CHAPTER TEN

CYBERINFRASTRUCTURE FOR KNOWLEDGE SHARING

John Wilbanks[1]

> Infrastructure never gets adequately funded because it cuts across disciplinary boundaries, it doesn't benefit particular groups. Infrastructure is a prerequisite to great leaps forward and is thus never captured within disciplinary funding, or normal governmental operations. We need to revise radically our conception of cyberinfrastructure. It isn't just a set of tubes through which bytes flow, it is a set of structures that network different areas of knowledge … and that is software and social engineering, not fiber optic cable. The superhighways of the biological information age should not be understood as simply physical data roads, long ropes of fiber and glass. They need to be structures of knowledge. The Eisenhower Freeways of Biological Knowledge are yet to be built. But that doesn't mean the task isn't worth starting.
>
> – James Boyle, William Neal Reynolds Professor of Law, Duke University Law School

KNOWLEDGE SHARING AND SCHOLARLY PROGRESS

Knowledge sharing is at the root of scholarship and science. A hypothesis is formulated, research performed, experimental materials designed or acquired, tests run, data obtained and analysed, and finally a publication. The scholar writes a document outlining the work for dissemination in a scholarly journal.

[1] Executive Director of Science Commons. This chapter was first published as an article in (2007) 3 (3) *CTWatch Quarterly* <http://www.ctwatch.org/quarterly/articles/2007/08/cyberinfrastructure-for-knowledge-sharing/>.

If it passes the litmus test of peer review, the research enters the canon of the discipline. Over time, it may become a classic with hundreds of citations. Or, more likely, it will join the vast majority of research, with less than two citations over its lifetime, its asserted contributions to the canon increasingly difficult to find – because, in our current world, citations are the best measure of relevance-based search available.

But no matter the fate of an individual publication, the system of publishing is a system of sharing knowledge. We publish as scholars and scientists to share our discoveries with the world (and, of course, to be credited with those discoveries through additional research funding, tenure, and more). And this system has served science extraordinarily well over the more than three hundred years since scholarly journals were birthed in France and England.

THE INFORMATION TECHNOLOGY REVOLUTION: MISSED CONNECTIONS AND LOST OPPORTUNITIES

Into this old and venerable system has come the earthquake of modern information and communication technologies. The Internet and the Web have made publication cheap and sharing easy – from a technical perspective. The cost of moving, copying, forwarding, and storing the bits in a single scientific publication approach zero.

These technologies have created both enormous efficiency gains in traditional industries (think about how Wal-Mart uses the network to optimise its supply chains) and radical reformulation of industry (Amazon.com in books, or iTunes in music). Yet the promise of enormous increases in efficiency and radical reformulations have to date failed to make similar shattering changes to the rate of meaningful discovery in many scientific disciplines.

For the purposes of this article, I focus on the life sciences in particular. The problems I articulate affect all the scientific disciplines to one extent or another – but the life sciences represent an ideal discussion case. The life sciences are endlessly complex and the problems of global health and pharmaceutical productivity such an enormous burden that the pain of a missed connection is personal. Climate change represents a problem of similar complexity and import to the world, and this article should be contemplated as bearing on research there as well, but my topic is in the application of cyberinfrastructure to the life sciences, and there I'll try to remain.

Despite new technology after new technology, the cost of discovering a drug keeps increasing, and the return on investment in life sciences (as measured by new drugs hitting the market for new diseases) keeps dropping. While the Web and email pervade pharmaceutical companies, the elusive goal remains 'knowledge management': finding some way to bring sanity to the sprawling mass of figures, emails, data sets, databases, slide shows, spreadsheets, and sequences that underpin advanced life sciences research. Bioinformatics, combinatorial drug discovery, systems biology, and an innumerable number of words ending with '-omics' have yet to relieve the skyrocketing costs and increase the percentage of success in clinical trials for new drug compounds.

The reasons for this are many. First and foremost, drug discovery is hard – really, really hard. And much of the low-hanging fruit has been picked. There are other reasons having to do with regulatory requirements, scientific competition, distortions in funding, and more. But there is one reason that stands out as both a significant drag on discovery and as a *treatable* problem, one that actually can be solved in the short term: we aren't sharing knowledge as efficiently as we could be.

FORGET 'WEB 2.0' – WHAT ABOUT 'WEB 1.0' FOR SCIENCE?

Much of the functionality we take for granted on the Web comes from making the choice to make sharing information easier, not harder. A good example is the way that Google interacts with the scientific literature.

With few exceptions, we rank the importance and relevance of scientific articles the way we always have, with citations and 'impact factors'. Citations are longstanding and important. Impact factors – the number of citations to the articles in a journal – are the dominant metric for journal quality. And for a long time, citations were clearly the best, and perhaps the only, statistical measure of quality of a journal. In a print world, a world without hyperlinks and search engines and blogs and collaborative filtering, citations are a beacon of relevance.

But we live in a different world now. We have the ability to make connection after connection between documents, to traverse easily from one page to another page. Hyperlinks are cheap and they're everywhere. It was a conscious design decision made by Tim Berners-Lee to allow this functionality. Other competing systems thought it insane that the WWW would let just anyone link to just anything else – those links might be broken, leading to the dreaded '404 not found' – and that would obviously kill the WWW! It hasn't worked out that way. The choice to allow users the right to make hyperlinks, to make hyperlinking easy and fast, not only did not kill the Web, it is a big part of what makes Google searching so powerful.

Google ranks pages by downloading enormous chunks of the Web and running software that analyses the linkages between Web pages. The system quite literally depends on there being lots and lots of links, many of them perhaps useless on their own, but which in aggregate provide hints of relevance. Thus, the number one Google search on the words 'Science Commons' is the Web page analysed with the words 'Science Commons' that has the most links pointing to it. There's more complexity, obviously, but that's a big part of the idea.

If those Web pages were private, the page ranking system wouldn't work. The Web pages themselves are part of the infrastructure on which Google operates, on which millions of start-up dreams are founded. In a world where every page was locked, where every Web designer had to ask permission to make an inbound link … we wouldn't have the sprawling value creation we associate with the Internet. It would look a lot more like Prodigy looked a long time ago: a closed network that can't compete in the end with the open networks.

Put another way, we have far more efficiency brought to bear on accelerating our capability to order consumer products than we do on accelerating our capability to perform scientific research. Biological reagents and assays are re-invented and reverse-engineered by readers of 'papers' – years of laboratory work, data, living DNA and more compressed down to the digital equivalent of a sheet of dead tree.

We need the Web to work as well for science as it does for other areas. The capabilities now exist to integrate information, data, physical tools, order fulfilment, overnight shipping, online billing, one-to-one orders, and more. If we are to solve the persistent health problems of the world, of infectious disease in the developing world and rare disease in the developed world, the 'Web 1.0' efficiency is an obvious benefit to bring to the life sciences.

But these advances we take for granted in daily life, like Google's relevance based search of the entire Web, eBay's many-to-many listing and fulfilment, Amazon's one-click ordering, won't come to science accidentally. There's a significant collective action problem blocking the adoption of these systems and preventing the network effects from taking over in discovery.

But it's not just the Google issue, which simply forces us to forego existing technology and focus on citations as we have always done. Citations carry more constrictions as a search metaphor. You are likely to enter the citation search ranked world when you know what to search. But you might not know what you're looking for. You might not know how to say it in the nomenclature of a related, but distinct, discipline.

It goes on. Citation linkages between papers are subject to enormous social pressures. One cites the papers of one's bosses, of course. Review articles can skew impact factors. And of course, a tried-and-true way to get a heavily cited article remains to be horrifically, memorably wrong.

And over the long term, the lack of more complex and realistic interconnections between articles – a web, a set of highways, an infrastructure connecting the knowledge – is that we can't begin to integrate the articles with the databases. That's because the actors in the articles (the genes, proteins, cells and diseases) are described in hundreds of databases.

And if we could link the articles not just to each other by a richer method than citations, but to the databases, we can inch closer to the goal of a Rosetta Stone of knowledge, the small element upon which we can begin to have truly integrated, public knowledge spaces. That would in turn allow us to begin automatically indexing the data that robots are producing in labs every day, to meaningfully extract actionable information from the terabytes of genomic data we are capable of producing.

You get those virtues only where you are dealing with the knowledge claims themselves, not the sub-component of them the people in the field thought it worthy to expose. Only a better infrastructure gets you there, just as the modern highway system in the US allowed for better efficiency than the evolved hodge-podge of state highways. Citation linkages are very useful (and a later version needs to cross reference them with these highways we propose – we didn't throw away the state highways, after all!). This is simply a different set of tasks, and one that can be accomplished if enough smart people have enough rights and time to work on the knowledge.

But sadly, no one – no one! – has the right to download and index with scholarly literature without burning years of time and money in negotiations. Google has spent years asking for the right to index a lot of the scholarly canon for its Scholar project, but that's not some open land trust for any researcher to work on. It's just for Google. And the fact that Google alone has the right to index articles for such a service means that the *next* Google, the next set of genius entrepreneurs with a taste for search coding away in the halls of the local university, can't apply their skills to the sciences.

Though we have the capability to drastically increase the sharing at a much lower cost through digital distribution, search, and more, the reaction has been instead to segregate knowledge behind walls of cost, technology, and competitive secrecy. The net result is that we're doing things the way we always did, but only somewhat faster. If we want to bring both efficiency gains and radical transformation to the life sciences, getting more knowledge online, with the rights to transform, twist, tag, reformat, translate, and more, is going to be part of the solution. We have to start allowing the best minds of the world to apply the newest technologies to the scientific problems facing us.

There isn't a single, open 'Web' of content to search – it's owned by a group of publishers who prevent indexing and search outside their own engines, and who use copyright and contracts to keep it locked up. There isn't any easy way to find the tools of biological science – it's a complicated social system of call-and-response, of email and phone calls, of 'are you in the club of scientists worth partnering with?' questions and answers. And there isn't a standard way to get your orders fulfilled, but instead a system in itself of materials transfer and ordering, university technology transfer, commercial incentives, deliberate withholding, and more. We don't have the Web working yet for science.

INFRASTRUCTURE FOR KNOWLEDGE SHARING: SCIENCE COMMONS

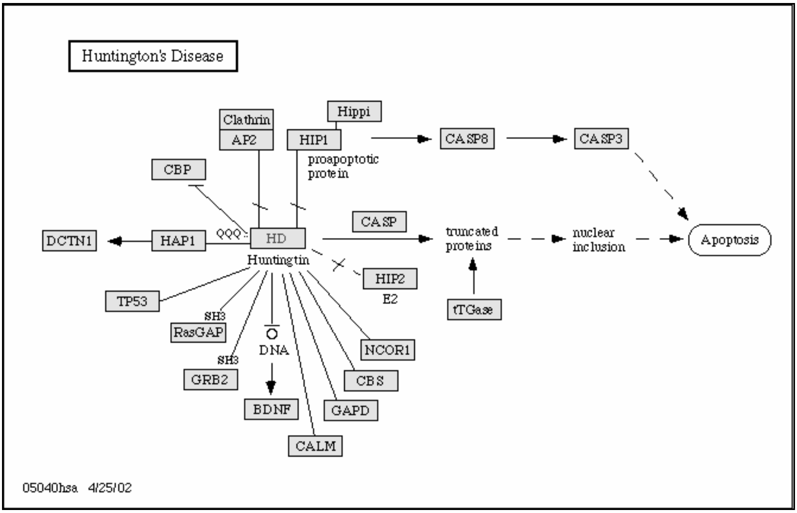
I work on a project called Science Commons – part of the Creative Commons (CC) non-profit organisation (CC is the creator of a set of legal tools for sharing copyrighted works on the Web using a modular set of machine-readable contracts. CC licenses cover more the 150 million copyrighted objects on the Web, including such high-impact offerings as BioMed Central, Public Library of Science, Nature Precedings, Hindawi Publishing, and the UniProt database of proteins. Science Commons is building a toolkit of policy, contracts, and technology that increases the chance of meaningful discovery through a shared, open approach to scientific research. We're building part of the infrastructure for knowledge sharing, and we're also deploying some test cases to demonstrate the power of investing in this kind of infrastructure.

Science Commons isn't alone. Sharing approaches that address a single piece of the research cycle are making real, but painfully slow, progress. Open Access journals are far from the standard. Biological research materials are still hard to find and harder to access. And while most data remains behind the firewall at laboratories, even those data sets that do make it online are frequently poorly annotated and hard to use. The existing approaches are not creating the radical acceleration of scientific advancement that is made possible by the technical infrastructure to generate and share information.

Science Commons represents an integrated approach – one with potential to create this radical acceleration. We are targeting three key blocking points in the scientific research cycle – access to the literature, experimental materials, and data sharing – in a unified approach. We are testing the hypothesis that the solutions to one problem represent inputs to the next problem, and that a holistic approach to the problems discussed here potentially benefits from network effects and can create disruptive change in the throughput of scientific research. I will outline how these approaches represent tentative steps towards open knowledge infrastructure in the field of neuroscience.

KNOWLEDGE OVERLOAD FOR HUNTINGTON'S DISEASE

Figure 1. Biological pathway for Huntington's Disease.



Source: Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y., "KEGG for linking genomes to life and the environment." *Nucleic Acids. Res.* 36, D480–D484 (2008).

Above is the biological pathway for Huntington's Disease. This pathway is like a circuit – it governs the movement of information between genes and proteins, processes and locations in the cell. This one is a relatively simple pathway, as far as such things go. More complex pathways can have hundreds of elements in the network, each 'directional' - not just linked like Web pages, but typed and directed links, where the kind of relationship and the causal order are vital both in vitro and in silico.

In this pathway, the problem is the HD gene in the middle of the circuit - if that gene is broken, it leads to a cascade that causes a rare, fatal disease where the brain degenerates rapidly. Although the genetic element has been understood for a long time, there is no cure. Not enough people get the disease for it to be financially worth finding a cure, given how expensive it is to find drugs and get them to market.

That's cold comfort to the tens of thousands of people who succumb each year and to their families who know they have a 50% chance of passing on the gene and disease to their children. But that's the reality.

Years of research have led to an enormous amount of knowledge about Huntington's. For example, a search in the U.S. government's free Entrez web resource on 'Huntington's' yields more than 6000 papers, 450+ gene sequences, 200+ protein sequences, and 55 000 expression and molecular abundance profiles. That's a lot of knowledge. The papers alone would take 17 years to read, at the rate of one paper per day (and that's assuming no new papers are published in the intervening years). Yet Huntington's actually provides a relatively small result. One of the actors in the pathway is called 'TP53'. That brings up another 2500 papers, but also brings up (in an indirect link to a page about sequences for this entity) that it has a synonym: 'p53'. Entrez brings back 42 000 articles from that search string – 115 years to read!

It goes on and on. And having all of this knowledge is wonderful. But there are more than a few problems here. The first is something you might call 'cognitive overload'. Our brains simply aren't strong enough to take in 500 000 papers, read them all, build a mental model of the information, and then use that information to make decisions - decisions like, what happens if I knock out that CASP box in the pathway, with 27 000 papers?

The other problems stem from the complexity of the body. In what other circuits is each entity in the pathway involved? What about those tricky causal relationships above and below it in the circuit? What are the implications of intervention in this circuit on the other circuits?

Some of these entities, the boxes in the diagram, are metaphorically similar to the airport in Knoxville, TN. Knocking out that airport doesn't foul up a lot of air traffic. But some of these - P53 for example - are more like Chicago. Interfering with that piece of the network reverberates across a lot of unrelated pieces of the network. That's what we call side effects, and it's one of the reasons drugs are so expensive - we know that we can impact this circuit, but we don't realise how badly it affects everything else until we run the drug in the only model available that covers all possible impacts: the human body.

And this is just the papers. There are thousands of databases with valuable information in them. Each of them has different access privilege conditions, different formats, different languages, and different goals; wasn't designed to work with anything else; and is maintained at different levels of quality. But they have vital - or potentially vital, to the right person asking the right question - information. And if we could connect the knowledge around these knowledge sources into a single network we just might be able to leverage the power of other technologies built for other networks. (Like Google – but maybe more like the next Google, something as dramatically better and different and radical as Google was when we first saw it in the late 1990s.)

There are two problems to be addressed here. One is the materials that underpin this knowledge, these databases and articles. Those materials are 'dark' to the Web, invisible, and not subject to the efficiency gains we take for granted in the consumer world. The second is the massive knowledge overload that the average scientist faces. I'll outline two proofs of concept to demonstrate the value of investment in infrastructure for knowledge sharing that can address these problems.

PROOF OF CONCEPT: E-COMMERCE FOR BIOLOGICAL MATERIALS

The Biological Materials Transfer Agreement Project (MTA) develops and deploys standard, modular contracts to lower the costs of transferring physical biological materials such as DNA, cell lines, model animals, antibodies and more. Materials represent tacit knowledge – generating a DNA plasmid or an antibody can take months or years, and replicating the work is rarely feasible. Gaining access to those materials is subject to secrecy, competition, lack of resources to manufacture materials, lack of time, legal transaction costs and delays, and more.

There is significant evidence that the transfer of biological materials is subject to significant slowdowns. Campbell[2] and Cohen[3] have each demonstrated that materials are frequently denied. Legal barriers are part of the problem – more so than patents – but the greater problem is frequently the competition, secrecy, and incentive systems involved.

[2] See Eric Campbell and David Blumenthal, 'The Selfish Gene: Data Sharing and Withholding in Academic Genetics', *Science*, 31 May 2002

[3] See Wesley Cohen et al, *Where Excludability Matters: Material v. Intellectual Property in Academic Biomedical Research* <http://siepr.stanford.edu/programs/SST_ Seminars/walsh.pdf>, which illustrates the benefits of self-archiving.

This is why we brought in funders of disease research and institutional hosts of research from the beginning – this is the part of infrastructure that is social engineering, not software. The secrecy and competition do not maximise the likelihood of meaningful discovery coming from limited funding, and thus funders (especially of rare or orphan diseases) have a particular incentive to maximise the easy movement of biological materials to maximise follow-on research.

The MTA project covers transfers among non-profit institutions as well as between non-profit and for-profit institutions. It integrates existing standard agreements into a Web-deployed suite alongside new Science Commons contracts and allows for the emergence of a transaction system along the lines of Amazon or eBay by using the contracts as a tagging and discovery mechanism for materials.

This metadata driven approach is based on the success of the Creative Commons licensing integration into search engines and further allows for the integration of materials licensing directly into the research literature and databases so that scientists can 'one-click' inline as they perform typical research. And like Creative Commons licensing, we can leverage the existing Web technologies to track materials propagation and reuse, creating new data points for the impact of scientific research that are more dimensional than simple citation indices, tying specific materials to related peer-reviewed articles and data sets.

The MTA project was launched in collaboration with the Kauffman Foundation, the iBridge Network of university technology transfer offices, and neurodegenerative disease funders. It currently includes more than 5000 DNA plasmids covered under standard contracts and is available through the Neurocommons project described in the next section.

PROOF OF CONCEPT IN KNOWLEDGE SHARING: A SEMANTIC WEB FOR NEUROSCIENCE

In collaboration with the W3C Semantic Web Health Care and Life Science interest group, we are integrating information from a variety of standard sources to establish core interoperable content that can be used as a basis for bioinformatics applications. The combined whole is greater than the sum of its parts, since queries can cut across combinations of sources in arbitrary ways.

We are also providing an operational knowledge base that has a standard, open query endpoint accessible by Internet. The knowledge base incorporates information marshalled from more than a dozen databases, ontologies, and literature sources.

Entities discussed in the text, such as proteins and diseases, need to be specifically identified for computational use, as do the entities' relationships to the text and the text's assertions about the entities (for example, a particular asserted relationship between a protein and a disease). Manual annotation by an author, editor, or other 'curator' may capture the text's meaning accurately in a formal notation. However, automated natural language processing (including entity extraction and text mining) is likely to be the only practical method for opening up the literature for computational use.

We were only able to process the abstracts of the literature as the vast majority of the scientific literature is locked behind firewalls and under contracts that explicitly prevent using software to automatically index the full text where it is accessible. Although most papers run more than five pages, the abstracts typically were limited to a paragraph.

For tractability, we limited the scope to the organisms of greatest interest to health care and life sciences research: human, mouse, and rat. We are also providing the opportunity for interested parties to 'mirror' the knowledgebase and we encourage its wide reuse and distribution.

In combination with the data integration and text processing, we are also offering a set of analytic tools for use on experimental data. The application of prior knowledge to experimental data can lead to fresh insights. For example, a set of genes or proteins derived from high throughput experiments can be statistically scored against sets of related entities derived from the literature. Particular sets that score well may indicate what's going on in the experimental setting.

In order to help illustrate the value of semantic web practices, we are developing statistical applications that exploit information extracted from RDF data sources, including both conversions of structured information (such as Gene Ontology annotations) and relationships extracted from literature. The first tools we hope to roll out are activity centre analysis for gene array data and set scoring for profiling of arbitrary gene sets, donated to Science Commons by Millennium Pharmaceuticals.

Taken together, we call these three efforts the Neurocommons – an open source, open access knowledge management platform, with an initial therapeutic focus on the neurosciences. And we hope to use the Neurocommons both as a platform to facilitate knowledge sharing and to secure empirical evidence as to the value of shared knowledge in sciences.

CONCLUSION: CYBERINFRASTRUCTURE FOR KNOWLEDGE SHARING

The Neurocommons project is a very good start. It shows the potential of shared knowledge systems built on open content. And it has the potential to explode through horizontal downloading, editing, and reposting, as the Web exploded. The idea of connectivity via 'viewing source' is an explicit part of our design methodology, and our tools have already been picked up and integrated into such systems as the Mouse BIRN Atlasing Toolkit (MBAT), which was built from the combined efforts of groups within the Mouse BIRN (Biomedical Informatics Research Network, a distributed network of researchers with more than $25 million in U.S. Government funding).

But the Neurocommons is, at root, a proof of concept. And from it we are learning some basic lessons about the need for infrastructure for knowledge sharing. Science Commons is on a daily basis forced to create namespaces, persistent URLs, and line after line of 'plumbing code' to wire together knowledge sources.

If we are going to get to the goal stated above, of dramatic increases in efficiency and radical transformation of outmoded discovery models, we are going to need a lot of infrastructure that doesn't yet exist.

We need publishers to look for business models that aren't based on locking up the full text, because the contents of the journals – the knowledge – is itself part of the infrastructure, and closed infrastructure doesn't yield network effects. We need open, stable namespaces for scientific entities that we can use in programming and integrating databases on the open Web, because stable names are part of the infrastructure. We need real solutions about long-term preservation of data (long-term meaning a hundred years or more). We need new browsers and better text processing. We need a sense of what it means to 'publish' in a truly digital sense, in place of the digitisation of the paper metaphor we have in the PDF format. We need infrastructure that makes it easy to share and integrate knowledge, not just publish it on the Web.

None of this is easy. Much of it is very, very hard. But the current system is simply not working. And the reward of pulling together what we already know into open view, in open formats, where geniuses can process and exploit it, could be a world in which it is faster, easier, and cheaper to find drugs and cure disease. This is possible. We just have to have the vision and courage to build the highways.