

Introduction to Science Commons



John Wilbanks
Executive Director, Science Commons
James Boyle
William Neal Reynolds Professor of Law,
Duke Law School

August 3, 2006



www.sciencecommons.org

Imagine a Brazilian postdoctoral student driven to cure malaria. She knew she would not be able to do her work in Brazil with the same impact she would have in the United States or Europe (she wouldn't have the resources, or the level of access to journals, tools, and collaborations) so she joined the legions of expatriate scientists in Boston. She is ridiculously talented, and very lucky. She gets a prestigious grant and finds a position at Harvard.

She is working on a protein called glycoporphin A. It's a key part of the way malaria infects blood cells. She checks the major literature repository and finds nearly 2000 papers with a glycoporphin A search. Her 50% overhead from her National Institutes of Health grant, combined with the grants of the other researchers there, is enough to pay for an elite library with subscriptions to all the journals. So at least she can read them. Yet behind her stand thousands of other scientists and potential scientists from around the world who cannot get access to this material and who thus are lost to her and to us as potential collaborators.

There are many problems other than access to scientific journals and research to be dealt with, some of them more fundamental. But even **after** the inequalities in access to basic and scientific education, and after eliminating research problems that require hugely expensive technical infrastructure, we still effectively "discard" minds we might need to solve problems because they do not have full access to the research texts they need. Given the rising cost of scientific publications and research services, this group is not confined to the developing world. It is a global problem.

Stay with our main character. If she reads all the papers at the rate of one a day, it will take her five years to process the relevant knowledge about her target, much less the dozens of related entities in the cell that are involved in malaria. And this is just the documents. There are

hundreds of databases to access, and thousands of data sets. The digital knowledge is simply overwhelming. This too is a global problem, one that is faced by commercial and academic researchers from every country. It is one that is actually exacerbated for the "information rich."


© Robert Cudmore; licensed to the public under Attribution-ShareAlike 2.0.

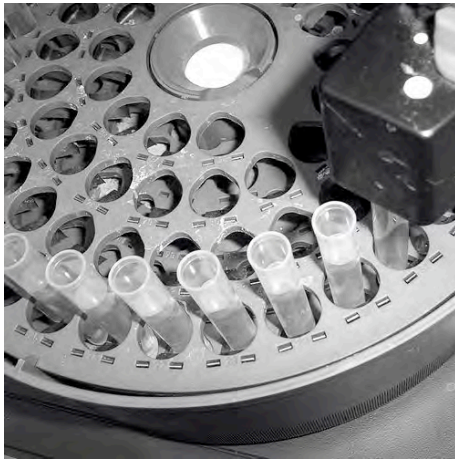
Of course, in theory, computers could help us mine the



wealth of data that computers have made available to us. Our researcher could use some advanced technology to help her. She could use software tools to extract the facts from the literature, to find new connections in the existing knowledge, to tie datasets and journals together and tag the information so that it could be found by others in the future. Unfortunately, the contracts that Harvard signed with the publishers often make that illegal, and digital rights management technologies enforce those contracts.

©3rd Coast Chick; licensed to the public under Attribution-NonCommercial-ShareAlike 2.0.

	Introduction to Science Commons	www.sciencecommons.org
	John Wilbanks and James Boyle	September 7, 2006



If she builds a collaboration with the inventor of the World Wide Web to try out his new "semantic web" technologies on the articles and data she needs, she puts Harvard at financial risk for breach of contract. [The semantic web is explained in more detail in the penultimate section of this paper.] And she's not allowed to email copies of key papers to her collaborators, either, so she can only really work with other scientists who have access to wealthy libraries. There are software companies that serve the pharmaceutical industry that might be able to help but their software costs \$100,000 a year, more than twice her salary.

So she uses the services available to her - free text search, Google, the free digital resources published by the United States National Institutes of Health, some biology driven desktop applications, Microsoft Office - and she narrows down to a few key papers. By necessity she has thrown away the vast majority of information that might be relevant but is separated by the accident of an inapposite or unlikely keyword, or a source in an apparently unrelated scientific process.

She reads in a paper published by a prestigious journal that glycoporphin A is a key mechanistic part of malaria. She needs to get some "research tools," actual physical stuff this time – cell lines with and without glycoporphin A - to verify the published results and start looking at potential ways to understand the mechanism in the context of glycoporphin A. Her grant covers this, to a certain extent.

So tools are available, but she needs the actual tools that were used in the paper she read to reproduce the result she is interested in. To get access to those is hard. She has to track down the contact information of the key authors, call the lab, discover that the tool actually came from the fourth author in another lab, call him, ask him to assign a student to create a supply of the cell line and mail it to

her at Harvard. All this takes time. And even after finding the fourth author, and finding him willing to share the materials, there are more hurdles.

Sending the cell lines from his institution requires the execution of a contract called a Materials Transfer Agreement. And everyone involved in science agrees that these can be a problem. Wendy Streitz and Alan Bennett, of the University of California Office of Research Administration and Technology, capture the problem eloquently from the scientist's perspective: "One of your colleagues at BigAg, Inc. (or at BigAg University) says that she'd be happy to send you her transposon insertion lines that saturate the right arm of chromosome 9; you'll just need to have a material transfer agreement (MTA) signed by your institution. Six months later, the terms of the agreement are still under negotiation, you've missed the field season, your grant has expired and there is now a better resource that's been developed at LittleAg University--and if you start negotiating an MTA now..." (2)

Of course there are other reasons things might not go well. The scientist with the cells simply won't share at all, perhaps out of fear of being scooped, perhaps out of competitive spirit, perhaps because it is a diversion of his laboratory's scarce resources to generate materials for another researcher. The Journal of the American Medical Association published a study in 2002 describing a world where 47% of academic geneticists had been rejected in their efforts to secure access to data or materials related to research by other academics.

This represented an increase from 34% who had been rejected in a previous study in the mid 1990s. There were multiple causes involved in this pattern but the leading one was the effort required to produce and transfer the materials, effort to which the MTA negotiation process frequently adds. So our scientist is not alone when she

finds it exceedingly hard to verify the results claimed in the paper.

She presses on. She spends her grant money on the commercial tools, or tools that are similar to the tools she is looking for, and is able to verify some elements of the research, though it is a second-best approach. She decides to invest her postdoctoral time on looking at potential mechanisms for malaria, based on this glycoporphin A work. One year in, the results are promising. Two years in she finds a paper, published years earlier, related to glycoporphin A's activity in a totally unrelated field - cancer. It was published in an obscure journal and it wasn't very well indexed at the time, so it would have been very hard to find. Even if it had turned up in her searches she would probably have ignored it in her attempt to narrow the field. But it contained a nugget of knowledge that would have saved her a full year of money and a full year of progress towards the end goal of curing disease. And it means that her key result has already been published, though not in the context she was exploring, which makes her paper much less likely to help her get tenure, or another grant.

Sometimes, of course, that nugget of information is necessary for the science to progress and, though it is out there in the archives, it is never found. Sometimes it never even gets into the literature, because the materials necessary to do the experiment cannot be acquired. Sometimes the experiment is not even attempted because scientists with talents and good ideas do not have practicable access to the literature. And this holds true whether the research is on a drug for a neglected disease, like malaria, for which the commercial market is in doubt and which will probably need alternative sources of funding, or research on a drug for a disease that has a thriving commercial market, such as diabetes or heart disease.

We do not know how many cases like that there are. We do not know how much fuller our faltering drug pipeline would be if at every stage of the process described, we had managed to lower even a few of the economic, legal and technical barriers to scientific journal research, data mining and linking, materials acquisition and testing. We do not know what would happen if we could eliminate some of the legal and technical barriers to building a "semantic web" for science.

Perhaps the result would be dramatic; some fairly impressive scientists and computer scientists believe so. Perhaps it would be more modest. But where it is practicable to do so, lowering those barriers is clearly a good idea. It might be a **great** idea. That is the idea behind Science Commons and it would be surprisingly cheap - by the standards of science funding - to make the idea a reality.

History of Science Commons

Creative Commons was formed to deal with a problem of access to materials caused by the conjunction of technological developments - computers' increasing capability to store and process data vastly enhanced in effect by interconnection via the World Wide Web--and legal change. Creative Commons enables creators to select among various copyright license options to make their work available to the public on generous terms. The licenses are designed so that they can be understood not merely by lawyers, but also by ordinary people and even by computers - the license terms are expressed in an easy to understand "commons deed" complete with icons, but also in "metadata" so that one can search not only for the content of the work, but also for its degree of legal openness. (Give me calculus textbooks that are available for non commercial use and modification, say.)

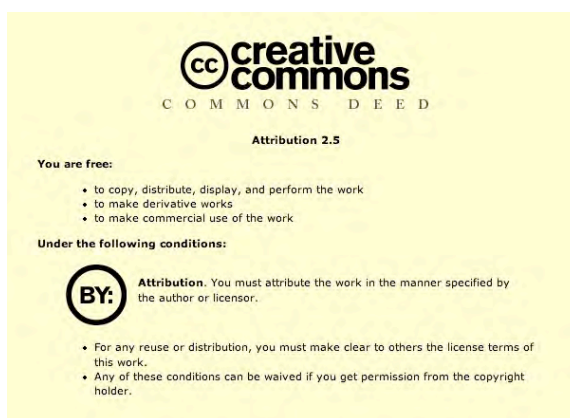
Creative Commons' charge initially was entirely in the cultural and copyright realms - in the world of music, texts, blogs, pictures, films and so on. Nevertheless, at the first board meeting, the founding board members

expressed strong interest in the possibilities of developing the creative commons model in the scientific area.

Several times, in fact, board members expressed the feeling that the Creative Commons approach might be more of a "killer app" in science than in culture.

Recognizing that developing open pathways for scientific research would be complex and contentious, the Creative Commons board did not feel that at that point we had the expertise or the technical capability to enter this field.

Creating an open regime of sharing and reuse in the sciences is a complicated proposition. Though copyrights guard the final published documents in peer reviewed journals, patents protect inventions (some more unique than others) and a web of handshakes and contracts guard the tools, materials, datasets, databases and informal knowledge transfer of day-to-day science. What works for a biologist will likely fail for a physicist, neither of



whose solutions will perfectly solve the legal problems of the anthropologist.

In some fields, especially the life and health sciences, the commercial opportunities are as immense as the risks - hundreds of millions of dollars in research and testing costs bet against the possibility that a drug will make it through clinical trials and perhaps become a billion dollar blockbuster. Thus, any sharing regime for science must

	Introduction to Science Commons	www.sciencecommons.org
	John Wilbanks and James Boyle	September 7, 2006

be flexible, adaptable, contemplate copyrights, patents, and contracts and more. From the beginning, it must be compatible with commercial innovation as well as the academy. Ill-conceived intervention that makes commercial development more difficult will hurt rather than help.

Adding to the complexity of the pure legal and policy work is the sheer size and variability of the stakeholders. Science requires universities, funders, companies, researchers, publishers, consumers, technicians, librarians and more. Each stakeholder represents an opportunity to inject control into the scientific process, especially as each one moves into the networked culture - and some of that control is beneficial or necessary. To find which barriers to sharing are unnecessary is a problem that demands both interdisciplinary and practical investigation. To remove the unnecessary barriers requires an ability to produce consensus among disparate parties, and even more, a large degree of humility: neither the problems nor their solutions might be predicted by reigning academic theory.

Creative Commons returned to science in early 2005 with the launch of Science Commons. Millions of creative works were already on the Web under Creative Commons licenses (the current count is 140,000,000 - ranging from music, films and political blogs, to textbooks and MIT's Open Courseware) and we had gained significant experience in open licensing approaches, complex negotiations, and community building. We had the ambition of achieving for the world of science and data, what Creative Commons had begun to achieve for the world of culture, art and educational material: to ease unnecessary legal and technical barriers to sharing, to promote innovation, to provide easy, high quality tools that let individuals and organizations specify the terms under which they wished to share their material. Scientific American seemed to like the idea.

olds to punish music pirating. In this environment, the introduction of Creative Commons's middle path of "some rights reserved" is surely a welcome arrival.

THE EDITORS editors@sciencemcommons.org

The first six months of Science Commons revolved around building the right set of people to run the project and conducting a broad survey of the various discipline-specific efforts in open science.

John Wilbanks, an entrepreneur and former bioinformatics CEO with experience at Harvard Law School's Berkman Center and the World Wide Web Consortium, came in to lead the effort as Executive Director.

We built an advisory board composed of two Nobel Laureates, Sir John Sulston and Joshua Lederberg, a Berkeley scientist and leading expert in "open access" publishing, Michael Eisen, the distinguished innovation economist Paul David, and the prominent intellectual property academic Arti Rai.

Four board members of Creative Commons join this group and act as a steering committee: James Boyle, from Duke, Mike Carroll an expert on intellectual property and scholarly publishing from Villanova, Hal Abelson, a renowned MIT computer scientist, and Eric Saltzman, a lawyer, filmmaker and former Director of Harvard's Berkman Center.

Science Commons hosted a series of private meetings covering research funding, drug patent licensing, biological materials transfer, and access to scholarly literature. Wilbanks made a tour of different communities: biology, chemistry, archaeology, geospatial, physics, geography and more.

We reached out widely and formed relationships with key players in discipline-specific efforts in agriculture, neuroscience, anthropology, information technology, and

	Introduction to Science Commons	www.sciencecommons.org
	John Wilbanks and James Boyle	September 7, 2006

more. We forged working relations with funders of research, universities, technology managers, software companies, standards organizations, and libraries. We were delighted by the reception that we received. From the beginning we were guided by a set of principles. Like Creative Commons, our proposals use coordinated private action, not public fiat, to lower barriers to research and sharing. This makes these proposals both much cheaper and faster to implement than solutions which require Congress or other regulators to act.

Wherever possible our solutions were based on both empirical and interview-based investigation of the problems. We tried to discard preconceptions; when we formed the organization, for example, we expected to spend more time on patent pooling. While we do not rule that out, we found Materials Transfer Agreements to be a more important area on which to focus initially. We tried to come up with projects where success was not an all-or-nothing proposition - selecting issues where **any** alleviation of the problems we identified was a good thing. We picked projects that played to our strengths and to the considerable experience that Creative Commons had acquired in negotiating standard form agreements

```
<rdf:RDF xmlns="http://web.resource.org/cc/"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  <Work rdf:about="http://example.org/gnomophone.mp3">
  <dc:title>Compilers in the Key of C</dc:title>
  <dc:description>A lovely classical work on compiling code.</dc:description>
  <dc:creator><Agent>
  <dc:title>Yo-Yo Dyne</dc:title>
  </Agent></dc:creator>
  <dc:rights><Agent>
  <dc:title>Gnomophone</dc:title>
  </Agent></dc:rights>
  <dc:date>1842</dc:date>
  <dc:format>audio/mpeg</dc:format>
  <dc:type rdf:resource="http://purl.org/dc/dcmitype/Sound" />
  <dc:source rdf:resource="http://example.net/gnomovision.mov" />
  <license rdf:resource="http://creativecommons.org/licenses/by-nc-nd/2.0/" />
  <license rdf:resource="http://www.eff.org/IP/Open_licenses/eff_oal.html" />
  </Work>
  <License rdf:about="http://creativecommons.org/licenses/by-nc-nd/2.0/">
  <permits rdf:resource="http://web.resource.org/cc/Reproduction" />
  <permits rdf:resource="http://web.resource.org/cc/Distribution" />
  <requires rdf:resource="http://web.resource.org/cc/Notice" />
  <requires rdf:resource="http://web.resource.org/cc/Attribution" />
  <prohibits rdf:resource="http://web.resource.org/cc/CommercialUse" />
  </License>
</rdf:RDF>
```

among disparate communities, merging legal and technical solutions, making deals comprehensible to non lawyers, and using metadata and the semantic web to produce "usable openness" and machine-readable contracts. Finally, we sought places where all sides agreed there was a problem and where many stakeholders would benefit from its removal.

Sample metadata from a Creative Commons license

Out of this research, we discovered a mix of legal, cultural, and technical controls - at least one of which bore down on the scientific process at each step, preventing the realization of the promise of new technologies like the Semantic Web. Some of these controls were necessary, of course, but many were not.

Some of the problems came from fractured contract regimes which created high transaction costs and confusion, preventing the emergence of smooth electronic transfer systems for knowledge and research materials. Other problems were technical: digital controls designed to prevent widespread copying of entire articles, which also prevented the extraction of key facts from papers for publication in new web languages. Some were based on legal mistake; overbroad claims of copyright in unoriginal databases, for example. Still others were a matter of institutional policy, practical difficulty or scientific culture. For example, commercial publishers can hardly be blamed when even those scientists who have the right to "self archive" their articles do not make their work freely available online. Sharing between laboratories is inhibited more by a complex mixture of transactional, practical and prestige obstacles, than it is by overbroad patents. And so on.

We realized that we could and should tackle each class of problem individually, but that our overall goal should be to bring the projects together so as to enable the true possibilities of open science in a networked world.

	Introduction to Science Commons	www.sciencecommons.org
	John Wilbanks and James Boyle	September 7, 2006

Proposals

In 2006, we began to act on our conclusions. We targeted three areas; scholarly publishing, licensing policies, and the realization of the "semantic web" for science. In each we have been running "proof of concept" projects and we now have early-stage efforts in scholar's copyrights, biological materials transfer, and the intersection of semantic web with Open Access content in neuroscience. Our projects are designed to intersect to yield evidence of the benefits of the overarching Science Commons vision of open, networked science, but also to stand on their own as worthwhile efforts in their own right.

Scholarly Communication

Scholarly communication in the sciences primarily involves three kinds of information:(1) data generated by experimental research,(2) peer-reviewed journal articles explaining and interpreting the data, and(3) metadata that describes or interprets articles or their underlying data.At each of these levels, the Internet and associated digital networks create a range of opportunities and challenges for changing the nature of what information is gathered, stored and communicated as well as how and when such information is shared, identified and located.

The Science Commons Publishing Project promotes effective use of digital networks to broaden access to all three types of information. Science publishing is obviously an area that has attracted a great deal of attention. There are many stakeholders already engaged in attempts to make scholarly publishing more open, and a variety of strongly - some might say "religiously" - held beliefs about which approaches work best. Science Commons approach has been extremely pragmatic and "non denominational." We have identified a series of places where opportunities were not being fully seized, where absence of collaboration was preventing

innovation, or simply where our specific expertise could add value.

i.) **Pragmatic Open Access Publishing:** Some publishers of peer reviewed science journals are employing a new, Open Access business model where the authors grant generous rights in their articles to the public under Creative Commons licenses. These licenses make clear to the public the broad range of uses they may make of the articles, without further permission or fee.

The goal of open access is to broaden the dissemination of knowledge about the natural world to researchers and other readers who can put this knowledge to use. But for this goal to succeed it is vital that readers easily grasp what rights they are granted under the license; a traditional Creative Commons concern. Publishers that have adopted this approach, and who are using our licenses to implement it, include the Public Library of Science, BioMed Central and Springer's OpenChoice program. (It is notable that this group includes commercial, non-profit and government-funded publishing efforts.)

ii.) **Enabling Self-Archiving:** It is increasingly common for scholarly authors to be given rights to "self archive" their work in institutional repositories. Some journals explicitly give these rights, while others are willing to give them only if asked. The rights vary as to the versions of the paper that may be posted and the timing of the post, leading to confusion among researchers. Worst of all, perhaps, even where the rights do clearly exist, they are used only infrequently, at least partly because of the perceived practical difficulties involved in the process. Self-archiving could be an incredibly valuable way of achieving freer access to scholarly materials. Science Commons has analyzed all the impediments to it and is working to minimize or remove them.

We have already developed "Author Addenda" - a range of short amendments, with varying degrees of openness, that authors can attach to the copyright transfer form agreements from publishing companies. The Addenda ensure, at a minimum, that scholarly authors retain enough rights to archive their work on the public Internet. We are spreading the word in the scholarly community and finding that there is considerable interest from institutions who have good reasons to want to ensure that their researchers retain enough rights to enable self-archiving.

In the Fall of 2006 and Spring of 2007 we plan to release

- A Web-based tool that will enable faculty authors to generate the Addendum of their choice with all form fields automatically filled in.
- Layperson-readable versions of the Addenda (similar to the Creative Commons "Commons Deed" copyright documents).
- Machine-readable versions of the Addenda to enable advanced software usage of the Addenda, database tracking, and empirical evidence gathering. This builds on our pre-existing metadata partnership with SPARC.
- We are also developing an application that will sit on the scientist's desktop and enable "drag and drop" self-archiving to an appropriate repository. The Internet Archive has agreed to host CC licensed material for free permanently, and numerous institutional repositories - using tools such as D-Space - are also available.

iii.) Facilitating the Use of Metadata:

Within its Publishing Project, Science Commons has

convened a working group comprised of publishers, librarians, and researchers to explore ways of better associating research articles with research data and for standardizing the metadata associated with both.

Licensing

Science Commons' Licensing Project aims to simplify licensing so as to speed science. We have been working on the creation of a "research commons" for neglected or orphan diseases (so that funders can simply specify that funded research must be available to all researchers in the field).

We have also been approached by one of the world's largest pharmaceutical companies with the idea of forming a "tox commons" that allows all researchers to pool toxicity data from failed commercial drug attempts, in a pre-competitive process of sharing. The idea is simple. While a successful drug application results in open data - the FDA requires publication and review - every failed drug results in secrets and obscurity. So a tempting target, tried and again and again, can mean repetition of failure. It's as if each company has just a few pieces of the treasure map, and each company beaches on a different set of rocks on the way.

Enter Science Commons.

Take a drug target for which compound after compound has failed in the clinic due to toxicity concerns, a graveyard of over \$10,000,000,000 in sunk costs and uncounted years of now-hidden research. Extract all the relevant facts about the target and its toxicity, its mechanisms, interactions, annotations, and more, from the literature and databases.

Attach annotations and data from the internal files of pharmaceutical companies who have tried, and failed, to get drugs to the market. Integrate the relevant descriptions of biological materials and public data sets.

Broker a set of contracts for access and recontribution to the data. Then let the scientists go after the combined knowledge, free of clickwraps and free to exploit \$10,000,000,000 of previously private, unintegrated, inaccessible, invaluable knowledge.

In this introduction, though, we will concentrate on one project that seems to exemplify the Science Commons approach - the attempt to streamline of the process of acquisition of research materials.

Biological Materials Transfer

Research materials are essential to the practice of modern life sciences' experimentation. Cell lines, model animals, DNA constructs, and screening assays each represent a tool for testing and validating hypotheses of biological function and human health. Each offers a perspective into biology that cannot be replicated without access to the material.

Research materials are developed in multiple environments: university laboratories, startup companies, biotechnology companies, hospitals, and non-profit research clinics. Some of the materials are patented; many are not. These tools are frequently licensed out to other institutions through material transfer agreements" (MTAs). Thousands of MTAs are signed each year in the biological sciences, covering such diverse materials as genes, proteins, chemicals, tissues, model animals, software, databases, "know-how" and reagents.

Although "standard" material transfer agreements exist (the Uniform Biological Material Transfer Agreement, or UBMTA, was developed in 1995) empirical research confirms that the licensing of materials remains a problem. A complex set of interlocking licenses covering dozens of different materials imposes significant transaction costs simply to gain the opportunity to begin research.

The long-term impact of this complexity is severe. University technology transfer offices can become clogged with requests that ought to be routine. Scientists must waste time trying to negotiate agreements. Commercial researchers find it hard to obtain materials. The end result benefits no one - we get less research, less innovation, less diffusion of knowledge.

Discussions with stakeholders reveal a number of recurring problems. Supposedly uniform agreements are actually "customized" in time-consuming negotiations, although all players would benefit if they could bind themselves to restrict choices to a more limited set of standard options. Even the "short form" version of agreements are perceived as too long and too complex. The agreements themselves are hard to interpret and scientists often find them mystifying, (or ignore them altogether as a result.) Finally, there is no connection between efforts to streamline the **legal** process for clearing materials, and efforts to streamline the **practical** process of actually fabricating and transferring the materials themselves.

It would be hard to find an area more perfectly suited to a Creative Commons-type solution. It is Creative Commons' **raison d'être** to analyze creative communities to find out which are the most common terms under which rightsholders are willing to make their works available, to generate licenses through a simple and intuitive radio button interface that allows a range of those choices to be expressed, (see Figure 1, page 5).

These licenses are expressed on three layers - lawyer readable contracts, human readable Commons Deeds, (see Figure 2, page 5) and machine readable meta data. (See Figure 1, page 6) This is exactly what is needed for a more rational Materials Transfer system, particularly if the process of building consensus around such a system can be led by a trustworthy third party - neither a funder,

nor a research unit, nor an academic institution nor a for profit company.

More ambitiously, in the relatively near future the material that is now covered by some Materials Transfer Agreements will be capable of being synthesized directly by DNA synthesizers. One could literally "print out" one's research material, or more likely order it from a third party specializing in such work. The cost at the moment is about \$2 a base pair, but it is dropping. As MIT Professor Drew Endy points out this could revolutionise the process of hypothesis formation, testing and experiment. (Science Commons has been working with Professor Endy on dealing with such issues in the emerging field of synthetic biology.)

Science Commons is exploring the implications this could have for the MTA process. The "blue sky" idea beyond streamlining of the MTA process (itself hugely valuable) is of a simple procedure by which a researcher reading of a development in the literature, could merely "click to get the cell line." Materials, or the information that allows them to be synthesized, would be automatically deposited with intermediaries or clearing houses, accompanied by metadata-expressed licenses that clearly expressed the uses to which those materials might be put. Clear licenses with, clear machine readable terms would allow quick, perhaps even automated, matching of institutions or activities with the restrictions on a license.

At the very least, simple licenses with iconic representations of their terms would allow researchers to know which materials were available. It is hard to overstate the advantages in streamlining that such a process could offer. And in some areas at least, one might be able to click right from the description in the literature of an experiment using a DNA sequence, to a cheap "print out" of that sequence ordered online from a low cost intermediary, applying the terms of the standard MTA. This sounds like science fiction, but some of the

experts with whom we have talked argue that for some materials it is scientifically practicable now. (MIT's repository of standard biological parts is an example.) The principal obstacles are not scientific, or a matter of computer science, or metadata expression. They are a matter of law, social engineering and institutional commitment.

Of course, the difficulties of procuring MTA's are not the only, not even the main reason that it can be hard to procure research materials. Competitiveness, secrecy, and the sheer hassle of producing and shipping the materials all play a part. The prevalence of these tendencies also varies from one scientific area to another. But the overall problem of obtaining materials is a huge one. The literature indicates it may be the single largest reason for the abandonment of promising lines of research.

Even if the legal transaction costs made up only 15% of the total impediments, it would be well worth reducing them. If, in doing so, we could make it easier to set up streamlined systems for obtaining "pre-cleared" materials from institutional and commercial repositories, the effect would be remarkable. And for some scientists it might actually affect the increasing tendency towards possessiveness and secrecy. Studies on social sharing networks indicate that one's willingness to help others is directly related to one's experience of receiving such help in the near past.

These beneficial spillover effects are possible but, the licensing effort does not depend for its success on the achievement of such ambitious goals. If we could simply streamline the MTA, or make it easier for Foundations working on orphan or neglected diseases to create a "research commons" for all such research, we would have achieved something extraordinarily important.

Beyond that tangible set of metrics for success, the licensing project does express a larger vision - one central

	Introduction to Science Commons	www.sciencecommons.org
	John Wilbanks and James Boyle	September 7, 2006

to Science Commons. The point is that **the social and legal engineering of science has largely lagged behind the technical engineering and investigation that it seeks to facilitate.**

Science Commons' licensing project attempts to use some of the developments in computers and metadata, together with more traditional legal and consensus building skills, to make the process of legal clearance and practical availability move at a pace closer to that of science itself.

Data

Introduction to the Semantic Web

In the course of this paper, we have several times used the term "semantic web." The phrase may be unfamiliar to some, but the idea behind it is quite simple. "The Semantic Web is about two things. It is about common formats for interchange of data, where on the original Web we only had interchange of documents. Also it is about language for recording how the data relates to real world objects. That allows a person, or a machine, to start off in one database, and then move through an unending set of databases which are connected not by wires but by being about the same thing.

A different way to put it is that right now we mainly use network searches that look for **words** - say the word "glycophorin." But of course, such a search would pull up this paper (of little use to our Brazilian researcher) as well as a paper that was actually talking about the biochemical process she wished to investigate.


The semantic web allows searches by function, or meaning. "Show me all the statements in the literature which deal with X interaction between glycophorin A and the malaria disease process." This is accomplished by "tagging" information with metadata. One simple example that is familiar from another context, is to tag a bibliographic record with an "author," "title" and "date"

field. When I search for Bronte within the author field, my time is not wasted with articles or books **about** the Bronte's, nor with maps of a place called Bronte. But metadata tagging can do much more than this. The semantic web holds extraordinary promise for science.

At its most ambitious, it would allow seamless integration between scholarly articles, the data those articles refer to, and to cross references with other articles dealing with similar processes in different areas of science. But the process of mining, linking, tagging and cross-referencing that the semantic web requires faces extraordinary difficulties. Some of those difficulties are financial. Tagging takes time and costs money. Some of the difficulties involve the coordination of standards and formats for metadata, something that Creative Commons has considerable experience in.

Perhaps the single greatest obstacle to the semantic web, however, is that the process of integration it requires is now impeded by multiple barriers. The journal article is copyrighted, and sits behind a digital fence. The data to which the article refers cannot be integrated because it too, is protected by licensing agreements, assertions of copyright (some of them unfounded), and technical controls. These legal and technical restrictions may be aimed at preventing very different activities than those necessary for the semantic web. (Stopping wholesale copying and transmission of the text of journal articles, say.) But their negative effect is real.

To solve these problems, one needs an organization with considerable experience in law, publishing, computer architecture and metadata. And those of course, are the central focii of Creative Commons, and of the people who run Science Commons. (John Wilbanks actually came to us from the World Wide Web Consortium (W3C) initiative on the semantic web for science, and MIT computer science professor Hal Abelson serves on

	Introduction to Science Commons	www.sciencecommons.org
	John Wilbanks and James Boyle	September 7, 2006

the Creative Commons board and Science Commons steering committee.)

Science Commons is pursuing a number of projects aimed at enabling the semantic web for science. The most fully developed at the moment is the Neurocommons.

Neurocommons

The Neurocommons project, a collaboration between Science Commons and the Teranode Corporation, is building on Open Access scientific knowledge to create a Semantic web for neurological research. The project has three distinct goals.

- To demonstrate that scientific impact is significantly related to the freedom to reuse and technically transform scientific information without violating the law. In short, that a large degree of Open Access is an essential foundation for innovation.
- To establish a framework that increases the impact of investment in neurological research in a public and clearly measurable manner.
- To develop an open community of neuroscientists, funders of neurological research, technologists, physicians, and patients to extend the Neurocommons work in an open, collaborative, distributed manner.

Today's life scientist faces a dizzying array of knowledge sources. Peer reviewed journal articles, online repositories of sequences and pathways, robot-driven data collection, all must be integrated into experimental design and analysis. Many scientists spend as much time on Google and PubMed as they do at the bench; the difference between success and failure in the lab or clinic can be the judicious and timely utilization of information. But this is all local knowledge utilization. As we pointed

out earlier, the logarithmic explosion of information in science overwhelms any one individual's ability to store and model all the relevant science in her head.

The result is a "scalability problem" in life sciences: while methods for generating information have gone digital, methods for using that information remain stolidly analog. Technology can help. Bandwidth, processing and storage are cheap. Machines can transmute from a string such as "aaattcaggagattacagta" to a physical molecule of DNA - and back again, making genetic information truly fungible, something that can be shared via the Web. Advances in language processing and ontology development allow for the construction of machine-readable and interpretable representations of scientific information. Logic and reasoning engines can crawl across massive data sets and come back with suggestions on causation.

As we suggested in our introduction to the semantic web it is neither cheap nor easy to seize the moment and use technological advances to solve these human and scientific problems. Legal and economic factors have to date muted the impact of new technologies on the life sciences: copyrights and contracts intertwine with software-enforced restrictions on reusing and republishing knowledge in a more usable format.

The Neurocommons Project rests on the hypothesis that there is enough information on the Web, in the form of taxpayer-funded databases and openly licensed scientific literature, to demonstrate the utility of a legally open, technically standardized approach to knowledge. In doing so, we wish to sow the seeds of a massive change in how scientific knowledge is licensed and reused.

The life sciences represent an ideal test case for the semantic web. Semantic Web technologies make the most sense where there is a certain set of conditions.

1. A massive amount of data: We certainly have that: Clinical images, robot-arrayed "gene chips", machines that can sort materials cell-by-cell, gene sequencers and massively high throughput chemical screens. There are hundreds of public databases, from flies to humans to plants, each potentially able to inform a decision or experimental design.
2. Rapidly changing knowledge: Every journal article, every paper, every experiment in the lab creates new knowledge about our bodies and the world we live in. This makes it very hard to apply traditional computational approaches or even integrate the data. We know what goes into a car - engine, tires, wheels, axles, fenders - and thus we can create a fairly fixed representation of a car for a computer, for model building and more. But we don't have anything resembling consensus to items as fundamental as "what is the role of the non-coding DNA in the human genome?"
3. Distributed knowledge and expertise: the nature of modern life science is specialization. One scientist is an expert on the genetics of Huntington's Disease (a rare neurodegenerative disease) another an expert on the impact of protein folding on Alzheimer's Disease. The two both work on the brain, on many of the same genes and proteins. But they attend different conferences and are pressed for time to study the refereed literature outside their own disease. Possible synchronicities between the researchers are at a minimum because their knowledge can't interoperate without distracting them from the lab.

For problems which have these features, the semantic web is a natural fit. Like the Web itself, the semantic web is intended to "scale" - to be capable of dramatic expansion in size, mission, and reach - through a process of decentralization rather than centralization, and an emphasis on information reuse, not recreation. It is a means to capture and network the relationships implicit in high volume data sets, or the outputs of sophisticated analytic software. It can relate anything to anything, as long as that anything has a unique name. Data-driven relationships can attach to the descriptions of related genes and proteins, and to the knowledge about those genes and proteins as described in the scientific literature.

The semantic web does not require that the picture be complete. If the relationships between one gene and another change as our knowledge changes, the technical burden is no lower than adding another hyperlink between web pages. And the concept of integration around unique names makes it easy to create serendipity between researchers: instead of bumping into a colleague in the hall at the right time, a scientist can see the ecosystem of knowledge around a particular gene expression in the brain. Whether that knowledge comes from her work on Alzheimer's or a distant colleague's work on Huntington's makes no difference. It all gets published to the semantic web.

The legal and economic problem

If these potential advantages are real, why do we not already have a vast semantic web for life sciences? The technological and standards problems are being solved. The National Institutes of Health has invested in the national centers for biomedical ontologies, language processing technologies are evolving in leaps and bounds, and public databases are investing in machine readability and open licensing. The problem is simpler. Despite what

appears to be an information overload, the sparse availability of truly **machine-readable** scientific knowledge has prevented robust testing of the Semantic Web. The barriers we described earlier - legal, technical and digital - have prevented easy aggregation of data, and thus have denied us the ability to test rigorously whether this approach will indeed be as productive as it promises to be.

The Strategy

Rather than complain about the problem that machine readable knowledge is sparse, the Neurocommons Project is taking as its focus those areas where we do have truly open access information and thus **can build a test case**. We have formed an initial community of neuroinformaticists, practicing neuroscientists, Semantic Web experts and language experts to ensure our work is accurate and scientifically valid. The first stage is underway:

- Using automated technologies, we are extracting machine-readable representations of neuroscience-related knowledge as contained in full-text Open Access literature, free text such as the PubMed abstracts, and legally open databases
- We then assemble those representations into a semantic web for neuroscience publish the resulting "graph" freely and
- Assemble a standard software implementation to store, update, and manage the changes to the graph as knowledge evolves.

Plans for stage two of the project involve the deployment of additional software infrastructure, the development of operational manuals so that interested parties can "port" the entire Neurocommons approach into new scientific domains without involving Science Commons, new

publishing techniques to automatically add knowledge to the Neurocommons graph, and active community development.

Again, Science Commons has sought partners who can provide credibility and competence. Teranode, a for profit company, provides direct financial support to the Neurocommons project as well as in-kind donations of software and services.

Jonathan Rees, formerly in charge of the curated protein-protein interactions database at Millenium Pharmaceuticals and a veteran of MIT's project MAC, leads the project on a day to day basis as a Science Commons Fellow.


The project is deeply involved with the World Wide Web Consortium's Health Care and Life Sciences Interest Group as well as MIT's Computer Science and Artificial Intelligence Laboratory (which hosts Science Commons).

Conclusion

We have tried to pick projects where "even if you fail a bit, you still succeed." In our most heady and optimistic moments, we imagine a very different landscape for science. No longer would the price of access to scientific literature act as such an impediment to research. Our Brazilian researcher would be joined by potential collaborators from poorer countries and institutions, and could share information freely with them.

Whether the material was obtained from an invigorated practice of self-archiving, from commercial or non profit open access journals, or from journals which make their work openly available after a fixed period of time, the world of scholarly literature would be more open - and more open on standard and interoperable terms, easily understood by all participants.

What's more, literature searches would be transformed, as the technology of the semantic web cut through the

	Introduction to Science Commons	www.sciencecommons.org
	John Wilbanks and James Boyle	September 7, 2006

information glut that overwhelms scientists, allowing the kind of cross-discipline, and cross-disease insights we can only imagine right now. When hypotheses were formed, the researcher would be able to click to obtain research tools and materials, according to truly standard, machine **and** human readable Materials Transfer Agreements. In many cases, those materials could be obtained automatically and at low cost from depository institutions.

The results of this research, in turn, would be fed back into the web of scientific knowledge. The universe of science would be enlarged, participation rendered more egalitarian, commercial exploration of drug targets easier, the drug pipeline fuller and so on.

Scientists may be justified in retaining privileged access to data that they have invested heavily in collecting, pending publication — but there are also huge amounts of data that do not need to be kept behind walls. And few organizations seem to be aware that by making their data available under a Creative Commons licence (see <http://creativecommons.org/license>), they can stipulate both rights and credits for the reuse of data, while allowing its uninterrupted access by machines.

Nature editorial, November 2005

That is the utopian vision, and we genuinely believe it has real chance of succeeding, at least in part. But assume that we fall short of such lofty aspirations, what happens? At worst, scientific publications are made more accessible, on more easily comprehensible terms, more researchers self-archive, and finding the result of that self-archived material is easy. We form examples of Semantic Webs for science and test their validity. We help researchers on orphan and neglected diseases build research commons, and help companies to pool their knowledge of discarded drug candidates. We simplify the process of Materials Transfer, and cut down on the truly crushing burden that it imposes on all participants. And, in the process, we learn from our mistakes and come back with a better plan to try again.

A Final Note on Funding

Science Commons has achieved a remarkable amount despite the relatively modest amount of start-up funding it has received. In part that is because it has been able to draw on Creative Commons' resources and on **massive** amounts of highly skilled volunteer labor from its Board, Advisory Board and pro bono lawyers. In part it is because the problems are simply ripe for solution and we have found many partners willing to work with us and leverage our efforts. But now the first stage of our plan is almost complete and we need significant new funding to realize the promise of the projects we describe here, all of which are already under way.

In each of these three areas we believe we have a high probability of success. The communities around the projects are the best indicator. The major stakeholders agree there is a problem and that we are a good vehicle for discussing - and, we hope, creating - the solution. Even in publishing, where the tension between corporate publishers, universities and open access advocates can break through the surface, we maintain strong relations with major publishers such as Nature and Springer. Indeed, Springer even uses Creative Commons licenses as part of the Open Choice alternative for authors. Also, as you can tell from this document, we are particularly excited about the prospects for the licensing and data areas.

In the short term, (4-8 months) we need \$500,000 to build the staffing and institutional framework necessary to continue the projects described here, to fund the meetings and conferences that will attract new partners, and to pay for the expensive legal advice that will be necessary even beyond the generous pro bono contributions we already receive.

In the longer term, we estimate that we need \$5 million to \$6 million over the 3 years beyond that to bring all of these projects to fruition, and to build on the benign feedback they will generate for each other. (We would be happy to provide a more detailed budget, of course, explaining precisely where the money would go.)

The amount of work necessary is staggering, but the goals are concrete, achievable and worthwhile.

We have tried in this document to describe accessibly and for a general audience our goals, techniques, strategies and institutional resources. We hope you found it of interest and would be delighted to outline our projects more precisely, and rigorously, as well as to go into details about the projects not covered here. We ask for your feedback and, in particular, your suggestions as to possible funding sources.

1. John Wilbanks is Executive Director, Science Commons.

James Boyle is William Neal Reynolds Professor of Law at Duke Law School and faculty co-director, the Center for the Study of the Public Domain.

This is a draft introduction aimed at readers with a wide range of backgrounds prepared for the Science Commons Funders Meeting, Duke Law School Aug 3rd , 2006.

Please do not circulate without permission.

The projects described in these pages were made possible by seed funding from the HighQ Foundation and Creative Commons. Science Commons receives additional funding from the Omidyar Network and the Teranode Corporation; Edwards Angell Palmer & Dodge LLP provides pro bono legal services.

2. Material Transfer Agreements: A University Perspective

Plant Physiology 133:10-13 (2003)

<http://www.plantphysiol.org/cgi/content/full/133/1/10>

3. Campbell et al,

Data Withholding in Academic Genetics, JAMA Vol. 287 No. 4, January 23, 2002.

4. W3C statement on Semantic Web activity.

<http://www.w3.org/2001/sw/>