Interview of Peter Murray-Rust for OpenForum Academy, 17 March 2015

Maël Brunet : Hello Peter and welcome to this first installment of 'meet the fellows'. Can you start by briefly presenting yourself for people who don't know you and talk a little bit about what your current research interests are?

Peter Murray-Rust : Hi there I'm Peter Murray-Rust, I'm a scientist, I'm a chemist who for the last 20 or 30 years have worked particularly with information in chemistry, and now I've moved on to other sciences as well. So what I believe is that information, publicly available to anybody is at the core of scientific research. It's not just what you do in your laboratory, it's also what other people have done and how that can be integrated into a better view of science, technology and medicine. What I found over 20 years is that it's often incredibly difficult to find scientific information that ought to be available. Sometimes this is due to laziness, sometimes it's due to lack of funding, but often it's because people want to keep this information to themselves, and aren't prepared to release it – and that's true of both academics and companies. So my current goal is to liberate scientific information, on a massive scale, by doing this with technology. I've been very fortunate for the last year that I've been funded by the Shuttleworth Foundation, and as a Fellow of that organisation, I and my team are building software which is going to automatically read published science from journals and other source and turn it into semantic form so machines can understand it and then make it available to everybody.

MB : Great, thanks for this introduction. So, how would you say that this area – open access if you call it this way – relates to other areas of openness ? This is a topic that we're interested in at OpenForum Academy, to see how different Fellows, each in their respective fields, have a common interest.

PMR : That's one of the really exciting things, because when I came to Cambridge, 13 years ago, very shortly after that I met up with Rufus Pollock, who founded the Open Knowledge Foundation, now known as just Open Knowledge, which explores the question of what is open, why is open valuable, and how can we use open to change the world. Open Knowledge has covered a wide range of things like copyright, the value of the public domain, through to Open Government, how can we make government processes transparent, open medicine, open clinical trials, and my own area of open science, which is how can we make science available to everybody. Open isn't just the permission to read something, it is the permission for everybody to use and re-use that, and increasingly that means machines, so I'm looking at how machines can read scientific information, digest it, and then help humans make best use of it. But further upstream is the question of how the process is governed, how the materials are created and what the process is. So open isn't just a passive thing, it's not just about consuming information, it's how you carry out the operation, whether it be scientific research, whether it be government, whether it be medicine, so that we know what's going on, we have transparency, we can measure that transparency, we have an input when things go right or wrong, and we can get the information as soon as it's made available, which is often at a much earlier stage in the process than simply publishing the whole lot at the end in some dense set of documents.

MB : You talked about liberating scientific information using software. This is often referred to as text and data mining, but I think you prefer to call it 'content mining'. Can you expand on that distinction and how for example mining a text database might differ from mining other types of information ?

PMR : Historically, science and much other communication has been through print and science is still communicated through objects that look like printed pages. Let's call it 'e-paper' rather than being full semantic information. So people get PDFs, which is a format which is reasonably developed for sighted human beings, but is no use for blind people and is no use for machines. So we've developed technology which can read everything that is published, not just the PDFs but we're also very interested in images and other types of data. So there's been too much concentration on the paper, textual representation, which is where the phrase 'text and data mining' comes from, and not on images and audio and things like that. I should say that I work on images, I don't work on audio which is much harder but I campaign for the rights for other people to extract information from it. Now the problem with this is that technically this is covered by copyright. Copyright is one of the most complex and I think broken systems that we have at the moment. There's often a special role given to images as being creative works and they have more protection in some cases than text. Now I believe that scientific images are often the only way of communicating something, so if you photograph a cell, that is a primary scientific information, and I would call that data rather than a creative work. Similarly there are expressions within text that can only be understood in conjunction with the diagrams in the document, and the diagrams are also fundamental science. So we've developed the term 'content mining' rather than simply text and data mining because we wanted to make clear that there's no line to be drawn between scientific text and scientific diagrams and images, that all of this is legitimately minable.

MB : Let's come back a moment to what you said about PDFs not being of any use for machines because of the lack of semantic information. Are there any other tips or best practices that you would like to highlight for people who are publishing content online, whether in the form of formal journals, 'grey litterature', or individuals just publishing their findings in a blog post?

PMR : When people publish something, they still have a very strong metaphor for the sighted human who can read it and the thing they consume is the photons that come from the screen rather than the semantic information. You can represent information at a variety of levels : the worst is when you only got a photograph, or a scan of a piece of text, and then all you've really got are pixel images, and those pixels may in some cases be rather blurred, and although the human will understand it, the machine or the blind person has virtually no chance of understanding anything. Then you come to characters and PDF in most cases transmits characters but the problem is that PDF does not structure the document. What it does is it describes the page, so if I've got something that reads say 'Cambridge' as 'C, A, M, B, R, I, D, G, E', I say that in that order but in PDF you can put those characters on the page in any order you like and it's where they are on the page rather than the order that is important. That makes it difficult for a blind person or a machine to understand it. Now, I and other people have developed software which can turn many PDFs into structured sequential text but it can't be done for all of them. It's a wide range of different qualities, the worse being scanned paper or photographs, and the best being completely semantic. And when we talk about completely semantic, this can be done in a well-created HTML, which is why Tim Berners Lee developed it in 1993, the tragedy is that people don't use it in the proper way. So if they create good HTML, it's a good practice for linking things in, for adding other types of information. Then we have a hugely better form of communication. I've played my part ind eveloping the tools for chemistry, something called the Chemical Markup Language, which is able to represenyt chemistry in a way that machines can read and understand and process. So if everybody used HTML5 so that it was compliant, we would be massively further than if people used PDF.

MB : Can you give some specific examples of ways in which people use HTML in an improper way, that is without respecting the technical specifications?

PMR : For example, if you take character sets, the unicode character set will hold all characters, fallbacks of characters which will hold everything we want to in the modern world. But there are many systems which don't use UTF8 and unicode encoding, and they're often very difficult to interpret and you will get corruptions, so if you have a character set with the greek letter µ, which is used for microgram, then some systems will turn that µ into 'm' when it becomes milligram. That's a thousand times heavier, and it may be absolutely catastrophic if you're talking about the dose of a drug or something like that, you could kill somebody! So it's this problem that we have reliable ways of sending information so we can verify it, but people don't use it. So it's encoding, they're not properly used. Many people don't support diacritics, so for example your name has a diactric, but many systems would destroy that. We also find it en masse in technical language, people often use things which are wuite innapropriate, so instead of a degree character (°) or a degree Celsius (°C), people will use a superscript 'o' (º). It looks the same but it's wrong, and that can be misunderstood. And that can be technically solved, it's easy to do but people just don't do it because they can't be bothered to upgrade their system.

MB : You touched already a little bit upon your project at Content Mine and what you're trying to achieve with this. Can you talk about how people who might be interested to get involved can participate?

PMR : Absolutely. Yesterday we had a wonderful meeting in Oxford at the Cochrane center which analyses clinical trials. For any new treatment, there are often ten different trials. Some will get positive reports of the treatment, some neutral, some negative. The Cochrane Center analyses this very carefully and then comes out with a considered judgement as to what the most appropriate treatment might be or decide that the treatment actually has no beneficial effect or even a negative effect. The problem with this is that many trials are not properly published, or they may be published in obscure private places, and there are people who may not want negative findings of files to be published. If you are a pharmaceutical company who is selling a drug then you don't really want negative trials to impact on your market. So you will often find that when you look at the distribitioon of this you get right-skewed distribution so only the positive results are present, and that's an indication that the negative ones haven't been published. Cochrane wants to be able to process huge amount of information. It's almost all in paper form or electronic paper form, and our technology will help read these documents and find out what the trial was about, what the outcome was, the number of patients (which matters a lot), how the randomisation of the trial was done, and many other things which will tell you whether it's a trial worth looking at or not. They are very excited about applying this technology to filtering out those trials which are likely to be valuable in their studies and also discovering ones from places where they might not otherwise be able to look. So that's just one example but it also applies to many different subjects of potential interest, so for example we're looking at species and we're monitoring the daily mention of species in the litterature and people interested in bio-diversity would vbe interested to know the diverse places in which the specie was mentionned so that they would have a better collection of the information about it.

MB : So if someone is interested in collaborating with this project, how can they reach you? Are you looking for any specific skill sets or input that you're looking for at the moment?

PMR : What we're looking for are people who have a well described problem they need to solve, they need enthusiasm and they need commitment to keep going at it as we need them to be working on it in 6 months time – something like that. We've got approximately five areas that we're starting with. The first is clinical trials, and as I said we're very excited about the potential and the commitment there. Then one of my Shuttleworth colleagues, Rory Aronson, is looking at collecting public information on growing plants, his project is called OpenFarm, and we're looking to see if we can extract information from agronomy journals and other things about the best growing conditions for plants and things like that. I have a growing collaboration with crystallographers, I'm a crystallographer myself, and we're looking to extract crysographic data from the litterature and make it publicly available. Then we've got a project that has been funded by the Biotechnology and Biological Sciences Research Council (BBSRC) which is looking at philo-genetics, which is how species evolve, and there are huge numbers of papers published on that and they all contain what is essentially tree of life diagrams so they're diagrams about how the specie branches and how it evolves. We're able to examine those diagrams and actually extract machine-readable information from them in less than a second per diagram, and that's with the university of Bath. The fifth one is my own particular interest which is what is the relationship between the evolution of plants and the evolution of chemicals that they emit : many chemicals emit odours, they emit pheromones for insects in it, defense molecules and so on, and we wanted to see does the chemistry evolve in the same way as the other measurable features of the specie. So that's five projects that we are starting with. We want people who are enthusiastic. Not everybody needs to be tech-savvy, we need people knowledgable in the domain but we also want general Unix hackers or people with a Unix mindset who will understand this – so any hacker is very valuable in these projects. And like all open projects it's a community, different people will find different things they're interested in. We are working very closely with Wikimedia, so we plan that our data is, when we've extracted it from the litterature and we've evaluated it we will put it in Wiki databases.

MB: Let's jump to a different topic. You mentioned copyright - are there any legal impediments preventing the take up of content mining, and if so, what would be your policies recommendations?

PMR: The answer is yes. Copyright has been described by MEP Julia Reda as incredibly complex and she is right, it is something which has evolved over the years and in many peoples' view is no longer fit for purpose and it actually seriously stands in the way of using modern technology for creating and disseminating knowledge. The problem is that most operations dealing with electronic information involve copying at some stage. Either downloading from a site on the Internet or transmitting it to other people in your community and technically if that information is protectable by copyright you have to ask the question: "Am I violating copyright in carrying out this act of copying?". Let's say I have a journal article, an article I have the right to read because it's either in the public domain, it's covered by an open licence, or, most problematically, if it is something that I subscribed to on a regular basis; most scientific information is only available in published form as journal articles through subscriptions and I can read it because I am in Cambridge University which has a large number of subscriptions. But even in downloading it from the publishers website I am making a copy and that is allowed by the contract with the publishers because otherwise I couldn't download it. But if I am now going to mine it that involves using that copy, and the publishers have, I think, very narrow-mindedly and counter productively decided that mining this information requires their permission. There is nothing in the law that backs this up, but the law does support copyright in copying materials so there is a grey area here in that I can copy it and read it but that the publisher says that I can't copy it and run my software over it. I have produced the mantra: “The right to read is the right to mine” in other words, if you have the right to read something you have the right to mine it. A lot of people support this idea, most recently it has been supported by LIBER, the association of European research libraries and we have come out with a declaration from a meeting in the Hague just before Christmas where among other things we have asserted that the right to read is the right to mine. Unfortunately, a large number of conventional publishers have opposed this, saying that you require addition permission and in some cases additional funding to be able to mine this content. The position is that we are in conflict with the publishers and in the UK, the government has taken a very pro-active action and has created an exemption for copyright which was proposed by professor Hargreaves so I will call it the Hargreaves exemption and in June 2014 it came into law in the UK as an additional statutory instrument. What it says is that a number of actions in copying are now legal in the UK without the permission of the copyright owner and they include things like copying for archives or by libraries, copying for format shifting, copying for parody so you can copy and release a parody copy and in the particular case of my community, copying for data analytics interpreted as text and data mining. We are allowed to do this for research purposes and for non-commercial use. This has been passed in UK law, it hasn't been tested in court but I believe that this legitimises what I do and we are going to go ahead and do it. A number of European governments and organisations are pushing for the European Parliament to legislate in a similar manner and this is what Julia Reda proposed earlier this year in a very balanced and valuable paper on copyright and the European Parliament where she proposed a number of actions which are similar to the UK government, some of them go beyond and one or two don't go far enough but it's basically very similar legislation. There is now fairly intense lobbying in Brussels from both sides on this issue and she said she has had numerous invitations to diner from rights holders who are lobbying against this reform.

MB: You mentioned the Hargreaves exemptions, and of course as you said this is also being discussed at the EU level. Some of the policy proposals would be to go for a wide exemption for text and data mining – or content mining – and some others are probably more aligned with the UK current framework to restrict exemptions exclusively to research and non-commercial purposes. In other areas, the non-commercial clause especially has generated much discussion and some concerns. I was wondering if you thought that this also applied to your domain and if there are any issues with distinguishing commercial from non commercial uses and pure research from other purposes ?

PMR: That is a very valuable question. My understanding is that the reason that the UK government chose the non commercial researches was that they could then enforce this change through a statutory instrument which didn't need a full act of Parliament whereas if they were to get a complete exemption for commercial use as well it would have to be debated in the House of Commons and then in the House of Lords. The current legislation went through just with ratification from just the House of Lords. I think it was a pragmatic political action, I clearly don't think it goes far enough. The first point is that non commercial can be very restricting and I am sure you know the recent case in Germany with whether teaching was a commercial use or not and early last year, the German Court found that the use of material for teaching was commercialthe end of last and I heard that at year a higher Court overturned that decision. Clearly, if the courts themselves have to debate about if something is of commercial use or not, it is going to be very difficult for ordinary people to do it. The uncertainty is extremely restrictive, different people will take different views and the more that there is uncertainty here the less change there is because people are afraid. The second thing is that the UK government has given us a paradox in that the reason for doing this copyright exemption was to generate greater digital wealth in the country and to promote industries which were going to be wealth producing. Now if you only allow this for non commercial purposes, it's very difficult to create an industry which is going to be able to rely on that and not end up in court at the first sign of any conflict with other rights holders. I am taking the view – and I say this very carefully – that I am doing research, personal research, and as a good scientist I have to read the whole literature to find out what might be pertinent and I also have to publish all the facts that I extract because only then can I give a complete scientific record of what I've done. I plan to do research and publish the facts as the correct, ethical procedure. Lots of people will challenge that and I don't know when or where these challenges are going to come from but – I'll call it the traditional publishing lobby, by which I mean the large commercial publishers have been lobbying against it with a variety of FUD, saying that nobody wants to do it, there's no call for it and that as soon as people start doing it, it will overload their servers. This is all rubbish, it's not going to overload their servers, studies by PLOS (the Open Access publisher) have shown that it is something like a ten to the minus seven of the daily load will come from text and data mining. So it's a complete red herring but it's a sort of deliberate misinformation that it's being put out there.

MB: You said some publishers would like or currently require mining to comply with specific permissions. Is that for the act of mining itself or for the publication of the result of the analytics process?

PMR: It's for the purpose of mining and probably the most likely use of this will be by the pharmaceutical industry. So if you are a pharmaceutical company and you subscribe to a scientific journal, the publisher will very probably have a clause that says you can't carry out text and data mining without a licence and you should seek an additional licence which may require a fee.

MB: So that's irrespective of whether the results of this exercise are published in any way ?

PMR: Absolutely. It's the act of copying the content from the publishers site into the pharmaceutical site which is almost certainly closed and nothing is likely to be published directly. The publishers have put in clauses to licenses for many years forbidding this and you are not allowed to spider their site, you are not allowed to crawl it, you're not allowed to extract stuff, etc. Librarians in universities and almost certainly in companies as well have signed these clauses. The Hargreaves legislation says that you can ignore these clauses, in other words, if your organisation has signed away the rights and if you want to mine stuff and you are doing it for personal, non commercial reasearch purposes you can ignore these clauses which is what I will be doing. But if you are a pharmaceutical company, clearly you can't ignore these clauses because the law doesn't give you any rights, there's no way pharmaceutical use can be seen as non commercial and so they are no better of than they were before.

MB: Are there any additional points that we haven't covered up and you would like to discuss?

PMR: I don't think so, other than to say that it is very important that people engage on this because if we do not assert our rights they are likely to disappear by default and they will be increasingly hard to recover. There are technical control measures here so that the publishers often cut off people for downloading too much even if it is otherwise allowed in the contract. So the publishers can control what is downloaded, and we have to fight against this. I'd also say that publishers have started to make APIs available where you get permission to access this in a way that they tell you is the best to do it. First of all I dispute that, and I've shown that. But also that means that they can monitor every use of this, so that they end up with a very large amount of metadata as to who mines what and for what purpose. They are potentially violating privacy and they can also use the API to control what you have access to, what types of information and how comprehensive it is and so forth. Most publishers will only allow you to look at the text and not at the figures, for example. In finishing, I would like to say how valuable it is for me to be a Fellow of Open Forum Academy and how very critical and valuable the organisation is.

MB: Thank you! Just to finish, are there any papers or online resources or plugs that you would like to make, or recommended reading for our readers?

PMR: One of the important papers is the one we published with you about two years ago on content mining, and that gives a reasonable overview of content mining. More generally I have a number of slides on this topic and if you go on slideshare.net, Peter Murray Rust, then you will find a lot of my presentations there, many of which are well-suited to discussing this. If you go to contentmine.org we are organising the slides there as well and there's an excellent presentation by Charles Oppenheim which he did with us last September about what the Hargreaves allows, it's very clear and balanced.

MB: Thanks a lot Peter!