



PLoS Mulls Hosting Software amid Growing Crossover between Informatics and Publishing

August 07, 2009

Byline: Vivien Marx

Newsletter: [BioInform](#)

[BioInform - August 7, 2009](#)

As the boundary between databases and journal articles blurs, researchers and publishers are exploring methods to make data and software more readily available with papers, as well as ways to make the growing body of online literature more "database-like" to help users find and share results more easily.

In the latest example of this trend, the *Public Library of Science* is planning to launch a software section for its *PLoS Computational Biology* and *PLoS One* journals that may include a function for authors to deposit their software when they submit their papers for publication.

The section, which PLoS expects to launch some time this fall, will only accept open source software, Phil Bourne, founding editor-in-chief of *PLoS Computational Biology* and professor at the Skaggs School of Pharmacy and Pharmaceutical Sciences at the University of California at San Diego, told *BioInform* last month at the Intelligent Systems for Molecular Biology conference in Stockholm.

Bourne said that Carnegie Mellon University's Robert Murphy will be the section's editor.

"We'd like to have software deposited with the article," Bourne said, but noted that the team is ironing out details, such as whether to create a repository like SourceForge to host software that is not yet live at the time of article submission.

Some reports will be about software "well-used" by the scientific community. The slated section's kickoff paper will be on version 3 of HMMER, Janelia Farm researcher Sean Eddy's protein sequence analysis tool.

Bourne disclosed PLoS's plans during the BioLink special interest group meeting at ISMB co-organized with the ISCB Publications Committee and *PLoS Computational Biology*, which included several sessions on the future of scientific publishing.

Not Just Software

During the session, other publishers discussed ways that journals could improve the dissemination of large data sets that support a paper's findings.

Publishers often treat "data as an afterthought" and "rarely" put it in a form that is readily re-usable, BioMedCentral's publisher Matthew Cockerill said, adding that BMC is working on techniques to put data in context. BMC is also starting a new journal in this area called the *Journal of Biomedical Semantics* with Dietrich Rebholz-Schuhmann of the European Bioinformatics Institute and Goran Nenadic of the University of Manchester as editors.

Nature's database publisher Matthew Day agreed that publishers do a "pretty poor job" of handling data and said that large data sets tend to remain "unpublishable."

Publishers should explore ways to help manage research data such as from genome-wide screens, he said. He added that publishers could also help by annotating data or by finding ways to give credit to widely used data sets, but did not elaborate on any particular *Nature* projects.

Other publishers are experimenting with ways to better integrate existing databases with journal articles.

For example, Elsevier last April kicked off an initiative to add "structured digital abstracts" to some articles for its journal *FEBS Letters*. In partnership with the Molecular Interactions Database, Elsevier extends the abstracts for protein interaction papers with machine-readable abstracts that include a series of sentences explaining relationships between two biological entities, which include identifiers that point to the appropriate database entries containing the full details of the interaction.

Anita de Waard, a researcher in the Disruptive Technologies Division at Elsevier Labs, said at ISMB that the journal has so far published 117 articles with such abstracts, which are published along with the text-based abstracts and stored in a database.

Elsevier is also involved with the OKKAM consortium, a project funded by the European Commission to develop the Entity Name System, a "global, public infrastructure to foster" creating and re-using unique identifiers for information.

De Waard explained the OKKAM group is exploring how to create semantically enriched texts. The group is testing a Word plug-in prototype that provides "entity identification," in this case, proteins, in order to semi-automate the creation of structured digital abstracts.

A similar project, led by UCSD researcher Lynn Fink and Phil Bourne in collaboration with Science Commons and Microsoft and announced this March [BioInform, March 20, 2009], is developing the Ontology Add-in for Word 2007, which identifies text as it is typed, and tags identifiers and ontology terms. Microsoft has made the source code available.

DeWaard said that the UCSD plug-in is "much more mature" than the OKKAM prototype.

Elsevier is also exploring ways to computationally identify novel, experimentally verified content, for example parsing "lines of argument," verb tenses, and "discourse segments"

in text, to help with database curation. As an example, she said that a hypothesis plus supporting evidence might be a "citable entity" of interest.

Reading Machines

During the panel, David Shotton, who heads the image bioinformatics group at the University of Oxford's department of zoology, presented work published earlier this year in *PLoS Computational Biology* on the "semantic enhancement" of peer-reviewed articles in order to make them machine readable.

Reactions have been positive, Shotton told *BioInform*, while some scientists have said "we haven't gone far enough."

In their paper, the Oxford team said that PDFs were "antithetical to the spirit of the Web" and set out to make the article easier to read by a computer, with the consent of the scientists who authored the original paper.

The original article, on the bacterial infection leptospirosis, is published here and the enhanced version is here.

The team added digital object identifiers, applied JavaScript for interactivity, and used a Google Maps API to mashup data with other publicly available information — such as superimposing onto a figure a satellite photo of the region in question. In addition, they added the author's raw data to images and graphs in order to make them "actionable."

As part of the enrichment, they highlighted Infectious Disease Ontology terms and added Citations in Context — a functionality that puts supporting claims from a reference in a hover box above the citation.

The work of semantic enrichment "is going to be a collaboration," Shotton said. "The authors will have to do some," and publishers can be active as well. Publishers and archivists he declined to name have contacted him about this work, he said.

Shotton has launched a new project called Semantic Publishing for Infectious Disease Epidemiology Research, or SPIDER. The idea is to bring together stakeholders such as scientists, IT firms, and publishers, as well as organizations such as CrossRef, a consortium of several publishers, to form partnerships on semantic enhancements that stand to "change the face of publishing," in many fields, he said.

Taste It

"Once people get a taste of it, they are really going to want more," Bourne said of semantic enrichment. Publishers now need to figure out how to integrate the enrichment process into the journal production workflow while keeping up the necessary speed of biomedical publishing, he said.

Oxford University Press' journals senior publisher Claire Bird said she and her colleagues are working to add semantic technologies and Web 2.0 technology to journals and looking into the role Oxford can play in post-publication, such as rating systems.

Since online communities can bypass traditional journals altogether, she said she feels

publishers must show their value by continuing to "orchestrate quality control" in research and ethical practices as well as long-term archiving, helping to make data and scientific results more "discoverable" and "accessible to data-mining."

Bird said that she is intrigued by projects such as [OpenWetware](#), a site promoting open lab notebooks, or [WikiGenes](#), which "allows clear authorship tracking," but she questioned whether scientists have "the time and incentive" to contribute to such sites "when they are not being chased" by journal editors.

PLoS's director of publishing, Mark Patterson, said that even though publishers can nowadays broaden the meaning of article "usage" beyond citations, "surprisingly not many publishers are doing much in this arena."

This spring PLoS began offering "a range of metrics" at the article level, he said, to include post-publication usage statistics, social bookmarking activity, media, and blog coverage. The team plans to take the approach beyond a "scorecard" to let scientists download and analyze the data and to promote adoption of this functionality by other publishers, he said.

Cambridge University Press' David Tranah, publisher of computer science, physics, and mathematics journals, said that trends in biomedical publishing approaches have "very wide ramifications" in publishing since biomedical journals comprise around 50 percent of science journals.

The open access movement caused a "mini-crisis" and "revolt" against traditional publishers, he said, and has also fostered debate about whether a journal's impact factor is an adequate metric of a paper's influence.

Cambridge is currently working on how to "build tools into papers that capture various kinds of information" he said.

Furthermore, electronic publishing has brought on a "problem of versioning," Robert Campbell, CrossRef's chairman and senior publisher at Wiley-Blackwell, said, since papers held in institutional repositories or on researcher's web sites are often amended.

The organization is launching [CrossMark](#), an authentication logo. The thinking is that there is no "final version" but rather a "publisher-maintained" version of a journal article, which will, for example, include retractions or corrections, Campbell said.

Just a Number

In addition to marking up texts so that researchers and computers can find and share information more readily, some publishers are also looking into new methods for tracking authors and their work.

UCSD's Bourne, for example, advocates a unique identifier for researchers that will lend each scientist a scholarly identity and deliver "better metrics" of output. The number can also help with disambiguation — a feature that he can personally use, he said, since he is often mistaken for computer scientist Stephen Bourne.

He told *BioInform* that reactions to the idea have been "mixed," with more favoring the concept, which he laid out in an article in *PLoS Computational Biology* [last December](#).

When a scientist's software or data is downloaded hundreds of times it is not measured by current metrics, Bourne said. "A [digital object identifier] for people" could be tagged to different types of scholarly output, he said, to include papers, blog contributions, curation efforts, peer reviewing, or software development.

During the panel discussion at ISMB, CrossRef's Campbell expressed concern about the feasibility of a cross-publisher author identifier.

Bourne pointed out that the CrossRef initiative has already led to cross-publisher citation linking. "If it can happen once, surely it can happen again," he said.

The National Center for Biotechnology Information is working on a functionality called My Bibliography, for scientists to pull together their PubMed citations, and ISI Web of Knowledge has identifiers that, for example, address misidentification, Bourne said. "If one big publisher adopts it, the rest will follow."

Elsevier has its own proprietary system of researcher identifiers, de Waard said, though she added that she is "very interested" in exploring systems that can straddle "a mix" of proprietary and open identifiers.

Genomeweb system

These settings are generally managed by the web site so you rarely need to consider them.

Issue Order: 3

