This is a transcript of an interview between Monica Duke (UKOLN) and Myles Axton, editor of Nature Genetics and collaborator on the JISC-funded SageCite project,that took place on 02/08/2011. 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 authors submitting 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 proposal fit in with current publisher workflows? What are the implications for peer review and reviewer workflows?

2. What are publishers' views with regard to their responsibility for 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 (to articles or to each other?) 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 to 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 guide 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 agreement 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?

MD = Monica Duke
I = Interviewee: Myles Axton, editor, *Nature Genetics*

MD: I don't know if you've had a chance to get more of an overview of what the technical part of the project has achieved?

I: I haven't really got to the way in which subparts of a model can be referred to.

MD: That's not something that we have tackled in the timeline of the project. I think it took a lot more time than we expected at the start.

I: Thinking back to our visit to Seattle[1], it took a lot of time to understand what the workflow actually involved.

MD: That's right. And after that Peter and Brig [Peter Li, researcher on the SageCite Project, and Brig Mecham, researcher at Sage Bionetworks] followed up with a number of conversations, emails, Skype calls simply for Peter to take one of the workflows and to reproduce it as a workflow in Taverna[2].

I: Ok, let's get to the questions because they are very, very wide ranging.

MD: Yes. So I guess, in one sense what we are saying with that question there is that it is all very well assigning the DOI to the network, but you think there is scope for further work in looking at how to look for subparts of a network... is that what you were getting at?

I: I think that as soon as you can refer to the network, people will want to be able to refer to versions of the network and parts of the network, and that will be the next challenge. Just as we now have a DOI for the article, we will eventually have DOIs for each figure and each part of the figure.

MD: Ok.

I: So, your question: “How could journal authors submitting to PLoS and Nature journals interact with the journal submission process?” A DOI assignment to the network model is factored in. That begs the question: who is paying for the DOI? Because it seems me in the Sage Bionetworks[3] model that you would want to be able to assign the DOIs within the Sage Bio network and use them even before publication. And Sage Bionetworks might even become its own form of database publication. But the first model would be that the DOI is only assigned to the Sage Bio network and a stack of DOIs is available from the grant funding of the Sage Bio. And when anyone creates a model they just avail themselves of the DOI and they use that. The DOI in that case will point to the Sage Bio network and not to the publisher. The next question is: “Would a publisher want to buy a stack of DOIs for data?” I think that's a very nice model. My feeling is that you could have DOI assignment upon submission, and then if the paper was rejected, there's an export file of XML which might include the DOI. I think that would not be popular with publishers, but I think that assigning it upon acceptance of the paper is possible. But that has a lot of attendant problems of how you make things back and forth in order to get everything right at the point of acceptance. So on the whole, I am favouring a DOI assignment within Sage Bionetworks, then an honouring of that by PLoS and Nature Genetics.

MD: Thanks, that's raised some useful thoughts on at what point do you assign the DOI. As you've explained it can be assigned by the organisation, it could be assigned on submission or it could be assigned on acceptance.

I: Also, the whole problem is fast knowledge turns. You want to be able to use the knowledge that's in the Sage Bionetworks knowledge base. And so having the DOI right at the point of creation of the network maximises the authority of the network and it maximises the citations. You might have a year's worth of citations before the paper even gets published.

MD: Ok. I think that starts to move on to the next question as well, which is on publishers' responsibilities on the part of submissions. So clearly if an organisation has already used an identifier should the journal carry on with using that identifier?

I: Yes, I come from the genomics field and we use accession numbers extensively. I think it's established that those belong to the community who work with those databases. Remember my survey which I

1 See http://blogs.ukoln.ac.uk/sagecite/2010/11/18/talking-to-the-users/
2 http://blogs.ukoln.ac.uk/sagecite/demo/
3 Sge Bionetoworks http://www.sagebase.org/ collaborators on the SageCite project.

contributed to the finished SageCite project? I did it for the Human Variome Project to find out what publishers are capable of citing. It includes grant numbers, author roles and accession numbers, DNA sequences and all these kinds of things. And it turns out that traditional publishers are not very good at enforcing any requirements. And it was really only *Nature Genetics*, *PLoS, Genes and Nutrition* and *Human Mutation* and maybe only a couple of other journals that were able to really push the use of human gene nomenclature and DNA accession numbers and so on. What we ended up with was a group of eight journals who were able to agree that wherever there is an accession number that can be linked to where the database is stable, generally used and is capable of linking bi-directionally so you can get credit to the journal as well - then the journal would use the accession number to the data. And we would do as much as we could to ensure bidirectional links.
Now the practical problem, that may be one of the later questions, but the practical problem is you need to get the accession number (field?) into your DTD[4]. When you do your mock up, you need to recognise that this is an Ensembl[5] accession, or this is a GEO accession or this is a GenBank[6] accession, so you need to make sure that you have in your ontology the types of data that you are going regularly link to. And I think that's the limiting point for the publishers. A lot of the publishers, old publishers which are now moving to online first workflows are building their DTD around the NLM, (National Library of Medicine) DTD and then they are adding any other markup that they are going to require, and that's kind of a moving area. But that's not how I interpreted this next question. I got off into the area of supplementary information and what is the journal's duty through handling data. First of all there's the status quo, which is we allow supplementary information at author discretion, so we are data repositories but we don't do much to format the data. We just take tables and PDFs from the author. That's the status quo. Then there's the proposal from publishers that we can charge to process, index and store, thereby becoming data publishers. I think that would be good, we could move the supplementary information behind the pay wall and it could become part of the business model. We could charge authors to process the tables properly and we could make the data more accessible. It's a very low level activity, and I'm not sure that Nature Genetics would be very keen to get involved in doing that. It's something we could outsource, I'm sure. But if it was done badly it wouldn't help our reputation at all. And my approach is actually a third line, which is to try to abolish supplementary information in favour of linking and indexing accessions. So we would end up hosting tables of accessions and indexing them at a higher level between papers and among papers and we would actually work with the community, so the databasing and storage would be a community function. My only concern about that is that Nature is a journal of record of 140 years or so and following that tradition we at Nature Genetics are committed to storing absolutely everything that is essential to the conclusions of the paper and making it available in perpetuity. I'm not that that's true because we've noticed the SRA (short read archive) for DNA reads is no longer supported and so there's a lot of worry in the community that the place that they store their data where they get their accession numbers may not be as stable as the journals. But provided we get our business model right we could become data storage places, but my preference is for indexing rather than storage. That doesn't cover our responsibility, that just covers our commitment to storage. There is another responsibility which is to enforce community standards and I think we're much better at that than we are about coming up with solutions to storage[7].

MD: So that ties in with the issue of guidelines and policies from publishers explaining to authors: for example, you mentioned the stability of the discipline-based databases, which gives publishers the confidence to use their accession numbers, although you followed that up with an example where that stability seems to be in question now. So do you think that publishers might want to set up some criteria, which links up with the survey you've made of the types of identifiers which are accepted? Rather than name specific acceptable repositories, they might say that these are conditions of storage that we would like to have met in order to have the confidence to use the link to that storage place?

I: Yeah, one of the problems that we have is that most databases don't like to use “linked data” links for accession numbers. They like to have a link to the database and then they have some kind of relational structure and search which gets you to the accession. That allows them to be quite free about where they store things. They're not happy with the idea of providing the unique linked data URI that takes you exactly to the accession and which gives you the information of what the database is, what the authority owning the database is and where the accession is and which version of the accession you're dealing with.

MD: So I would say there are two issues related to that, so first of all this relies on a human being following

---

4 DTD – Document Type Definition: a set of mark-up declarations that declare a document type http://en.wikipedia.org/wiki/Document_Type_Definition

5 Ensemble http://www.ensembl.org/index.html

6 GenBank http://www.ncbi.nlm.nih.gov/genbank/

7 See also Axton, Myles. (2011) No second thoughts about data access. *Nature Genetics* Editorial 43, 389 http://www.nature.com/ng/journal/v43/n5/full/ng.827.html

the reference to the data and being able to go to a database and input some search parameters...

I: Yeah, well they tend not to, but I spoke recently with somebody who is responsible for metadata links and they say it really doesn't matter. Each publisher will have their lookup table, and so it really comes back to the DTD again. You know the ontology of a lookup table and if something is a GenBank ID then the publisher's lookup will take you to the GenBank. I find that very unsatisfactory because I believe there should be a semantic web where everything is referred to uniquely. But at the moment what we're doing is local reference and look up, and I think that's creating problems for the future, but it's what we do now.

MD: And I think if we get to this environment where we are trying to reward people based on perhaps citations of their data, we do need a way to be able to count citations to perhaps one specific entity in the database, so we do need to be able to refer to it uniquely.

I: Yes, so the problem that I see that we're building up for ourselves is by having local references to an accession, then you create a responsibility for the publisher to count the citations locally and export them, and that's never going to happen. You need a much more open way to do it.

MD: And that's eroding the idea that citations might help towards motivating data sharing.

I: Yeah, possibly. I mean, my original idea when I was talking about microattribution is that the database would count citations and so would the journal and it would use them in different ways.

MD: Ok, so imagine exporting that information as one way for somebody, like a funder, who wanted to make counts of citations across materials ... ?

I: That's one of the later questions, but what I'm working on right now is the sustainable open access model for journals like *Nature Genetics*. This is somewhat business confidential, but this is something that I talk about with people in the field because everyone understands that the top publishers have trouble making an open access, author-paid model work and still maintain the quality of the value we add. So my solution is really that we do more for the funder, and one of the things would be to build a dashboard for the funder which gives data citation metrics, author contribution metrics and so on. And it tells them how their grant is being used and so on. So I think if we do more for the funders, we might have a model whereby we could sustain open access publishing, because the Nature Publishing Group is committed to a hybrid model, offering open access. I just don't know quite how to get there, so I think we're trying to develop bridges between our current subscription models, licence models and the open access future.

MD: So one of the questions we have here is about a facilitated workshop for publishers to come together - maybe if there are some agreements that need to be built? And, in the discussion with Phil Bourne, he mentioned that a meeting amongst the repositories or the databases would be useful as well in terms of having some agreement about how their information can be accessed.

I: If we get the leading databases. I mean EBI[8] and NCBI[9] cover quite a lot, and then you have a couple of large databases in Japan and China and for genomics that's pretty much it.

MD: So do you think there is scope to reach these types of agreements at least within certain disciplines?

I: I think that it's not so much a matter of policy and agreement it's the matter of the gritty implementation of things. We have a problem creating bi-directional links. For example, in *Nature Genetics,* accession numbers to microarray datasets drive traffic to the micro array databases, but because of the asymmetry of the publication date, often the accession number in the database never links back to the published paper, because the links to the accession were set up before publication and then it requires somebody to build a workflow that lets the database know what the DOI of the published paper is and create that link back to the paper. I think it would be in the interest of the database to point back at the journals, but I don't know of any journals that routinely do well at bi-directional links to more than one database. So that's something we could get better at. At the moment there are not any incentives for us, so maybe the data citations would help to clear up that problem.

MD: Ok, I'm trying to carry on with the questions. I think we've more or less covered the next two down.

I: First class citation objects – can data be cited? Yes, of course they can. What about: A database entry

8 http://www.ebi.ac.uk/
9 http://www.ncbi.nlm.nih.gov/

can be the equal of a paper – absolutely. I've seen database entries that are much better cited than some papers. Reference lists – that's not a problem online, *Nature Genetics* has unlimited references. We also allow unlimited URLs. So there's no problem there. What policies will be required? Consistent policy will happen. Basically, CrossRef has built CrossMark[10], which is an amazingly flexible display of any metadata you want to put in it, and publishers will want to show off what they are doing, so as data citation becomes popular it could happen through CrossMark. Author contributions can be displayed through CrossMark, whether something is peer-reviewed or not, whether it is in a data repository or whether it has a publisher – all of that is very nicely handled in CrossMark. I think we have the forward capacity to build all of that information in. We also have ORCIDs so we know who all the people are. One of the problems is identifying institutions, I think that's one area of precise attribution that needs to be worked on. So reaching an agreement, I think we ought to do it with the publishers who get it first, and *Beyond the PDF*[11] was a good example of the kinds of discussion that you can have. I'm not actually a cartel person... I'm very much business-minded and I think what we need is a minimum standard and I think the PubNet Central standard or the NML-DTD is a good kind of basic standard. If we decide through CrossRef or CrossMark for some kind of publisher agreement to have an open standard or minimal metadata, I would really like that and I would think that some very basic author role information would be one of the things we could push towards. But I don't want to tie anybody's hands if they want to do better and they want to make business out of it. I don't know what we should be mandating. I think essentially a fair open platform that allows commercial innovation is what we're looking for.

MD: So do you think it is more of a case that if someone starts doing something innovative and useful, then the others will follow, just because they need to be competitive?

I: Well there's two ways for doing it. I think Thomson Reuters very much wants to develop an official data citation measure that will be as useful as the Impact Factor. But we've seen where that goes because they had an author ID and that became the open ORCID[12] which is actually more useful to Thomson and to everybody else as an open standard. So I hope that they can be persuaded to turn the data citation efforts into another ORCID project.

MD: Ok, shall we move on? I'm keeping an eye on the time.

I: So, the critical mass of agreement in certain disciplines …. at *Nature Genetics*, whenever we get a new paper we have three questions which we ask ourselves before sending the paper to review - doing due diligence to the principles of data access. If there's a genome sequence, we check to see if it's actually accessible, whether it's not just the raw sequence reads of the genome, but the scaffolds of the genome assembly; if there are microarray datasets we make sure that those are accessible as well, that they have the proper accession number; and for DNA variants we need human genome (HGNC and HGVS) nomenclature – is there a reference sequence for the accession to correctly cite its position within the reference sequence and is it in the proper nomenclature? Is the gene named properly? Is the reference sequence present? And are the co-ordinates within that correct? If any of that is incorrect, we contact the author for further details and we never accept a paper without those details. I would say in genomics, that's a good example of three areas where there is widespread community agreement and publisher enforcement.

MD: Thanks.

I: So, the barriers to publishers for implementing a data citation framework. I think the main barrier is inertia because no one sees any point to it. The drivers for it would be giving the funding bodies better metrics and the demonstration projects that various groups are doing – there's the NSF/NIH STAR METRICS[13] project which is to see how the grants are being used, then there's Harvard University where Amy Brand[14] I think is doing some studies to look at the productivity of faculty for promotions and tenure, and the Wellcome Trust is doing one as well, and then there's the VIVO[15] project – but I think the incentives come from funders wanting metrics and institutions wanting to make more rational promotion decisions and from publishers wanting to be more helpful and make more of their stored metadata. So I think what we are doing is moving into an area where publishers can see the possibility of making products for institutions and funders from the metadata of the papers they are already publishing. I think that will overcome some of the inertia as we realise that that could actually give us sustainable funding for opening up some of our database files. So I'm

10 http://www.crossref.org/crossmark/index.html
11 http://sites.google.com/site/beyondthepdf/
12 http://orcid.org/
13 http://nrc59.nas.edu/star_info2.cfm
14 http://beyondthebookcast.com/whats-in-a-name/
15 http://vivoweb.org/

quite hopeful about a proper business-driven commercial application of data citation in order to keep the funders and universities interested.

MD: That's great, thanks.

I: Lessons to be learnt from other domains... yes, well at one of the early data citation meetings that I went to there was a lovely example from OECD of economic modelling datasets where you can get the GNP of Vietnam broken down into sectors and that's all written by teams of analysts and if you want to use a slice of that data, you have a sub-DOI that refers to, for example, the tennis shoe exports or on the shrimp farming sector of Vietnam and you can just write your review article on shrimp farming in Vietnam or even aggregate it with shrimp farming data from other datasets. But that credit then accrues to that team that assembles that year's econometric data. So I do think there are models in other fields that are attractive to us.

MD: Ok, I wasn't aware of that example in the economic modelling, so thanks for that suggestion.

I: And then... The future of data publishing by traditional publishers... you know, we're free from the storage limitation. I think we could either rent cloud storage for the data or we could index the data. I think the challenge for high quality traditional publishers is that every paper we publish is supposed to be different to other papers, yes you can get a vein of genome-wide association studies, but that's going to be in the hundreds, and you can get microarray studies and that's going to be in the low thousands, but that's not the way the editors and publishers think of the papers. They don't think of them as being a big chunk of data, they think of them as being something qualitatively different or some conceptual advance that justifies the paper. We're not buying a truck load of sand every time we publish a paper. So I think that limits our ability to build standardised data repositories, because we don't know what to prepare for. The volume is tiny compared to, say, accountants storing people's retirement plans, where there's going to be millions of them. It's worth building your data repository for that kind of data because it's going to be the same stuff over and over again. But I think in scientific publishing it has to be very flexible and we're never going to store things in standard ways. That said, there are some elements that we get again and again, like tables – tabular data – so it would be worth us building a really good table and then forcing authors to use that table but storing it in such a way that the tables are human and machine readable. I think there is a future in traditional publishers becoming that kind of data manager. And I think the high-level indexing of data is something we have to get into or vanish, otherwise it could all be done at the database level. I think that the job of selection and promotion of publishing requires us to go to the high level and concentrate on becoming better indexers of other people's data in a non-exclusive but extremely useful way.

MD: Ok, I'm looking at the last two questions. … What interesting and encouraging developments have you seen in the last year in data publishing and citation? Maybe some projects you have been involved with or what you've seen...

I: Data publishing and citation projects that I've been involved in. Yes. I was involved in a model for incentivising the community annotation of the human genome, under the auspices of the Human Variome Project. We took the human haemoglobin database, which is HbVar, and we assigned Thomson Reuter author IDs to each of the participants, and we gave credit in the database and in a *Nature Genetics* review analysis article[16], to everybody who had data – whether it was published or unpublished. We ensured that during the data collection exercise, all the gene variants were described in the proper nomenclature with the proper co-ordinates and people got credit for putting them in the database. We got new biological information out of it and we also got a much more complete record of the literature. We got the users to correct mistakes in the record. The readout of citation is not that impressive. We gave credit in a table to all the contributors and we also made them authors of the paper so that's how we gave credit to the individuals. The individual pieces of data can be referred to by their unique variant ID, so they do have a unique identity and an existence in an independent database, so they are semantic objects that you can link to. So then we mined the paper for meaningful triples of the form "gene variant has frequency" or "gene variant is associated with disease" and those triples were stored by the NBIC Leiden by Barend Mons and Erik Schultes' group under the Open Pharmacological Space Initiative[17] they're building a triple store. This was one example of how the dataset in the database can be mined or the supplementary information table can be mined or the text of the article can be mined. And we compared the efficiency of making RDF triples from the database from the supplementary information and from the paper, and then those triples can be used going forward. But we didn't look to see whether we could make triples from the contributor IDs and then

16 Giardine, B., Borg, J., Higgs, D.R., Peterson, K.R., Philipsen, S., Maglott, D., ... Patrinos, G.P. (2011). Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach. *Nature Genetics*, 43, 295-301. doi:10.1038/ng.785

17 http://www.youtube.com/watch?v=qB1ljMvGxXg

give numerical credit for the number of triples created by particular individuals. I think that would be the next step and that's what I'm interested to start to do. But that's one initiative. The other initiative that I'm doing right now is to try to make more of author contributions information in our own papers at *Nature Genetics* and to give a cross-cutting view of the papers where you can find, for example, the Wellcome Trust Case Control Consortium (WTCCC) and find out who published in *Nature Genetics* under the auspices of the WTCCC? Which grants did they bring to the table? How many of those were Wellcome Trust grants? How did they describe themselves? And so with all that valuable information, that I would regard as data citation for nontraditional precise citation of exactly the same sort, so although we are not pointing at individual datasets, what we're looking at is fractional contributions to an overall paper and the underlying metadata to do with funder provenance, institutional affiliations and everything else. So I think that way of handling our content will also allow us to look at fractional contributions to data, which is one of the questions that you have here. So those are the two things I'm concentrating on right now.

MD: And have there been any other major events or initiatives that you would like to mention?

I: Yeah, I mean along with the Sage Bionetworks project, we've been working with Susana-Assunta Sansone and BioSharing[18] at Oxford and what we've developed is the *Nature* data standards website, which is currently called Nature Precedings, (it started as a preprint archive). We're not encouraging preprints any more because they don't get as much traffic as the new content. The new content has been data management plans (DMPs) [19], which are explicit plans that are developed in advance of data release, which say how the data will be generated, which repositories they'll be stored in, how you're supposed to cite them and whether there are any rules or restrictions or embargoes associated with them. So we have some of those from private industry and we have some from NIH's NHGRI Human Microbiome Project[20] doing bacterial community sequencing in various areas of the human body in health and disease. So that's been good. It's been very successful, we've been getting a lot of citations to the DOIs of the data management plans even before any data has been released. And so I think it's fairly useful to have that resource. In addition we have some standards papers – some of them from the Sage group[21], some of them from other groups – on how data should be released, how data should be processed and how particular areas of research should consider their strategies. So community standards[22] is a very popular area of the *Nature* data standards site. And the third area is funder policies[23] from NERC, BBSRC, Wellcome Trust, and the D.O.E., so they have been interested in telling resource-generating projects how and where to put their data. And so we've been hosting the funder policies on that site as well. So I think that what we collect at *Nature* data standards will be a very helpful adjunct to the information that we mine out of our papers, and that together we will be able to combine that information to give funders and institutions a lot more metrics about what data is being generated, how they're generated and how they should be stored. And it's helped *Nature Genetics* because we've managed to resolve all kinds of concepts to do with data release and data storage, and we've managed even to get access agreements from private companies to allow people to use the data, so I'm very very happy to have participated in that, and we're fully engaged with BioSharing, and their index brought us some data management plans (DMPs) and all the other data access initiatives. I think that if we're going to do another SageCite or something like it, we should look to work with BioSharing because they have a very nice technical team, they have a nice interface and they are very, very well connected but they don't waste much time agreeing things or holding meetings or taking different people's views into consideration. They just set everything up and put a good connection to it so everyone can use it. They're not trying to reorganise or do anything difficult the way we're doing. They are simply just trying to index and point to things. That very lightweight activity turns out to be quite useful and complementary to what we've been doing.

MD: That's great, I think we've covered all the questions. Thank you very much for your participation in the SageCite project!

ACKNOWLEDGMENTS
We thank Kirsty Pitkin of TConsult www.tconsult-ltd.com for her services in transcribing the interview, and especially Myles Axton for giving generously of his time. The SageCite project was funded by JISC under the Managing Research Data Programme: Strand A. http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx

18 http://www.biosharing.org/
19 http://precedings.nature.com/documents/type/marker_paper_data_plan/revisions
20 http://precedings.nature.com/collections/human-microbiome-project
21 e.g. Derry, J et al (2011) Developing Predictive Molecular Maps of Human Disease through Community-based Modeling. Nature Precedings http://precedings.nature.com/documents/5883/version/1
22 See http://www.nature.com/ng/journal/v42/n11/full/ng1110-915.html for an editorial perspective.
23 http://precedings.nature.com/collections/biosharing-policies