Best Practices for Data: A Biologists View

Philip E. Bourne
UCSD
pbourne@ucsd.edu
My Perspective is Drawn from Being:

- A data producer and a data user*
- An overseer of data curation efforts
- A database provider (PDB & IEDB)
- Suspicious of workshop reports, data standards bodies …
- A supporter of data publication
- An open access journal founder
- Opinionated
The RCSB Protein Data Bank

- It's a global community
- It's 42 years old
- It's a diverse set of users
- Single international archive for all information about the structure of large biological molecules
- 300,000 unique IP’s per month
- 300 simultaneous users
- 7 structures (data objects) downloaded every second
PDB Data Are Important

Structure Summary page activity for H1N1 Influenza related structures

- 3B7E: Neuraminidase of A/Brevig Mission/1/1918 H1N1 strain in complex with zanamivir
- 1RUZ: 1918 H1 Hemagglutinin

* http://www.cdc.gov/h1n1flu/estimates/April_March_13.htm
Agenda

- Trends we are seeing
- What we have learned
- What needs to change re data*
- What we are planning to facilitate change

* Personal opinion
So What Are the Trends?
## Trends – Data Access

<table>
<thead>
<tr>
<th>Trend</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant demand for better performance (damn Google)</td>
<td>Attention to SEO, memcache, denormalization, query interfaces</td>
</tr>
<tr>
<td>Use of Web services increasing (dropped SOAP and now all RESTful)</td>
<td>All data exposed</td>
</tr>
<tr>
<td>Limited use of widgets</td>
<td></td>
</tr>
</tbody>
</table>

Will Widgets and Semantic Tagging Change Computational Biology?

*PLoS Comp. Biol. 6(2) e1000673*

So what are the trends?
Guiding Users to Specific Results

Search Suggestions

- cAMP-dependent Protein Kinase (PKA)
- Src Tyrosine Kinase
- Cyclin-dependent kinase 2 (257)
- Mitogen ... protein kinase 14 (194)
- Nucleoside diphosphate kinase (82)
- cAMP-dependent protein kinase catalytic ... (131)
- Thymidylate kinase (57)
- Aurora kinase A (64)
- Sequence

Molecule Name

- L12: INHIBITOR OF P38 KINASE
- 274: MET KINASE INHIBITOR
- BI3: INHIBITOR ... PROTEIN KINASE-1

Chemical Name

- Has exact structure C1c1cccc1
- Has sub-structure C1c1cccc1
- Is very similar (95%) with C1c1cccc1
- Is similar (70%) with C1c1cccc1
- Super-structure of C1c1cccc1

Smart Suggestions

- Very significant (E Cut Off:0.001) to GSTFIVITAGADGLAT
- Significant (E Cut Off:0.01) to GSTFIVITAGADGLAT
- Includes Insignificant (E Cut Off:1) to GSTFIVITAGADGLAT
- Extended Search (E Cut Off:10) to GSTFIVITAGADGLAT

Sequence

- SMILES
  - Has exact structure C1c1cccc1
  - Has sub-structure C1c1cccc1
  - Is very similar (95%) with C1c1cccc1
  - Is similar (70%) with C1c1cccc1
  - Super-structure of C1c1cccc1

Mouse over for more details
Trends - General: Mobile Devices

- Mobile devices well suited to deploy educational materials (*Molecule of the Month*)
- Mobile device users expected to exceed desktop users by 2014

So what are the trends?

Trends PDB: Support for PDB Mobile

Available on the App Store

Browse
Molecule of the Month articles

Simple search

iPhone

Interactive 3D viewer*

Browse search results

iPad

*NDKMol, Takanori Nakane, Kyoto University

So what are the trends?
So what are the trends?
Trends: Reduced Footprint and VM

- Legacy Hardware – Deploy on the Open Science Grid

So what are the trends?
Trends: Global Load Balancing

- Why is it important?
  - Enables us to better serve our users by providing increased reliability and quicker results

- How will it be done?
  - By more evenly allocating our resources at Rutgers and UCSD
  - By directing users to the closest site

So what are the trends?
So what are the trends?

We manage to handle increased volume and complexity at a lesser cost

Usage increases, community broadens, applications expand

Trends: Complexity Begats Use
So What Have we Learned that is Relevant to RDA?
Its All About Trust

Trust in the data is perhaps our biggest achievement

So what have we learned?
Its All About Trust

- Trust is like compound interest
- Comes from listening
- Comes from engaging the community in every aspect of the process
- Comes from data consistency and level of annotation
- Comes from responsiveness
- Comes from the quality of the delivery service

So what have we learned?
Data Quality Begats Trust

- About 25% of our budget has been spent on data remediation

- Support for versioning hence the copy of record

- Our ontology/data model has been a critical component of our workflow and data accuracy

- Until recently the same data model was too complex to facilitate wide adoption by others that use our data
Its All About People
Curators are the Unsung Heroes

"We pay homage to these special individuals who are dedicated to making our research endeavors a success."

- They really should do more to promote themselves
- Institutions must do more to respect their efforts

So what have we learned?

http://collections.plos.org/ploscompbiol/biocurators.php

20
Its All About People
The Users

- Constantly striving to have the user distinguish raw from derived data

- All data are not created equal but the user thinks so
Its All About People
The Global Personalities

So what have we learned?
Its NOT All About Institutions

- As far as I am aware no data standards body has directly influenced anything we have done in 15 years of running the PDB

- The structural biology community created a very successful data sharing plan long before funding bodies did
It is About Openness

- There are no restrictions on the usage of the data beyond attribution
- The PDB runs exclusively on open source software
- We maintain and contribute to the Biojava repository
- We need to be transparent about data usage
So What Needs to Change re Data?
That All Data Are Created Equal Must End

- We need to understand how data are used
- Sustainability is not more money from the funding agencies it's about business models
- Reductionism is not a dirty word
- We need to do more with the long tail

On the Future of Genomic Data
Science 11 February 2011:
vol. 331 no. 6018 728-729
Institutions That Generate Data Must Play a Greater Role

- We need institutional data sharing plans
- We need data scientists to be better recognized by institutions – it’s not all about papers – this implies new metrics
The {Lack of} Distinction Between Data and Knowledge Needs to be Better Appreciated

- The PDB paper has been cited 14,000 times
- No one has ever read it
- Some PDB datasets have 1,000’s of downloads
- These data are not associated with publications

So what needs to change?
We Need to Think More About the Questions We Wish to Answer Rather than Simply Being Able to Retrieve The Data

- Data are stove piped – Broad questions are difficult to answer

- Usage alone can be used to answer questions
What Will We Contribute Going Forward?
What will we contribute? Research Data Alliance March 19, 2013

The Knowledge and Data Cycle

1. User clicks on content
2. Metadata and webservices to data provide an interactive view that can be annotated
3. Selecting features provides a data/knowledge mashup
4. Analysis leads to new content I can share

PLoS Comp. Biol. 2005 1(3) e34

1. A link brings up figures from the paper
2. Clicking the paper figure retrieves data from the PDB which is analyzed
3. A composite view of journal and database content results
4. The composite view has links to pertinent blocks of literature text and back to the PDB

0. Full text of PLoS papers stored in a database
PLOS Data Papers

- Individual repositories register their metadata which includes access statistics, commentary etc. – DataCite is a beginning
- Identify identical data objects and their respective metadata for comparative analysis
- Funders support registration
- Publishers support registration

What will we contribute?
Data Paper Workflow (Part 1)

1. Author
   - Primary Data
   - Author fills out Web form(s) to submit Primary Data, create Metadata and submit / generate draft Data Paper

2. Author
   - Draft Data & Resource Paper
   - Draft Metadata
   - Primary Data
   - Author reviews and submits to PLoS

3. Data Editor
   - Draft Data & Resource Paper
   - Draft Meta Data
   - Primary Data
   - Data Editor reviews and interacts w/ Author to finalize.

   NOTE: If data has already been, or should be, submitted to a specialist DB such as PDB, only the Metadata, i.e. cataloging information, is captured by PLoS, with cross-references to the specialist DB. (2) If data has not been submitted to a specialist DB, but should be, Data Editor requires proof of this submission (identifiers or accession #s) from Author prior to proceeding.

4. Data Editor
   - Final Data & Resource Paper
   - Final Metadata
   - Primary Data

   a. Column headings are normalized where necessary.
   b. Data Editor submits Primary Data and Metadata to PLoS Dataverse.
   c. DOI(s) are minted and attached to Final Data Paper.

   Normalization examples:
   - Gene, protein, organism names
   - male/female
   - "missing value" codes

Research Data Alliance March 19, 2013
The PDB as an App Store

- **The App model**
  - Think of it operating on the PDB content base rather than a mobile device

- **Advantages**
  - Enriches the questions the PDB can answer
  - Creates an even greater sense of community
  - Simple and consistent user interface
  - Needs to pass some quality control
  - Has a rating and reward system
Acknowledgements

- www.force11.org
  - Tim Clark
  - Ivan Herman
  - Paul Groth
  - Ed Hovy
  - Maryann Martone
  - Cameron Neylon
  - David Shotton
  - Anita de Waard

- www.plos.org
- Beyond the PDF
- Many others

Funding Agencies:
NSF, NIGMS, DOE, NLM, NCI, NCRR, NIBIB, NINDS, NIDDK

Research Data Alliance March 19, 2013