Thoughts on the Future of Scientific Dissemination

Philip E. Bourne
University of California San Diego
pbourne@ucsd.edu

"So Much Done and So Much to Do"

Winston Churchill, upon being accused of being slightly tipsy by a member of the temperance league

“So Much Done and So Much to Do”

Winston Churchill, upon being accused of being slightly tipsy by a member of the temperance league

Motivational Questions

- Have publishers have yet to realize the power of the Internet?
- Is a database really different from a journal? *PLoS Comp. Biol*. 2005 1(3), e34
- Does a scientific society meet the needs of its members?
- Are scientists taking full advantage of the digital world?

Motivational Questions

- Have publishers have yet to realize the power of the Internet? NO
- Is a database really different from a journal? NO *PLoS Comp. Biol*. 2005 1(3), e34
- Does a scientific society meet the needs of its members? NO
- Are scientists taking full advantage of the digital world? NO

Background

- As a computational biologist this is a view from the trenches
- This is a view from the life sciences – while disciplines have similarities there are also differences
- I do not actively follow developments in eResearch
- As an EIC of a major biological journal and a developer of a major biological database I have a certain perspective (bias)

Agenda

- Introduce a research vision
- Diagnose that vision
- Illustrate what we are doing to address this vision
- Invite comment

The Vision…

Prior to leaving home the graduate student syncs her IPOL with the latest video 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. She notices that her colleague has also discovered the same paper and they IM annotating the results. By the time the bus stops she has recomputed the results, proven the anomaly and written a rebuttal, included a podcast as a letter to the Editor 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
- IM is prevalent but not for scientific discourse
- Video and podcasting is prevalent but not for scientific discourse
- Why should the way we do science not change in the next five years?

What is Missing to Make the Vision a Reality?

1. Seamless integration between the data and the publication upon which that data are based
2. Seamless integration of the authoring and publishing process
3. Notion of traditional publications being associated with podcasts and video
4. Professional networking akin to social networking

What are the Catalysts for Change?

- Open access publishing
- The emerging generation of digital scientists
- The increased ease of working with digital media, notably sound and video

What is Missing to Make the Vision a Reality?

1. Seamless integration between the data and the publication upon which that data are based
2. Seamless integration of the authoring and publishing process
3. Notion of traditional publications being associated with podcasts and video
4. Professional networking akin to social networking

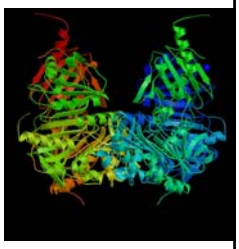
1. Database and Journal Integration- The Test Bed

Journals

http://www.wwpdb.org/

RCSB PDB PROTEIN DATA BANK

WORLDWIDE PDB PROTEIN DATA BANK



Database

The PLoS Corpus

- Established in 2000
- Identified as a high quality publications (*PLoS Biology* impact factor 14.7)
- Currently 8 journals with healthy growth
- Open Access – free to all

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)

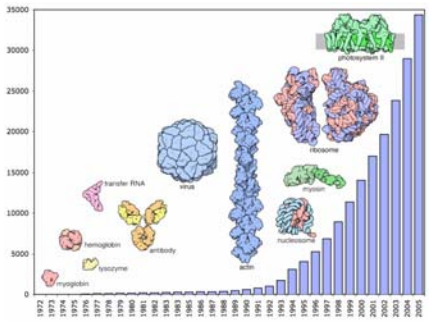
1. All published materials available on-line free to all (author pays model)
2. Unrestricted access to all published material in various formats eg XML provided attribution is given to the original author(s)
3. Copyright remains with the author

The catalyst

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 – 5000 citations

The Protein Data Bank

- 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

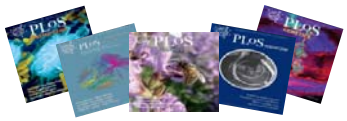
http://www.pdb.org

1. Seamless Integration between Data and the Literature – What Does That Imply?

- Improving semantic consistency in the literature – best done at the point of authoring
- Post processing to establish semantic content
- New forms of visualization and interaction at the presentation layer

I argue:

This really should not be that difficult since journal processes and content are not that different to database processes and content



RCSB PDB PROTEIN DATA BANK

Similar Processes Lead to Similar Resources

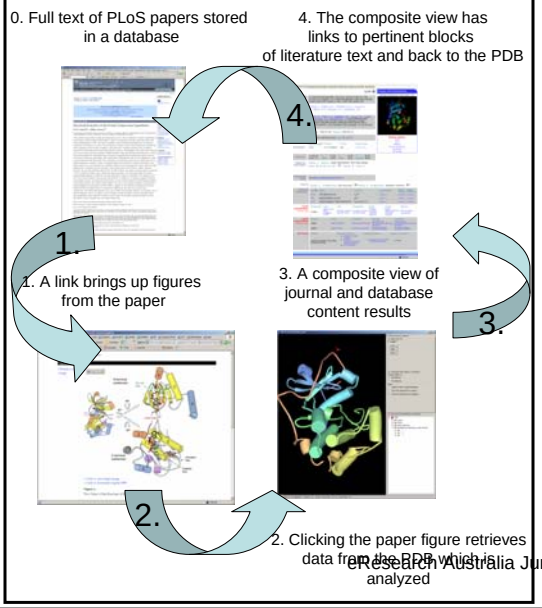
Author Web Submission	Depositor Web Submission
⇩	⇩
Syntax Checking	Syntax Checking
⇩	⇩
Review by Scientists & Editors	Review by Annotators
⇩	⇩
Corrections by Author	Corrections by Depositor
⇩	⇩
Publish – Web Accessible	Release – Web Accessible

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

<u>BioLit:</u> Tools for New Modes of Scientific Dissemination

The Knowledge and Data Cycle

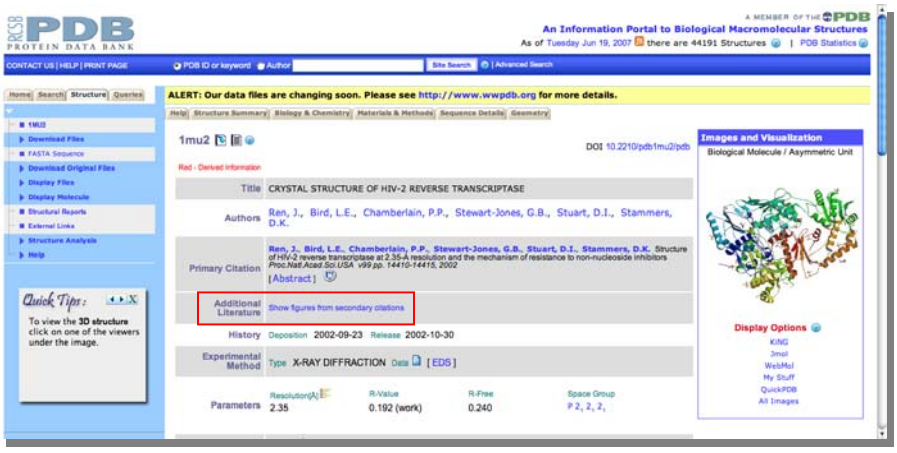


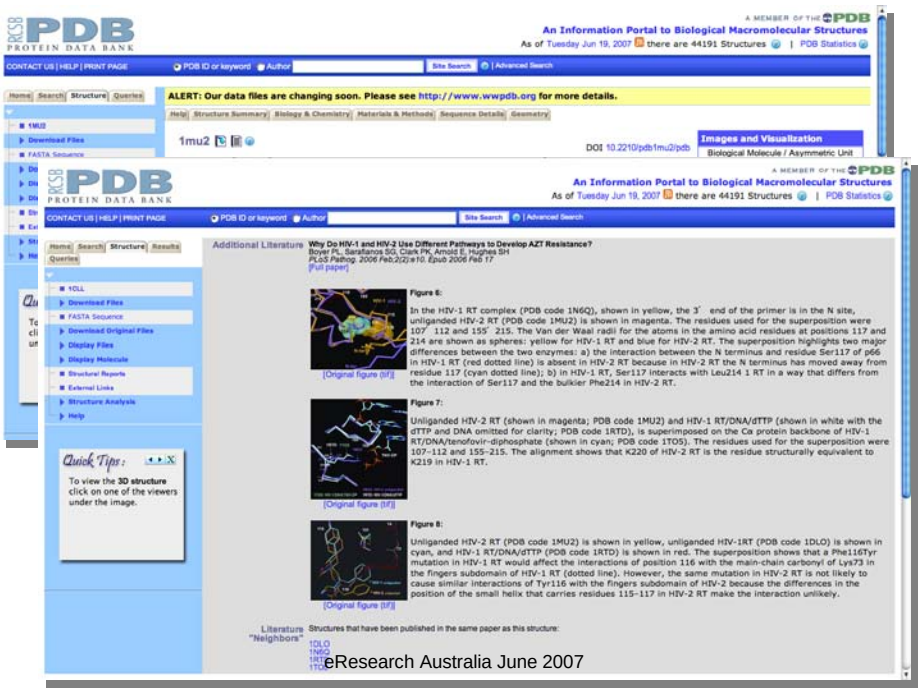
- **Biolit integrates biological literature and biological databases and includes:**
  - **A database of journal text**
  - **Authoring tools to facilitate database storage of journal text**
  - **Tools to make static tables and figures interactive**

PDB-PLoS Journal Integration – What has Been Done so Far?

Lynn Fink

- Retrieve full XML files of all articles in PubMed Central
- Parse into database
  - Extract DOIs, captions for figures
  - Extract PDB IDs from article and figure text
  - Use NLP algorithms to find PDB structure properties (i.e., mutations)
  - Extract MeSH terms, EC numbers





What is Missing to Make the Vision a Reality?

1. Seamless integration between the data and the publication upon which that data are based
2. Seamless integration of the authoring and publishing process
3. Notion of traditional publications being associated with podcasts and video
4. Professional networking akin to social networking

Microsoft

- Collaboration with Microsoft to develop Word 2007 plug-in
  - Semantic mark-up using ontologies and controlled vocabularies
  - Facilitate/automate referencing to PDB (and other resources) from manuscript
  - Conversion of manuscript to NLM DTD for direct submission to publisher

Semantic Mark up Using Ontologies

"Also, it has been shown that APH(3')-IIIa has some protein kinase activity, providing a functional link between the APHs and TPK."

gene_ontology.obo

```
id: GO:0004672
name: protein kinase activity
namespace: molecular_function

def: "Catalysis of the transfer of a phosphate group,
usually from ATP, to a protein substrate." [GOC:jl]
xref_analog: EC:2.7.1.37
is_a: GO:0016301 ! kinase activity
is_a: GO:0016773 ! phosphotransferase activity, alcohol
group as acceptor
```

Structural Evolution of the Protein Kinase-Like Superfamily, PLoS Comput Biol. 2005 Oct;1(5):e49, Scheeff ED, Bourne PE.

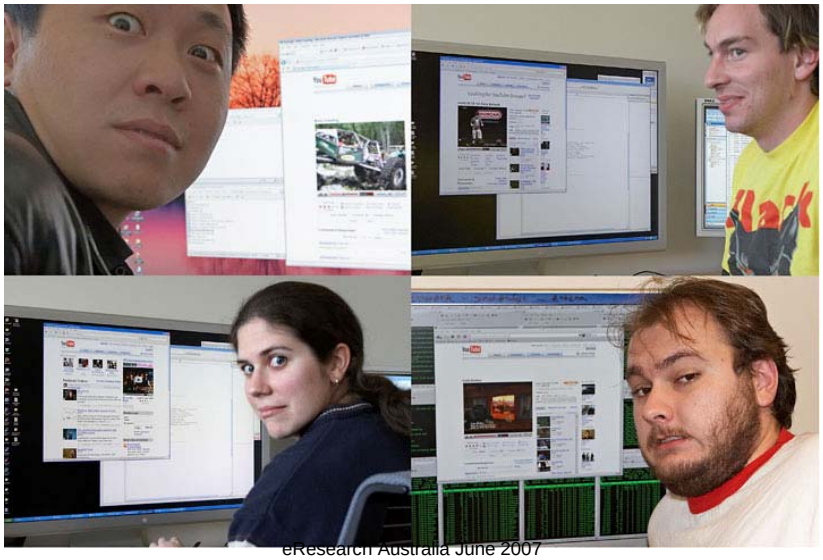
Link to Data Repository

```
Human PEX5 [PDB:<ext-link ext-link-type="pdb" xlink:href="1FCH">1FCH</ext-link>]
```

What is Missing to Make the Vision a Reality?

1. Seamless integration between the data and the publication upon which that data are based
2. Seamless integration of the authoring and publishing process
3. Notion of traditional publications being associated with podcasts and video
4. Professional networking akin to social networking

Motivation



Motivation

- Scientific understanding requires we digest ever more and diverse information – 16,000 papers are being added to PubMed every week
- The graduate students and postdocs of today are the leading scientists of tomorrow
- This generation is comfortable with the short video clip and sound bite format
- Scientific publishers have been slow to adopt and disseminate sound and video

What is Wrong with Text Only?

- Often times text does not capture the excitement of the work i.e., impersonal, boring
- Time consuming to digest – the abstract is too limiting and the full text too consuming
- Video conveys information not possible with text e.g. a simulation or experimental

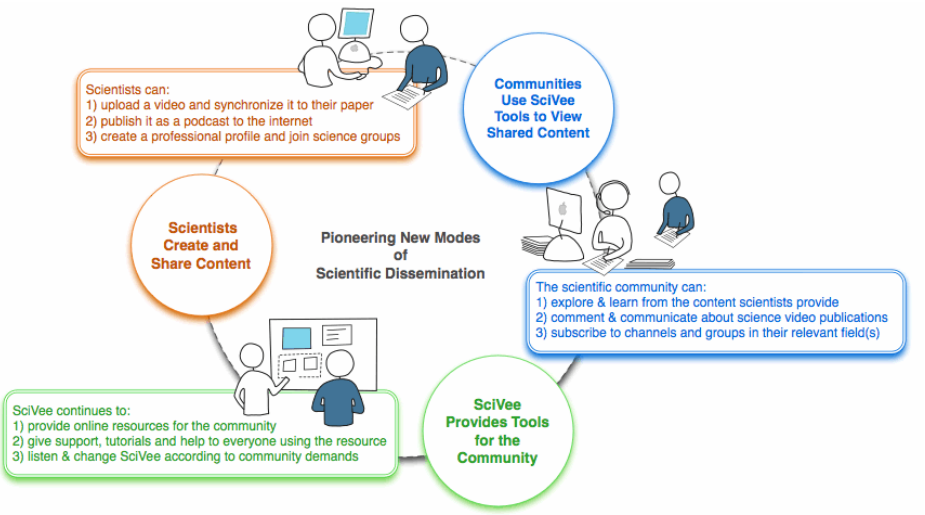
Concepts

- Use video and podcasts to enhance dissemination and comprehension
- Leverage existing technology
- Cater to the YouTube generation
- Partner with a respectable scientific publisher known for innovation
- Have the author present only what is in the paper

Product

- Video clips and podcasts of presentations by authors of peer reviewed scientific papers – duration 3, 5, 7, 10 minutes

- Later – Other forms of scientific video material – conferences, K12 education etc.

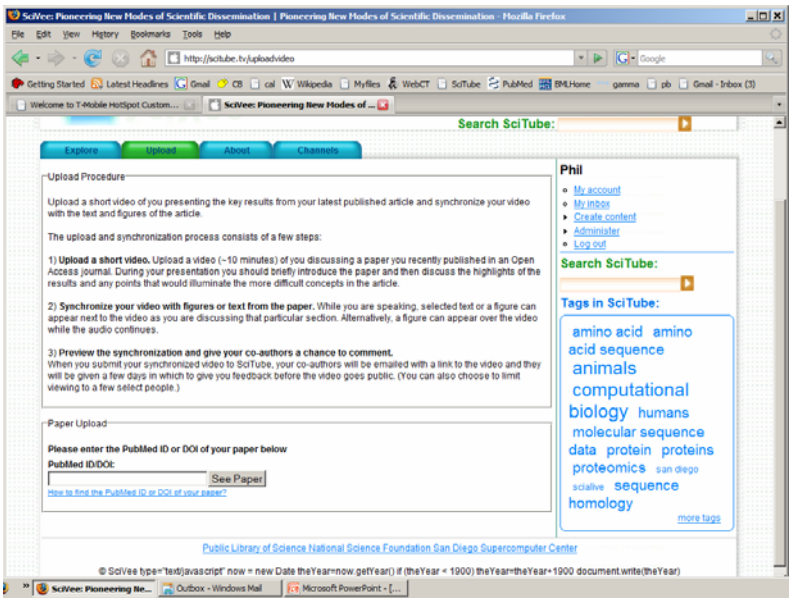


Developmental Phases

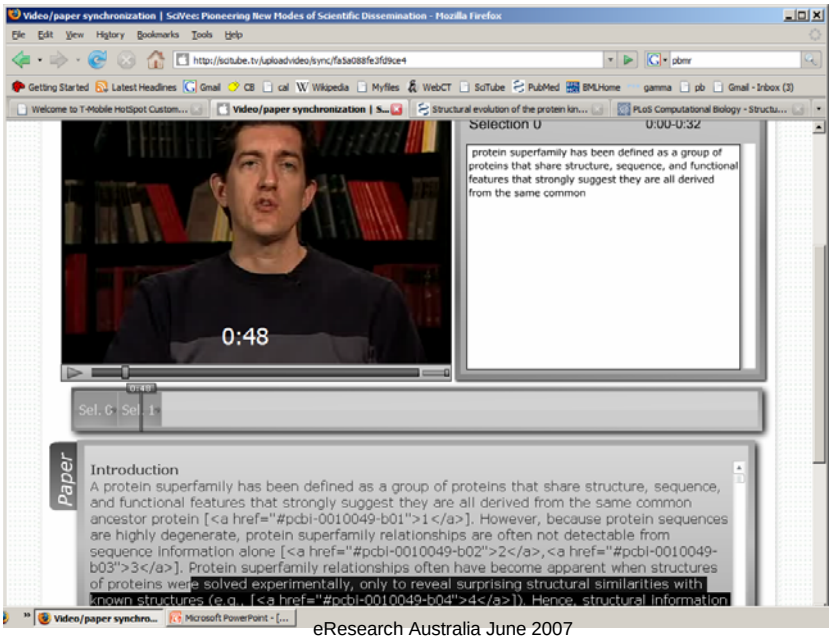
- **Phase I** (One Year) – Invite authors of papers published in PLoS journals to upload a video or podcast to *SciVee.tv* describing the motivation, key results and major conclusions of the published study. Establish linkage between literature and video – source of metadata etc. – September 2007
- **Phase II** (Years 2 3)- Scrape PubMed on a daily basis and extend the invitation to authors of all papers in the life sciences; develop video authoring server; provide ratings and virtual community comment
- **Phase III** (Year 4)- Extend to other scientific disciplines

eResearch Australia June 2007

eResearch Australia June 2007



eResearch Australia June 2007



eResearch Australia June 2007

What is Missing to Make the Vision a Reality?

1. Seamless integration between the data and the publication upon which that data are based
2. Seamless integration of the authoring and publishing process
3. Notion of traditional publications being associated with podcasts and video
4. Professional networking akin to social networking

Issues and Solutions

- Science is hierarchical – let the graduate students and post docs be heard – if not be their peers by each other
- How to get profiles to start? – PLoS author/reviewer/editor base
- How to organize? – around labs, organizations, topic areas

Acknowledgements

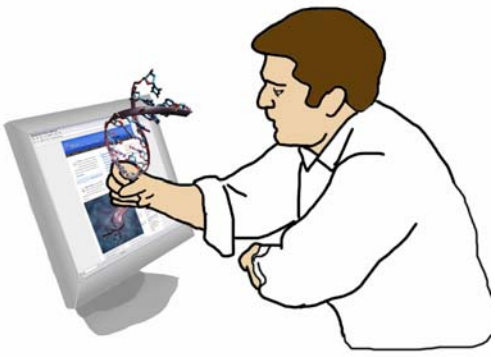
PLoS for being PLoS

BioLit

Lynn Fink

National Science Foundation
WHERE DISCOVERIES BEGIN

- SciVee Team
  - Apryl Bailey
  - Diane Baxter
  - Leo Chalupa
  - Louie Coffman
  - Caroline Drakely
  - John Matherly
  - Alex Ramos
  - Willy Suwanto
  - Kevin Walsh



Questions?

RCSB PDB
PROTEIN DATA BANK

- 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.

Deshpande et al., Nucleic Acids Research. 2005 33: D233-D237