Protein Data Bank
CoreTrustSeal Certification:
A Community Biomedical Archival Data Repository Example

John Westbrook RCSB PDB
Outline

- Brief overview of the PDB resources and services
- PDB support of the CTS principles
- Some challenges supporting CTS
PDB Overview
Protein Data Bank History

- PDB 1st Open Access digital data resource in all of biology
- Founded 1971 with 7 X-ray structures of proteins
- Single global archive for protein and DNA/RNA experimental structures
- Today, Open Access to >150,000 structures
- wwPDB collaboration US (RCSB PDB), EU (PDBe), Japan (PDBJ), and BMRB

Some of the earliest structures in the PDB
Structure Data Contributes to Fundamental Biology, Biomedicine, and Energy

Fundamental Biology
- Nuclear Cell Biology
- Molecular Evolution
- Molecular Transport
- Cellular Signaling
- Protein Folding
- Enzymes
- Nanotechnology
- Molecular Infrastructure

Biomedicine
- Explaining T-cell of Immunology
- Zika Virus
- Cancer
- Precision Medicine
- Anti-Microbial Resistance
- Vaccine Development
- Ebola
- Type II Diabetes

Energy
- Illuminating Biological Energy
- Biofuels
- Renewable Energy
- Cyanobacteria
- Biotechnology
- Crop Sustainability
- Methane Production
- Hydrogen Gas
Impact of PDB Data on Drug Approvals

210 NEW DRUGS approved
2010-2016

>$100 BILLION of NIH funding contributed to these approvals (>95% on targets)¹

2000-2016

>6,000 PDB Structures contributed to 183 of these drug approvals

RCSB PDB Services Support the Full Structural Biology Data Life Cycle

Global Data
1. Deposition/Biocuration
   - Data processing and validation
   - >31,000 Structural Biologists/PDB Data Depositors

2. Archive Management/Access
   - Integration of external annotations

3. Data Exploration
   - RCSB.org

4. Outreach/Education
   - pdb101.rcsb.org
     - Educational Resources

Global Knowledge
- >400 Different Scientific Resources
- >1 Million Unique Users per Year
- >621,000 Unique Users per Year
PDB Support for CTS:
Motivation and Process
PDB Incentives for CTS Certification

- Strong commitment and tradition within our scientific community for support of data and process standards
- Expectations of both our repository contributors and users to adopt and maintain best practices in archiving and data management
- Reinforces our strong commitment to FAIR practices in concert with Increasing focus of funders on supporting FAIR data management
- Certification documents the resource investment required to responsibly manage the full life cycle of archival data
- Relatively low barrier and modest effort certification process
- Good balance between rigor and certification effort
The CTS Certification Process

- Straight forward application process with a variety of examples to frame your input
- The majority of the required information was already in public view or in existing project documents
- The required/expected level of detail is a bit ambiguous
PDB Support for CTS
How we tackled the requirements
Introductory Materials

- Repository type
- Repository community (contributors and users)
- Overview of activities
  - Biocuration policies and practices
  - Preservation of primary data artifacts
- Repository usage
- Repository organization and role

This information provided this at level of detail of our latest grant application and progress report.
I. Mission and Scope

- wwPDB maintains a single archive of macromolecular structural data that are freely and publicly available to the global community

- wwPDB maintains these organizational details on the wwpdb.org resource web site
II. Licenses

- PDB primary data are free of all copyright restrictions and made fully and freely available for both non-commercial and commercial users
- This PDB license pre-dates contemporary open source licenses
- Some additional conditions on adaptation of data protect authenticity of repository data files
- Compliance issues with primary data are rare
- Other PDB software and educational materials are provided under standard open source licenses (e.g., Apache and Creative Commons)
III. Continuity of Access

- 40+ year track record of funding support in US
- wwPDB organization provides for continuity of data and service access if a regional partner site should become unavailable

Regional partners responsible for data from:
- RCSB PDB (US): Americas and Oceania
- PDBe (UK): Europe and Africa
- PDBj (Japan): Asia and Middle East
IV. Confidentiality and Ethics

- Aside from a brief embargo period at deposition time, PDB data are open.
- Personal identifying information (PPI) maintained on contributors is limited to the minimal contact information required to conduct the operations of the archive.
- Usage statistics are presented in aggregate and appropriately anonymized.
- Management of PPI data deemed GDPR compliant.
- No other PPI flows out of the wwPDB.
V. Organizational Infrastructure

- Regional wwPDB partner data centers
- Global load-balancing and failover of deposition services
- Complimentary data access services

Regional partners responsible for data from:
RCSB PDB (US): Americas and Oceania
PDBe (UK): Europe and Africa
PDBj (Japan): Asia and Middle East
## VI. Expert Guidance

**wwPDB Method-specific Community Task Forces**

<table>
<thead>
<tr>
<th>Task Force</th>
<th>Meeting</th>
<th>Chair(s)/Membership</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDBx/mmCIF Working Group</td>
<td>2011 -</td>
<td>Paul Adams (LBL) 13 members</td>
<td>Regular virtual meetings and workshops</td>
</tr>
</tbody>
</table>
VII. Data Integrity and Authenticity

- Community metadata and data standards
- Metadata and data change management policies
- Maintaining consistency through retrospective repository remediation
- Repository snapshotting
- Explicitly versioned data repository
- Expert biocuration
- ORCID identification for depositors
VIII. Appraisal

- Pre-deposition validation services
- wwPDB OneDep Deposition System
- Expert biocuration
- Data delivery in well-defined community data formats
IX. Documented Storage Procedures

- Conformance with OAIS Archive Reference Model

X. Preservation Plan

- Reference public documentation describing the full PDB data life cycle (wwpdb.org)
  - All primary data deposition requirements
  - Deposition, validation, and biocuration policies
  - Transformations during data processing
  - Accessioning, versioning and release processing
  - Post release remediation
  - Repository archiving procedures and repository management
XI. Data Quality

- All PDB deposition, validation and biocuration tools support and enforce Community data standards
- PDB provides validation reports describing compliance with Community data quality standards
- Validation reports tailored for depositors, editorial reviewers, and general users
- PDB validation reports required by most scientific journals describing 3D structure determinations
XII. Workflows

- Review workflows across the PDB data life cycle
- Describe workflow representation and implementation
- Extensibility to increases in data volume and data content
- Workflow change management
XIII. Data Discovery and Identification

RESTful Web Services & GraphQL
XIV. Data Reuse

- Data and metadata requirements for deposition
- Content and format extensibility
- Maintaining repository content and format consistency through retrospective biocuration
- Repository metadata and data content documentation (mmcif.wwpdb.org)
XV. Technical Infrastructure

- Data reference standards and ontologies in use
  - Lengthy and requiring consolidation from many sources
- Full software development and deployment process
- Managing community software tools
- Infrastructure management
- Capacity monitoring and management
XV. Security

- Service availability, redundancy, disaster recovery
- Institutional security protocols and resources
- Application security protocols
  - Coding standards
  - Code review
  - Testing and deployment protocols
  - Version control
Outcomes and Challenges
Supporting CTS
General Benefits of the Certification Process

- Requires an audit of the full life cycle of the repository data pipeline
- Uncovers implicit knowledge of processes that may lack proper documentation
- Useful exercise to identify systematic weaknesses and potential areas for improvement
- Provides an opportunity to explore how other disciplines are addressing similar data management challenges
- Provides a useful benchmark for resource and capacity planning
- Provides an excellent learning experience
Some Certification Outcomes for PDB

- Harmonized practices and documentation across our regional data centers
- Improved alignment of our documentation with FAIR/FACT objectives
- Introspection helped focus our long-term plans to improve availability and disaster preparedness
- Explored some new approaches for schema registration, exchange and data discovery
- Certification beneficially contributed to our funding reapplication
Some CTS and Certification Challenges

- Supporting CTS requires diverse expertise in data science and engineering as well as in the target domain.
- The long time horizon of some CTS objectives are difficult to support with typical 3-5 year competitive funding cycles.
- Addressing long term objectives is similarly complicated for leased or cloud deployed infrastructure.
- The resource burdens for robust CTS support may not be:
  - fully accounted in the scope of current program offerings
  - fully appreciated by grant reviewers
- Meetings and workshops like this will be important in providing the broader education to address some of these challenges.
Funding

RCSB PDB is funded by a grant (DBI-1338415) from the National Science Foundation, the National Cancer Institute, the National Institute of General Medical Sciences, and the US Department of Energy

Management

RCSB PDB is hosted by:

Rutgers University
UC San Diego
UCSF

RCSB PDB is a member of the Worldwide Protein Data Bank partnership (wwPDB; wwpdb.org)

Follow us  

Facebook  Twitter  Youtube  Instagram