WESTMINSTER HIGHER EDUCATION FORUM

Next steps for Open Access and Open Data research policy
22nd November 2016

Seminar supported by

SPRINGER NATURE

CONDITIONS FOR USE OF TRANSCRIPTS:

This document is intended to provide a timely reference for interested parties who are unable to attend the event to which it refers. Some portions are based on transcripts of proceedings and others consist of text submitted by speakers or authors, and are clearly marked as such. As such, apart from where it is indicated that the text was supplied by the speaker, it has not been possible for the transcript to be checked by speakers and so this portion of the document does not represent a formal record of proceedings. Despite best endeavours by Westminster Forum Projects and its suppliers to ensure accuracy, text based on transcription may contain errors which could alter the intended meaning of any portion of the reported content. Anyone who intends to publicly use or refer to any text based on the transcript should make clear that speakers have not had the opportunity for any corrections, or check first with the speaker in question. If in doubt please contact the forum first.

Contents

Please be advised that speakers' PowerPoint presentations are included within the transcript itself, just beneath the relevant speaker's text. Please note that not all speakers are able to grant permission for us to include their slides.

Making the most of Open Data
Louise Corti, Associate Director, UK Data Archive

Thank you very much for inviting me.

I've been asked to talk on making the most of open data, so I thought I'd tell you a little story about preparing some open data and then taking it through to an app challenge, so I'm just going to give you a little quick review of that.

So, I wanted to talk here about high quality open data, not just open data, because there's so much around and I think we do need to focus down on what we mean by open data, and how we get to high quality open data that we can use and trust. So, I'm going to tell you a little bit about our various pathways to access that we use at our data service to make sure that the data is going through the right pathway. I'll tell you a little bit about Open Data Institute Platinum Certification, which we managed to get for some of our published open data. And finally, about running our app challenge, so I'm going to whizz through those three things.

Okay, sorry, so my service, I work for the Data Archive at Essex University, which has been running for 50 years, set up by the Economic and Social Research Council very, very early on with the intention of sharing surveys from academics. We have 50 years of creation to build on and I've been there 25 years myself, so half of that period, so I've seen a lot of change in the digital data landscape.

So, we have 7,000 collections spanning social science and 256 of those are open. Now, five years ago none of them were truly open, but these ones have Open Data licences and that number is increasing, but the majority of them, because we're talking about data from people, organisations, businesses are not truly what I would call open. We support the Economic and Social Research Data and Policy, they've had a policy since 95, they've been very forward-thinking and we are the data police for that organisation. And we work with thousands of data creators to help them think about how they can share data well, and what open data means, and what holding onto data and giving permission means, so a lot of that dialogue with quite complex varieties of data.

We hold all of the best Government surveys, the Health Survey for England, the British Crime Survey they all come through us and we manage that through a formal concordat with ONS, so the data comes to us after six months. We also hold sensitive data too, so a lot of the really, really high quality UK survey data we hold and I think the UK actually has some of the best survey data in the world. And any of you who work in Government organisations will know that our data collection standards are very, very high.

So, we use three devices, three vocabularies for actually the pathways for data, open, safeguarded, controlled and that vocabulary has been used quite a lot. Open it does mean truly open, so managed through proper Open Data licences like Creative Commons open Government licence. Safeguarded means there's some barrier put in front probably because of disclosure risk, or probably because the depositor doesn't want stuff to go truly open and that's a whole variety of having basic registration through to whole vetting procedure. And then finally, we do work with quite a lot of controlled data where we make available data that does a disclosure risk and that's managed by secure labs. We work very closely with ONS, HMRC, Justice Lab to make sure that we're using the same protocols to deliver data.

And a little bit about them, we use these kind of three pillars for enabling safe access to data, informed consent, protection of identities and de-identification and regulated access where needed, so those three things put together that enable us to deliver the portfolio of data from open to safeguarded.

I'm not going to talk too much about it, but we used the five stated principles, which is a really nice protocol for looking after data what we developed with ONS. And there's a really nice little video with an avatar of a researcher about what he does and how he's actually getting to use data, if you'd like to go and look at it.

And basically, the safe data, safe people, safe projects, safe settings, safe outputs is a very nice way to conceptualise how you look after data. So, today we are talking about data that is safe, the open data has been made safe, it doesn't present any disclosure risk so we can use it.

So I'm going to move on a little bit to, well just as part of that, to say how do we get so much data in, how does it work? It's all about using very standardised protocols for doing everything and because we've had the luxury of 50 years to build these, is all about using one depositor licence not hundreds. It used to be that every time somebody wanted to deposit data for republication they'd say, here's our data sharing agreement we can't use yours and it gets very complex, I think you'll all understand that having very, very standardised templates and agreements actually makes the process much easier.

We have standardised metadata, I'm going to talk about that in a minute, a really robust metadata scheme that we use in the social sciences and all the other data archives in the world use the same ontologies and the same descriptive documentation and standards. Timetables are very important, if we say we're going to release data in two months, we need to do that. And finally, we do a lot of training capacity building on how to share data.

So, just a bit about how this came about, we actually had an innovation fund and we want to try and reach out to communities who weren't currently using our data, which is mostly education, policymakers, charities who wanted to see what the open data world was doing because clearly they love data and use a lot of data. So, we decided to have an app challenge with a company that's called App Challenge that actually helps promote data to developers. That meant creating a really well documented data set that was completely non-disclosive and we used a topical data set, it was two years ago called European Quality of Life from 2011 back to 2006, so looking at change in what European citizens felt about their life and households and things. We went through the Open Data Institute to certify our open data and we ran an app challenge, and I'm just going to tell you a little bit about the data.

So, we did get the first platinum certified data set and it actually took six months working with the ODI to do that and we're very proud that we have got that, because it means that it is a truly open data set. And it was quite a painful process to go through I have to say, so I'll tell you a little bit about why that is and why people coming from an academic data world have a slightly different view on the open data world, and bringing the two together in terms of standards has been very important. I've blogged on it if you want to go and look about our experience of doing that and I was actually nominated for a Women in Data Award from ODI two weeks ago on the back of this, because they felt it was a nice piece of work, so it's nice to see all the women in the room who love data as well.

So, just to say the process involves 90 questions and every single question they ask about quality, provenance, ethics, documentation, communication all has to resolve in a URL, so every single thing you say about the data, whether it's your privacy impact assessment, has to be URL. And we're lucky at the Data Archive that we document our procedures very well, so actually everything pretty much did have a URL. The real focus here is on machine actionable metadata, which actually a lot of archives, a lot of catalogues don't go to the extent to which you can really actually grab data programmatically.

There was a tension between preservation standards, which is making sure that things are available in the longer term, all the kinds of preservations you know formats stuff like that we do very well. The link to open data challenges are quite different because in order to get platinum certification you need URIs in your data, now that actually goes against the preservation experts feeling about digital data and we had huge arguments with our preservation people about you cannot put URIs in your data because they're not permanent, you can't do that they might break. So, we did and what we actually did is use something that was more robust like ISO country codes, hopefully will be around in the years to come, but there's not many published ontologies out there that are really stable, so I think we do need more of them and Geoffrey talked about them too, we need to be able to point to these things so we can safely put URIs on our data.

Also then we delivered some of this data via an API, now that's a challenge because developers don't read the documentation for out data sets, so we produce all this lovely data documentation and they don't read it they

just go and grab the data and expect it to be perfect and clean and of course, much open data has issues and problems with it. And most importantly, when you run a survey there's bias in the survey, there's non-response error, there's selection criteria that need to be weighted and balanced against the population to give true estimates, so we had to add weights to the data whilst they're being published. Now normally we actually publish the survey data set and people apply weights themselves doing analysis, so this was quite different things that we learned.

We ran this competition and we had quite a few entries from all over the world, young people, older people, a whole variety of people doing different things, very keen to develop some apps based on this European data. A series of judges judged them and most of them cluster around these four things which I thought was quite interesting, some are kind of quizzes, they were asking quizzes about who is more worried about crime in the neighbourhood in Poland or England, so those sorts of quizzes that they thought people might engage with, kind of social facts I think we like to call them. Then, some linked off to news items on crime, because that was quite easy to do. Some, by some of the younger programmers so some 16-18-year-olds actually wanted to run challenges, so you came up with a statistic like 80% of people think their neighbourhood is dirty. So they wanted to set a challenge for their school to go and find out how to improve the cleanliness of the neighbourhoods, which might seem quite a strange thing to do, but actually within schools looking at these social facts and then turning them into action, it was felt to be something that they really fancied doing.

And then finally, there was quite a lot of visualisation of some of the data with nice pie charts and superimposed onto maps, so quite a nice group of things that we would never really be able to do ourselves, so actually crowd sourcing the ideas out based on really good data is a nice thing to do. And there are quite a lot of other app challenges that go on, but I think it was nice you know from an academic point of view to see what people are doing.

So, the one that won you'll be pleased to know you can have a look at it, it's called EU Quizzes and what it's done is embed some of the social facts from this survey into other facts about countries. Most of them were not fully developed apps, they were probably about 50%-80% ready, so we had to spend time and effort actually getting them ready to publish. There is now on iStore and Google Play two apps called Data Quiz where we've constructed various quizzes based on some of these findings, together with other facts about countries and we'd like to populate them more. But just to say it was a very interesting pathway for us to work with people in the open innovation developer community, a community that we hadn't necessarily worked with and to look at what the Open Data Institute are doing as compared to what we do in our maybe cosy preservation format world, you know the people who are involved in data management we don't often look to see those standards, so it was very interesting to look at the connections there.

Louise Corti's PowerPoint presentation can be downloaded from the following link:
http://www.westminsterforumprojects.co.uk/forums/slides/Louise_Corti_Open_Access.pdf

Margaret Sharp, Baroness Sharp of Guildford: Thank you very much indeed. I think I'm quite excited by all the things that you are doing. Right, now we've got Iain Hrynaszkiewicz who is going to be talking to us, he is the Head of Data Publishing at Springer and he's going to be talking to us about delivering open data, the aims, the challenges and the next steps.