This is a transcript of an interview between Monica Duke (UKOLN) and Philip E. Bourne, editor of PLoS Computational Biology and collaborator on the JISC-funded SageCite project, that took place on 11/07/2011 – the discussion loosely followed the following set of questions:

Here is a link to a demonstrator (which has been developed in the SageCite project) that shows the processes whereby a network model is assigned a DOI.
<link to video : TBA please see attached slides>

1. How could journal author submitters to PLoS and Nature interact with the journal submission process so that a DOI assignment to the network model is factored in. How does this fit in with current publisher workflows? What are the implications for peer review and reviewer workflows?

2. What are publishers' views with regards to their responsibility on data submission (data archiving and access) in addition to journal articles? Both in the short-term and the longer-term.

3. Can data citations be treated equally by journals? Can data citations be first-class citation objects? i.e. all the references to data be included, and represented equally with other cited objects in the reference lists? For example, are there issues with limits on the number of items in reference lists? How can multiple contributors be handled? What about supplementary data submissions that are 'in addition' to supporting data?

4. What guidelines and policy are required from publishers to authors regarding submission of data supporting publication and supplementary to publication? Would a consistent policy across publishers be possible and useful? Is all the data required as part of the evidence supporting the article?

5. What is required for publishers to reach such an agreement? Would a facilitated workshop for publishers be helpful?

6. Can you suggest areas where critical mass of agreements can be achieved? e.g. within certain disciplines

7. What are the barriers for publishers to implementing a data citation framework? What would the drivers be?

8. Are there lessons to be learnt from other domains?

9. How do you see the future of data publishing for traditional publishers? What are the interesting and encouraging developments have you seen in the last year in data publishing and citation?

10. Do you see a future for traditional publishing in publishing data?

PB – Fire away!

MD – Regarding the demonstrator – have you had time to look at it?

PB – I have looked at the slides[1] [Editor's Note: the video was not available at the time of the interview] The slides looked interesting.

[1] http://blogs.ukoln.ac.uk/sagecite/demo/

MD – Our project focused on how processes are captured, and the assignment of DOIs - how does this mesh with current models of publications? Does it fit in with publisher workflows? For example say someone submitted to PLoS ....

PB – My main experience comes from PLoS, where I have learnt that changes to the publishing work flows are difficult, from a technical point of view for PLoS, and it is also quite difficult from a sociological point of view. Publishers can no longer ignore the need, the desire, of the community, which I'd say is becoming increasingly vocal, to do something with data. I would break data down into two categories: that which is associated with publications, and that which is not. Right now we are talking about that which is - PLoS is faced with dealing with this, we are just about to have a Skype call with Thea Bloom about this, they are about to embark on a project with DRYAD[2]. PLoS is not the only publisher looking at ways of dealing with it that minimises changes to their workflow – at least the assignment of a DOI is something that would be embraced, that is something that will come from DRYAD: the data will be there and there will be some kind of resolution to get to the data from the paper. That's about as much as can be done in the short term, but does not begin to address what should be happening, in my view.

MD – OK! Next question on publisher views regarding responsibility on data submission, archiving, and access..... can you speak about your own view, and do you perhaps have a feel for what directions the rest of the community might be heading in?

PB – Data Journals are emerging. BMC has one, more are emerging, I was surprised there are more out there in different disciplines than I was aware of. I made a posting recently on the Beyond The PDF discussion list, about data journals.[3]. In the main it is not the main-stream publishers who are dealing with this, not those with the deep pockets, or those with different publishing models like PLoS– but mainly small publishers trying to leverage this idea to get more into the business, I guess, trying to get more exposure within the business. On the other hand, some of the big publishers like Elsevier are doing interesting things that don't necessarily at this point embrace data, but clearly if they did that would be enormously powerful. I think what is happening with Sciverse[4] which is opening the corpus to the community, at one level …. are you familiar with this … they are inviting the community to deliver apps that sit on top of the Elsevier corpus and allow you to use it in some way – it is, again, a way of getting the community to do all the work, effectively using it. What it would do, if it is successful, would be to have different apps e.g. for semantic tagging, some kind of indexing, some graphical view of the content of subject areas, whatever those apps might be, that the community develops. The community can use the apps for free, but the moment you dig down into individual article level, you are back to the pay wall and the pay model. It is really about getting the community to do a lot of work, to make the scientist aware of what is available, at some level, which of itself is positive. If people can get to content in ways they cannot do currently, even if the content is still closed access, it really sets the open access model back. Really what needs to happen is that the open access folks really need to start developing applications equivalent to this, apps that sit on top of open access content. Beyond what is available through PubMed central – which is really very little – there's not much out there, there's certainly nothing that operates on BMC or PLoS corpus that I am aware of – these kind of things really need to be developed. Otherwise it will be a step backwards for Open Access. The bind is of course that Elsevier have money to develop this stuff, whereas open access publishers or PubMed operate on a shoestring.

MD – So on the one hand we had the people with the closed content building the apps and open content people are not building apps. Are these currently separate communities?

[2] http://datadryad.org/

[3] http://groups.google.com/group/beyond-the-pdf/browse_thread/thread/971581832e5ecf93

[4] http://info.sciverse.com/

PB – Sciverse can also operate on open access content, so effectively they cover the majority of science, whereas open access content applications will hit publishers' firewalls very quickly. Elsevier don't currently have access to, say, Taylor Francis or Wiley content – but they could come up with schemes for sharing that would be interesting.

MD – I have seen Sciverse demonstrated at Science Online 2010 - Peter Murray Rust (PMR) was there, and commented that the APIs are open but the content is not open, therefore it is not an open solution.

PB – That is correct, but notwithstanding, this is happening, and it is very, very clever, and there is nothing equivalent around open content. If you bring data into it, and if under this umbrella the apps being developed for you also included the data, it would be very powerful. Where I'm going with all this, after some debate and discussion, is I see a prototype developing, that hopefully some of the bio-med open access publications would get behind. If stuff goes into DRYAD – that is fine because that's open, it could be accessed here or elsewhere. But what we really need to do is to say, here's a few disciplines or a sub-discipline, and here is the data they are putting into this resource, then we need to use the same model, a well defined open API, that accesses the content, first of all from the point of view of validating what is in the archive, in some way, perhaps annotating the data archive, and provides those tools that operate on top of that archive. You could actually get another layer of development on top of the data archive, that starts to create more uniform and easily accessible apps that operate on the data than we currently have.. This could be a point of coalescence. Those sorts of things are going to be very important. The point is that the data citing is the first step of that. I'm speaking as a database developer/publisher, but if I speak to you as a scientist, here's what I'd worry about in all this. If you ask people in my department I don't think they care if they get a DOI for their data or not, that is not enough of an incentive or reward – you've really got to have a fully fledged data journal where people start to include citations. If you could see a paper with a list of references, where one or more of those references in the list is a formal data citation referring to a data journal, refers to a year, a volume number, and then to the DOI. We need something recognisable as a more generic form of publication in its own right, with associated bibliometrics: to see how many people have downloaded it, this particular data set, how it has been rated etc. – this is getting towards an impact factor for data.

MD – You have just given suggestions on how the citation needs to be dealt with in journal articles to make it useful, the citation needs to be recognisable and we have just started talking about metrics. Are there issues with how data can be cited? It is often said that data are not treated as first class citation objects e.g. authors are limited with regards to the number of items in reference lists.

PB – Those are definitely issues. The value of data citation at the moment is highly variable, it is true most don't have the value of paper citations. On the other hand there will be cases where the paper is only cited by the authors, and they also cite a data set that has been downloaded by 100 different people – which is more valuable? If 100 people downloaded the data and used it, published stuff related to it or built upon it, it is more valuable than the publication only cited by the people who wrote an associated article. (although these are rare cases). This goes hand in hand with the change in how we measure the value of traditional publications.

MD – These are clearly issues for those who are putting a value on data contributions within the community e.g. people deciding tenure. Is there a role for publishers in the way these citations are presented or accepted, that may help to make the data citation acquire more value?

PB – You can start to see it happening in a small way, e.g. the PLoS article level metrics, and the refusal to effectively acknowledge the impact factor, is having a small effect. On the other hand, PLoS' point of view as a publisher is the impact factor has no value, but, when you talk to the editors of the PLoS journals including my own, almost the first thing they ask is 'what is our impact factor?'. I would say this kind of change is proceeding very slowly. I wrote an editorial about this 'How to get ahead as a computational biologist in academia'[5] The first thing to do is to educate the tenure committee (or whoever is reviewing your file) of what they should be looking for in what you‘ve done, that has value. And that of course includes what data you have deposited in public archives or made available to the community - that does not even get on the radar of a lot of experimentalists. In terms of digital data these are very important contributions. We need to be educating review committees to take this kind of thing into account. It happens slowly, in its own right as, increasingly, those same review committees, the people on them, slowly but surely are starting to use the data from other people. When they do so, they start to appreciate its value. In the life sciences the change is slow but is brought about, in part, by the evasive nature of public data into the whole research enterprise. You ask about lessons learnt from other domains – maybe areas like high energy physics and astronomy have something to teach us. This has been the norm for a long time in those communities, so those data sets have effectively frequently been out there, and the community for a long time has appreciated their value, in ways which has only just started to happen in biology or bio-medical sciences.

PB – Let's see what else we've got ... having a workshop – several publishers have contacted me since making that posting on data journals – including Nature, PLoS & Elsevier, I'm having discussions with all of these – so a workshop for publishers around data would be good – This is all complicated by other factors, although publishers are the focus, there are also institutional repositories that offer homes for data. They are being set up left, right and centre, but it is not clear to me that most of those work very well at all. There is a great article where a librarian calls them roach motels [Dorothea Salo][6] – you check in but you don't check out! So you can get data in, but you can never get any data out! There are certainly lessons to be learnt in that whole process. Getting a facilitated workshop would be useful not just for publishers but also for people doing institutional repositories, maybe that's something we could do next January, in San Diego? I could maybe bring this up at the Dagstuhl meeting in Germany[7].

MD – Would that be both for publishers and Institutional Repositories – for people to explore cross over, and share lessons learnt, to learn from each other, or do the two groups need to consider different issues?

PB – No, it's more about who does what, and how they interface with each other. When life sciences went on line, there were lots of databases. For many years, they popped up like mushrooms. Some thrived, many died. And they were separate entities. Now there is more effort going into interoperability and getting information from one to another. It is history repeating itself, to some extent - maybe we can short circuit the process by virtue of the technology that is available today, and changing awareness. If data is public, it can be referenced by a paper that is published by any number of publishers, and some common tools that allow you to go to any paper and immediately browse and get a sense of the data sets that are associated with it, would be useful. I'm making this up as I go along!

MD – If the publishers are not going to be solely holding and curating this data, then the data may be held in different repositories. Are you suggesting what publishers want is a way of interfacing with those different repositories, through something common, so that one could provide a common tool. Tool builders don't want to have to interface with each repository via different means ...

PB – Yes, in fact the technology has improved. For some time we were struggling with Technology that was cumbersome for working across databases – I'm thinking back to CORBA which was a nightmare – now there are RESTful services and web services that can make this straightforward. It requires each resource to have an API and some commonality within the API. Things don't seem to be going on at that level ...

[5] http://www.PLoScompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002001

[6] Salo, Dorothea. Innkeeper at the Roach Motel. Library Trends. 57 (2). 2007. http://minds.wisconsin.edu/handle/1793/22088

[7] http://www.dagstuhl.de/en/program/calendar/semhp/?semnr=11331

MD – So do we need the repositories to meet and agree 'we are going to support common features or standards in the API' because of a demand from the publishers saying 'if you all agree to this we start linking to the data in your repositories'? Is this what we're getting at here?

PB – Agreed. That could be one way to go. What happened in response to my posting (which is depressing in some ways) is that the responses probably contained around 15 different technical solutions suggested by various people, from people very passionate about what they do and who are doing good stuff, but no regard, in my view, to interoperating with others.

MD – Would you say that if the publishers agree and declare they will link to data in repositories, this would provide the motivation for the repositories to agree to standards?

PB – It would help, but I don't think the publishers are going to do that. I don't think so anyway. Publishers are very cautious. It requires authors to put data in these repositories. I know from my colleagues, if there are two journals of equal stature and one requires data deposition and the other doesn't, they are likely to go to the one that doesn't. It is only when the community starts self-policing and says *we should be depositing our data,* then it happens. If it's a question of revenue, by virtue of having articles published and so on, they will be reluctant to change the system that works for them.

MD – Do you see the publishers, if they were to reach some agreement ... would it put pressure to change the system or would they not have the motivation to change?

PB –If you take standardised data, when a community says *you should not accept papers that do not deposit the data about those papers,* then once one journal does it and its driven by the community, the others will do it. There is general agreement across many journals that sequenced data, structured data and array data and other things appear within these resources. That's a good thing, but that is being driven from the bottom up. If the community wants it and once one publisher is doing it then the other publishers will do it. The driver has been the community, not the publishers. Maybe it is time that could change, I don't know?

MD – Thinking back to holding this workshop for publishers, what would we be hoping for as an outcome?

PB – A sense of interoperability and accessibility between traditional published information and the data from which that information is derived. This brings into question not just the data, but also the methods used on that data to get to the knowledge that is in the paper –at that point it essentially becomes publishing workflows. There's different levels here. As we're talking about this I can see this as something that might happen, and will become much clearer at the Dagstuhl meeting in August in Germany.

PB – It is the usual suspects organising it, Anita de Waard, Tim Clark and Gully Burns, it has been restricted to around 30 people. The title is the Future of Research Communication[8]. What else have you got there....

MD – We were just talking about the agreement that has to come from the community rather than the publishers. Looking on to next questions... How do you see the future of data publishing for traditional publishers? What are the interesting and encouraging developments you have seen in the last year in data publishing and citation?

PB – I am not up on what DataCite is doing; people seem to be talking very much about Dryad, but when I look in there, there is not much in there. Clearly there is a lot more buzz around data, from all interested

[8] http://www.dagstuhl.de/en/program/calendar/semhp/?semnr=11331

parties: publishers, scientists, database providers, scholarly communicators, institutional repositories. Unfortunately often these seem to be competing rather than collaborating. So if you were a scientist wanting to put your data somewhere, it is all very confusing. You have to start with the customer – take the customer view - a lot of these things are not even asking the customer. What does the customer want and why do they want it. One of the things that happens in the PDB[9] is that customers come to get their data back – that is probably one of the biggest uses of being a data repository. People who originally generated data can get it back 'cause they lose it and they want to use it again. That's a service in itself. And that is just for the people who originally deposit it. Let alone anyone else who might want to use it. You need to ask what those people want. I can tell you what I want …. I want a place in the cloud to provide data that can be cited, and used by others and retrieved by me, if I want to. It has to be done. I accept it needs to be done, with a good ontological description of it, but it better be easy to do … the barrier to entry needs to be pretty low, or I won't deposit it!

MD – Going back to something from the publisher perspective - what are the prerequisites for the publisher to start citing data, are there questions around the stability of the repository where the data is kept?

PB - That's one of the pushes for something like DRYAD or DataCite - they have some semblance of stability and some guaranteed level of funding for some period of time.

MD – and that gives the publishers more confidence?

PB- That's a good question. PLoS are thinking of doing a trial with DRYAD – what's their plan if in 3 years DRYAD says, we are going out of business, and you have 150 data sets, half a petabye of data in here, from PLoS articles. What are you going to do with it? In other words, what is the contingency plan if the archive goes off line? One of the things scientists have always trusted about journals is that they have information in perpetuity, and if you start saying you are supporting data, then really, you are implying you are supporting that data forever as well.

MD – Do you think that the publishers do not want to take that responsibility on themselves?

PB – I do not think most of them are geared up to do that right now. This is why DRYAD is an attractive option - they are they a group of people who think a lot about data.

PB – Lets finish off with developments I have been involved in recently: I've already referred to the posting I made on Data Journals, which covers data citation, on the Beyond The PDF discussion list. This had lots of feedback, probably twenty messages or more, which is worth following. And it is certainly worth taking a look at the upcoming meeting in Dagstuhl, where data citation will be one of the topics under discussion.

ACKNOWLEDGMENTS
Thanks goes to Sarah Hext of UKOLN for transcribing the interview and to Philip E. Bourne for generously giving of his time.

The SageCite project was funded by the JISC under the Managing Research Data Programme: Strand A
http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx

[9] http://www.wwpdb.org/