The Open Access Interviews

Richard Poynder talks to John Wilbanks, Vice President Science Commons

The goal of the Open Access (OA) movement is to have the peer-reviewed literature made freely available on the Web. That is, to remove research from behind the financial firewalls imposed by the traditional publishing model — which charges users (or their institutions) a fee to access scholarly articles, usually by means of a subscription.

"Freeing the refereed literature," argues OA advocate Stevan Harnad, is both optimal and inevitable, since in the age of the Internet most of the distribution costs of scholarly communication go away, and so continuing to restrict the number of people who can access it (by imposing artificial barriers) is to hobble science without just cause. After all, science is essentially the cumulative process in which people develop ideas, and make new discoveries, by building on the work of others. Clearly, they need to be able to access other researchers' work in order to do this.

Rather than treating research as a scarce resource that needs to be rationed, therefore, Open Access advocates argue that we should aim to maximise the number of eyeballs that can read it.

As the Open Access debate has developed, however, it has become increasingly clear that maximising eyeballs is just the first step. Open Data advocates like Peter Murray-Rust, for instance, argue that research papers also need to be accessible to machines — an argument he put to me recently with some passion. The problem today, says Murray-Rust, is that even where papers are freely available on the Internet it is difficult, if not impossible, to automatically extract the data contained in them, for both technical and legal reasons.

John Wilbanks, VP of Science Commons, has an even broader view of the role the Internet has to play in science. Like Murray-Rust, Wilbanks believes it is essential for research papers to be machine-readable. Likewise, he believes we need to develop an appropriate legal infrastructure to facilitate this. He also believes it is essential that science databases are freely available, and that these databases are interoperable — not just with one another, but with research literature.

In addition, Wilbanks believes the Internet should be viewed as a platform for facilitating the free circulation and sharing of the physical tools of science — cell lines, antibodies, plasmids etc. In a sense, he wants to see these tools embedded into research papers — so if a reader of an Open Access paper wants more detailed information on, say, a cell line, they should be able to click on a link and pull up information from a remote database. Should that researcher then want to obtain the cell line from a biobank, they should be able to order it in the same way as they might order an item on Amazon or eBay, utilising a 1-click system available directly from the article.

To make this possible, points out Wilbanks, we need to build the necessary technical infrastructure. This, he says, will require creating new ways of automating the collection, aggregation and discovery of scientific information, as well as the construction of an effective ecommerce system for the physical materials of science. And the best hope for achieving that, he adds, is by helping to create the so-called Semantic Web.

The end game, explains Wilbanks, is to make the research process as seamless and frictionless as possible. This implies that the scholarly paper is no longer simply an article to be viewed by as many eyeballs as possible, but also the raw material for multiple machines and software agents to data mine, a front-end to hundreds of databases, and the launch pad for an ecommerce system designed to speed up the process of research.

In this light, Open Access is by no means an end in itself, but the necessary precondition for a complete revolution in the way that science is done.

Some Open Access advocates believe that it is too early to be thinking about such matters. If — fourteen years after Harnad's seminal Subversive Proposal — the Open Access movement has still only succeeded in freeing around 25% of the peer-reviewed literature, they argue, we should remain firmly focused on freeing the other 75%, not fretting about what we do with it once it is free.

But that, surely, is too short-sighted a view. Besides, points out Wilbanks, we do not have the luxury of time, and so cannot afford to wait until the peer-reviewed literature is all available before starting to build the tools to exploit it.

If science is to continue benefiting mankind, he suggests, radical change is needed, and it is needed quickly — because science has reached the point where traditional methods are no longer able to deliver the goods.

In short, we are approaching the point where we will not be able to develop new life-saving drugs, or devise solutions to complex problems like global warming, without the kind of dramatic change in the way we do science that Science Commons envisages; for science is now so complicated that we will soon be unable to crunch the data quickly enough, or effectively enough, unless we embrace the kind of machine-driven, network-centric approach envisaged by the Semantic Web. As Wilbanks bluntly puts it, "The fact is that the complexity involved in studying a living system is such that even Pfizer — with $4 billion a year in R&D — can't handle it."

Wilbanks shared his thoughts with me during a recent telephone conversation; a conversation I had been trying to arrange ever since attending a fascinating presentation he gave at the Oxford Internet Institute one snowy February morning last year — a presentation I only succeeded in attending after digging my car out of a snowy hill in the Cotswolds!

A smallish, fresh-faced man with glasses, Wilbanks is undeniably very bright. He is also extremely knowledgeable about the life sciences. However, what I found most striking was the contrast he presented to most other Open Access advocates that I have interviewed. Open Access supporters are usually passionate and opinionated, and invariably argumentative. Wilbanks appears to have none of these characteristics.

Indeed, my overwhelming impression was of a quiet, unemotional and dispassionate man. Certainly I found it hard to envisage him getting into a heated argument about Open Access or, in fact, about much else. He seemed far too rational for that. But then, as Wilbanks himself pointed out to me, he is an "entrepreneur manager", not a lawyer, not a scientist, and not an activist.

And this is precisely what an organisation like Science Commons requires. In times of revolutionary change there is always a need for tub-thumpers able to inflame the passions, and inspire people to act. What is often forgotten, however, is that it is equally important to have effective organisers who can oversee and manage the transition. That, surely, is the role we can expect to see Wilbanks play going forward.

The interview begins...

RP: Can you start by saying something briefly about yourself and how you came to work with Science Commons?

JW: I was at the Berkman Center for Internet & Society at Harvard Law School in the late 1990s, when Larry Lessig was then professor, and when many of the early discussions took place that led to Creative Commons. That was also the time when discussions were talking place about *Eldred v. Reno*.

RP: Which became Eldred v. Ashcroft — a US Supreme Court case that challenged the constitutionality of the 1998 Sonny Bono Copyright Term Extension Act, and for which Lessig was lead counsel. Sadly, the case did not succeed. Tell me: are you a scientist by background?

JW: No, I'm more an entrepreneur manager than a scientist, a lawyer or a technologist. I was running the administrative centre at Berkman. And while I was there I started a software company.

RP: That would be Incellico I think?

JW: Yes, it was a company focused on analytics for the pharmaceutical sciences. Then when I finished with my company I spent a year in Boston consulting with the World Wide Web Consortium on the Semantic Web for life sciences.

RP: When you say finished with your company you are referring to its acquisition by Genstruct in 2003 I think?

JW: Yes. The bioinformatics market was far from robust and has remained far from robust since the technology bubble broke in 2001, although we had a good run compared to many companies. We made software, sold it, installed it onsite for customers, where it worked and did what we promised it would do — which is actually not that easy.

But our investors were ready for something more biology and less software, so we put together a deal with Genstruct.

Anyway, that was around the time when the Creative Commons board was preparing to launch Science Commons. I had known several of the board members when I was at the Berkman Center, and when I was consulting at the World Wide Web Consortium at MIT I saw some CC people around Boston. One thing led to another.

RP: What is the relationship between Science Commons and Creative Commons?

JW: Science Commons is part of Creative Commons. Basically, it's a distinct project within Creative Commons like ccLearn. We use a lot of the infrastructure of Creative Commons, and we build on the Creative Commons licences and experience, but our task is to evaluate how Creative Commons might work in spaces other than culture, and more generally to explore the idea of how the "commons" can enable innovation beyond the remix.

So we are part of the CC family but we live in a different office — we are on the East coast of the Unites States, Creative Commons is on the West coast — and we have different needs and different staff in some areas.

RP: Why does the world need Science Commons?

JW: Well from the very beginning of Creative Commons, during the first meetings at the Berkman Center, there was an understanding that the commons might have a really useful impact in science. Here, by the way, I mean the commons broadly construed: something built on the public domain and which uses contract to tilt the system towards permissiveness and sharing.

But as the board and the founding group didn't have a science background, they decided to test the concept in the space of culture, which is where Larry and the others were most comfortable.

RP: The idea of its relevance to science was there from the very beginning then?

JW: It was. And as Creative Commons exploded beyond everyone's expectations there was a growing desire at board level to look at CC in a couple of different places.

CC Learn was the effort that went into open educational resources, and Science Commons was very much driven by the emergence of the Open Access movement.

RP: The aim of the Open Access movement is to ensure that scholarly papers are freely available on the Web.

JW: That's right. The thinking was that Creative Commons had had some success, and gained some social capital, Open Access publishers like the Public Library of Science and Biomed Central had started using the CC licences, and the Open Access movement was growing rapidly, so it seemed natural to go ahead and get a formal science project going.

RP: The aim was to supplement the Open Access movement then was it?

JW: Yes. That was our first goal, and the wellspring of the Science Commons. But let me give you some background.

RP: Please do.

JW: When we started we spent a fair amount of time analysing the Open Access movement and the commons in science, and we came to a series of conclusions about what we thought the problems were that weren't being addressed. We didn't want to duplicate other people's work, and the legal work was already done because the CC licences were already in place and being used by Open Access publishers.

We decided, therefore, to work with the Open Access movement as an advocate for change, but to work on the broader context of access to digital knowledge. That includes Open Access to the literature, but also ensuring legal and technical access to databases — which is effectively the technical work that can be done on the back file and emerging corpus of Open Access — and thinking generally about the way in which the published information isn't currently taking full advantage of the electronic opportunities. Because even though it is now electronic, scholarly publishing is still very much paper driven. PDFs and papers, for instance, aren't cross-linked into databases today.

Copyright

RP: Science Commons has a number of projects underway today. Can you talk me through them?

JW: Sure. We have four projects at the moment, and we have one more that will be announced shortly. The first is what we call our Scholar's Copyright Project. This is really a very broad look at the shift in publishing from paper to digital, both from a legal and technical perspective. And we have created a series of small tools to help.

So there is the Copyright Addendum generator, for instance. We have also created some author journal agreements for our Open Access Law project. These are essentially technical fixes. The addendum generator is a great way to comply with the NIH mandate or with university policies, and it is starting to really get some attention since Harvard's big vote.

RP: These are very much legal fixes to help researchers, rather than, say, publishers, to embrace Open Access.

JW: True and the biggest thing we have done in this space is the Open Data Protocol (ODP), which we put out in December.

RP: This is focused on databases rather than journals I think?

JW: Yes. The goal was to address the issue of Open Source data integration, which is a really big problem. For instance, in molecular biology alone you have 1,000 primary databases with 1,000 different data policies, and there isn't a lot of understanding of how these database terms and conditions propagate if you integrate them into a single knowledgebase and then try to redistribute that knowledgebase.

So the Open Data Protocol is the result of almost eighteen months of work, and represents our conclusion that a single licence approach doesn't work for databases.

RP: The issue you are seeking to address is the legal barrier imposed by science databases using a range of different licences, which make it hard to aggregate them, or make them interoperable.

JW: So we had to do some important creative legal work trying to figure out what can you actually do in terms of building standardised terms and conditions for databases.

RP: How do you set out to do that?

JW: We began with the CC licensing suite, but gradually had to pare it away — because the terms used in licences like Share Alike and Non Commercial can have some very negative impacts if you use them in the data world. Share Alike, for example, can actually create intellectual property rights where they didn't exist before, so we concluded that that would have a negative impact. We also didn't want to end up propagating things like the European Database Directive, which we think is a bad law.

RP: Because the Database Directive introduces a new* sui generis *right that allows database creators to claim a 15-year monopoly on the data. As I understand it, your end point was that databases require some form of public domain terms and conditions?

JW: Yes. That was really what we came down to: that the only terms and conditions that really satisfy the needs of data integration, and data federation and propagation, are either to use the public domain, or some form of a contractual reconstruction of the public domain.

RP: Like the new CCZero licence?

JW: Yes, but we didn't want to force everyone to use our licence, or any one tool even — because there is a lot of stuff already in the public domain. Consequently, the approach we took was to issue the ODP and then certify different terms and conditions as conformant — so long as they met the conditions expressed in the ODP.

RP: That allows you to endorse as compliant with the ODP many of the terms and conditions that already exist for public databases?

JW: Precisely. So the terms and conditions of the US National Centre for Biotechnology Information for the human genome, for example, qualify under the protocol. In fact, most data in the United States that's fit for the public domain qualifies.

We also decided that this was the best way to let lots of different terms and conditions grow. Amongst other things, this allows locally relevant implementations of how to do the public domain to flourish.

RP: So you have used a similar model to the one adopted by the Open Knowledge Foundation, which drafted the Open Knowledge Definition. There is a meta licence — which is how one might describe your protocol — that delineates what can and cannot be required in order to comply with that definition, and then any number of downstream licences can be developed, so long as they meet the requirements of the meta licence? To this end you worked with Jordan Hatcher who drafted the Public Domain Dedication & Licence (PDDL), the first new ODP-compliant licence to be released?

JW: Right. The PDDL is a *legal tool* that conforms to the *Data Protocol*. The ODP is akin to the Open Knowledge Definition, and indeed it has itself been certified by Open Knowledge Foundation.

So you are correct: there will be many legal tools that conform to the ODP, and that we will certify. The PDDL is just the first one to carry that certification. I would add, however, that the ODP is a more stringent requirement than the Open Knowledge Definition, because it is really aimed squarely at data and databases.

RP: So that is the copyright project?

JW: Indeed, and that is obviously where we have our roots. But as we dug into the sciences it become clear that when you think about the overall research process in the sciences, the digital knowledge part — the copyright and stuff — is just one piece of the process. It is the instantiation of the publishing of knowledge part of it.

But there are a other elements that could be viewed as commons issues, although not necessarily in the domain of intellectual property. And as you look at them, these commons issues start to tie together into a theory about how the commons itself can support new models for innovation and collaboration.

The physical tools of science

RP: What kind of things are we talking about?

JW: Well, one place we looked at was the physical materials that underpin science, because our fundamental goal is to make science easier and faster. We want science to be able to translate basic research into discoveries that are meaningful and useful for people — like drugs or climate change policies, and so on — and as quickly as possible.

RP: And physical materials are important in this regard.

JW: Absolutely. The physical research material of life sciences, and of many other sciences such as physics, chemistry, and anthropology, depend on the shipping of materials from laboratory to laboratory, and from researcher to researcher.

Moving physical goods around the world quickly and easily is something that we take for granted in our day-to-day lives, thanks to the network, but the sciences don't do it very well. There are a lot of complex and interconnecting reasons for that, the least of which is the law.

RP: So this is not necessarily about intellectual property?

JW: No. When we looked at it we found that institutional lack of standardisation, and individual completion, turn out to be much more significant drags on the sharing of physical tools than intellectual property — in fact, in most cases the stuff isn't actually intellectual property, but physical property. It is, if you like, the theatre, not speech.

RP: In an article you wrote for CTWatchQUARTERLY last year you talked of the need to create a "transaction system along the lines of Amazon or eBay". What you're talking about is creating an ecommerce system for the physical materials of science?

JW: Yes. What we have been looking at is how to put together a fully functioning system to facilitate the rapid movement of these research tools around labs and scientist. And contracts are just part of the mix.

RP: What kind of physical materials are we talking about?

JW: We have chosen to focus in on life sciences and on Recombinant DNA materials — because these provide the easiest test case. So we are talking about things like cell lines, probes, DNA and antibodies. As I say, these things are more like speech than other things; more like what economists call "non rivalrous" goods.

RP: Which implies that once you have one it is very easy to replicate it presumably?

JW: Indeed. So from a batch of cells, for instance, you can grow more cells — which means that you and I can both have the cells, as long as someone gets paid to do the manufacture and distribution work.

We are also working with a set of Biobanks, which are sort of greenhouses for these tools that can store materials and fulfil orders — because there is a growing understanding that these are relatively early stage tools that really need to be moving around.

RP: One would think the issue would have been addressed before now.

JW: Well, there are policies in place that in theory mandate the availability of these tools, both government and journal based. But no one has really put it all together.

There have also been attempts to standardise the contract infrastructure around the movement of these research tools and a number already exist, including the Uniform Biological Materials Transfer Agreement (UBMTA) and the Simple Letter Agreement (SLA).

But these have all been single-contract approaches, and one of the characteristics of the current system is that the contracts are constantly modified, and so stop being standard. This slows the process down because the minute a lawyer modifies one word in a standard contract the opposing lawyer has an obligation to review every word all over again.

RP: How have you set out to address this problem?

JW: What we did was to look at the NIH-recommended policies for the movement of these tools, and then build a CC-style suite of contracts out of those recommendations.

RP: What do you mean by a CC-style suite of contracts?

JW: I mean a suite of contracts that guarantee a core right, with modular elements you can add to modify that core right. So where the core right in a CC contract is a right to copy things, in the materials transfer space it is a right to do research. That is the core right that is granted.

And where with CC copyright contracts you can then apply various conditions to the core right to copy, in the Science Commons materials transfer agreements you can add a set of conditions to the core right to do research.

RP: Basically you provide a contract template that people can, as it were, take off the shelf and use as is, or they add various conditions to it?

JW: And like the copyright licences it is a binding contract. However, it is not based on intellectual property. Importantly, it also allows for two major changes to the existing contract models.

The first thing that is different is that it isn't based on a naive idea that your tax status determines whether or not you have a need for a materials transfer contract — because the existing contracts regime doesn't contemplate the idea that research is done inside companies.

RP: Pharmaceutical companies have a need for these tools too?

JW: Yes. The reality of the modern pharmaceutical world is that a significant amount of research happens inside companies. So there is a disconnect in the current contract regime — a disconnect between the $30 billion a year in public funded research in the United States, and the $30 billion a year that is privately funded.

The second change is that our system assumes it is important to be able to reuse the research tool without having to handcraft it.

RP: Can you expand on that?

JW: It is sometimes hard for people outside the sciences to understand this point, because there is a belief that if you just share the data then everything is going to be fine. But for a simple and relatively highly quality-controlled and standardised experiment — using a microarray — you need a minimum of seven pages of annotations for every data file you share. This is called the MIAMI standard — Minimum Information About a Microarray Experiment — standard.

RP: The principle here is that a scientific paper should contain all information required to replicate an experiment.

JW: Exactly, but those seven pages don't even tell you what the experiment was about; they just tell you the actual settings of the machine at the time. And doing this for non-standardised assays, which form the vast majority of science, is next to impossible right now.

The upshot is that it is far better to give away the tools and let people run the experiments themselves, rather than trying to force everyone to spend all their time annotating data.

RP: When you talk about tools here you are talking about giving people the cell lines, the DNA and the antibodies etc.?

JW: Yes. Think of it this way: Imagine if every time you needed to make a curry dish in the kitchen, you had to hand-grind the spices and hand-make the oils. Now multiply that by several orders of magnitude and complexity, something that would take months and years, and you've got the situation in materials.

RP: And so ultimately the aim is to speed up research by allowing researchers to buy these tools rather than have to constantly remake them themselves. This I think is the point you make in your **CTWatchQUARTERLY** ***article when you say, "Materials represent tacit knowledge — generating a DNA plasmid or an antibody can take months or years, and replicating the work is rarely feasible." Rather than grow it, researchers should be able to buy it? Who do you expect to use these materials transfer contracts?***

JW: I think we are primarily talking about institutional use for now. So it is going to be funders of research who want to see the output of their funding reused. After all, if you give money to a laboratory to work on Alzheimer's disease and the output of that is a single paper, then you are not getting a great return on your investment. Or rather — since we are talking about publicly-funded research — the taxpayer isn't getting a great return.

So if that paper required the creation of five distinct tools — DNA plasmas**,** cell lines and other important research tools, then those tools should ideally be funnelled into a system in which plasmas go to a plasma repository, cells go to a cell line repository and so on.

And if in the process they are given digital identifiers they can be integrated into Open Access journal articles.

RP: How do you mean?

JW: I mean that when you are reading a paper you can 1-click and order the tools. Moreover, you can buy them for no more than the cost of manufacture and distribution. That is the world we hope to get to.

RP: These research tools would be available to anyone would they?

JW: To anyone who meets the terms of the contract, and is prepared to abide by those terms. The principle is the same as the click-wrap licences used in e-commerce.

But this is going to take a lot of social hacking. It requires technology transfer offices — which currently do a lot of artisanal lawyering — to give up some of that work; it requires that universities give up the idea that they can make money on every transaction involving a for-profit company; it involves companies being willing to accept contracts that give them research rights, but not commercial rights; and it requires scientists letting go of a world in which they use control over the physical tools to promote their laboratory, at the expense of science.

RP: How hopeful are you that you can succeed in your aims?

JW: Well, as with Open Access, the funders hold the reins.

RP: Are the signs positive there?

JW: Fairly. But I think we will have to let the legal code float a little.

RP: How do you mean?

JW: We may not be able to get from today's world to a CC-style world in one jump. Consequently, we are looking at ideas in which we can encourage the actors to post their own contracts and let a 1,000 flowers bloom.

RP: Can you expand on that?

JW: If you can get people posting one-to-many offers that say, "If you sign the contract I will do the deal with you" that is much better than it is today. At least you get the ecommerce benefits of this. You still have a contract propagation problem, but you can begin to have a discussion about standardisation.

RP: Instead of research institutions putting their laywers together to work out a bespoke solution each time, you suggest that they post their contracts publicly?

JW: Yes. We think it may be easier to let some of these actors get into the game by saying, "If you need to draft your own materials transfer agreement, that's fine, but we are going to use metadata and we are going to allow people to tag the contract with something akin to the CC metadata.

RP: The Science Commons website talks about using the licensing as a discovery mechanism for materials. What you are saying implies that some of these tags could include opinions on the

contracts, reviews of them even. The aim then is to encourage people to tag the licences Flickr-style — to use folksonomies if you like?

JW: Yes, and this allows you to identify people who put up toxic contracts versus people who put up liberal contracts. That could be a good way of leveraging a piece of the tagging world to create a market in favour of liberalisation of the contracts. However, that is something that we are still exploring.

RP: It sounds very subversive.

JW: I like to think of it more in terms of transparency. Our goal is not so much to have people adopt our contracts, but to encourage the emergence of a system in which these materials can flow. If one of the impediments today is artisanal contracting then anything that gets us out of that culture is a win.

That means that if we end up with 1,000 actors using 500 contracts, but each one of those are binding contractual offers to the world, then we have come a long way — because at least we have digital identifiers, and we have standard contracts. From there is it's simply a matter of beginning to rate those contracts and allowing users to choose the ones that have the terms they like.

Anyway, that is what we call our Biological Materials Transfer Project. And at this point there are about 6,000 materials in the system through various partnerships. We have also developed a set of contracts for extending our network of biobank partners, and we are now starting to go to funders of research — both foundations and states.

And what we are saying to them is that the next step beyond Open Access to the digital information is Open Access to the research tools. We believe, by the way, that in many ways this is much more important than access to researchers' data.

RP: Why?

JW: Well, there are two classes of data. There is fundamental data, by which I mean databases — things like the genome, or federated databases about protein structures — which absolutely need to be open, and which are useful for almost all researchers. And there is the kind of data that comes out of a machine in someone's laboratory.

We believe that Open Access to the former is more important if you want to speed up the research process. The latter is good stuff, but as I noted earlier, the lack of annotation really kills the utility of most lab-generated data. It's easier to get the tools and just run your own, at least for now.

The Semantic Web

RP: Can we move on to your third project?

JW: Ok. So if you think about the three core projects there is the copyright project (the digital knowledge stuff), there is the physical knowledge (the tools), and then there is the Semantic Web.

RP: The Semantic Web is important because today's web isn't adequate for what you want to achieve.

JW: Correct. So the Semantic Web is about creating the necessary infrastructure. One way to think about this is that after you get all this digital knowledge online, and you get all the physical tools online, what is still missing is the core infrastructure for lightweight collaboration.

RP: What we are saying is that Open Access is just about making research papers freely available online. What is also necessary is to integrate interoperable public databases and the physical tools of science with those papers, which means also creating the necessary infrastructure. As you put it in* CTWatchQUARTERLY*, we need more complex and realistic interconnections between articles ("a web, a set of highways"), the knowledge in those articles, and the subject matter of those papers: the genes, proteins, cells and diseases. You added, "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." In other words, making research a seamless process of clicking from a research paper, to the details of a particular gene or cell discussed in that paper (information that will be held in a remote web-accessible database), and then clicking on another link to order the cell line or DNA plasmid discussed in the paper?

JW: Exactly. Today we take this for granted in the cultural space — because we have web browsing, we have the domain name system, we have blogging platforms, we have Ubuntu, we have Wikipedia, and we have things like Flickr, tagging, and microformats.

That infrastructure makes it very easy for anyone to participate in culture, but there is very little of that for science. Moreover, the cultural infrastructure is not robust enough to deal with big science. Tagging and microformats don't quite cut it.

For that reason a big part of what we do at Science Commons is to invest staff time and money in order to participate in the development of the Semantic Web, both from a standards and software perspective.

RP: Can you expand on that?

JW: For instance, Jonathan Rees is on the technical architecture group of the World Wide Web Consortium, Alan Ruttenberg is the chair of their web ontology working group, and we've all been involved in the W3C Health Care and Life Science Group. So we all spend a lot of time working on arcane early stage Semantic Web standards.

RP: Why you? As you said, you are not a technologist.

JW: We have two goals: one is to ensure that the infrastructure meets our needs; the other is to ensure that the core infrastructure remains open.

If you think about it, we are at the same point with the Semantic Web as we were with the Web itself in the early days, when TCP/IP was being developed. These are early, early days. Today, for instance, there is no equivalent to the DNS for life sciences. So, for instance, we don't have a unique name for any given gene. Those mappings have yet to be done.

RP: Which requires things like ontologies I guess. Unlike what you call the cultural space of the Web — which is about people talking to people, or people talking to machines — Big Science requires a Web in which machines can talk to machines.

JW: Sure, and a big part of what we are trying to do is create open structures, both name space structures as well as standards, so that the system can't really be captured.

RP: You mean preventing it from being made proprietary by commercial interests?

JW: Exactly. If it is valuable it will be captured; that's what business does. That's not bad. It is just what business does.

So the Semantic Web stuff is the third leg of Science Commons.

Neurocommons

RP: Ok, let's move on to the fourth project, which must be the Neurocommons.

JW: The Neurocommons is a relatively well-established project. As I said earlier, we also have a new one coming down the pike which I can talk about in broad terms. What these projects have in common is that both are, if you like, implementations of all three of the core areas of work that we have discussed.

RP: What then is the Neurocommons?

JW: So the Neurocommons is the first proof of concept of our core areas. That is, we are putting together the digital copyright project, the physical materials project, and the Semantic Web infrastructure.

At its heart the Neurocommons is a knowledge base. What we did was to hire some people from Millennium Pharmaceuticals who had been working on computational biology. These guys have an enormous amount of experience in integrating public databases into a single knowledgebase, and then using that knowledgebase to help scientists answer complicated questions.

This, by the way, is a real advantage that the pharmaceutical scientist has over a regular non-pharmaceutical scientist, and it really creates a distinct division in the ability to make meaningful use of the public information space.

RP: Presumably the advantage you refer to is that there are a lot of public databases in the pharmaceutical area?

JW: Yes. At the last count there were about 970 databases**.**

So I set these guys off to integrate as many of these big public databases as we could, and currently the Neurocommons has about ten of the most important major molecular biology and neuroscience databases indexed in a single resource.

RP: Ok, so this speaks to your aim of aggregating databases. You said Neurocommons is a knowledgebase. Does that mean that the end product is a database in its own right?

JW: It's a few things. At one level it is a database in its own right, and it is already being mirrored around the world, because we are redistributing it without any restrictions. At the same time it is a knowledgebase of hundreds of millions of relationships that we have extracted from these ten databases and reformatted and put into a single structure.

RP: So you are replicating and aggregating these ten databases. In addition you are plotting relationships between the data in them?

JW: Well, we aren't actually replicating the databases. We are indexing them. So it's a giant index of all these databases. In addition, it allows you to write queries using something close to SQL, but a web version of SQL that lets you treat them as a single database. So you can ask complicated questions and get precise answers. This is a basic principle of information retrieval, but applied to a distributed set of scientific databases over the Web.

The other piece of this is that we are also providing the toolkit to allow other people to create their own knowledgebase. In other words, we also provide the scripts that create the relationships and do the indexing. In addition, we are putting the descriptions of the openly available physical materials in there too.

RP: Which allows the 1-click ordering you mentioned earlier?

JW: Yes. So from any given query you can just right click and it will show you the materials that are available, which you can then order using a 1-click process.

Finally, there is also text-mined content coming into it from the Open Access literature.

RP: So Open Access is assumed as a given. In that sense it is very much a next-step project?

JW: Neurocommons is based entirely on content that is open, and it is built on the infrastructure of the Semantic Web. So it leverages all the technologies that can currently be leveraged, including 1-click access to the physical materials.

That is why I say that it brings together all three of our core projects.

RP: Can you give me an example of the kind of uses that could be made of Neurocommons?

JW: As I said, it can give researchers precise answers to research questions. So, for instance, a life scientist might say: "Find me genes that are associated with importing stuff into the ribosome that are active in cancer cells."

If you ask Google that question it brings back a few hundred thousand web pages, and research papers. But you don't really want that; what you want is a list of genes.

RP: Neurocommons can provide a list of these genes?

JW: Correct. Now, the point to bear in mind is that the methodology for doing this is relatively straightforward. But it is so high throughput that very few humans have the time to do it. If you did it on a human basis you would never do anything else.

RP: Why would it be so time intensive?

JW: Because you would have to take the US National Library of Medicine's Medical Subject Headings and then find all the papers that are associated with cancer cells. Then you would have to do a cross reference of all of the identifiers for all of those papers, and all the genes associated with those papers. After that you would have to take that list of genes and filter it by the gene oncology category for ribosome or protein imports.

So what we have created is a system that can automate these types of queries, or "graph traversals." The relationship index looks like a big graph of nodes and edges, and a big part of writing queries is how to move around the graph**.** In addition, once these queries have been written to represent the HTTP bit commands they can be turned into clickable links.

RP: So it becomes a cumulative process: every time someone queries the system that query is linked with the answer by means of a hyperlink?

JW: So that is one level at which Neurocommons helps: It is a very rapid way to write queries and, because of the way it is written, it allows you to leverage all the power of HTML-style approaches.

As you say, after you have written a query once you can represent it as a link. Moreover, if someone wants to change the query they simply have to go in and edit it — which is much easier than writing the query itself. In that regard it's reminiscent of HTML in the mid 90s.

At the individual level, then, it is a much more powerful search engine for public domain information. In addition, at the general level there are several other benefits that we hope to see emerge over the long term.

RP: Such as?

JW: We hope, for instance, to see fewer and fewer resources aimed at continuous reintegration of the public domain. Inside biotech and pharmaceutical companies everyone does this over, and over, and over again.

RP: So the objective is to create an environment in which this only has to be done once?

JW: Because we want to get to a world in which every database that comes out is not only legally available under something like the ODP, but an end point is made available for this type of query using open standards and open links. Then everyone's database is going to snap together more easily.

RP: By propagating your licence, and encouraging openness, you hope to be able to shift the world of science away from multiple proprietary databases to one in which all databases are both open and interoperable.

JW: it's not just about multiple proprietary databases. There will always be proprietary databases. It's about making sure the public domain is formatted in a way that supports integration and meaningful queries, because then the best decision for everyone is to start formatting in the same way as the public domain. That's the best use of resources and it lets people start to compete on the questions they ask, not the information they control.

RP: Ok. So it's about the interoperability of public databases. As you put it in* CTWatchQUARTERLY*, "There are thousands of databases with valuable information in them. Each of them has different privilege conditions, different formats, different languages, and different goals" and each is "maintained at different levels of quality." So when you talk about making an end point available you mean a single search interface — like Google — that can search on all the relevant databases?

JW: Right, and it would mean that every database that comes online is more likely to contribute to a positive network effect in terms of knowledge creation.

RP: Because every new database automatically becomes part of a much larger "virtual database".

JW: It also means that we are able to publish things legally and technically and every new database contributes to the larger effort. What we have right now is a major knowledge gap in the life sciences. Our hope is that by providing bait to begin to do things the right way technically then within a couple of years as new databases are created they come online under open technical terms, and users have an end point that allows for structured queries.

RP: Each new database just slots into all the other databases automatically?

JW: That is one piece of it. The other piece is that right now there are probably only around 15,000 biologists in the world that have access to this kind of tool — and those are the biologists inside Big Pharma.

RP: Which goes to my point about proprietary databases: Only those working in Big Pharma can afford access to all the proprietary databases?

JW: And that means that there are a lot of very smart people that are working with tools that are ten years old — from an information retrieval perspective.

If we can make those same tools available to everybody, then as the cost of data generation drops, and as the ability to reuse data drops — which is what we expect to happen as the annotation comes out for each of these kinds of assays — then we start building towards a world in which the only limitation is in being smart enough, and being able to get data.

Open Source culture

RP: Can we clarify what we mean when we talk about annotations here? I assume this is the same thing as the "relationships" you talked about earlier. I note the Science Commons website says: "The long-term elements of the Neurocommons revolve around the mixture of commons-based peer editing and annotation of the pilot knowledge project and the creation of an Open Source software community around the analytics platform." As I understand it some of these annotations (relationships) will be automatically generated when people put in queries, some will come from users — who are being asked to contribute RDF from texts that they have mined themselves (some of which may have been mined in proprietary databases), using the software toolkit that you plan to release under an Open Source BSD licence, and some will be hand-edited annotations? This implies a combination of automated annotations and human-generated annotations. Have I understood correctly?

JW: I think so. What we're doing with our index work is machine annotations. And we have a lot more to do there. But we also want to see a world in which it's part of the science culture to annotate, both in one's own lab and on the larger web. I think this is likely to require some kind of prize models. If we could give away 100 MacBook Air computers to the 500 postdocs who did the most annotation, something tells me we'd get a lot of annotations.

But clearly there is a social and ethical aspect to this project as well as a technical one. Right now we live in a world in which there is a social division between the haves and the have-nots in terms of the tools to do analytics, and the tools to make meaningful queries in the information space. By creating and giving away this kind of knowledgebase we hope that people will start to write code to it.

We also hope that pharmaceutical companies will start to release software under open source licences. Millennium Pharmaceuticals has gone ahead and done that — they've given us their award-winning pathway analytics software under the BSD, which we plan to release as soon as we have it cleaned (that's taking a lot longer than we thought it would!)

Our hope is then that we can create a nucleus around which a nascent Open Source culture can develop in information management in the life sciences.

RP: Which will all add to the network effect you hoped to create. Additionally it will, as you say, reduce the cost of data generation — all of which will help close the knowledge gap in the life sciences that you referred to.

JW: Precisely, because there is just too much stuff for anyone to handle. The fact is that the complexity involved in studying a living system is such that even Pfizer — with $4 billion a year in R&D — can't handle it.

RP: The premise of Science Commons is that the best way of dealing with the growing complexity of science is to share the load.

JW: Right. And the promise is if we can get lots of small relevant efforts — where the local complexity is not overwhelming — to snap together both legally and technically we have a much better chance of having breakthroughs in our understanding, which is what it is going to take to get the cost of drugs to drop.

The reality is that all the artisanal legal work in the world doesn't really matter if you still don't know how to get drugs into cells, or how to understand toxicity.

So this a place where the commons, broadly construed, says that to the extent that it is possible all of the outputs of research can snap together legally and technically.

RP: Presumably this has implications for Open Access. Today most OA papers are deposited on the Web as PDF files?

JW: Correct. Instead of being published as PDF, papers need to be published as XML with hyperlinks. Likewise, instead of being created as bespoke one-shop systems with java script query interfaces, databases need to be exposed to a structured query endpoint, and to not restrict transformation and extraction of the data.

If you add to this the digital exposure of the research tools, and you make them seamlessly available within the system, and if the web infrastructure supports this, then we are talking about one of the only non-miraculous ways of increasing the odds of getting miracle breakthroughs.

Virtualising the drug development process

RP: You said there was a fourth project in the works?

JW: Yes, we haven't named it publicly yet, but we have been approached by a number of entities — organisations and companies — who are interested in the virtualisation of the pharmaceutical development process.

RP: What do you mean when by "the virtualisation of the pharmaceutical development process"?

JW: Well if you think about it, the modern pharmaceutical development process it is so byzantine and complex that it ends up that you need a company the size of a city state in order to go through it. That, in fact, is pretty much what a modern multinational pharmaceutical company is — it's more like a city state than it is like a business, right down to the five-year plans that get promulgated when they talk about their pipelines.

RP: It doesn't need to be like that?

JW: No. If you look at the mythical drug discovery development process, which is artificially linearised, you will see that it is actually hundreds and hundreds of separate transactions.

So, for instance, it is early-stage screens of the genome to find genes associated with, say, arthritis; it is the study of the proteins associated with those genes; it is the optimisation or selection of a set of proteins (or a single protein) as the core target for drug discovery; it is the screening of tens of millions of potential drugs against that target to see what sticks to it; it is the narrowing down of the fifty thousand hits against that to a set of "leads"; it is the process of timidly tinkering with those leads for several years; it is the process of testing them in rats, mice, dogs, and monkeys; and then it is the process of testing them in people.

Now when this takes place inside the city state of the pharma company you don't have to have any standardised way of moving materials and data from, say, a laboratory in Connecticut to a laboratory in Michigan — because you own both of them.

But if you want to break this apart and say, "Well there's actually a forest of vendors out there who do all of these various processes and can support biotech and pharma", then in theory you should be able to assemble a relatively virtualised process in which you order genomic screens from one company and proteomic assays from a second company, and you order chemical hydroponic screening from a third company.

But to do that it means you have to pass the materials and the data around among the vendors and service providers, and that means that you need both legal and technical standardisation.

RP: The point here is that not even a City State pharma company can deliver the goods anymore, so it will have to be broken apart and virtualised. I guess the distributed model you envisage assumes much greater collaboration?

JW: It wouldn't be collaboration in the traditional sense. It would be more like, saying, "I am a foundation that investigates Parkinson's disease and I want to have a contract with a company that does one thing, which is gene screening. Then I want to pass the results of that screening to a company that does protein screening, because that will inform their work.

Now that is not collaboration in the traditional sense, where everyone gets together and agrees to collaborate; it is collaboration in the sense that each of the actors are a different piece of a process that I am interested in.

Essentially what you would be creating is something more like how movies get made nowadays. When it makes a movie Hollywood will put together the talent for the project, and then everyone goes away after the film has been made. This is very different to the way the old Hollywood system worked, when the film company owned the actors.

RP: This raises some challenging cultural issues I guess?

JW: Absolutely. But remember it wasn't us who came up with the idea. What happened was that when the people who are working on this idea looked around they realised that they absolutely needed the Science Commons solution kit.

Because when the data set moves from one solution provider to another solution provider, for instance, there will need to be common naming standards, so that the data means the same thing to both providers. And you need to have the web infrastructure that makes it as seamless and painless as possible without having to lock into a single vendor. And you need to have the contracts to enable the movement of materials from academic labs to corporate research labs, and then back to academic labs.

RP: In short, all the things that Science Commons is currently working on?

JW: Right. The important point here is that any re-imagining of the pharmaceutical business model is going to have to be built on a commons. That is why we have been approached by these people to work on the project. We hope to announce details of this in the coming weeks.

RP: I guess we are talking about a long-term project here?

JW: We are. And it would require going much further than we have gone until now with our materials transfer work.

RP: How do you mean?

JW: Up to now we have only really dealt with the biological side of things — both in terms of the information technology and the materials. In Big Pharma that part of it is just the first couple of years, and a relatively small amount of money.

To take this further you have to get all the way into how you share highly proprietary compound libraries. And once you get into that you might find the solution is not to use a digital commons but something more like an insurance model, or an escrow agent model.

You also have to think about how you share failed clinical trial data, and under what technical terms and legal terms you put all this together. You also have to think about how you make data generated on a bespoke basis inside one of these companies valuable to everyone.

So you have to deal with things like that. It's a challenge, but we look forward such a challenge.

Given enough eyeballs

RP: In short, you believe that if science is to make further progress in the future the research process is going to have to look very different, and it will need to be based on the Semantic Web?

JW: Well the Semantic Web is a specific technology platform. The goal is to create a model where it is easy to leverage what we know, and in a way that allows us to overcome the transactional difficulties imposed by either the law or the technology.

The Semantic Web may or may not work. It may not be what we hope it is. However, right now it is the only candidate that lowers the transaction costs enough over time to get to the world we want to get to. I don't see any other technology even in the ball park that begins to allow the stitching together of databases for instance.

You know the Semantic Web gets a lot of abuse, some of it deserved. But if you think about it, what we are really looking for is a technology that allows us to federate and integrate databases, and to be able to query them from single end point. Essentially, the aim of the Semantic Web is to make data talk to data.

If you put it like that I think people will begin to understand the point of it. Perhaps a useful analogy here is HTML. HTML was a very weak markup language, but in the end its very weakness was its strength. So was its openness — because that allowed people to come in and put cascading style sheets on it, and it allowed someone to come along and stitch YouTube on to it etc.

My hope is that we are in the process of stitching together the early stages of a data web. And with luck people will come along and build amazing artificial intelligence systems on top of it.

RP: The take-home point of our conversation, I guess, is that scientific research has become so complex and challenging that collaboration is essential. Perhaps a good analogy is Eric Raymond's assertion that computer programs are now so large and complicated that software companies struggle to produce good software. As a result, they are increasingly having to turn to outside developers, often on a distributed basis. As he put it: "given enough eyeballs, all bugs are shallow." And as the Open Source movement realised, once you are in the business of distributed collaboration you really need open technologies and open processes?

JW: Precisely. And actually that's better than I said it, so I might have to cite you and attribute you. That is exactly what I am trying to say.

RP: Feel free to cite me! Thanks very much for your time.

If you would like to republish the interview on a commercial basis, or have any comments on it, please email me at richard.poynder@btinternet.com.

Please note that while I make this interview freely available to all, I am a freelance journalist by profession, and so make my living from writing. To assist me to continue making my work available in this way I invite anyone who reads this article to make a voluntary contribution. I have in mind a figure of $8, but whatever anyone felt inspired to contribute would be fine. This can be done quite simply by sending a payment to my PayPal account quoting the email address *richard.poynder@btinternet.com.* It is not necessary to have a PayPal account to make a payment.