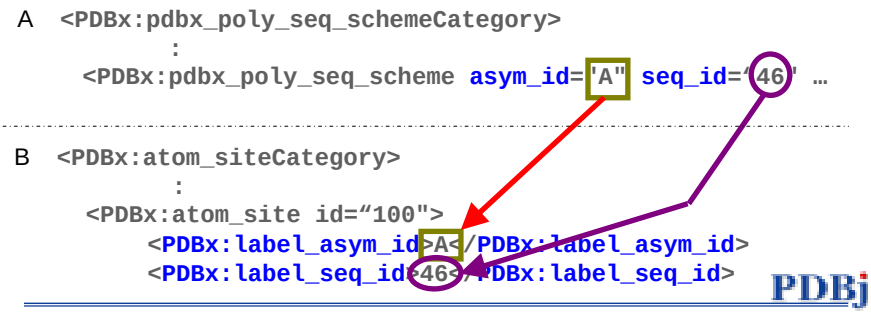
PDBMLの構造(4)～データ型～

- アイテムのデータ型を厳密に定義
  a. 文字列
  b. 数値(整数、単精度実数、倍精度実数)
     数値の範囲も厳密に定義
  c. 複雑な型(日時、URLなど)
- XMLバリデーション時に、定義型と違うデータはエラー出力

PDBj

PDBMLの構造(5)～データ参照～

- AからBのデータ参照がある場合、正しくBのデータを参照しているかをバリデーション時にチェック



PDBj

ＰＤＢＭＬの構造(6)～外部ファイル化～

- 座標データ：情報の大部分を占めるが検索の使用頻度は低い
  →外部ファイル化して本体を扱いやすく
- 外部ファイル内の情報もフォーマットの簡略化などでファイルサイズを極力低減

PDBj

```
<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>THR</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>17.047</PDBx:Cartn_x>
           <PDBx:Cartn_y>14.099</PDBx:Cartn_y>
           <PDBx:Cartn_z>3.625</PDBx:Cartn_z>
           <PDBx:occupancy>1.00</PDBx:occupancy>
           <PDBx:B_iso_or_equiv>13.79</PDBx:B_iso_or_equiv>
           <PDBx:auth_seq_id>1</PDBx:auth_seq_id>
           <PDBx:auth_comp_id>THR</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>
```

```
<atom_record id="1">ATOM 1 A A 1 1 ? . THR THR N N N 17.047 14.099 3.625 1.00 13.79</atom_record>
```

PDBj

外部ファイル例

```
<?xml version="1.0" encoding="UTF-8"?>

<datablock datablockName="1crn"
    xmlns="http://deposit.pdb.org/pdbML/"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://deposit.pdb.org/pdbML/
    http://deposit.pdb.org/pdbML/pdbx-v0.904-alt.xsd">
    <category_atom_record>
        <atom_record id="1"> ATOM 1 A A A 1 1 ? . THR THR N N N 17.047 14.099 3.625 1.00 13.79</atom_record>
        <atom_record id="2"> ATOM 1 A A A 1 1 ? . THR THR C CA CA 16.967 12.784 4.338 1.00 10.80</atom_record>
        <atom_record id="3"> ATOM 1 A A A 1 1 ? . THR THR C C C 15.685 12.755 5.133 1.00 9.19</atom_record>
        <atom_record id="4"> ATOM 1 A A A 1 1 ? . THR THR O O O 15.268 13.825 5.594 1.00 9.85</atom_record>
        <atom_record id="5"> ATOM 1 A A A 1 1 ? . THR THR C CB CB 18.170 12.703 5.337 1.00 13.02</atom_record>
        <atom_record id="6"> ATOM 1 A A A 1 1 ? . THR THR O OG1 OG1 19.334 12.829 4.463 1.00 15.06</atom_record>
        <atom_record id="7"> ATOM 1 A A A 1 1 ? . THR THR C CG2 CG2 18.150 11.546 6.304 1.00 14.23</atom_record>
        <atom_record id="8"> ATOM 1 A A A 2 2 ? . THR THR N N N 15.115 11.555 5.265 1.00 7.81</atom_record>
        <atom_record id="9"> ATOM 1 A A A 2 2 ? . THR THR C CA CA 13.856 11.469 6.066 1.00 8.31</atom_record>
```

PDBj

PDBML応用サービス

- XML-DB_PDBjデータベースの構築
- Webサービスの提供
- XML Nativeデータベース化
- XQueryを用いたXML検索サービス

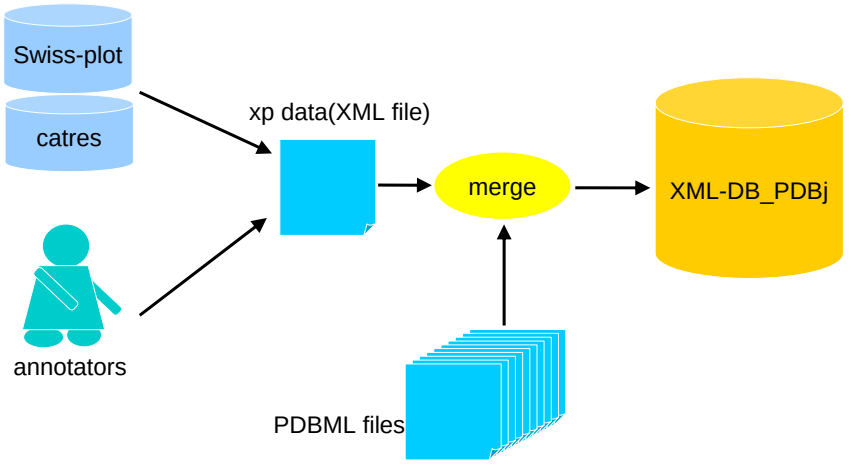
PDBj

XML-DB_PDBjデータベースの構築

- 以下の情報をＰＤＢj独自で収集/登録
  1．　他データベースからの機能情報
    ①　swiss-plot
    ②　catres(EBI)
  2．　ＰＤＢ(mmCIF)に欠損している情報を論文等各種文献から抽出

PDBj

XML-DB_ＰＤＢj作成シーケンス



PDBj

XML-DB_PDBjへの追加情報数

Total number in PDBML	21394
欠損情報追加 (by annotators)	3329
機能情報(from SwissProt)	8109
機能情報(from CATRES)	178

(2003/6/21現在)

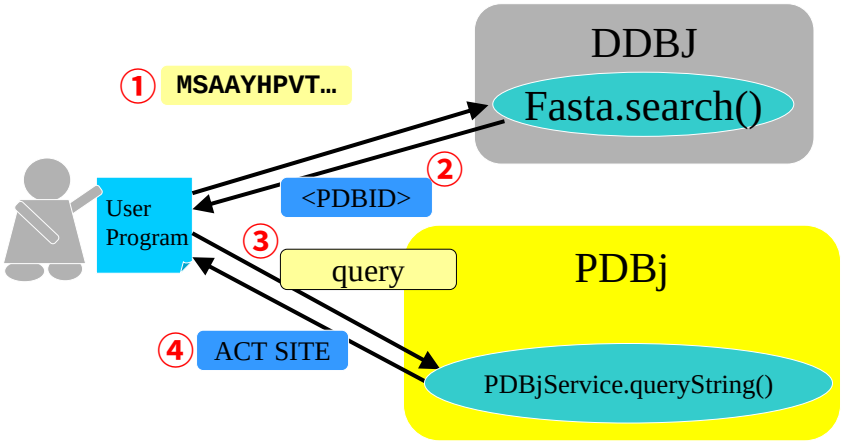
PDBj

Webサービスの提供

- SOAP(Simple Object Access Protocol)によるWebサービス
  - 通信にXMLを用いた技術
  - ユーザプログラムからPDBMLデータベースに対する操作が可能に
  - SOAPによってファイアフォール外からのアクセスも可能
  - 以下のようなサービスを提供予定
    1. PDBMLファイルダウンロード
    2. XQuery検索サービス(後述)

PDBj

SOAPを用いたWebサービス



PDBj

SOAPプログラム例

```
public class PDBjXPath {
    public static void main ( String args[] ) {
        try {
            String wsdlURL = "http://www.pdbj.org/wsdl/PDBjSoapService.wsdl";
            String nmspace = "http://www.pdbj.org/soap";
            String srvname = "PDBjSoapService";
            String fncname = "queryString";
            String query   = args[0];
            Integer rstnum  = new Integer( args[1] );

            QName serviceQN = new QName( nmspace, srvname );
            QName portQN    = new QName( nmspace, srvname );
            Service service = new Service( new URL(wsdlURL), serviceQN );

            Call call = ( Call )service.createCall( portQN, fncname );
            String result = (String)call.invoke( new Object[] {query,rstnum} );
            System.out.println( result );
        }
        catch ( Exception e ) {
            e.printStackTrace();
        }
    }
}
```

[java + apache axis使用]

PDBj

XML Nativeデータベース化

- XML構造のままデータベースに蓄積
- XML標準の検索言語XQueryやデータ位置指定表記XPathを用いてXMLの構造を意識した検索が可能
- 従来のRDBMSでのテーブル設計がXML構造設計に相当
  - 設計が容易に
  - 既存XMLの場合は設計不要
- フリーウェアは規模・速度面で苦しい
- 商用ソフトウェアの導入

PDBj

XQueryを用いたXML検索サービス

- 簡単なXQuery (XPath) 検索
- PDB Searchfieldと同様のQuery
- 複雑な検索

PDBj

簡単なXQuery (XPath) 検索

- 全ての蛋白質のPDBIDを表示

```
/datablock/@datablockName
```

- 中村春木の解析した蛋白質を検索

```
/datablock[citation_authorCategory/
  citation_author/@name="Nakamura, H"]
    /@datablockName
```

- 10残基以上のヘリックスを持つ蛋白質を検索

```
/datablock[struct_confCategory/struct_conf/
  pdbx_PDB_helix_length>="10"]/@datablockName
```

PDBj

Author情報(citation_authorカテゴリ)

```
<PDBx:citation_authorCategory>
        <PDBx:citation_author citation_id="primary" name="author名"/>
        <PDBx:citation_author citation_id="1" name="author名"/>
        <PDBx:citation_author citation_id="2" name="author名"/>
</PDBx:citation_authorCategory>
```

[例]

```
<PDBx:citation_authorCategory>
       <PDBx:citation_author citation_id="primary" name="Teeter, M.M.">
       <PDBx:citation_author citation_id="1" name="Hendrickson, W.A.">
       <PDBx:citation_author citation_id="1" name="Teeter, M.M.">
       <PDBx:citation_author citation_id="2" name="Teeter, M.M.">
       <PDBx:citation_author citation_id="2" name="Hendrickson, W.A.">
</PDBx:citation_authorCategory>
```

PDBj

Helix情報(struct_confカテゴリ)

```
<PDBx:struct_confCategory>
        <PDBx:struct_conf id="conf情報識別子">
           <PDBx:pdbx_PDB_helix_id>Helix識別子</PDBx:pdbx_PDB_helix_id>
           <PDBx:beg_label_comp_id>開始残基のアミノ酸表記(3文字)</PDBx:beg_label_comp_id>
           <PDBx:beg_label_asym_id>開始残基のチェインID</PDBx:beg_label_asym_id>
           <PDBx:beg_label_seq_id>開始残基の残基番号</PDBx:beg_label_seq_id>
           <PDBx:end_label_comp_id>終了残基のアミノ酸表記</PDBx:end_label_comp_id>
           <PDBx:end_label_asym_id>終了残基のチェインID</PDBx:end_label_asym_id>
           <PDBx:end_label_seq_id>終了残基の残基番号</PDBx:end_label_seq_id>
           <PDBx:beg_auth_comp_id>著者による終了残基のアミノ酸表記</PDBx:beg_auth_comp_id>
           <PDBx:beg_auth_asym_id>著者による終了残基のチェインID</PDBx:beg_auth_asym_id>
           <PDBx:beg_auth_seq_id>著者による終了残基の残基番号</PDBx:beg_auth_seq_id>
           <PDBx:end_auth_comp_id>著者による終了残基のアミノ酸表記</PDBx:end_auth_comp_id>
           <PDBx:end_auth_asym_id>著者による終了残基のチェインID</PDBx:end_auth_asym_id>
           <PDBx:end_auth_seq_id>著者による終了残基の残基番号</PDBx:end_auth_seq_id>
           <PDBx:pdbx_PDB_helix_class>Helixクラス番号(種別)</PDBx:pdbx_PDB_helix_class>
           <PDBx:pdbx_PDB_helix_length>Helix長(残基単位)</PDBx:pdbx_PDB_helix_length>
           <PDBx:details>本情報に対する詳細情報</PDBx:details>
        </PDBx:struct_conf>
 </PDBx:struct_confCategory>
```

PDBj

[例]

```
   <PDBx:struct_confCategory>
        <PDBx:struct_conf id="HELX_P1">
           <PDBx:conf_type_id>HELX_P</PDBx:conf_type_id>
           <PDBx:pdbx_PDB_helix_id>H1</PDBx:pdbx_PDB_helix_id>
           <PDBx:beg_label_comp_id>ILE</PDBx:beg_label_comp_id>
           <PDBx:beg_label_asym_id>A</PDBx:beg_label_asym_id>
           <PDBx:beg_label_seq_id>7</PDBx:beg_label_seq_id>
           <PDBx:end_label_comp_id>PRO</PDBx:end_label_comp_id>
           <PDBx:end_label_asym_id>A</PDBx:end_label_asym_id>
           <PDBx:end_label_seq_id>19</PDBx:end_label_seq_id>
           <PDBx:beg_auth_comp_id>ILE</PDBx:beg_auth_comp_id>
           <PDBx:beg_auth_asym_id>A</PDBx:beg_auth_asym_id>
           <PDBx:beg_auth_seq_id>7</PDBx:beg_auth_seq_id>
           <PDBx:end_auth_comp_id>PRO</PDBx:end_auth_comp_id>
           <PDBx:end_auth_asym_id>A</PDBx:end_auth_asym_id>
           <PDBx:end_auth_seq_id>19</PDBx:end_auth_seq_id>
           <PDBx:pdbx_PDB_helix_class>1</PDBx:pdbx_PDB_helix_class>
           <PDBx:details>3/10 CONFORMATION RES 17,19</PDBx:details>
           <PDBx:pdbx_PDB_helix_length>13</PDBx:pdbx_PDB_helix_length>
        </PDBx:struct_conf>
```

PDBj

PDB SearchfieldQuery(1)

- ReleaseDateによる検索

```
/datablock[database_PDB_revCategory/
 database_PDB_rev/date_original > “2000-01-01”]
    /@datablockName
```

- CitationAuthorによる検索

```
/datablock[citation_authorCategory/citation_author/
  [@citation_id=“primary” and @name=“Nakamura, H“]]
    /@datablockName
```

PDBj

PDBリビジョン情報 (database_PDB_revカテゴリ)

```
<PDBx:database_PDB_revCategory>
      <PDBx:database_PDB_rev num=“リビジョン識別子">
           <PDBx:date>登録/修正日時(REVDAT)</PDBx:date>
           <PDBx:date_original>初回登録日時(HEADER)</PDBx:date_original>
           <PDBx:author_name>サブミット責任者名</PDBx:author_name>
           <PDBx:replaces>置き換わった新しいPDBID</PDBx:replaces>
           <PDBx:replased_by>置き換えた古いPDBID</PDBx:replaced_by>
           <PDBx:mod_type>変更種別番号</PDBx:mod_type>
       </PDBx:database_PDB_rev>
<PDBx:database_PDB_revCategory>
```

PDBj

［例］

```
<PDBx:database_PDB_revCategory>
        <PDBx:database_PDB_rev num="1">
           <PDBx:date>1981-07-28</PDBx:date>
           <PDBx:date_original>1981-04-30</PDBx:date_original>
           <PDBx:replaces>1CRN</PDBx:replaces>
           <PDBx:mod_type>0</PDBx:mod_type>
        </PDBx:database_PDB_rev>
        <PDBx:database_PDB_rev num="2">
           <PDBx:date>1981-12-03</PDBx:date>
           <PDBx:replaces>1CRNA</PDBx:replaces>
           <PDBx:mod_type>1</PDBx:mod_type>
        </PDBx:database_PDB_rev>
        <PDBx:database_PDB_rev num="3">
           <PDBx:date>1983-09-30</PDBx:date>
           <PDBx:replaces>1CRNB</PDBx:replaces>
           <PDBx:mod_type>1</PDBx:mod_type>
        </PDBx:database_PDB_rev>
        <PDBx:database_PDB_rev num="4">
           <PDBx:date>1985-03-04</PDBx:date>
           <PDBx:replaces>1CRNC</PDBx:replaces>
           <PDBx:mod_type>1</PDBx:mod_type>
        </PDBx:database_PDB_rev>
</PDBx:database_PDB_revCategory>
```

PDBj

PDB SearchfieldQuery(2)

- ContainChainTypeによる検索

```
/datablock[entity_polyCategory/entity_poly/
      type="polypeptide(L)"]/@datablockName
```

- ECnumberによる検索

```
/datablock[entityCategory/entity/
            pdbx_ec="1.1.1.1"]/@datablockName
```

PDBj

複雑な検索

- DNA結合蛋白質でDNAチェイン情報を持つ蛋白質を検索

[Xpathによる検索クエリ]

```
/datablock[/datablock[struct_keywordsCategory/struct_keywords/pdbx_
     keywords='DNA BINDING
     PROTEIN']/entity_polyCategory/entity_poly/
         type='polydeoxyribonucleotide']/@datablockName
```

[Xqueryによる検索クエリ]

```
for $a in /datablock
  [struct_keywordsCategory/struct_keywords/pdbx_keywords=
                              "DNA BINDING PROTEIN"]
 where $a/entity_polyCategory
   [count(entity_poly[type="polydeoxyribonucleotide"])>0]
 return $a/@datablockName
```

PDBj

PDBj XMLDB公開実験

- XML DB＋XQueryを用いた検索サービス
- 期間：6月末～9月末
- バイオインフォマテックスに携わる研究者・技術者にどれだけXMLDBが有効であるかの検証
- 速度面・運用面・検索パターンデータなどのデータ収集→データは解析して公開
- アクセス先アドレス情報、具体的な数値条件
  XMLDB製品名などは公開しない。
- 協力：NEC、三井物産、三井情報開発、
  ビーコンIT

PDBj

PDBjホームページ

- **PDBj Top page**
  http://www.pdbj.org
- **PDBj FTP Server**
  ftp://ftp.pdbj.org
- PDBj XML**データベース公開実験サイト**
  http://www.pdbj.org/XML-DB_PDBj

PDBj