science commons

The Control Fallacy: Why OA Out-Innovates the Alternative…

John Wilbanks
VP, Science Commons

Lost in too much of the debate over Open Access (OA) is the relationship between access, control, and innovation.

Too often, the OA discussion is one of radical polarization. Much of this comes, in my opinion, from the focus of the debate on economics and business models. While the money side of this is clearly vital - peer review needs to be paid for, after all - it's also the issue that often leads to the least constructive debate. Money has a way of bringing out the arguments. But is it the right place to have the discussion about OA? Perhaps not.

If we go back to the beginning of the OA movement, which Peter Suber traces to 1966, it's clear that the concept comes from education and science: the founding of the Educational Resource Information Center and the launching of the Medline service[1]. OA's wellspring is the idea that innovation in education and science is best served by access to information. That's often buried in the debate over prices and profits, business models and legislation.

In the end, OA isn't even about money. It's about innovation. It's always been about innovation.

It's also about a lot of other things. Author self-interest in increased impact, information justice or fairness, and so on. And it bears mentioning the publisher's response, which is their initiative to improve access to the literature in the developing world through HINARI and AGORA[2]. But, not only do these initiatives provide only partial access to these audiences but these initiatives also do not respond to perhaps the least understood and perhaps most important argument in favor of OA - the argument for innovation.

We've learned over the past decades of technological development that open systems create the conditions for exponential innovation. The Internet and the Web are obvious examples. But it is worth remembering that there was private competition for these systems, premised on control, and that the systems that won were those that embraced access over control.

The Web is part of our lives. We take it for granted, we use it on our phones, we read journals and book restaurants. But it faced a lot of competition at the time it was created.

For just one, let's look at DynaText. At the time of the creation of the WWW, DynaText was a powerful system with a lot of the functionality that CERN needed. But it came with a price, both in terms of economics and control. You had to pay to make copies, pay to

[1] http://www.earlham.edu/~peters/fos/timeline.htm last retrieved 30 march 2008
[2] http://www.who.int/hinari/en/, http://www.aginternetwork.org/

make changes, and to ask permission over and over again.[3] The WWW was a weaker system, but it was open – anyone could build a web site without permission, anyone could link to another site without permission, and anyone could look at the underlying site code to jumpstart their own.

The choice between control and access had other downstream implications. Imagine a world where hypertext linking was controlled – where you had to ask permission to do it. Modern search engines might never have come into existence. Google copies and indexes web pages, and ranks them based on the trillions of hyperlinks connecting them[4], and it'd be a miserable life for them if they had to ask permission from every site, every time. Access is why our entire day-today experience of search exists.

On the network, releasing control turned out to be a good design choice: empower the users, and they will build – and extend – the network for you. They'll start businesses you never imagined, and draw more people to the network than you ever dreamed.

We see this in science as well. The Human Genome Project, with its open stance, not only ended up benefiting the private competition, Celera (which used enormous amounts of public sequence in its own assembly[5]), but also out-competed it in the end[6]. There's a simple reason for this: more people can use and make citations to an open database than a closed one, and, crucially, more people can annotate an open genome than a closed one. This is a vital point: the genome itself is just the A's, T's, C's and G's – akin to the 1s and 0s of binary digital code. Annotation is the process by which we add meaning to the code. And you can't annotate what you can't access.

Again, openness was a good design choice: empower your scientists, and they will annotate the genome for you. The network lets the smart folks in the Distributed Annotation System bring together the distributed contributions of the many[7], creating a whole much greater than the sum of parts you could create with a closed genome.

These examples are not accidents. Good design choices are at the heart of powerful network effects. And when dealing with fundamental information resources, there aren't many more important choices than access versus control.

We are facing that precise choice in the scholarly literature right now. The biomedical journal articles of the past 30 years capture the core knowledge of the biotechnology revolution: the detailed relationships between genes, proteins, and diseases, the experimental protocols – *everything*. This is the fundamental information resource for new biological discovery.

[3] http://en.wikipedia.org/wiki/World_Wide_Web#History last retrieved 31 March 2008
[4] http://www.google.com/technology/ last retrieved 31 March 2008
[5] http://www.sciencemag.org/cgi/content/full/291/5507/1304
[6] http://www.nature.com/nature/journal/v435/n7038/full/435006a.html
[7] http://www.biomedcentral.com/1471-2105/2/7

It's the knowledge genome[8].

And we have an emerging knowledge infrastructure that could really make the information sing. It comes from the finally-emerging Semantic Web[9], with its controlled vocabularies and ontologies, and it comes from the explosive power of the grassroots tags and folksonomies of Web 2.0[10].

Funders understand this. The innovation argument has been central to the way in which the US National Institutes of Health have thought about and implemented the policy mandated by the US Congress. It is worth pointing out that the deposit requirement has not been completed until the PI signs off on the XML-formatted version in PubMed Central. The whole reason NIH is spending resources reformatting these articles and making sure the PI certifies the accuracy is to make the articles more machine-accessible and linkable. Access is the key to increasing innovation through digital technologies.

It allows us to apply the full power of technologies to index and sort, slice and dice, annotate and tag the literature. It allows us to make links between databases and journal articles, to build big graphs of relationships between papers and other papers, and to go even deeper, to build giant graphs of the relationships those papers describe between the genes and proteins[11]. It allows us to use machines and the power of crowds to figure out what those relationships mean, to avoid repeating past mistakes or conduct experiments that are doomed to fail because a piece of existing knowledge is just out of reach.

That's the power of what we might call a "knowledge web[12]," built on a knowledge infrastructure. Just to be clear, here's what I mean by a knowledge web: it's when today's web has enough power to work as well for science as it currently works for culture. That means databases are integrated as easily as web documents, and it means that powerful search engines let scientists ask complex research questions and have some comfort that they're seeing all the relevant public information in the answers. A knowledge web is when journal articles have hyperlinks inside them, not just citations, letting systems like Google do their job properly.

A knowledge web is predicated on access, and not control, of knowledge. There will never be a competition to provide the best single-point query to the full-text of journals without access- unless the journals all merge down into one company. That's the only way a controlled system covers the whole world, through monopoly. There will never be a knowledge web where the entire backfile is hyperlinked to databases for relevance based indexing without access. Scientists won't get to use the newest and best

[8] http://bioie.ldc.upenn.edu/

[9] http://www.w3.org/2001/sw/

[10] http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

[11] http://sw.neurocommons.org/2007/kb.html

[12] http://www.ctwatch.org/quarterly/articles/2007/08/cyberinfrastructure-for-knowledge-sharing/

technologies until those companies that control knowledge decide to adopt those technologies. Control is the enemy of testing the newest technologies, of building one's own system to suit one's own needs. We have to have access to build a knowledge web, at least if we hope to replicate the success of the regular Web and the Internet.

But unlike in the early days of the Internet and the Web, where the core design decisions were made before anyone realized the economic impact, and when the instinct to control could more easily be overcome, we're trying to build a knowledge web post-hoc. And there's already an extraordinary control culture that is pervasive at several layers of the knowledge network.

The most obvious layer of control is the use of copyrights to control the articles themselves. The articles are creative expressions by the authors, transferred to the publishers – that part is easy to understand. But those copyrights on expressions are being used to control and *limit the impact of the ideas* that are contained in the articles[13]. This is an inversion of the original conception of copyright – it was never supposed to restrict the movement of ideas, and certainly not to restrict the movement of scientific facts like the one cited above.

Those scientific facts are core pieces of the knowledge web, but they are sitting on the Web with few links, and few hooks to the emerging infrastructure. Think for a second about how densely each sentence in a scientific paper is packed full of linkable knowledge. We can zoom in on one:

> IGFBP-5 plays a role in the regulation of cellular senescence via a p53-dependent pathway and in aging-associated vascular diseases[14]

A quick set of queries reveals that in just this one sentence, we could make five hyperlinks to databases: "IGFBP-5" is a protein[15], "regulation" is a kind of activity with an ontological URI[16], "cellular senescence" is a concept defined at multiple websites, including Wikipedia[17], there is a set of p-53 dependent pathways aggregated online[18], and the US government uses vascular diseases as one of its terms to classify the medical literature[19].

[13] http://network.nature.com/blogs/user/wilbanks/2008/03/26/creative-works-copyrights-and-publishing
[14] http://www.molbiolcell.org/cgi/content/abstract/E07-03-0280v1?ck=nck
[15] http://www.ihop-net.org/UniPub/iHOP/gs/89377.html
[16] http://amigo.geneontology.org/cgi-bin/amigo/go.cgi?view=details&search_constraint=terms&depth=0&query=GO:0050789&session_id=9014b1206992334&show_associations=list
[17] http://en.wikipedia.org/wiki/Senescence
[18] http://p53.bii.a-star.edu.sg/aboutp53/pathway/index.php
[19]

http://www.nlm.nih.gov/cgi/mesh/2008/MB_cgi?mode=&index=14093&field=all&HM=&II=&PA=&form=&input=

You can repeat the above example pretty easily, and people like Cliff Lynch have been tireless advocates for this kind of re-use[20]. PubMed lists about 16,000,000 abstracts, all of them chock full of linkable knowledge. But precious few of them are available in full text – and if you start to think about the number of sentences that you can't see, the number of entities to which you cannot link, and the implicit knowledge network that remains dark to all of us, you can start to comprehend the scope of the problem that the control culture represents. More than ninety percent of our modern biomedical knowledge heritage – including nearly everything since the birth of modern biotechnology in the 1970s – is essentially off-limits to anyone but the publishers who own it[21].

Now, those publishers are happy to rent access to the knowledge heritage. Rent is the key word here, though. When the scientific publishing industry went online, they stopped selling journals to people and started renting them. If you've ever rented an apartment, you know that rentals come with a lot fewer rights than ownership. In this context, the users lost a slew of rights – remember, you can legally resell a physical copy of a book or CD, but you can't legally forward a PDF from the newest issue of Science. You don't have the right to share things like journal articles when you rent them.

Many of the controls that a publisher can impose are built on top of that copyright. So even if you have rented access to the full text of articles, the license agreements you've signed with the owners frequently make it illegal to use software to index and mine the literature. Elsevier's copyright rental agreements are a good example – make sure to read through to page 5[22].

This control culture is not the result of bad people making evil decisions. It's simply an antique system. It made sense when it started, and it actually made sense until the Internet came along and changed everything. But the control culture is a powerful drag on innovation when you're in a networked reality.

Now, it's wrong and unfair to talk about "the publishing industry" as some sort of monolithic control-beast. Hundreds of journals deposit their full text into PubMed Central[23]; hundreds of online repositories host self-archived papers[24]. More than 500 journals are licensed under the open Creative Commons copyright system[25] – including this one. And many publishers are making good faith efforts to understand and explore

[20] http://www.cni.org/staff/cliffpubs/OpenComputation.htm

[21] As of June 2007, PubMed Central contained 1,000,000 articles of over 16,000,000 abstracts indexed, for just over 6% of the total. That number is growing at about 7% a year, so it is safe to say that at the time of writing, PMC has yet to exceed ten percent of the total.

[22] http://orpheus-1.ucsd.edu/acq/license/cdlelsevier2004.pdf

[23] http://poeticeconomics.blogspot.com/2008/03/over-400-journals-participating-in.html

[24] http://archives.eprints.org/

[25] Compiled by counting the journals listed by BioMed Central, Hindawi, Public Library of Science, and other journals using Creative Commons licensing.

this new world, to balance their own cost needs and the demands of their newly-empowered networked customers.

These ideas are far from anathema to publishers who understand the networked world. BioMed Central and Hindawi publish immediately to the Web under open copyright licenses. Between the two they brought in almost $15,000,000 in revenues last year.

This is not just a feature of the newer OA publishers. Nature Publishing Group, among the traditional journals, has made extraordinary efforts to engage with innovation on the Web[26]. Nature Precedings is an open effort aiming to replicate the success of the physics arXiv[27] pre-print server in the life sciences (under the most permissive Creative Commons license available). Nature makes papers describing the full genome sequence of an organism available under a Creative Commons license too.

That's empowering the user. And it's smart. It dramatically increases the odds that a scientist comes along and innovates on the content. It doesn't limit the universe of innovators through a series of controls and contracts and invoices.

This is in the end the fallacy of knowledge control. The power of the closed system is rooted in coherence, consistency, quality control – things that are vital and important in the right context. But these are powers that frequently fail to scale when you're dealing with a problem of great complexity.

Complexity challenges coherence. Complexity overwhelms consistency. Quality control can only scale as the people scale, and in closed systems, all of those people must somehow be paid by the same paymaster. Closed systems and cultures of control simply don't work as well as open systems in complex, rapidly shifting environments. And is there a more complex, more rapidly shifting information space than life sciences research?

16,000,000 papers. 30,000 genes. And that's the easy part. As one of my favorite bloggers noted, the complexity of the living systems is such that in comparison, the new Intel processor looks like the back of a shampoo bottle[28]. The systems under study are so complex that it is impossible for any one company to gather enough information in one place. None of us are smart enough on our own to figure it out.

But maybe all of us, together, are smart enough. This idea, inherent in the "wisdom of the crowds" philosophy of the Web, is incredibly important and powerful in the life sciences. This is the idea that if we can stitch together all the little pieces of knowledge the right way, we'll realize we know a lot more than we thought, that if we can organize and

[26] http://www.nature.com/launchpad/index.html

[27] http://arxiv.org/

[28] http://pipeline.corante.com/archives/2007/11/06/andy_grove_rich_famous_smart_and_wrong.php

manage our knowledge better, we can use the information to make better decisions, rather than letting it overwhelm us.

But this knowledge web, where all of the literature and databases are cross-linked and searchable from a single interface like Google, isn't going to happen by accident. Unlike when we built past information networks, we don't have the luxury of building the knowledge web before anyone knows it's valuable.

We have to build this web together. It's going to require commercial publishers – they have the backfile of medical knowledge. It's going to require hackers – they know how to do the hard technical work. It's going to require funders – they create the incentives to extract and reformat the knowledge. It's going to take users, like the pharmaceutical companies and the academics. It's going to take all of us to build a knowledge web, a web that truly supports the kind of complex queries required to get valuable answers out of a deluge of information.

That's how we're going to change the human health landscape and the drug discovery process, through disruptive innovation, through accelerating the process at which we make breakthrough discoveries about basic cellular systems. We need a lot of scientific revolutions and we need them fast.

That's why access is so vital. That's why it's vital to support the publishers that go OA, and the traditional publishers who are taking bold steps to foster innovation and knowledge creation. That's why it's important to focus on *access* and *rights* and not *price* – because giving knowledge away for free but without the rights to make it useful doesn't make the grade. Freedom here isn't about prices, but about rights.

It might be good to close with a little history. In September of 1995, the Clinton administration released a document on the relationship of intellectual property to the then-emerging National Information Infrastructure. This was commonly known as the "White Paper.[29]" To say it was controversial in the legal community would be an understatement. But broadly speaking, most folks weren't paying attention in 1995 – the Internet boom hadn't hit yet, and at least in the US, everyone was much more worried about the O.J. Simpson trial.

There's a great quote from the paper: "an information infrastructure already exists, but it is not integrated into a whole" – and the authors advocated sweeping expansion of intellectual property rights to ensure that the network would be integrated into a whole and populated with useful content. Creators, the paper stated, would never create unless a powerful set of controls was added to the network.

It didn't work out that way. The controls weren't all added, and those that were added got broken pretty quickly[30]. But the Internet and the Web not only survived, they exploded.

[29] http://www.uspto.gov/go/com/doc/ipnii/
[30] http://www.npd.com/press/releases/press_061220.html

This happened because of the triumph of access over control, indeed, despite the best efforts of control to take over the network[31].

The fallacy of control in networks is that in return for building in power today, you lose the ability to let your users make the system more powerful over time. That's why in the long run network systems built on access tend to out-innovate network systems built on control. Remember the fate not only of DynaText but also of Prodigy and the "walled gardens" of the early Internet – powerful for their time, but terminally unable to compete with the unruly but generative Web[32].

We're at a similar inflection point now, reminiscent of the moment when the White Paper was written. Most people aren't paying attention, though perhaps they've heard of the potential of Semantic Web or Web 2.0. We have a knowledge infrastructure emerging, but it's not integrated into a whole. And the dominant voices are the voices of control. It can sometimes feel like dark days.

But the good news is that over time, the more powerful networks are the open networks. The good news is that the creators of scholarly literature are also the consumers of scholarly literature, and they want those powerful networks to function as well for scholarship as they function for commerce.

If we focus on innovation – and its connection at the hip to access and users' rights – we can start to see the light at the end of the tunnel. The knowledge web is coming. We just have to keep building it, server by server, article by article, person by person. And we have to keep it open.

So what can you do? You can start by exercising your rights – even with the present culture of control, many journal publishers permit self archiving of the author's final manuscript, and many authors fail to act on this right. The first is to exercise this right and to contribute to the knowledge web right now. Second, there's a good chance you are members of scholarly societies. Your societies should be the leading the charge towards the knowledge web – are they? It's worth asking that question. And last, look to the example of the editorial board of this journal. Faced with the choice of embracing a control culture, they simply said "no more" – they chose access.

[31] http://www.cs.cmu.edu/~dst/DeCSS/Gallery/

[32] http://papers.ssrn.com/sol3/papers.cfm?abstract_id=847124