Open Science & Scientific Publishing: Open Access and the progress of science

OPENING SLIDE

Thank you.

John has just shown us how exciting the prospect of open science can be in the field of neuroscience. In the next fifteen or twenty minutes before we open things up for discussion, I'm going to share a couple of things with you about how libraries are working to make the vision of open science a reality.

OUTLINE SLIDE - 1

First, I want to share a couple of other examples about how open science and open access can advance the progress of science.

Then I'll share with you what research libraries like MIT's are doing to help make that promise a sustainable reality.

Libraries are working on two issues that have to be addressed if we are going to realize the promise of open science: those are permissions, and persistence.

Finally, I want to talk briefly about the importance of partnerships. Researchers and libraries are in this together.

SLIDE 2

PROMISE – more than the same old story

One of the obvious barriers to sharing scientific information is economic.

The high cost of library journal subscriptions means not only that library budgets are stressed, but also that researchers in many countries and at many institutions simply can't afford access to the research they need.

Most faculty and students are very aware of the high cost of research journals. Provosts at research universities certainly are.

This is an important story to tell. But it is not the whole story, and I'd like to talk about another part of the story.

SLIDE 3

Another part of the story is that the publishers not only control access to the **content** of scientific publications. They also control how that content can be used and repurposed. That means they control the ability of scientists to create new ways of using their

findings to advance science. And this is ironic, because there's a whole lot that scientists and technologists could do if publishers didn't have this control:

SLIDE 4:

The question that we face today is not only, how can we control the cost of journals, but also, what would be gained if scientists could create new ways of using the record of science?

John gave us a great illustration of what is to be gained if efforts like Neurocommons can use the scholarly record in the ways he described.

I'm going to talk about a couple of other examples of what the scientific record might look like and how it might work if publishers didn't have exclusive control over the use of scholarly articles.

SLIDE 5:

text mining

SLIDE 6:

One idea that we've heard about from John today is the idea of "computing on scientific literature" – not only indexing, but also computational analysis, abstraction, correlation, anomaly identification and hypothesis generation - often termed "data mining" or "text mining."

SLIDE 7:

CrystalEye is one such example:

CrystalEye is a project at Cambridge University to aggregate crystallography from supplementary data to articles at publishers' websites. It converts the data to Chemical Markup Language and generates pages for easy browsing with 2D and 3D renderings of the structures.

This slide shows a screenshot of the CrystalEye table of contents for an issue of Acta Crystallographica Section E.

SLIDE 8:

retaining information

SLIDE 9:

Another opportunity to create a record of science that can be mined effectively is to take advantage of the capabilities of semantically rich XML documents. Traditional presentations and file formats for information in articles are PDF and HTML.

These may fail to preserve or carry forward data that could in fact be preserved by linking to underlying values. Rich XML document representations would permit future

tools for viewing, annotation, and visualization of results. This slide shows a piece of XML that carries a lot of content, not just formatting information.

SLIDE 10:

Rich, flexible units of scholarly communication

There are other opportunities to create a richer, more interesting scientific record.

The August 2007 issue of Cybertechnology Watch Quarterly was a special issue on the coming revolution in scholarly communications and cyberinfrastructure.

Some of the ideas discussed in this are peer review of data sets, publication of data visualizations, and creating new "packages" of research outputs made up of related elements such as research models and data as well as research articles.

SLIDE 11:

For example, Herbert Van de Sompel at the Los Alamos (LANL) Library has been working on a new infrastructure for sharing scholarly communications, based on changing the basic unit of communication from articles, to "objects".

They call this project "Object Re-Use and Exchange (ORE)". They are developing standards that use and re-use what they call "compound scholarly communication units", which are aggregations that form a logical whole. The units can include datasets, simulations, software, dynamic knowledge representations, annotations, as well as articles. Here's a simple illustration:

SLIDE 12:

Here's how they illustrate how the compound object is "composed" by a resource map also published on the web.

SLIDE 13:

Here's another idea for enriching the scientific record: mash-ups of data & articles

SLIDE 14:

Phil Bourne is famous for his work in founding the Protein Data Bank (PDB).

Like John, he's concerned that the huge volume of publishing output in his field makes it prohibitive to use traditional methods of keeping current with the literature.

One of Phil's ideas is called BioLit. It's a Word plug-in designed to capture semantically rich metadata during authoring. Articles created with the plug-in can be linked together in new ways that highlight related uses of data.

As Bourne says,

"The literature will simply become another interface to biological data in a database, and the database can recall appropriate literature."

SLIDE 15:

Phil Bourne has come up another idea, he's partnering with Public Library of Science on it. He calls it "SciVee". It's like YouTube for scientists.

SLIDE 16:

It allows authors to upload an open access article they've published, along with a video or podcast they've made that describes the highlights of the paper. The author can synchronize the video with the text, figures, etc. of the article. They call the result a "pubcast":

SLIDE 17:

SO, those are a few of the ideas that people are coming up with to create a more richly interlinked scientific record - one that takes advantage of all the technological intelligence we can muster.

SLIDE 18:

2. PERMISSIONS & PERSISTENCE – Barriers & Library Actions

This is all very exciting, but anyone trying to develop new systems like these is going to run into two major barriers: permissions, and persistence.

Permissions:

Most scholarly publishers ask authors to sign over their copyrights to them – the entire bundle of copyrights enjoyed by authors.

SLIDE 19

Here is a typical copyright transfer form. It's from the American College of Neurology.

SLIDE 20

Here's an excerpt so you can read the transfer language.

SLIDE 21

When an article is open access, it can be used in all the ways we saw above.

When it isn't open access – when the publisher owns the copyright – it can't be, without the publisher's permission.

There are also other, more personal consequences of publishing without open access:

When an author's work isn't available **as open access**, it probably means that it won't

be accessed as often, or cited as often, as work that is open access. Recent research suggests that open access increases citation by 51% in electrical engineering and 91% in mathematics. (Antelman, C&RL, Sept.04)

SLIDE 22

Hal Abelson, John Wilbanks, and others associated with Creative Commons are doing great things to make it easier for authors to retain their copyrights.

Faculty and students, including Professor Brian Evans and the MIT Faculty Committee on the Library System, and student members of the Free Culture group here at MIT, are also working to increase researchers awareness of these issues and encourage researchers to use open access publishing channels.

SLIDE 23

I'd like to tell you about a few things the MIT Libraries are doing about the problem of permissions:

First, we have a program of outreach and education. The MIT Libraries Director, Ann Wolpert, regularly speaks and consults on these issues with MIT Faculty. Many of MIT's subject librarians also communicate actively with departments and individual faculty about open access.

A key part of the Libraries' outreach and education on this is Ellen Duranceau's position, as Scholarly Publishing and Licensing Consultant, This new position was funded by the provost last September.

Ellen maintains an active website on scholarly communication issues.

SLIDE 24

She also teaches workshops, such as a "Publishing Smart Workshop" for graduate students, co-sponsored with the Graduate Students Council this coming Friday, November 16.

SLIDE 25:

Another way the MIT Libraries are addressing the permissions problem is by supporting the "Amendment to Publication Agreement" developed by MIT's Committee on Intellectual Property. This amendment makes it easier for MIT researchers to retain their copyrights.

It's a simple form for authors to attach to the standard copyright transfer agreement. There's even an agreement-generating web tool.

The agreement enables authors to continue using their publications in their academic work at MIT, or to deposit them into open access repositories like NIH's PubMedCentral.

These activities at MIT Libraries are both exemplary and typical of the activism on the permissions problem at other research libraries and in the library profession at large.

SLIDE 26

Persistence:

In talking about the "promise" of science we are talking about the future. The unpredictable, distant future, and our ability in that future to access and use the research that's being done today. That ability will be nil if the record of science doesn't last – if it isn't, in the words of archivists, "persistent."

We all know how much of the open web is not persistent. It takes the heroic and visionary work of the Internet Archive and their "Way Back Machine" to keep even snapshots of it persistent.

SLIDE 27

For example, there is only one version of the MIT web site available from the year 1999, and thanks to Brewster Kahle and the Internet Archive, we can still see it. That's less than a decade ago, and we've lost 364 out of 365 of MIT's 1999 home pages!

That probably works OK for popular culture, but that ratio would be extremely problematic for research science.

The problem of persistence of the digital scientific record is really two-fold: Who controls access to the archive? And who is taking care of it?

Research libraries like MIT's are playing an active role in supporting archival arrangements that will ensure that, whatever happens to today's publishers, tomorrow's scientists will have a scientific record to work with.

SLIDE 28

At MIT, for example, the Libraries created DSpace, a widely-adopted open source system for storing and preserving copies of open access research created at MIT. Recently we've added new positions to support DSpace. One of them is our DSpace product manager, Sean Thomas, who's developing enhancements to DSpace that include easier ways for faculty to submit their work to DSpace.

SLIDE 29

MIT Libraries are also working on a project called "FAÇADE," for "Future-proofing Architectural Computer-Aided Design), to develop a preservation strategy for CAD and Building Information Models.

SLIDE 30

This is a two-year project and collaboration with Professor Bill Mitchell of the MIT Media Lab, and the architectural firm of Frank O. Gehry. The initial subject is the CAD files used in building the Stata Center.

SLIDE 31

3. Partnerships:

Realizing the promise outlined by John and other visionaries, and dealing with the issues of permissions and persistence, is going to take partnerships - among researchers, libraries, publishers, technologists, and research funding agencies.

I want to mention just two examples:

SLIDE 32

At CERN, the birthplace of the WWW, the head librarian of CERN, Jens Vigen, is putting together an extraordinary international consortium of research libraries.

He's proposing that they **pool** their collective subscription costs and use the money to pay a negotiated fee to all publishers of high energy physics literature. In exchange this literature would all be made open access to the world.

This project is called SCOAP/3, Sponsoring Consortium for Open Access Publishing in Particle Physics. (scoap3.org)

SLIDE 33

The other example I'll mention is Portico. Portico is a partnership of publishers and libraries that's creating a deep archive of the scholarly record published by Elsevier, Springer, Wiley, IEEE, and many other major publishers.

Libraries and publishers both pay Portico for what is essentially an insurance policy against unforeseen disaster. It's not an open archive, but it does offer some promise of long-term persistence for the digital record.

So those are some of the partnership strategies that are very promising.

SLIDE 34

I'd like to thank you now for your attention, and close with a quote from Harold Varmus, former head of the NIH, who has done as much as anybody to foster the progress of science through the promise of new scientific publishing models:

"The change will come when scientists understand that they are in control. The publishers need us more than we need them." – Harold Varmus (interview in Wired magazine, 2006)

SLIDE 38

Web sites