John Wilbanks

January 10, 2012

Response to Request for Information: Public Access to Peer-Reviewed Scholarly Publications Resulting From Federally Funded Research

There are two kinds of markets for the access and analysis of peer reviewed publications emerging from federally funded research.

One is the "mental" market, or the size of the readership base. This current market for the results of scientific research is limited, artificially, to those researchers who sit inside wealthy institutions whose libraries can afford to subscribe to the majority of scientific journals. This excludes researchers at many state educational systems, community colleges, middle and high schools, state and local employees, the American taxpayer, and the American entrepreneur. By implementing a robust public access policy for federally funded research outputs, each of these groups will have access to the literature and, if the policy is crafted correctly, the right to begin creating new knowledge and experimenting with new businesses atop it.

This leads to the second kind of market – the economic one. At the moment, there is at best a sputtering startup culture built atop the scholarly literature, with a few text-mining companies here and there, mainly in the life sciences. A small number of publishing houses exploit their gatekeeper function to impose prices on elemental services like abstracting that in the consumer world would cause revolt, and the American venture capital industry invests instead in social media. The lack of robust public access to the literature – and the relentless focus on asserting and controlling copyright – means that economically it remains a content industry and not a knowledge industry. We will not see meaningful job creation in secondary markets as long as the primary secondary use of digital literature is informal file transfer via Twitter (using the #icanhaspdf hashtag).

The scientific enterprise would clearly be better served through some creative destruction. We have replicated the analog production and distribution system digitally, realizing few of the cost benefits, few of the speed benefits, and none of the innovation benefits of the transition. iTunes came out more than a decade ago. Netflix, more than 15 years ago. Content industries are disrupted by technology, and should respond with innovation, creating new jobs that are durable against outsourcing. Yet we have seen none of this in the scholarly publishing industry, which given its enviable almost-monopoly on the outputs, has little incentive in the absence of policy to make the admittedly difficult transition to a knowledge industry.

The intellectual property interests of the stakeholders must be aligned with the scientific goals of the government and taxpayers, which is easily done through the use of open copyright licenses such as those provided by Creative Commons. Open copyright licenses protect the rights of the author or legal copyright owner while providing for conditional access to the public – for example, copying and republishing may be allowed, even for commercial purposes, but attribution back to the author and

original journal, including a link to a free copy of the paper, would be required and if not present the full power of copyright remedy could be brought to bear on the violator.

Open copyright licenses can also be phased in alongside an embargo in a way that both protects the economic interests of publishers and the long term public interest in access to research literature. For example, during an embargo period, no open license might be used, switching to a license like Creative Commons Attribution-Non-Commercial for a second intermediate embargo period, and then eventually decaying to a Creative Commons Attribution license that is fully compliant with community definitions of Open Access. One could easily imagine using real data about economic usage of the literature to set these times in a noncontroversial fashion, creating a truly open corpus of literature both in terms of technical access and legal rights, without an emotional argument unfounded in data or the reality of modern web-based copyright licensing.

The pros of a centralized approach to managing the public access are fairly straightforward. First, a single point of access to the research, with stable and common identifiers, radically decreases the cognitive burden to find and download the research. Second, the centralized approach raises the odds of common standards being applied to link the research to data (as we see in the vastly popular PubMed links to both internal and external data sources). And third, the centralized approach relieves the publishers of the need to perform these infrastructural functions, which should lower economic demands on the industry. However, it is important that a centralized repository be accompanied by open copyright licenses, so that additional copies of the open corpus can be maintained in libraries and research institutions, providing additional security to the preservation of the scholarly record. This mixture of a centralized resource with open licensing and standard technologies mirrors that of the internet itself, which runs on a small set of centralized resources (the domain name system, for example).

Centralization of resources also radically lowers the burdens on the researchers and their host institutions. A single interface to upload to learn, a single interface for libraries to manage, and the comfort of a persistent repository rather than the funding of local repositories at library after library, reduces the burden of compliance not just on the publisher but on the other key stakeholders in the process.

The cons of a centralized approach are also straightforward. It must be funded (and thus can be defunded in a crisis) and it takes a certain amount of control out of the hands of the publisher – but since the goal is to remove access controls, removal of control may in fact be a pro rather than a con.

To encourage interoperable search, discovery, and analysis capability (and the small business, venture-backed job creation that innovation in each of those spaces will bring) the federal government should make a commitment to clear standards in document format, metadata, structured vocabulary and taxonomy, and commit to using its procurement power to only pay for articles that carry the designated metadata. Standards building is a long and cumbersome process, and any standard that doesn't have adoption may be worth less than the (digital) paper on which it is printed. Having a stable customer for

metadata in the person of the government creates a defined and clear market for startup business to serve, and creates potential for top-line economic growth at more established publishers as well.

It is vital as well to ensure that the metadata associated with the research is itself public. While the copyright status of metadata has not been extensively tested in court, there is reason to believe (from cases involving medical procedure codes among others) that at least some metadata, especially vocabularies and ontologies, may carry copyright obligations. The federal government should authorize the use of open copyright licenses such as the Creative Commons licenses on metadata, and preferentially select vendors who use the most open of copyright licenses and tools.

While scholarly articles are the traditional focus, and should be the first order of business in a federal open access policy, book chapters and conference proceedings (and even perhaps more novel forms of communication, like blogs and wikis and social media) should be evaluated for inclusion in the policy. However, careful attention should be paid to the level of effort required to create the work, and different rules might be applied to works that require a bit less effort (a conference poster might be required to be open immediately, no embargo) compared to those that require significant effort (a book chapter might receive a longer embargo than an article).

About me:

I am a Senior Fellow at the Ewing Marion Kauffman Foundation, and a Fellow at Lybba. I am a Senior Fellow at the Kauffman Foundation, the Group D Commons Leader at Sage Bionetworks, and a Research Fellow at Lybba. I've worked at Harvard Law School, MIT's Computer Science and Artificial Intelligence Laboratory, the World Wide Web Consortium, the US House of Representatives, and Creative Commons. I also started a bioinformatics company called Incellico, which is now part of Selventa. I sit on the Board of Directors for Sage Bionetworks, iCommons, and 1DegreeBio, as well as the Advisory Board for Boundless Learning and Genomera. I have been creating and funding jobs since 1999.