Digital research/analog publishing – one scientist's view

Based on a paper presented at the 34th UKSG Conference, Harrogate, April 2011

The scientific workflow, from initial ideas to data generated during hypothesis testing, to conclusions drawn from those experiments, is increasingly in a digital form. The publishing process loses much of that digital context since while a PDF or HTML page is digital, it really is a reflection of an analog past in terms of what can be done with that content. While a PDF of a research article has merit as a concise description, it is only one view on that work and others are possible. I argue here that these other views could increase the rate of scientific discovery through provision of a digital continuum that includes the data and methods (where possible) used to reach the major conclusions of the work. Such changes are afoot being driven from the bottom up by scientists and the top down by publishers.

PHILIP E BOURNE
Professor
Skaggs School of Pharmacy and Pharmaceutical Sciences
University of California San Diego

I am a scientist who spends his time working in such esoteric areas as early-stage drug discovery and trying to understand evolution from a three-dimensional molecular perspective. What on earth am I doing writing an article for *Serials*, you might ask? The answer is complex and spans 12 years beginning in 1999 with my work in running a public database of biological information called the Protein Data Bank (PDB). This was scientific content that everyone owned, and around which, communities of scientists gathered and communicated – the equivalent of a scientific sing-along around the camp fire. Six years later, I co-founded an open access journal with the Public Library of Science (PLoS) in computational biology. At the time my motivation was more about providing a journal of the right scope for our discipline than any beliefs about open access. About a year into the journal being published, I had a revelation of sorts, which I defined in an editorial for that same journal titled 'Will a Biological Database Really be Different from a Biological Journal?'[1] My punch line was no, it will not. Both accept content through the web; both have a review cycle; both are ultimately endorsed in some way and subsequently made public on the web. What is different is the value and reward associated with their respective contents and who assumes responsibility for each. What this leaves us with today (at least in the biosciences) is a disembodiment of data (database) and knowledge about that data (research article). This made some sense when data were sent by post on large reels of magnetic tape and new knowledge physically appeared regularly on the current journals shelves of your institutional library. Those days are over, yet the distinction persists to the point of absurdity. Tell me it is not absurd that databases hire curators to read the literature and transpose content from those research articles into databases while at the same time journals accept more data as supplements; but data that are not standardized and not easily used. Both are a waste of intellect and money and can be assumed to slow the rate of discovery. Why not have the two integrated together[2] so that the reader can better explore the value of the work and have a better chance at reproducibility?

An interesting effect of bringing the two together – databases managed by scientists and journals managed by publishers – would be to change the relationship publishers have with scientists who publish. Currently, publishers are in no way part of the scientific endeavor until the very end, at which point the scientific author reformulates all the information they have that began with an idea or observation, led to a hypothesis, a subsequent effort to prove that hypothesis and an outcome to meet the needs of an analog device – the printed journal article. A digital version is then transmitted as a PDF and increasingly read on a digital device.

After the paper is accepted, the scientist is no way part of the publishing process beyond reviewing a proof – and so it has been for centuries. Don't get me wrong, there are some positive aspects of the process – a research article provides a common interface we all learn to use very early, the article itself can be a succinct view, but my points remain if the research article is just one view of the scientific endeavor. It would be interesting to consider a model where scientist and publisher had a different type of contract together and produced a different kind of product that, rather than go from digital to analog to digital, remains digital and can be manipulated in new ways that improve comprehension. About all you can do with a PDF is read it in a PDF reader or print it; whereas with a digital product you can share and build upon the experiments that the original authors undertook. In short, have a very different learning experience from simply reading a PDF.

To bring about such change requires a consideration of sociological, procedural and technical issues. Procedurally, as I have written before[3], means, 'the publishing process begins not at the end of the scientific process but at the beginning of the scientific process itself'. I suspect this is a scary thought for most publishers. It requires a new way of thinking, new technologies, new contracts between customer and publisher and a very different business-to-consumer model with an unknown outcome. But I don't think publishers can ignore it since it is my sense that there are enough drivers to make it, or something like it, happen. Here are a few of those drivers.

Chaos in the laboratory

Much of what we do in the laboratory involves digital content, however, our ability to successfully manage that content – catalog, store and retrieve – is woeful. We lose digital information at a time when funders are demanding data preservation and sharing policies, in other words, that we don't lose it and we make better use of that information. Eventually these policies will grow teeth and impact funding decisions, a very strong driver for doing better. Scientists need help from someone, but who?

The library

It should not go unnoticed that what is needed in terms of managing digital laboratory content to avoid chaos is what the library has been doing for generations: cataloging, storing and retrieving. Unfortunately, the foray that libraries and their respective institutions have taken in this direction already through institutional repositories appears to be, to put it politely, less than a success thus far[4]. Institutional repositories to date have been too much of a 'build it and they will come' effort, rather than 'let's figure out what our scientists really need, have committed scientists help drive the effort, and deliver it to the community if there is a sustainability model that can be found'. If institutions could make this work, it begs the question well why don't we just publish the content as well and bypass publishers? The prestige of established learned societies and their journals, based on domains of knowledge, might work against an institutional model, but institutions have prestige and are hungry to find alternative revenue models in these hard times of limited public funding. In summary, libraries and associated institutional repositories could be a driver of change if they could be made to work properly.

The app

This driver of change is a bit more hypothetical, but bear with me. Some newspapers and publishers, for example Elsevier, are doing something that I think, in time, will be a game changer. They are leveraging aggregated content by making the scientific community do the work – a brilliant strategy and one that has worked for centuries for the review process and might now work for the process of disseminating and analyzing scholarship. As an example, consider SciVerse as applied to an Elsevier hub, a system that provides aggregated content, from multiple journals and the web, on a modular platform and invites the crowd to come up with the ideas and the applications for how it can be used. It's a win-win for scientists and publishers alike. Some scientists get rewarded for apps; all scientists get rewarded by new knowledge discovered when using the apps. Through the apps, the publisher's content becomes more valuable. This becomes so intrusive – just like apps

on mobile devices – that the pressure to increase the content base to include data and other unpublished material will increase. When that happens, labs will effectively be able to 'publish' in the sense of making available with some form of automated peer review, or not, anything they choose. This includes the methodologies they used to generate that content in the first place. If all this were to come to pass, being a publisher would be very different than what it is today.

If you do not 'buy' what I am saying, then let me give you an example of a prototyped app that, as a scientist at least, I find compelling as a driver of change. It comes from my own discipline, but is a generally applicable concept. The scientific literature is full of references to gene names. For years we lived in chaos in trying to interpret just what was known about a single gene because it was referred to by a different name in different disciplines; names that the expert in that gene might know, but not the novice. Scientists, through common naming schemes, have done better and also, since that gene likely has data associated with it in a public database, there was a reference to a database gene identifier included with the published article. That did not mean much when I had to read every paper and make novel associations between genes based on my scientific knowledge. Either I was not smart enough or there was too much to read to make new discoveries. Enter a new app. The app scans the literature for references to database gene identifiers and represents each gene as a node in a network, as shown in Figure 1. If two genes appear in the same paper an edge (line) is drawn between them. The more this happens, the thicker the edge. Such networks often appear as incomprehensible hairballs, but sometimes properties emerge.

The emergent network on the left has a topology – two lobes with a single connection. In itself, this does not mean much, but if I start to overlay other properties on that network a different picture emerges. On the right is the same network with just the type of journal in which the gene associations were made added. What emerges is a single gene that appears in both the cardiac disease literature and in the immunology literature. It may well be that immunologists were not aware of the role of this gene in heart disease and vice-versa. The history of science is filled with such serendipitous discoveries. Let's have apps that accelerate the rate of such discoveries. This is only a hypothetical example, but I hope you get the idea. Imagine if I could then take that shared gene discovered in this way and delve into an organized view of what scientists had and had not published

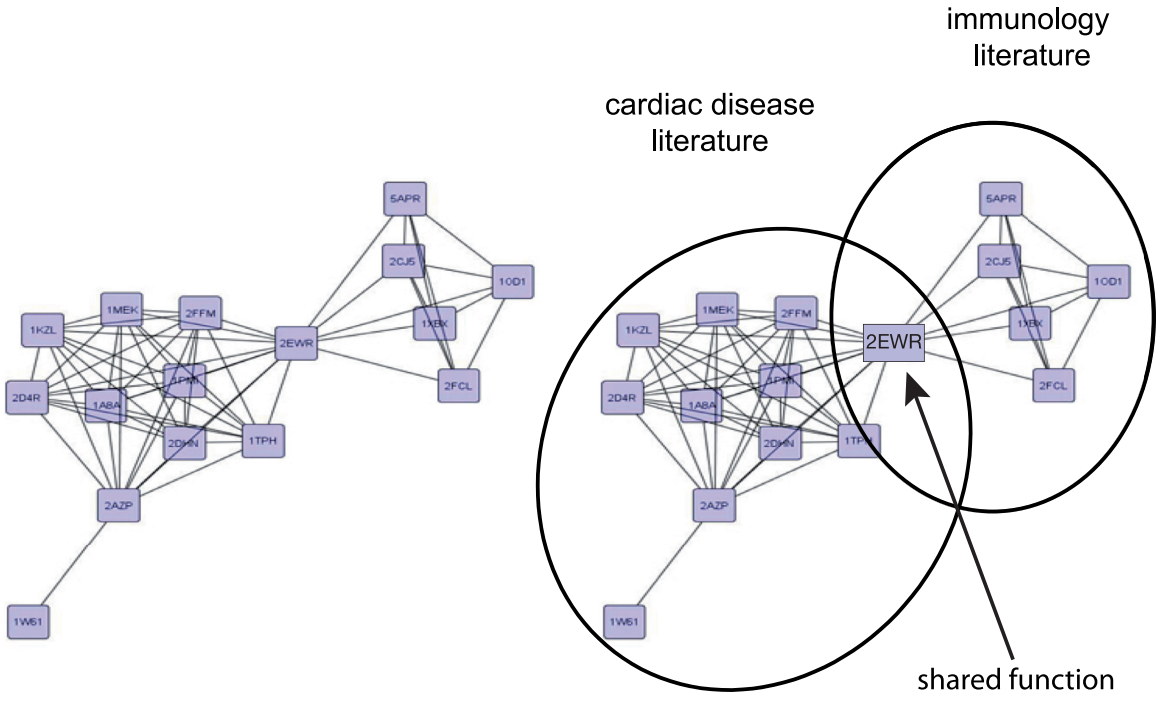

Figure 1. Left: a network of genes connected by references in the same research article; right: the same network overlaid with the source of the literature

concerning that gene – all theoretically possible in a digital medium. New scientific discovery, such as that imagined here, is a powerful driver of change that gets the attention of the publisher's customer base, i.e., scientists, particularly if those in the same field are making those discoveries using new tools and they are not.

If publishers and scientists accept that what I have described are indeed examples of drivers of change, and that this change is inevitable, what should we be doing now? In January 2011 there was a workshop in San Diego called 'Beyond the PDF'[5] in which a number of stakeholders – publishers, scientists, tool developers, librarians - got together to address this question. The answer was a well-designed and well-executed example of a change in scholarly communication that would make a difference in science that would become a poster child for further such grass roots efforts. The goal the group set itself was to make a difference in the understanding of a childhood disease called spinal muscular atrophy (SMA). A poster child in the true sense of the word since it is a childhood disease in which it is believed that progress can be made and for which some answers are already buried in the literature if they could only be recovered. If those answers were recovered, even with some difficulty, it would open the minds and pockets to what could really be accomplished if we truly treated science from the original idea through to dissemination as a digital continuum.

References

1. Bourne, P E, Will a Biological Database Really be Different from a Biological Journal?, *PLoS Comp. Biol*., 2005, 1(3), e34 (accessed 13 June 2011).
2. Prlic, A, Martinez, M A, Yukich, B T, Dimitropoulos, D, Beran, B, Rose, P W, Bourne, P E and Fink J L, Integration of Open Access Literature into the RCSB Protein Data Bank Using BioLit, *BMC Bioinformatics,* 2010, 11:220 (accessed 13 June 2011).
3. Bourne, P E, What Do I Want from the Publisher of the Future?, *PLoS Comp. Biol., 2010*, 6(5):e1000787 (accessed 13 June 2011).
4. Salo, D, 'Innkeeper at the Roach Hotel', available at: http://minds.wisconsin.edu/handle/1793/22088 (accessed 13 June 2011).
5. Workshop held on January 19-21, 2011 University of California San Diego: https://sites.google.com/site/beyondthepdf/

■ **Philip E Bourne PhD**
Professor, Department of Pharmacology and Skaggs School of Pharmacy and Pharmaceutical Sciences
University of California San Diego
Associate Director, RCSB Protein Data Bank
Co-founder and Editor-in-Chief, PLoS Computational Biology
E-mail: bourne@sdsc.edu

The DOI for this article is 10.1629/24119. Click here to access via DOI:
http://dx.doi.org/10.1629/24119
For a link to the full table of contents for the issue of *Serials* in which this article first appeared, click here:
http://serials.uksg.org/openurl.asp?genre=issue&issn=0953-0460&volume=24&issue=2