The Open Access interviews

Richard Poynder talks to Peter Murray-Rust about Open Data

Peter Murray-Rust is a committed advocate of Open Access (OA). He is, however, a disappointed one. He is disappointed not because so few researchers are willing to self-archive their scholarly papers on the Web, not because it is proving so hard to persuade funders and research institutions to introduce Open Access mandates, but because of a failing he sees within the movement itself. Out of his disappointment, however, has come a new movement: the Open Data movement.

As a Reader in molecular informatics at the University of Cambridge Murray-Rust is interested in scholarly papers less for their textual content, more for the raw data contained within them — the graphs and tables, the molecular structures, the spectral and crystallography data, the photographs of proteins, and all the other factual information that litters science papers.

As such, much of Murray-Rust's time is spent not on reading the scholarly literature, but mining it — using various software tools to automatically extract the "embedded data" contained in the tables, the charts, and the images in science papers, and capturing the "supplemental information" that invariably accompanies the papers. After aggregating all these data Murray-Rust will compare them, input them into programs, use them to create predictive models, and reuse them for a variety of different purposes.

In short, Murray-Rust is working at the frontline of what has been dubbed Science 2.0, an online interactive environment where a great deal of the information used is more likely to have been discovered, aggregated and distributed by software and machines than it is by humans; an environment where data are constantly used and reused — pumped through new tools like RSS feeds, and displayed in mashups, wikis, and the various other tools developing around Open Notebook Science.

Murray-Rust's ultimate goal is to create and exploit what he calls the chemical semantic web — a web that would assume most scientific information was unencumbered by proprietary interests, and able to be freely shared and exchanged. In practice, however, mining the scholarly literature remains a difficult and risky activity, explains Murray-Rust — not so much because the technology is still in its infancy, but because scholarly publishers routinely appropriate the content of research papers, and then lock it up behind financial firewalls and prohibit its reuse.

Assuming that the Open Access movement was committed to removing these barriers, Murray-Rust became an OA advocate. After all, as leading OA advocate Peter Suber puts it, Open Access implies scholarly literature that is "digital, online, free of charge, and free of most copyright and licensing restrictions". That, says Murray-Rust, is what is needed to build the semantic web.

But while the definition of Open Access agreed at the launch of the 2001 Budapest Open Access Initiative (BOAI) states that any paper made Open Access must be free of copyright and licensing restrictions, Murray-Rust discovered that in most cases publishers and authors still fail to provide the necessary permissions when making papers Open Access. Where a paper is flagged as being Open Access, reuse is often prohibited. And even where there is no specific prohibition, usage conditions are frequently not specified, effectively placing the paper into licensing limbo.

In many cases, says Murray-Rust, Open Access publishers don't even articulate to themselves under what conditions they are making their papers available on the Web, let alone provide an appropriate licence. As a result, third parties cannot know what usage is permitted. And where publishers do think it through, and attach a licence, the usage conditions are in any case often non conformant with the BOAI definition.

The legal status of papers that researchers themselves self-archive on the Web, or in their institutional repositories, is equally uncertain, and sometimes reuse is expressly forbidden.

What frustrates him says Murray-Rust, is that this confusion could have been avoided — had the Open Access movement emulated the Open Source Initiative (OSI) and developed customised OA licences. And having done so, he adds, the movement (again like the OSI) could have policed the use of the term Open Access, and publicised and sanctioned publishers who fail to use the licences, or who make false claims about Open Access. It should also have better educated researchers about licensing.

Further limiting what he can do, adds Murray-Rust, traditional subscription publishers like the American Chemical Society and Wiley explicitly forbid text mining of papers they publish. At the same time these publishers insist that authors not only sign over the copyright in the paper, but also ownership of the supplemental data, despite the fact that factual data are not subject to copyright.

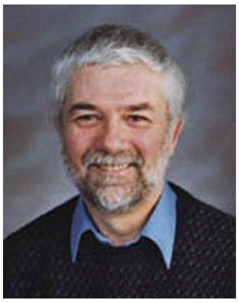
After failing to persuade Open Access advocates to hear his concerns, Murray-Rust began to direct his energies to what he calls the Open Data movement, for which he is now a leading advocate. While he remains an advocate for OA, he explains, he has come to believe that the issue of Open Data needs to be addressed separately. For where the Open Access movement is concerned only with ensuring that scholarly papers are human readable, the Open Data movement requires that they are also machine readable. And since Open Data implies reuse, it is vital that licences are provided that specifically permit this.

Fortunately, Science Commons stepped into the breach, and is proving a valuable ally, not least by developing the Open Data protocol and the recently-announced Public Domain Dedication & Licence (PDDL) — thereby providing the first component of the legal framework that Murray-Rust believes is needed to enable text mining, and helping in the creation of the chemical semantic web.

I had been keen to speak with Murray-Rust for some time, so I was pleased recently to be able to hook up with him on the telephone. I found his ebullient style, rapid delivery, and quick-fire mind both challenging and fascinating. Above all, the conversation offered me an interesting new perspective on Open Access, and confirmed suspicions I have long harboured that the Open Access movement would truly benefit from having an official body to represent its interests.

Murray-Rust is a vivid and rumbustious person who does not pull his punches. When I emailed the draft text of the interview to him, however, he asked that I stress the positive rather than the negative in this introduction. "Yes, I am angry, but not completely," he wrote. "I believe in the power of the bottom-up to change things and I am optimistic that we shall get change."

He also asked me to underline his appreciation for all that the Open Access movement has achieved, and requested I append this paragraph: "Although this interview highlights some of the shortcomings of Open Access movement I want to pay tribute to the many activists who have devoted and often courageously worked to make scholarly knowledge free for everyone. I'd particularly like to say something very appreciative about Peter Suber, and I'd like also to mention the Scholarly Publishing & Academic Resources Coalition (SPARC) and the Wellcome Trust — who in my opinion have probably been the largest force for change recently."



The interview begins...

RP: *Can you begin by saying something about your background and your research interests?*

P M-R: I am a chemist with a particular interest in crystallography. I began my professional career in academia, and helped to set up the chemistry department at the University of Stirling when it was a green field site. I then worked for the pharmaceutical company Glaxo for a number of years, in the drug discovery area. It was mainly involved in computational chemistry and structural bioinformatics.

I then spent four or five years at the University of Nottingham on a part-time basis (where I helped to set up virtual education), before moving to the University of Cambridge, where I am currently a Reader in Molecular Informatics in the University's department of chemistry.

Specifically, I am based in the Unilever Centre which, as the name suggests, has been sponsored by Unilever. The company has given us very generous support to investigate what needs to be done in the whole area of informatics in molecular science, and we are currently exploring what we think chemists are going to need in the future.

RP: *What should we understand by the term molecular informatics?*

P M-R: I see molecular informatics as creating a formal structure for describing information, and in such a way that it is possible to find it using as natural an approach as possible. So, for instance, we are looking at how to describe molecules formally using Chemical Markup Language (which is something that Henry Rzepa and I have developed over the last 13 or 14 years), and ways in which chemical information can be integrated into the web, not least so that search engines can best find information of this sort, and how it can be mashed up, and how we can add semantics to it.

RP: *This is what you refer to as the chemical semantic web?*

P M-R: Right. And one area we have been particularly active in is polymer informatics, developing new languages to support ways of searching for it.

RP: *Essentially, we're talking about using semantic techniques for the discovery and exploitation of chemical information?*

P M-R: We are. So much of what we are doing is aimed at the next generation of technology: Web 2.0, REST, markup languages, RDF (Resource Description Framework), RSS, along with the whole idea of the blogosphere and other types of virtual community. More recently we have also begun to look at things like Open Notebook Science.

***RP:** And it is because of your interest in these areas that you have become a leading advocate for what you call Open Data.*

P M-R: Correct.

Free of any restraint

***RP:** What then is Open Data? I note that the Wikipedia entry on Open Data — which I believe was mainly written by you — says, "In some cases open data can be considered as more properly open metadata."*

P M-R: Well, what I was trying to do with the Wikipedia entry was to review all uses of the term Open Data. So that particular statement refers to the kind of data that you send off to services like flickr.

***RP:** So the topic of Open Data is larger than chemistry, or science even?*

P M-R: Exactly. People like Talis' Paul Miller, for instance, are interested in Open Data in the context of what is generally called Library 2.0. Others, like Edd Dumbill, see it as part of a much wider development altogether. So while it is entirely reasonable to talk in terms of metadata when discussing, say, what people are doing with Web 2.0, that is not what we are talking about in the context of science.

***RP:** Ok, so Open Data can also refer to Web 2.0-type metadata. And as you say, Dumbill sees it in a much larger context. He talks, for instance, about what he calls open government data, and the kind of public data that companies like Amazon and Google make available. Let's note in passing, then, that Open Data encompasses far more than we need to discuss here, and focus in on the question of Open Data in science. How do we describe Open Data in that context?*

P M-R: I would say that Open Data are data that are free of any restraint on access and on reuse. If you want something more precise, I would refer you to the Open Knowledge Definition, which probably does a much better job defining it than I can do talking on the hoof.

But basically, I take Open Data to mean information that is made available to the whole human race, and done so in a form that everybody can see that it is openly available, and in a way that ensures that no one needs to ask permission to use it, or to reuse it.

***RP:** What does reuse mean in this context?*

P M-R: By reuse I mean that you can do anything with it — including transclusion, changing the format, translating it, cutting and pasting it, mashing it up etc. etc. However, people shouldn't be able to change it in such a way as to radically alter its meaning — by for instance inserting "nots" into it, or by taking it out of context.

***RP:** Does your definition of reuse also assume that third parties can collect Open Data and then sell it on commercially?*

P M-R: Oh absolutely, there is no question about that — so long as they do it on a nonexclusive basis. In other words, you must not sell it in a way that stops anybody else doing what they want with the data, including either giving it away free, or reselling it themselves.

RP: *What you are saying is that no one should be able to appropriate Open Data — by, for instance, becoming an exclusive distributor of it?*

P M-R: Exactly. The process of preparing Open Data for sale should not restrict other people from doing other things with it. Of course, it could well be — and here we get into the murky area of copyright — that the expression of the saleable data is copyrightable.

RP: *How do you mean?*

P M-R: You might, for instance, want to protect the fonts that the data is displayed in, or the look and feel of the page. I don't have a problem with that, but you should not be able to stop other people doing whatever they want with the raw data itself.

Fundamental infrastructure of science

RP: *Why does science need Open Data?*

P M-R: It needs it for a number of reasons, and I list some of these in a paper I have written for *Serials Review*.

First, data are part of the fundamental infrastructure of science. If we don't know the formula of a compound, for instance, we can't actually talk science about it. So if, say, somebody says that the formula for aspirin is copyrighted, secret or proprietary in some way, then I might not know what aspirin is. And if I don't know what it is, that stops me from doing science.

Second, the whole culture of Web 2.0 has shown that data are enormously enhanced in value when they are made freely available, and in a technically accessible form — by means, say, of a mashup.

RP: *Can you give me an example of how a scientist might want to do a mashup?*

P M-R: If, for instance, I find a set of biological reports on a receptor, and these reports talk about various compounds, I want to be able to immediately go off and find out what these compounds are, mash the whole thing up, and then maybe say: "Did you realise that all these compounds are chemically similar." Because I may have chemical software which can look at molecular similarity, where the original biological laboratory might not have access to that software.

RP: *In other words, you want to be able to put the original data contained in those reports on the Web, and combine them with the results from your software — allowing others to view the original data plus the data you have added to it?*

P M-R: Correct. And we all gain as a consequence.

The other point to make is that in the 21st Century we face the huge challenge of saving the planet. Sharing our knowledge is a necessary but not sufficient condition for doing that. Here, by the way, I am not just talking about global warming; I am also referring to how we save the planet from disease, from ignorance, and from all sorts of other things. In my view Open Data are a critical part of the cultural and political imperative attached to saving the planet.

RP: *When you talk about sharing scientific information I am reminded of the phrase coined by Eric Raymond to express the benefit of making the source code of software open: "Given enough eyeballs,* *all bugs are shallow." In other words, the more people working on a problem the quicker a solution can be found. But this requires that the relevant data is made freely available to everyone.*

P M-R: Exactly.

Merely facts

RP: *As I understand it, much of the data that you are interested in is data that is published in or alongside the papers in scholarly journals? What kinds of data are we talking about?*

P M-R: We are talking about anything that could be regarded as a fact in some way or another. I would regard a set of numbers as merely facts for instance. Likewise, I would regard a chemical formula as a fact. And I would also regard the tabulation of systematic scientific data as factual information.

RP: *Can you give me a concrete example?*

P M-R: Well if, say, someone had plotted carbon dioxide levels over time, I would say that those are facts; they certainly couldn't be called a creative work.

Bear in mind here that it is not possible to get the complete scientific message across other than by recording such facts, normally in a tabular or graphical environment. Tables and graphs, after all, are simply attractive and valuable ways of presenting factual information for a consistent series of facts that have been aggregated.

RP: *You mentioned mashups. What else might you want to do with such data: Put it in a database so that you can mine it?*

P M-R: Yes, we might want to do that, or we might simply want to extract the data from a paper that wasn't presented in tabular form.

For instance, the carbon dioxide levels I mentioned could have been published as a graph in a paper, but I might need them in a tabular form. In such circumstances the only way I can get that data into tabular form is to use a software tool that is capable of understanding images. This process is often necessary with things like chemical spectra for instance. The tools for doing these things are still fairly primitive, but they allow us to extract individual pieces of data, and also to repurpose them.

RP: *So the first problem you face is that the data you need are not easily accessible, because they are embedded within the text of papers. However, you are developing tools that are able to extract it, presumably using techniques like screen scraping. The second problem, I believe, is that some publishers prohibit text mining of their papers?*

P M-R: That's right. And remember I could say to a publisher: "I plan to extract this data by hand and type it up". Now I don't believe that any publisher would try to stop me doing that, and it is perfectly legal for me to do it. But what you are not allowed to do — perhaps not so much as a general reader but as a member of an organisation that subscribes to the journal — is to text mine the data electronically.

RP: *Do most scholarly publishers prohibit text mining?*

P M-R: I can't give you chapter and verse on the precise details, because I don't spend my time talking to the serials librarian, but I am confident that you would find that certain publishers do not allow text mining of the full text that they serve.

In effect, anything that comes through a closed portal is likely to be barred from text mining.

Inaccessible

RP: *In your paper you cite a specific example of this. Last year a University of Michigan PhD student called Shelley Batts copied a graph from a paper published in a Wiley journal on to her website. Wiley responded by threatening her with legal action. She was able to quickly solve the problem, however, by retyping the data.*

P M-R: And what a supreme waste of time.

RP: *Nor is it a solution for you, since you want to do this at an industrial scale.*

P M-R: Sure. So as you say, we face two problems with embedded information. One is that while we have various text mining methods to enable us to extract this data it is a far from perfect solution, and the tools remain somewhat primitive.

Second, there is a general view amongst closed-access publishers that the full text is sacrosanct. They would say, for instance: "We own everything in the PDF file containing the article, including the embedded data". In fact, they would probably not even understand what I was talking about were I to ask them for permission to extract the data.

RP: *And yet it is not clear on what legal basis publishers can prohibit text mining.*

P M-R: That's true. But it's a complex area. While chemistry doesn't use photographs very much if, say, you published a photograph of a gel in a bioscience experiment a publisher might claim that the image was copyrighted.

RP: *Because copyright treats photographs as creative expression.*

P M-R: Right. However, I would argue that that image is a fact — because it is the way in which proteins are reported. Indeed, there is no other way of doing it. For that reason it is not a creative work, it is a fact. So you may have four bands representing the proteins, and I would want to run software over that photograph to extract the intensity of each band and their distance apart.

I should add, however, that the situation is not even across the sciences. Some publishers, for instance, go to considerable efforts to make data technically available. That is the case with crystallography data for instance.

So at one end of the spectrum you have journals that help people get hold of the data — and for this reason the accessibility of CiFs, for instance, is on a par with genomic sequences, protein structures, and things of that sort. At the other end you have, say, chemical structures — which are published in the literature, but are often not in a technical form to enable them to be easily accessed.

RP: *So it is difficult to extract embedded data from scholarly papers, both for technical reasons but also because some publishers prohibit text mining of their journals. As I understand it, there is an* *additional problem. In your paper you talk about the problems of accessing "supplemental information". What is "supplemental information"?*

P M-R: It is information that is too large or "boring" to fit in the full-text, but which is necessary for a reader who wants to be sure that the experiment described in the paper has been carried out correctly, and that the right conclusions have been drawn.

***RP:** You say that many publishers insist that — as a condition of publication — authors have to assign ownership of the supplemental information that accompanies their paper to the publisher?*

P M-R: That's right. And again, the supplemental information files are simply a collection of facts. As such, they should not belong to anyone. In almost all cases we are simply talking about a record of the experiment, which will likely include temperatures, materials, and analytic results etc. Or it might be just a copy of the computer output of a simulation.

***RP:** In your Serials Review paper you say that you were first alerted to the issue of Open Data when you and Henry Rzepa submitted a manuscript to the Journal of Chemical Information and Modeling, published by the American Chemical Society. You were surprised to discover that the copyright transfer agreement stated that by signing it you were also transferring copyright in the supplemental information.*

P M-R: And since it was all factual information we assumed it was an oversight, since facts cannot be copyrighted. We discovered, however, that this was deliberate policy, and the requirement to sign it was only waived in special cases.

***RP:** And you did in fact get a waiver didn't you?*

P M-R: Yes, but I think Henry and me are the only authors who have managed to get a waiver. So in calling for Open Data I am trying to bring to the community's attention the enormous value of releasing this information into the public domain, and in a form where it can be re-used. The problem in chemistry today is that it is incredibly difficult to get hold of this data.

The fundamental problem is that while there are literally millions of molecular structures published each year most of this information isn't placed into the public domain. We want to see this information released into the public domain because we need to be able to use all the new Web tools for doing science. Not only do we need to be able to find this information, we need to reuse it too.

***RP:** As you say, publishers have no legal right to claim copyright on these data, since they are merely facts, not creative expression. No one can own a fact.*

P M-R: That's correct.

***RP:** This is not so much a copyright issue, but a contractual issue presumably. Publishers say to authors: "When you sign this contract with us you the author agree to assign ownership of these data to us as the publishers". One might, however, wonder about the legality of such a contract.*

P M-R: Sure, and I suspect that none of these authors — other than Henry and me — actually understands the situation. First, they don't realise that when they sign the transfer of copyright form they are also signing over ownership of the supplemental information, despite the fact that the public domain is being disadvantaged as a result. Second, they don't realise that it is a meaningless act legally.

So while there are technical issues, our greatest concern is that these data are implicitly covered by copyright. And that means that if we use them we have to live in fear of the publisher sending legal representatives after us.

The real attraction of supplemental information, by the way, is that in principle it is better structured for certain types of data and so it is easier to extract. Again the situation varies across disciplines, and among publishers. The Royal Society of Chemistry, for example, does not insist that supplemental information becomes the property of the journal in which it is published , and it makes the supplemental information freely available on the Web — which means that even if you are not a subscriber you can access it, and so no accessibility problem arises. In other cases there can be very real accessibility problems.

Two main problems

RP: *The implication of what you are saying is that even if a publisher doesn't claim to own the supplemental information, it may place these data behind a subscription firewall, making them inaccessible to all bar subscribers?*

P M-R: Exactly. There is no problem if the supplemental information is freely available on the web ("libre"). If, however, it is copyrighted by the publisher, or not made freely accessible, there is a problem.

RP: *You mentioned genomic sequences. I guess the first time that Open Data became a real issue for science was when the controversy erupted over the Human Genome Project — it was feared that the private company Celera would appropriate some of the data?*

P M-R: Absolutely right, yes.

RP: *The upshot was that the human genome data is freely available on the Web today. I wonder if perhaps the controversy over Open Data is essentially a by-product of the Web. Would it, for instance, ever have become an issue if the Web hadn't developed?*

P M-R: That's an interesting question. I think we can say that if information hadn't become electronic it would never have become an issue. The biosciences have had online information for a long time: we have had protein sequences and protein structures, and nucleic acid structures all available in electronic form since the early 1970s. That clearly wasn't anything to do with the Web.

RP: *You are talking about the period when proprietary online services like Dialog and DataStar were developed?*

P M-R: Right. You also probably wouldn't talk about the Web in the context of most of the 80s either. Yet during that period a great deal of science was being done by combining the different electronic sources that were becoming available. People would order tapes of the data for using internally, for instance, or later they would go to ftp and gopher sites, and download data from there.

So although the Web has made these data much more accessible in most fields, Open Data is an issue that predates the Web.

RP: *Ok, so the key change was the development of electronic databases, not the Web. And what was significant about this change was that the electronic medium made data much more malleable: It allowed different data sources to be seamlessly merged; it allowed data to be searched far more effectively, and randomly accessed?*

P M-R: Precisely.

RP: *And it is for this reason that publishers tend to be excessively proprietorial about their data. And this is presumably why Wiley threatened Shelley Batts with legal action. Once they are in electronic form, data can be so easily copied and limitlessly distributed by third parties.*

P M-R: Yes but remember that we are talking about factual data here. And as we said, if you retype the information from a paper you can copy it perfectly legitimately. By doing so you can avoid any claim that you are violating copyright, or violating contracts with publishers, suppliers and so forth.

RP: *Retyping requires far more effort of course!*

P M-R: Indeed. In fact, I have hearsay evidence of at least one company that has spent half a million pounds typing up the information in the ACS *Journal of Medicinal Chemistry*.

RP: *You said that not all publishers impose accessibility barriers. How widespread a problem is it?*

P M-R: Well, I haven't done an exhaustive survey, and so I don't know exactly. There is an awful lot of work still to be done simply in order to establish the details for each publisher.

RP: *I guess what is needed is an Open Data equivalent to Project RoMEO — which catalogued the different policies of scholarly publishers with regard to permitting researchers to self-archive their papers after they have been published in a journal?*

P M-R: That's right. So we still need to do a lot of research. And I'm sure that there are many publishers out there who simply don't appreciate that they are creating a problem and who, if told about it, would probably remedy the situation. So we certainly shouldn't tar them all with the same brush.

That said, quite regardless of any copyright issues, I am pretty sure that many of them put restrictions on text and data mining. We certainly know that some publishers are very restrictive about what people can do with their online services. We had a case here in Cambridge, for instance, that demonstrated just how closely they police user activity.

RP: *What case was that?*

P M-R: You can find details on my blog. What happened was that a student of mine was extracting terminology from an ACS publication. To that end he was reviewing 20 articles. In the process of doing so he mistakenly loaded all of them in to his Firefox browser at the same time. There is nothing illegal in doing that, but within a second he had received a notice from ACS saying that the whole of the University of Cambridge had been cut off from all of its publications until further notice.

RP: *ACS prohibits 20 papers being opened by one user simultaneously?*

P M-R: That's right. The system treated his action as evidence of serial download. So an action that could only be viewed as perfectly legitimate was judged by the publisher as an illegal act, and treated as evidence that Cambridge University had broken its contract with ACS.

***RP:** To recap then, Open Data in science faces two main problems. There is the issue of embedded data, which is not only technically difficult to extract from papers, but treated by some publishers as an illicit activity. And there is the issue of supplemental information, which is technically well structured but is frequently appropriated by publishers, who restrict access to — sometimes by directly asserting ownership, sometimes simply by placing it behind a financial firewall?*

P M-R: Correct.

Large subscriptions

***RP:** I'm wondering if perhaps it is no accident that it is a chemist that has taken up the cause of Open Data. There is, after all, a long history of selling chemical information, which attracts high prices. Nor is it surprising that ACS has become the most aggressively proprietorial chemical information provider. Although it is a non-profit organisation, ACS has a long history of extracting data from journals and then selling access to them via the electronic databases run by its Chemical Abstracts division. This surely explains why ACS is keen to acquire ownership of the embedded data and the supplemental information attached to the papers it publishes?*

P M-R: Yes.

***RP:** How did a learned society like ACS morph into an extremely wealthy organisation whose profits are primarily derived from selling chemical "facts"?*

P M-R: Well, I can give you my amateur historian's view of developments. The first thing to note is that during the middle of the 20th Century it became clear that there was a need to abstract and index information — and I think everyone would agree that in the chemical field Chemical Abstracts was genuinely innovative, and did a valuable job for the community — because at that time there was no other way of getting hold of the information.

Then, during the second half of the 20th Century, there was a rapid rise in industrial chemistry — particularly in the pharmaceutical industry — which was predicated on getting high quality rapid information about molecular structures. It turned out that the data that Chemical Abstracts had been extracting from the literature was absolutely fitted for that market.

The other point is that while other types of information are resold, they tend to require much more human input when it comes to formal extraction and annotation. If you want critical numerical data, for instance, it can take a lot of human effort to extract it from the literature.

***RP:** What kind of information are you referring to?*

P M-R: One example is the data for chemicals used in industrial processes. It's very important to know how the boiling point of a solvent varies with pressure.

But the answer to your question about ACS: Chemical information became absolutely critical to industry, so all the major businesses in the field were prepared to pay large subscriptions to get it.

***RP:** Presumably ACS was not the only company to develop chemical information services?*

P M-R: No they weren't. The Beilstein database was created around the same time. This too was developed by a non-profit organisation — the Beilstein Institute in Germany — but at one stage the rights to that database (not the database itself, but the rights to license it) were bought by Molecular Design Limited, a company that was subsequently bought by Elsevier. So today you have two important chemical databases: the Beilstein database and Chemical Abstracts.

Chemical Abstracts, by the way, is now often referred to as SciFinder, which is one of the ACS products that searches over the Chemical Abstracts database. SciFinder is very widely used by industry, and it has also been productised it for academia. Many people in academia now view SciFinder an essential tool for their work.

***RP:** Perhaps there is an analogy here with patent information. Historically patent information — which is public domain information — was very hard to extract from patent documents. Spotting a market opportunity, a number of patent information companies developed, and made a good business out of it extracting the information and selling it to business, initially in paper format, and then in electronic databases. When the Internet became available, however, patent offices began to make the data directly available on the Web themselves, which has posed a considerable threat to the commercial providers. Perhaps it is the same story with chemical information: Information that was difficult and time-consuming to extract from paper documents (journal articles) can now be directly distributed by the originators — scientists. And this threatens to make intermediary organisations like ACS increasingly superfluous, or at least poses a threat to their profitability?*

P M-R: Absolutely right, although we should distinguish between patent information and the patents themselves, which give inventors a legitimate monopoly.

***RP:** Oh sure. I am referring to the so-called "bargain" between society and patent applicants: Society awards patent owners with a time-limited monopoly that enables them to exploit their invention, on the understanding that they provide details of how their invention works. So while the patent owner gets a 20-year monopoly, he or she only gets this on condition that the details of that invention are placed in the public domain.*

P M-R: Exactly, so as with chemical information we are talking about public domain information that has been aggregated and resold by patent information providers. In the days of paper, companies like Thomson and Derwent did a valuable job in indexing and re-purposing this information, but now the patent offices can themselves distribute this information effectively online.

I have, by the way, been in conversation with the European Patent Office, which is keen to distribute the chemical structure information contained in patent applications in Chemical Markup Language. The secondary publishers, patent resellers, are obviously not very happy about that. But there is absolutely no reason in principle why patent information shouldn't be routed directly to those who need it.

***RP:** Does the Web allow researchers to directly distribute chemical information in the same way? Are we yet at the point where if someone discovers a new chemical structure they can put it on the Web for anyone to access? That, presumably, is the purpose of tools like Chemical Markup Language?*

P M-R: Within limits, yes they can. That is certainly one of the things we are working towards, and we have two or three projects with JISC right now that are looking at the best way to do this. We have a number of similar projects with other funders too.

As it happens, chemistry is particularly good for this because an awful lot of the information is very well formalised: Analytical chemistry hasn't changed in 150 years, and spectroscopy hasn't changed in probably 100 years, except for NMR which is about 60 years old. This means that chemists are producing information that is very easy to put into containers. This in turn means that it is technically easy to distribute chemical information over the Web. In other fields — ecological systems for instance — it is much harder.

***RP:** We said that for historical reasons chemical information has acquired considerable commercial value. Does that mean that, within the science space at least, Open Data is primarily a chemistry issue?*

P M-R: No. It is an issue in many sciences. It is an issue, for instance, in many aspects of biology, and it is an issue for materials science; and it is one reason why Science Commons has set up the Neurocommons for neuroscience data. So it is certainly not a problem in chemistry alone.

***RP:** What about the arts and humanities?*

P M-R: I have to admit to a little ambivalence about that question. I do think it is valuable to draw a line between science and the humanities, not because some of the principles don't carry over, but because not all of them do.

***RP:** What kind of differences do you have in mind?*

P M-R: In the arts and humanities there is more concern about things like context, and there is more concern with format. You might, for instance, want to protect the font in which something is created, where in science you would never want to do that — unless you wanted to resolve some technical ambiguity.

And generally what you find is that the closer you get to the social sciences, and to political and economic areas, the more you run into problems over who has the right to access data, particularly when it comes to issues like privacy. At that point you begin to bump up against issues about the misuse of data and so on.

So I would want probably want to draw a line at the biological sciences, and I would start to huff and puff a bit when it came to areas like psychology. Essentially, the closer you get to data on individual humans the more you start to run into problems; problems that I am certainly not capable of commenting on, although I know that they are issues that Science Commons is exploring.

Open Access

***RP:** Presumably Open Data has a great deal in common with Open Access. After all, both movements share the goal of making the contents of scholarly papers freely available on the Web. What is their relationship?*

P M-R: My view is that we need to differentiate the two issues, and keep them separate.

***RP:** Why?*

P M-R: Because I don't believe that the various definitions of Open Access — the so-called BBB definitions — meet the needs of Open Data. Here I am talking mainly about the Budapest Initiative.

I should add that while in theory the various definitions of Open Access should be adequate, in practice they aren't.

***RP:** Can you expand on that?*

P M-R: Well, the wording of the Budapest Initiative, for instance, says, that Open Access is where anyone can "read, download, copy, distribute, print, search, or link to the full texts" of scholarly articles. It also says that people should be able to "crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers." The only constraint, it adds, is that authors should have control over the integrity of their work.

Now I view that statement as a meta-licence. As such, it should control the way that Open Access is implemented. If that's right, then any papers or associated data made available on the Internet, and described as being Open Access, should permit me to "crawl them for indexing, or pass them as data to software." And if I cannot do that for any reason then those papers are non compliant with the Budapest Open Access Initiative, and so not Open Access.

The problem is that not everybody agrees. There are people like Stevan Harnad, for instance, who believe that it is only necessary that the information is available for human eyeballs to access; that so long as the information can be read it is enough to call it Open Access. As that does not provide for text mining it is not sufficient for my purposes.

***RP:** You have debated this issue publicly with Harnad. As I understood it, Harnad disagrees with your criticism. He makes a distinction between the publisher's version of a paper and the author's version. Since the author's version will be different to the published version, he says, it will not be subject to any transfer of copyright agreement between the researcher and the publisher. Consequently, regardless of any prohibitions imposed by the publisher, you are free to text mine the author's preprint in his or her institutional repository.*

P M-R: Sure, but there are many Open Access papers that are never placed in an institutional repository. A researcher may, for instance, have published their paper in an Open Access journal, which is then exposed on the Web by the publisher, and defined as Open Access, but never placed in an institutional repository by the author.

Even if the paper has been placed in an institutional repository you are often likely to find that it has been stamped as being the copyright of a publisher. Moreover, I have no guarantee that the publisher won't assume that I have indexed it on the publisher's site, and in a way that is contrary to the contractual arrangements he has imposed.

In both cases, therefore, I simply don't know what limitations have been placed on what can be done with the paper.

I am, by the way, not a great fan of self-archiving.

***RP:** Why?*

P M-R: Because of the way that scientists work. Scientists want to get rapid views of a large amount of data. For this reason we use PubMed as our primary source.[1]

[1] PubMed is a free search engine for accessing the MEDLINE database. It includes over 16 million citations and abstracts of biomedical research articles and other life science journals. The core subject is medicine, and it covers fields related to medicine, such as nursing and other allied health disciplines too. In addition, it provides

RP: *PubMed is not a full-text service. It is the NIH's database of abstracts of scholarly papers.*

P M-R: That's correct. And our robots can sift through half a million PubMed abstracts a day. After we have gone through those abstracts we will then want to see if we can get the full text. And I can tell you we are not going to go round trying to find whether somebody has put a paper on a departmental web site, or bunged it into some institutional repository without proper metadata. We simply want to get the full text.

RP: *As I understand it most institutional repositories use the OAI-PMH*, or *Open Archives Initiative Protocol for Metadata Harvesting, which allows harvesting services like OAIster to aggregate data from all compliant IRs, and treat them all as one virtual archive. Are services like OAIster not sufficient to your needs?*

P M-R: OAI-PMH is useful for exposing single documents but does not provide a mechanism for iterating through a systematic collection. We have, for example, put 200,000+ molecules into the Cambridge DSpace repository but while it is possible to access single instances there is no easy way to download all of them.

Major failing

RP: *For clarity's sake we should distinguish between self-archiving (the so-called green road* to Open Access), and *Open Access publishing (the gold road). Papers self-archived by their authors are likely to have been published in a subscription journal, and the publisher will have acquired copyright as a condition of publishing. And, as you said, where a paper has been published in a gold journal the paper will have been made available on the Web by the publisher itself. Would I not be right to assume that if an Open Access publisher has made a paper available on the Web you are free to text mine it?*

P M-R: Not necessarily. When it comes to Open Access journals I am concerned that there has been no real attempt to examine and catalogue the various licence conditions imposed by publishers.

Some of these conditions, by the way, are byzantine, and the terms tend to be spread over so many pages — using words like "must" and "shall" in strange ways. The end result is that they are almost impossible to understand; they are also often self contradictory, and so it is almost impossible to tell what the contractual obligations of a reader are, and what the contractual obligations of an author are.

RP: *So the problem is that Open Access publishers do not sufficiently explain what rights you have to text mine the papers they publish?*

P M-R: The problem is that the Open Access movement has made no effort to develop its own licences. And in the absence of licences embedded in the documents themselves, or explicitly stated on the web site, it is not possible to know what it means to call a journal Open Access.

RP: *You are critical of the Open Access movement in this regard?*

very full coverage of the related biomedical sciences, such as biochemistry and cell biology. It is offered by the United States National Library of Medicine at the National Institutes of Health as part of the Entrez information retrieval system.

P M-R: Correct. It has been a major failing of the Open Access movement that it has not developed the necessary licences. I feel very strongly about the lack of licences, and indeed of the lack of what I would call a meta-licence — that is, something like the Open Source Initiative definition [OSD], or indeed the Open Knowledge Definition.

***RP:** You earlier described the Budapest Initiative as a meta- licence. You also said you believe that in theory the definition of Open Access it articulated meets the needs of Open Data. That suggests that it is not so much a failure of the Budapest Initiative, but that people are inclined to ignore its definition of Open Access, or they disagree with your interpretation of the definition?*

P M-R: Well, I believe that my interpretation of the Budapest Initiative is a more correct interpretation than Stevan Harnad's, even though he is a signatory to it and I am not. So, yes, I believe that my interpretation is closer to what the words of the Initiative say, although obviously I can't talk about the spirit of the Initiative, because I wasn't there.

However, there is a second problem: while you could argue that the Budapest Initiative is a perfectly acceptable meta-licence, the problem is that it doesn't itself say anything about licences. If you compare it with the Open Knowledge Definition, for instance, you will note that the OKD states that all objects that claim to conform to the definition should carry a licence, and that that licence should be compatible with the definition.

***RP:** Presumably the OKD assumes some form of Creative Commons licence?*

P M-R: Right. It says that the CC-BY and CC-SA licences meet the Open Knowledge Definition, but CC-NC does not. The OKD, by the way, also says that the community will police the use of the definition. And it is a great problem that nobody has done this in the Open Access movement.

***RP:** As you pointed out earlier, there is no single definition of Open Access. Perhaps that is part of the problem. You are also saying that the Open Access movement should have done more to agree a universal definition of Open Access, and to have policed that definition?*

P M-R: Well, the fact of the matter is that the Open Access movement could have saved itself an awful lot of problems if six years ago, in the wake of the Budapest Initiative, people had started to investigate — as they did with RoMEO — who conforms and who doesn't conform to the definition of Open Access. And also why they conform, and why they don't conform.

As it is, the movement simply hasn't addressed these issues, and nobody at all seems to care about licences. And it is very, very much more difficult to retrofit the issue than to have addressed it upfront.

One result is that we have ended up with all these hybrid approaches — many of which in my view are counterproductive to the Open Access movement.

***RP:** Here I think you are referring not to Open Access publishers, but to traditional publishers who have begun to offer Open Access options — telling researchers that if they pay a fee their paper will be made Open Access. As I understand it, in some cases these publishers still expect authors to assign copyright to them?*

P M-R: Exactly. There are a number of publishers who offer free visibility in return for author charges, and most of the major publishers now have some sort of hybrid option. The problem is that most of these options cannot claim to be Open Access. In some cases the authors pay a large amount of money but get very little more than they would get with the traditional publishing model. It is no surprise to me therefore that there has been very little uptake of this model in chemistry.

RP: *Essentially, your criticism of the Open Access movement is that while the description of Open Access articulated in the Budapest Initiative implies that you should be free to text mine Open Access papers, the movement has failed to provide sufficient legal certainty to allow you to do this.*

P M-R: Absolutely, and Peter Suber made this case very clearly: it is about the law. That phrase sums it all up. If it goes to court, even if I think I am right, I am likely to be accused of breaking the law by some publisher.

Open Data Protocol

RP: *Nevertheless, most Open Access journals have adopted Creative Commons licences haven't they? Do these not provide a suitable licensing environment for science journals?*

P M-R: Actually, that is not correct. Many Open Access publishers have not addressed the issue of licences at all.

For instance, we recently started going through 50-something chemistry Open Access publishers that were labelled as Open Access in the Directory of Open Access Journals [DOAJ]. As we went though we discovered that some of them didn't have any licences at all, but simply said, "This is an Open Access journal."

And when it came to outlining what they permit you to do with the papers it was clear that many publishers hadn't given any real thought to the matter. In other cases, they had got as far as using a Creative Commons licence, but opted for one that is incompatible with the BBB definitions. Some, for instance, use CC-NC.

RP: *You don't believe that Open Access journals should use a non-commercial licence?*

P M-R: I don't. And to go back to your earlier question: No, I do not believe CC licences to be technically the best solution for science journals.

However, that is a minor concern compared with the other issues we have discussed. In fact, I am very grateful for Creative Commons, because without it we would have had no way of applying licences. People would probably have used a GPL licence, or something else totally inappropriate.

So Creative Commons is very useful, and I don't have a problem with CC-BY or CC-SA. If someone uses CC-BY, for instance, then everybody knows where they are: It says that the author wants to make their work openly available, and the publisher is happy with that. You can then put a stamp on the paper indicating that both parties have agreed, and readers will know what they are allowed to do with the work in question.

I am, however, pleased that Science Commons has recently developed a more appropriate approach for science.

RP: *You are referring to the Open Data protocol. As I understand it, this is a meta-licence intended, as the protocol puts it, "to conform to the Open Knowledge Definition and extend the ideas of the Budapest Declaration to data and databases."*

P M-R: Exactly. And it is also developing a range of licences, the first of which is the Public Domain Dedication & Licence (PDDL).

RP: *Presumably the new public domain licence developed by the Creative Commons — CC0 — would fit the bill too?*

P M-R: Yes it would.

RP: *An important difference between these licences and the traditional Creative Commons licences is that they are designed to put the data into the public domain, rather than provide a way for creators to give away some of their rights, but retain others. One potential problem with putting data into the public domain, of course, is that private interests can appropriate them, and take them back out of the public domain.*

P M-R: True, and that is a problem I ran into several years ago.

RP: *In what way?*

P M-R: I had a colleague at Glaxo — Roger Sayle — who wrote a program called RASMol. He published the code with a statement saying that he was putting the program into the public domain. A little later a software company took that software and put it into their own products, without acknowledgement. They then put their own licence on it, and since Roger had put it in the public domain they were perfectly entitled to do that.

RP: *One way of avoiding that, I guess, is to flood the world with copies of the data in question. Then if someone does seek to appropriate any one copy there are still plenty of free copies available. I believe that is the approach adopted by the Proteome Commons, which uses a p2p, hash-verified system. So long as the original data remain freely available no one can monopolise them.*

P M-R: Yes, as I understand it, the level of risk here is a factor of how well the public domain is defended. I'm not a lawyer, so I don't know how real the problem is. However, I do trust people like John Wilbanks, Paul Miller and Jordon Hatcher — the people who developed the Open Data protocol, and who understand these things better than I do. Certainly I would much rather people put things into the public domain than that they didn't make them available at all.

Permission barriers

RP: *It seems to me that the key difference between Open Data and Open Access revolves around the issue of reuse rights. As you said, it is enough for Open Access advocates that scholarly articles are freely visible to human eyeballs. Advocates of Open Data, by contrast, want to be able to reuse the data in them. To do that without risk they need legal certainty, and thus clearly-defined licences permitting such usage. Essentially, your dissatisfaction with the Open Access movement is that its main focus is on removing price barriers, rather than what Peter Suber calls "permission barriers".*

P M-R: Correct, although actually Peter Suber agrees with me on this.

But you're right: none of us are fighting about price barriers; we are talking exclusively about permission barriers. After all, if something is Open Access, and it can be accessed by anyone, then there are no price barriers.

RP: *Does that mean that if the Open Access movement had been as committed to removing permission barriers as it is to removing price barriers, you would not feel the need to differentiate Open Data from Open Access?*

P M-R: That is probably true, although there is a wider need for Open Data beyond Open Access. Certainly I had the algorithm in my head that anything that is Open Access also assumes Open Data. I discovered, however, that that is simply not the view of Stevan Harnad.

RP: *Although you said that Peter Suber agrees with you?*

P M-R: Sure, but while Peter Suber agrees that we need to remove permission barriers, he tends simply to say, "It is a pity that there are Open Access products that do not remove permission barriers, but since this is accepted usage within certain parts of the community I will not denounce it."

What this means is that Peter is prepared to let people promote journals as being Open Access even if they do not remove the permission barriers. Moreover, he may not be prepared to challenge them over this — other than to blog a comment saying something along the lines of, "This removes the price barriers, but it does not remove the permission barriers."

RP: *To summarise: while the aim of the Open Access movement has primarily been on ensuring that people can read scholarly papers, you view them as a kind of database, and want to be able to extract the factual data from them. We should perhaps point out, of course, that back in 1994 — when Stevan wrote The Subversive Proposal — concepts like Web 2.0, and mashups, didn't exist. And while the 2001 Budapest Initiative does talk about indexing papers, and passing them as data to software, the intention of that wording, I suspect, was to ensure that search engines would be free to index them, not to permit text mining. This perhaps tells us that the needs of Open Data were not ignored, but simply unknown when the Open Access movement first developed.*

P M-R: Oh, absolutely. And let me go on record as saying that Stevan has done a marvellous job in alerting the world to the whole Open Access issue, and I wouldn't want to take anything away from him in that respect.

His work has been remarkable and important. He has taken a very consistent, but a very simple, view of how to achieve his aims. The 1-click submission of your manuscript for instance, is a political slogan, and I don't detract from that, and I don't criticise it. He has done a great job there. The point to stress, however, is that Stevan's approach *is* a political one, not a technical one, and a political approach that hasn't always embraced the technicalities.

As an experimental scientist dealing with large amounts of data, I have a different view to a cognitive scientist, or someone like Peter Suber, whose background is the arts and humanities.

Official organisation

RP: *Do you think that some of the problems you see in the Open Access movement might have been avoided if there had been an official Open Access organisation — an organisation like, say, the Open Source Initiative (OSI). Such an organisation could, like the OSI, have taken responsibility for developing a set of licences, and it could also have policed their use?*

P M-R: Absolutely. And there is nothing in the laws of politics that says that such an organisation shouldn't exist.

RP: *You would be in favour of creating such an organisation today?*

P M-R: Well, if I move for a moment from my interest in Open Data to Open Access then, yes, as an Open Access advocate I would certainly be in favour of an official organisation, because there are many issues of Open Access that have been shuffled under the carpet, and right now we are not doing anything to enforce Open Access.

RP: *Do you think that there ought also to be an Open Data organisation?*

P M-R: Yes.

RP: *A separate body from any Open Access organisation?*

P M-R: Yes, they should be separate.

RP: *Do you personally have any plans to establish an Open Data organisation?*

P M-R: Are you offering me money?

RP: *I'm afraid not. But your response goes to the heart of the problem that the Open Access movement has faced I think: You cannot create an organisation without money, and without an organisation it is difficult to do the kind of things you believe the Open Access movement should have been doing.*

P M-R: True, but then Open Access is not at the absolute top of my agenda. Open Data is. And so far as Open Data is concerned I believe that Science Commons — combined with the Talis approach, and combined with the efforts that have come from the Open knowledge Foundation, is doing a pretty good job of getting us to the first milestone in the journey we have to take with Open Data.

It is also conceivable that the Open Knowledge Foundation could do for us what the OSI has done for Open Source software. That said, of course, the OKF goes beyond scientific data, since open knowledge relates to all sorts of other things that are not relevant to scholarly publishing, and which are not related to science.

So perhaps the best immediate hope is that Science Commons will take on that role, and promote Open Data. Indeed, I wouldn't be surprised if — behind the scenes — they weren't already looking at this sort of thing.

RP: *Could you envisage Science Commons taking responsibility for both Open Data and Open Access?*

P M-R: Do you mean Open Access in science or Open Access in general?

RP: *Open Access in science.*

P M-R: That would be a very interesting development. It could certainly catalyse things, and it would surely be a wakeup call to the rest of the Open Access community. I hadn't thought of it, but that sounds like a pretty good idea. However, I would still want them to treat them separately.

Either way, I see Science Commons as something that needs to be strongly nurtured, and I believe that is has got the track record, and the right people, to enable Open Data to succeed. Their co-membership of the OKF also means that at the moment all the people in this area are working in roughly the same direction — politically they are all aligned. So the difficulties we face are only technical, although very complex for all that.

Strongly aligned with Open Source

RP: *We are witnessing the growth of many different open and free movements today, with new ones being formed all the time. Where does Open Data fits into this larger picture?*

P M-R: I think Open Data's roots are very strongly aligned with Open Source. The phraseology, the technical approach, and the fact that people are looking at licenses at a very early stage, all suggests that it can draw a lot of strength from the Open Source movement. What this also means is that it doesn't need to explain many of its principles, or at least meta principles, to people — because they are enshrined in what Open Source has already done.

Those things don't really spill over from Open Access in the same way.

RP: *What do you view as the priorities for Open Data?*

P M-R: The first thing we have to do is to define what we are talking about. So, for instance, are we talking about maps, are we talking census data, are we talking about this, are we talking about that? The priority should be on working out exactly what we are talking about when we use the term Open Data.

RP: *A good starting point!*

P M-R: Yes, but it's a complex business. Open Access only really applies to scholarly publishing — in the way you and I use the term. So it is obvious what it applies to: it is something emitted by a scholarly publisher.

But with Open Data it is going to be quite difficult to define what exactly the term applies to, and we are going to have to come up with algorithms to do that. It won't be terribly easy, but it needs to be done.

Once we have done that I suspect that each sub-community will then need to work out what it wants to happen to its data. That too will vary. If you talk about open chemistry data, for instance, the situation is relatively straightforward. But when you start talking about, say, open genomic data, you need to consider whether that means that anybody can take data related to individuals and stick it in a government database.

RP: *This goes back to the privacy issue you raised earlier.*

P M-R: Yes. So the issues are different within different subject areas, and we are going to need to come up with domain-specific approaches. But then that is one of the features that runs over the whole of scientific data: different communities have different expectations.

This point was well put by Andrew Lawrence at the Science Commons meeting held in Washington about fifteen months ago. He pointed out that even within physicists there are three completely different approaches to Open Data.

So, for instance, particle physicists have very clear mechanisms for releasing their data, and it is all done on an industrial organisational scale. It's known who owns the data, when the data is going to be released, in what form it will be made available, and who is going to have access to it.

Astronomy, on the other hand, is one of the archetypal mashup subjects, and they do their best to share different types of data from day one. So you will have different wavelengths of the sky, different types of community observing it, and so on, but they all do their best to build a Web 2.0 approach to sharing the data.

Then you have condensed matter physicists, who take the approach: "This data belongs to our laboratory, and we are not going to let anybody have it. And when we have finished the experiment we are not going to archive it."

This brings us to something else that needs to be stressed: there is today a tension between the funding of research, and the availability of the data produced by that research. If you take a subject like condensed matter physics, or materials science, for instance, the funder will say: "Ok, do this work, and write up what you have done."

In the process of doing that research people will collect all kinds of data, and then publish a paper. However, as far as the funder is concerned, that is the end point, and the paper is seen as the sole property associated with the project.

***RP:** You think that funders ought to take a greater interest in the data that arises from research projects?*

P M-R: Exactly. Even though they fund huge instruments, and provide lots of money for a research project, they don't insist that the data created in the process is preserved. In the UK, the EPSRC is particularly lax in this. It simply says, "It is up to the research team whether they make the data available. We are not interested whether it is exploitable or protectable."

***RP:** As we discussed at the beginning of this conversation, your focus is on what Tim Berners-Lee calls the semantic web — a web in which much of the hard work in terms of discovering and processing information is done by machines, not humans. This requires extensive text mining, and then the re-assembly and re-aggregation of that data without having to worry about legal or technical obstacles. As more people think through the implications of the semantic web the issue of Open Data is attracting the attention of more and more people, not just Science Commons. For instance, the Open Data Commons was established recently. There is also a W3C Community Project called Linking Open Data. As part of the latter project, by the way, DBpedia plans to serve RDF for all 1.6 million concepts in Wikipedia. This is not a science project, but the overall aim is very similar to what you are trying to do with chemical information. The key difference, of course, is that all Wikipedia content is licensed in such a way that the data can be reused. As we've seen, the challenge you face is that most of the information you want to mine is not suitably licensed?*

P M-R: Actually, we are very keen to use the DBPedia approach, and are starting to convert Wikipedia chemistry into RDF. The attraction here is that it's relatively small, very high quality, and of great interest to many people both chemists and not. So maybe this start will catalyse the liberation of data.

eScience

RP: *I would expect that the growing interest in eScience will help too. Amongst other things, eScience assumes that scientists will share large data collections. Logically one would expect that these collections will be open rather than proprietary?*

P M-R: Absolutely. It is very difficult to do eScience without Open Data. I have been strongly involved with the UK eScience program over the last five years, and one thing we have learned is that when you have controlled access to data and resources, and when you therefore have complex authentication systems in place, you face very difficult problems — because as soon as you impose these things on scientists they just walk away.

RP: *Because access controls introduce too much friction into the process?*

P M-R: Yes. When scientists are surrounded by clumsy authentication systems, resource allocation, or legal constraints on what they can do with the data, they decide that they don't want to work with remote data and resources.

Of course, in some fields this is a necessary condition — in medicine for instance — but in nearly all other sciences, you find that the ones that flourish are the ones that haven't had to worry about whether the data are open. Bioinformatics, geodata, and oceanography, for instance, have all done well — because the data are open, and people are able to move them around. And where the data are not strictly open, there is a tradition that you can get hold of them. Nobody ever says, "If you want access to these data you are going to have to licence this information from here, and that information from over there, and you will have to pay so much per click for using it.

It is precisely for this reason that chemistry has been so spectacularly unsuccessful in eScience, and the US equivalent, the cyberinfrastructure . Because there isn't any Open Data, and chemists don't want to share data.

RP: *All the more reason for you, as a chemist, to push for Open Data!*

P M-R: Absolutely.

RP: *Ok, thank you 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.