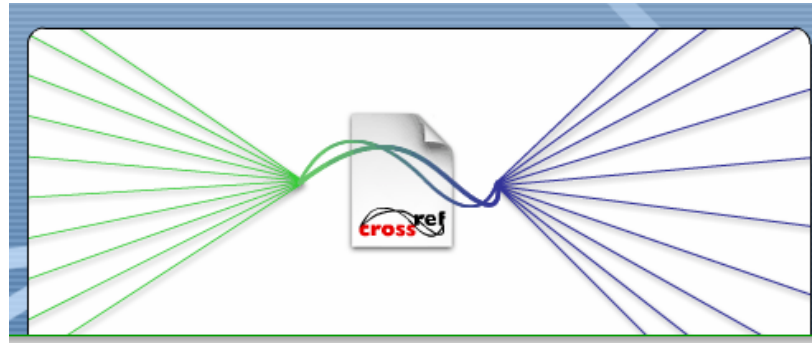


7th Annual Meeting CrossRef



1 November 2006
Cambridge, MA

DOIs for Biological Databases

Philip E. Bourne

University of California San Diego

pbourne@ucsd.edu

<http://www.sdsc.edu/pb>

The Vision...

Prior to leaving from home the graduate student syncs her IPOOL with the latest papers delivered overnight by the journal via RSS feed. On the bus she reviews the stream, selecting a paper close to her interest in HIV-1 proteases. The data shows apparent anomalies with her own work. By the time the bus stops she has recomputed the results, proven the anomaly and written a rebuttal including the revised data and sent it to the journal

Science Fiction?

- Five years ago Yes... Today No...
- Five years ago the idea of downloading data on a bus would have been absurd – not today
- Five years ago an IPOL would be absurd - not today
- Journals are providing RSS feeds today
- Why should the way we do science not change in the next five years?

What is Missing?

Seamless access to the data
associated with the scientific
publication

*This talk presents one approach to
overcoming this problem and highlights
why DOI assignments to data are
important*

Background & Disclaimers

- Research (computational biology) driven with a goal of maximizing the dissemination and comprehension of science
- Not an IT specialist
- Co-director of a major biological database, the Protein Data Bank used by 150,000 independent scientists and students per month (partial salary support)
- Editor-in-Chief PLoS Computational Biology (no salary support)

More Background

- Wrote an Editorial about an idea that had long been floating around in my head - *In the Future will a Biological Database Really be Different from a Biological Journal?* **PLoS Comp. Biol.** 2005 1(3), e34
- Received funding from the US National Science Foundation (NSF) for BioLit

BioLit - The Testbed



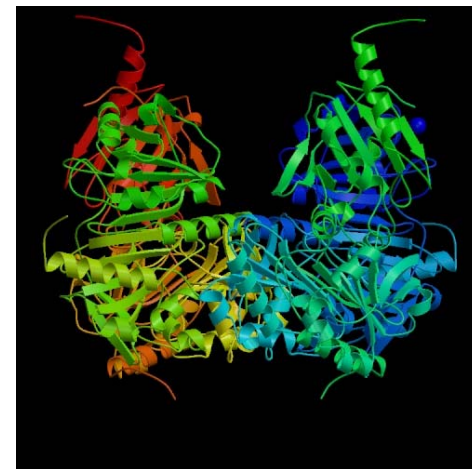
<http://www.wwpdb.org/>

RCSB
PDB
PROTEIN DATA BANK

WORLDWIDE
PDB
PROTEIN DATA BANK

8 November 2006

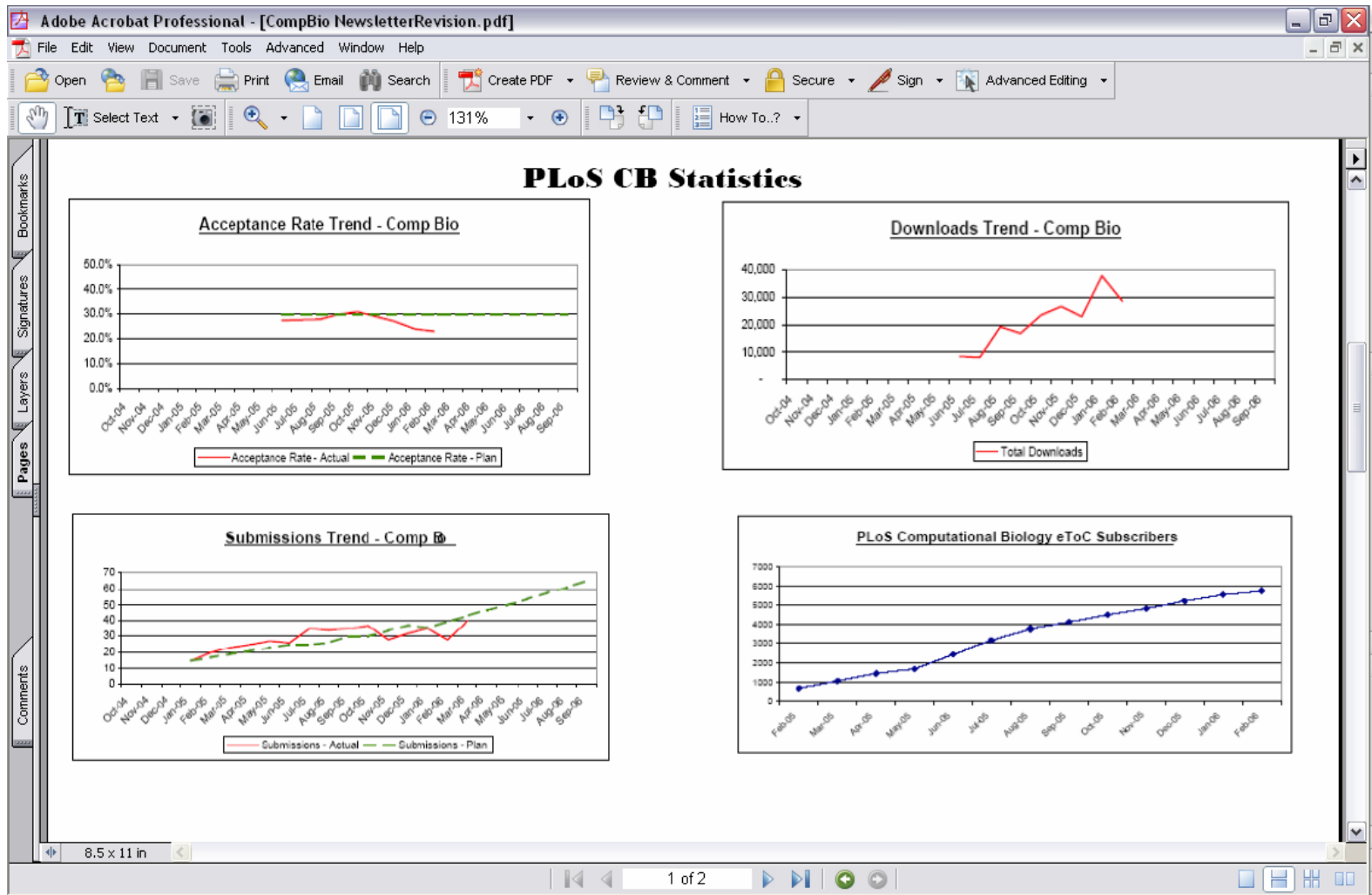
CrossRef, Boston



Published With a Society That Can Realize the Vision

- ~2000 member Society of computational biologists founded in 1995
- Previous “official journal” Bioinformatics (OUP) – royalty arrangement – *financial risk*
- Other journals offered at a discount
- FASEB Member
- Main activities: conferences, journals, education, SIG’s, regional meetings, Web portal

The Journal Thus Far



The PLoS Corpus



- Established in 2000
- Identified as a high quality publications (*PLoS Biology* impact factor 14.7)
- Currently 7 journals with healthy growth
- Open Access – free to all
- Assigns DOIs (of course)

It is the last point that makes this work a reality

Open Access (Creative Commons License)

1. All published materials available on-line free to all (author pays model)
2. Unrestricted access to all published material in various formats eg XML provided attribution is given to the original author(s)
3. Copyright remains with the author



Open Access (Creative Commons License)

1. All published materials available on-line free to all (reader pays model)
2. Unrestricted access to all published material in various formats eg XML provided attribution is given to the original author(s)
3. Copyright remains with the author

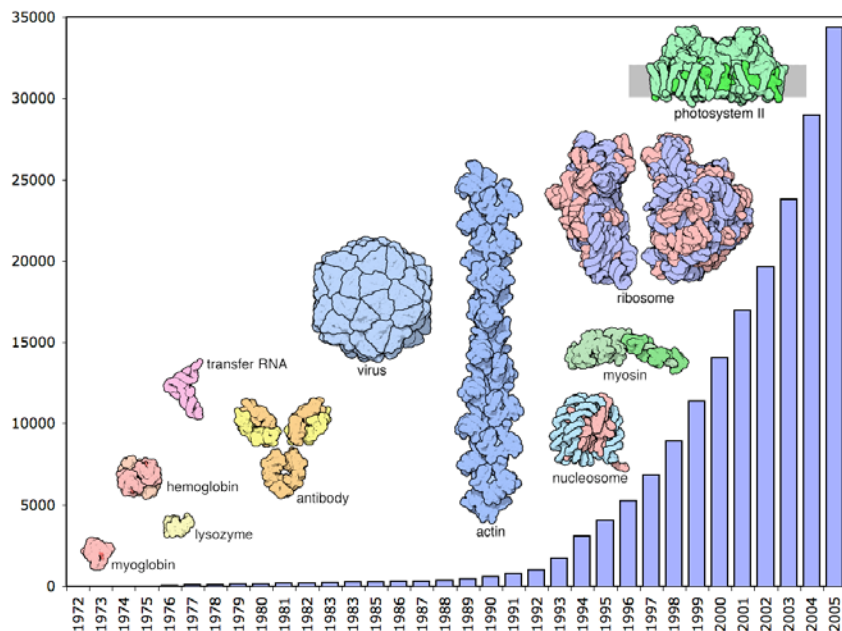
Critical to the success of this venture



The PLoS Corpus – Under the Hood

- Conforms to the NLM DTD – little markup of content
- Parallel development of Topaz – different emphasis – manuscript and content management with reusable software and backend infrastructure

The Protein Data Bank



- The single worldwide repository for data on the structure of biological macromolecules
- Vital for drug discovery and the life sciences
- Over 30 years old
- Free to all

Nucleic Acid Research 2000 **28(1)**, 235-242 – cited approx. 4000 times

The Protein Data Bank

RCSB PDB
PROTEIN DATA BANK

A MEMBER OF THE PDB
An Information Portal to Biological Macromolecular Structures
As of Tuesday May 02, 2006 there are 36344 Structures | PDB Statistics

Home Search

Welcome to the RCSB PDB

The RCSB PDB provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

The RCSB is a member of the wwPDB whose mission is to ensure that the PDB archive remains an international resource with uniform data.

This site offers tools for browsing, searching, and reporting that utilize the data resulting from ongoing efforts to create a more consistent and comprehensive archive.

Information about compatible browsers can be found [here](#).

A **narrated tutorial** illustrates how to search, navigate, browse, generate reports and visualize structures using this new site. [This requires the Macromedia Flash player download.]

Comments? info@rcsb.org

Molecule of the Month: Glucose Oxidase

Diabetes is a worldwide health problem affecting hundreds of millions of people. Fortunately, with careful management of diet and medication, the many complications of diabetes can be reduced. Part of this treatment includes the monitoring of glucose levels in the blood, so that proper action may be taken if levels get too high. The enzyme glucose oxidase has made glucose measurement fast, easy, and inexpensive.

[More ...](#)

[Previous Features](#)

NEWS

- Complete News
- Newsletter
- Discussion Forum

02-May-2006
RSS functionality at the RCSB PDB

An RSS (Really Simple Syndication) feed provides users with a list of newly updated structures as soon as they are available. RSS pushes information that can be read by client software (an RSS reader) that sits on your local computer. Rather than going to look for new PDB entries, they can come to you.

[Full Story ...](#)

25-Apr-2006
RCSB PDB Newsletter Spring 2006

18-Apr-2006
Education Focus: DNA Day

11-Apr-2006
Validating structures saves deposition time

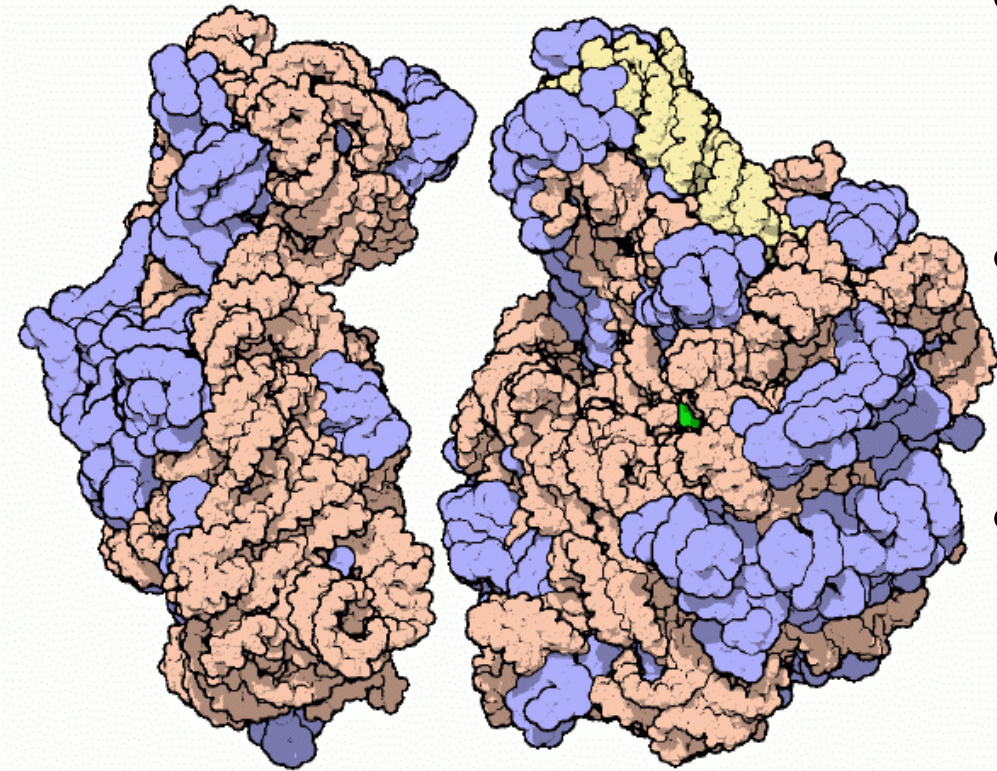
The RCSB PDB is supported by funds from the National Science Foundation (NSF), National Institutes of Health (NIH), National Institute of Biomedical Imaging and Bioengineering (NIBIB), and the National Institute of Neurological Disorders and Stroke (NINDS).

In citing the PDB please refer to: M.M. Bertram, J. Westbrook, Z. Wang, S. Gilliland, T.N. Bhat, H. Shindyalov, R.E. Bourne. Protein Data Bank. Nucleic Acids Research, 28 pp. 235-242 (2000).

http://www.pdb.org

- Paper not published unless data are deposited – strong data to literature correspondence
- Highly structured data conforming to an extensive ontology
- DOI's assigned to every structure

The Protein Data Bank



- The majority of structures have a primary citation
- There may be one or more secondary citations
- The previous mindset of data in the database, knowledge in the literature is *slowly* changing ...

The previous mindset of data in the database, knowledge in the literature is *slowly* changing ...

- An increasing number of structures are not published
- If they are it is in the form of a standardized report ie database like
- More knowledge is appearing in the PDB entry
- The natural conclusion is to assign DOIs to structures

DOI Assignments

- Each PDB structure has a unique identifier eg 4HHB
- Community bought up over 30 years to love PDBids – part of the language/religion of the field
- Newer generation do not care
- Web resolution of a PDBid not unique

Structure of a DOI Assignment

- 10.2210/pdb<pdbid>/pdb
- Eg 10.2210/pdb4hhb/pdb
- Approximately 150 new structures are registered per week
- Metadata includes:
 - Title
 - Contributors
 - Publication date
 - Description
 - Resource

What Do DOIs Buy Us?

- A unique reference in cyberspace
- Consistent lookup
- Potential for search forward – find all papers that have ever referenced this structure
- A step towards

So What is Needed to go from Vision to Reality?

- Seamless integration between papers reporting results and the data used to compute those results – public repositories of data and knowledge must become integrated in ways not possible today



Similar Processes Lead to Similar Resources

Author Submission via the Web



Syntax Checking



Review by Scientists & Editors



Corrections by Author



**Publish – Web Accessible
DOI Assigned**

Depositor Submission via the Web



Syntax Checking



Review by Annotators



Corrections by Depositor



**Release – Web Accessible
DOI Assigned**

Can it Happen?

- The data repositories have been available for a long time
- With open access the knowledge repositories are available – its more than just abstracts
- If the perception of the difference between data and knowledge is lowered
- The technologies are there

BioLit: Tools for New Modes of Scientific Dissemination

The Knowledge and Data Cycle

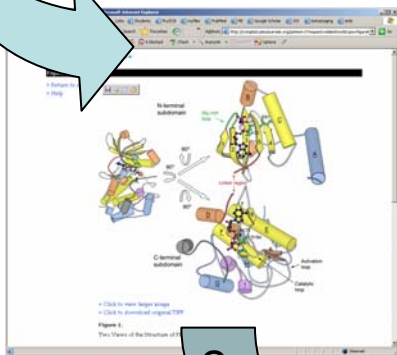
0. Full text of PLoS papers stored in a database



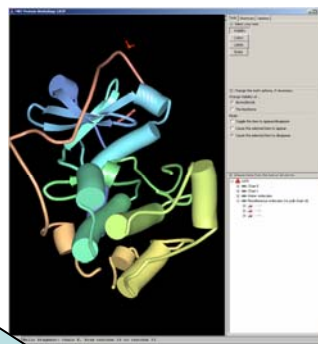
4. The composite view has links to pertinent blocks of literature text and back to the PDB



1. A link brings up figures from the paper



3. A composite view of journal and database content results



2. Clicking the paper figure retrieves data from the PDB which is analyzed

- **Biolit integrates biological literature and biological databases and includes:**
 - A database of journal text
 - Authoring tools to facilitate database storage of journal text
 - Tools to make static tables and figures interactive



- Underlying Postgres relational database
- Fedora (Flexible Extensible Digital Object Repository Architecture).
- Support for RDF metadata
- Ajax front end
- applications will access the repository's data by means of the four APIs by which Fedora is exposed: management, access, search (exposed via HTTP or SOAP) and the OAI provider API (exposed via HTTP).
- 3 Tier Architecture
 - SQL-92 Database (mySQL)
 - J2EE Application Server (JBoss, hibernate)
 - Web Client, Web Services Client, Webworks
- Lucene indexing engine
- Standards compliant using available open source tools. Intention is to make the site more mobile while offering easy access and integration to the diverse community.

Being Implemented Now – Making Associations

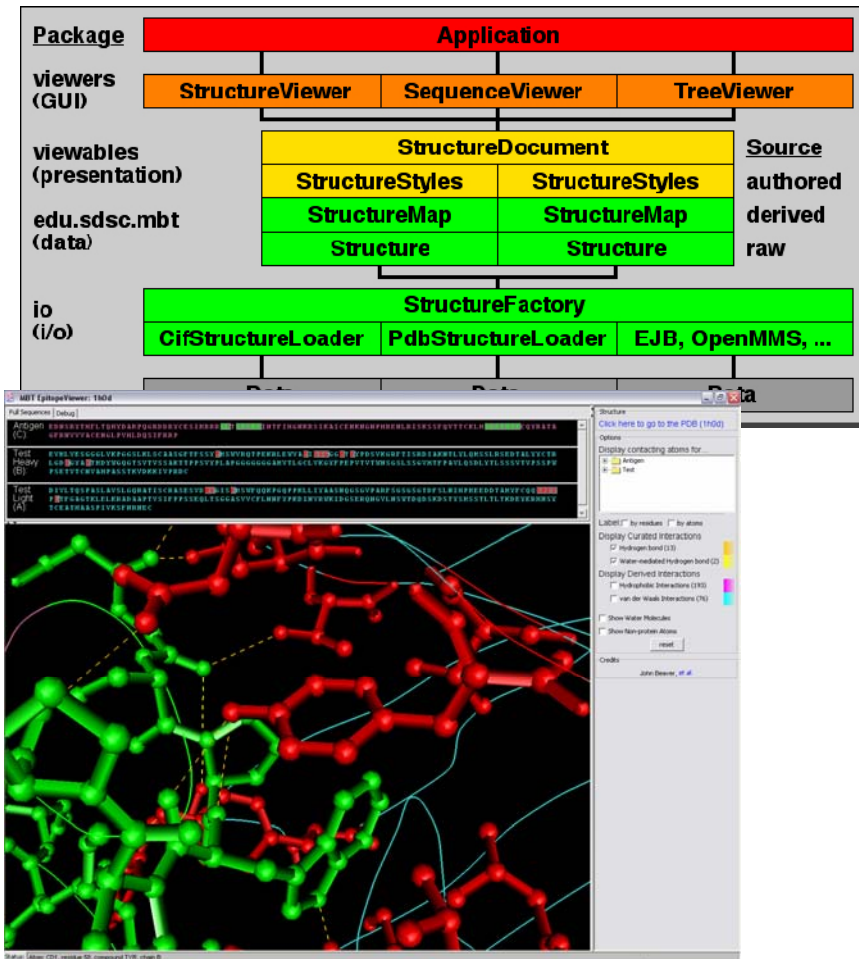
- Tools to find associations to PDB structures in the PLoS corpus and make contextual associations eg:
 - “catalytic triad” in the specific structure (easy)
 - All references to “catalytic triads” (easy)
 - All relevant references to ‘catalytic triad’ (difficult)

Being Thought About Now – The Authoring Process

- Previous prototype – BioEditor
<http://bioeditor.sdsc.edu>
- Plug-in for MS-word (cf Endnote) for formalizing vocabulary and creating suitable metadata to make needed linkages

Yang et al., Bioinformatics 2003 19(7) 897-898.

Done - Suitable Display Tools for Live Journal Content – <http://mbt.sdsc.edu>



- Metadata stored with the on-line paper provides a renderable view of an otherwise static image – new level of comprehension
- Requires seamless Web delivery on multiple platforms

Future Possibilities

- Part of the semantic web
- Data accessible and analyzable from the paper
- The paper is just one interface to the associated science
- We have MyPDB why not MyScience – each reader has their own interface to the content – right now we have two – full content and synopsis
- New understanding about how knowledge is transferred – *knowledge bubbles*
- Comprehension is evaluated and indeed found to have improved

Acknowledgements

- PLoS Staff for their willingness to embrace these concepts
- NSF DBI
- Wayne Townsend Marino and Jeff Marino Ott – The PDB software engineers
- John Moreland, Apostol Gramada and John Beaver for mbt and associated applications