Economics and Impact of the Protein Data Bank (PDB) Archive

R. Andrew Byrd
Chair, wwPDB Advisory Committee
Chief, Structural Biophysics Laboratory,
National Cancer Institute, NIH
Protein Data Bank

- First open access digital resource for biology data (est. 1971 with 7 entries)
- Single global archive of experimental 3D structures of biological macromolecules (>121,000 entries)
  - Primary data => structural biology, computational biology, drug discovery, …
  - Complements GenBank and UniProt sequence database
  - Data Management Plan for all biomedical grants in US
- All data freely available (scientists and educators – world-wide)
- Global archive of experimental macromolecular structure data central to biomedical research

ABL tyrosine-kinase inhibited by Imatinib for treatment of chronic myeloid leukemia (CML).

HIV-1 reverse transcriptase complex with DNA and nevirapine
Organizational Structure/Funding

- Partners share “Data In” responsibilities
  - Biocurate new depositions
  - Define deposition and annotation policies
  - Resolve data representation issues
  - Implement community validation standards

- Partners independently funded by each region

- Overseen by a wwPDB Advisory Committee

- Partners compete on “Data Out” resources
Impact Metrics

- ~11,000 new structure depositions/year
- Biocuration responsibilities distributed by geographic location
- ~1.5 million data files downloaded/day
- Pharma Industry use PDB archive behind company firewalls daily
Sustainability

wwPDB established in 2003

Goals: (1) protect PDB archive and prevent fragmentation
      (2) enable global cooperation on:

- Increased “Data In” productivity:
  - common OneDep system for deposition/biocuration/validation
- Geographical Distribution-Load Balancing of “Data In”
- Preparations to extend the wwPDB Franchise to
  - Consideration for sites in PRC, South Asia, South America
Evaluation of ICSPR* Funding Models

- Only 1 of the 8 funding models evaluated was deemed acceptable for wwPDB to ensure:
  - Economic Stability/Long-term Sustainability
  - Global Open Access
  - Equity for Data Depositors
  - Equity for Research/Teaching Institutions

- Infrastructure Model!

*Inter-university Consortium for Political and Social Research
Sustaining Domain Repositories for Digital Data: A White Paper
What is the Infrastructure Model?

- Funding agencies commit to direct payment of the costs of archiving experimental data/metadata generated with the research support they provide.

- Data Resource funding comes in the form of strategic, long-term infrastructure investments (divorced from typical 3-5 year grant cycles).

- Ensures Economic Stability/Sustainability for an Open Access Data Resource Ecosystem with Equity for Data Depositors and Consumers.
Infrastructure Model and the wwPDB

- wwPDB partners endorse the Infrastructure Model (i.e., a model in which research funders reserve a percentage of annual expenditure for digital data archiving and preservation across the sciences)

- Estimated annual cost ~1-2% of the cost of data generation
  wwPDB estimates for archiving experimental macromolecular structure data/metadata in the Protein Data Bank

- Conservative cost of replicating the PDB archive (assuming average unit cost of US$100,000) equals
  \[ \text{US$12 billion} \]

- Impacts >80% of biomedical research grants