# Data on environmental attitudes and behaviours – with Richard Kingston

## Transcript - The NERC Digital Solutions Programme: making 40 petabytes of environmental data available to users across the UK

Full resource (and video): <https://www.ncrm.ac.uk/resources/online/all/?id=20834>

Richard: Hello, everyone, and welcome to this webinar about the NERC Digital Solutions Hub. My name’s Richard Kingston. I’m Professor of Urban Planning and Geographic Information Science here at the University of Manchester. So, what I wanted to talk to you about today is this NERC Digital Solutions Programme. It’s a five-year, £8 million investment to build what UKRI, UK Research and Innovation, call a national facility. So, this Digital Solutions Hub is in phase 1 at the moment. It’s funded by NERC, the Natural Environment Research Council, and it’s supported by NERC’s five datacentres across the bottom of the screen there. If you search for the NERC Environmental Data Service, you’ll be able to then access data from these five datacentres.

 One of the interesting things about what this programme is doing is our core focus is to work closely with stakeholders outside of academia, so across the public, private and third sector. We also think there’s a lot of academics and researchers that are not in environmental science and climate science, who are NERC’s core users really that would benefit from what we’re doing. So, partly this is all about the fact that these five datacentres, across the five of them they curate over 50 petabytes of environmental data. This data is mainly used by climate scientists, environmental scientists, physical geographers, but there’s real value in this to others outside of those kind of domains, particularly when we look at and think of organisations within government, whether that’s national, regional, local, but also environmental sector, whether that’s in industry or third sector. So, there’s huge potential.

 So, the main primary aim of what we’re doing here is to make better use of this data that sits in these five datacentres. Now, when we proposed what we were going to do to NERC, we said, well, whilst this data is a great resource across those five datacentres, actually where you really get benefit from this data is when you combine that with a whole range of other social, economic, health and other environmental data. And what we’re doing at the moment is just focusing on the UK, and that’s the whole of the UK, all four nations, from a national level, down to country level, down to regional, local authority level, and much lower, down to neighbourhood scales, depending on who the users of this hub may be. Obviously, some of the data that NERC hold goes outside the UK. It’s global, in fact British Antarctic Survey, a lot of their data holdings cover, well, not surprisingly, Antarctica, but also Himalayas and a lot of things around Greenland and the Arctic around ice melting and things like that.

 The other important thing about what we’re doing is this isn’t just some sort of huge data portal. It’s not just a place where you will go and find catalogues of data. This blue image in the top right-hand corner, that’s NERC’s supercomputer, called JASMIN. So, the hub will sit on top of JASMIN and what that allow is something quite unique that a lot of our potential users haven’t had before, and that’s high-performance computing to do analytics and modelling on this data. We currently run until 2025 and then the idea is hopefully, if we do this correctly, that we will then get additional funding for another five years to keep this facility running.

 So, one of the reasons why we were awarded this, I don’t know if any of you... it’s quite an old film now, but if you’ve ever seen the Kevin Costner film Field of Dreams and they talk about this idea of if you build it, they will come. Well, we’re not doing that. Our overall approach around if we built it, i.e. we built some kind of digital solution, this digital solutions hub, and then we go out to the community and say, “Look, we built this great thing, you should really do it”. Well, no, no, that is not the way you do this. We could have spent four years building this out of the five years and in the final year go out to all of these users and say, “Look at this great thing, it can do A, B and C”. But what users then end up telling us, and we know this from previous work is, “Well, great, it can do A, B and C, but we also want it to do X, Y and Z”. And at that point you run out of money and you can’t respond to that requirement. So, we’ve shifted our approach and decided actually we need to put users at the front and centre of this.

 So, we spent a large part of going back into, when we were all in lockdown, a lot of my time was spent talking to all of these organisations, these are sort of the big players at the national level across the UK, but also lots of other smaller organisations, particularly local authorities and smaller environmental consultancies, the kind of organisations that could really benefit with using NERC’s data and integrating it in with their own data. So, lots of conversations with lots of organisations, but it will have funding gaps.

[0:05:16]

 So, as I said earlier, rather than building something first and then going out to potential users and saying, “Look at this great, thing,” we turned it the other way round and said, well, actually we’ll first of all start off with some context scoping, map the landscape of stakeholders, understand, and that was again some of those initial conversations during lockdown, understand what some of those users want. And then we had a really kind of intensive phase from September 2022 through to February 2023, where we went round the whole of the UK doing face-to-face workshops. So, we targeted very particular organisations. We were open for anyone to come, and I’ll say a little bit more shortly about the kind of people that came along. So, we did all of that work last winter. Then through March through to July, all of the knowledge and information that was generated from those user workshops around the UK, that was all analysed and brought together and we developed what are called user personas and identifying different scenarios of what users want to do with this data. That report is available through our website, and there’s a link at the bottom corner of the screen there.

 And now what we’re doing is going back to those users, the ones who ticked a box on their paperwork, saying they were happy for us to follow up with them. And we’re actually digging much deeper into their requirements, so two of my colleagues, two of my postdocs are doing one-to-one interviews. They’re starting in the next few weeks. They’re happening between now and middle of December, where we really dig down into, “Well, you told us this. What do you actually mean by that? What exactly do you want this hub to do within the context of what you’re doing?” In parallel to that, we’ve obviously got a lot of software development ongoing. We’re testing and trialling different approaches, with potentially up to 40 petabytes of data available to users. We can’t second guess exactly what datasets users are going to be wanting to use, but that creates a few challenging issues about how you ingest that data, how you extract it, transform it. How do you load it into this hub? At a time when NERC are also thinking about net zero and not wanting to shift vast amounts of data from a data store into some other environment. And then the idea is from the second half of next year again, my little face in the corner is covering this up, but the second half of 2024, we’ll go out doing user acceptance testing with those users from the workshops, and then we’ll move on to rolling that out to much wider, broader users during 2025.

 So, these are some of the headline outputs from what users have told us and some of the problems they face. They really struggle with the fact that data is held in lots of disparate places, whether that’s within their own organisation or lots of other organisations, national government, things like gov.uk. Data also is not held in format, so on systems that make it easy to search and find. Often if you don’t know exactly what you’re looking for it’s difficult to find relevant data. It’s not always obvious what the purpose of different platforms is and the variety of data that they contain, so over the last decade or so there’s been a pretty big growth in the number of data platforms, not just at national level in this country but other countries. And for a lot of the users we spoke to, they just said they felt inundated with so much data they didn’t really know where to start. So, it’s hard to keep up with the sheer range of different portals, different hubs, different places where they could go to find data. They also found some of these platforms rather clunky, difficult to learn, difficult to use. It will take them multiple clicks to get to the actually the point, to the data that they wanted, and they just found it a rather arduous and cumbersome registration process and often found things where... we find this with some of the tools we develop, sometimes I’ll get emails from people in local government saying they can’t access the tools outside of their organisation because of the firewall restrictions and things. So, a lot of practical problems that we’ve not really thought about that they were raising through these workshops.

 In terms of some of the headline key requirements that they were asking for, users want sufficient access to the data to be able to quality assure it and make sure that we can clean it and transform it into suitable formats for their particular use. So, what this comes down to is what they don’t want is to not be able to first, well, yeah, preview the data really to get an idea of what that data is, what’s the metadata, where has it come from, what’s its provenance. They will have spent a lot of time getting this data then doing things with it to actually then realise it’s not quite the dataset that they wanted, so certainly reviewing a sample of the data to help them determine whether or not it’s suitable in a far easier way than is possible at the moment. Normally you’d have to download, access a full dataset, when they may only want to preview a subset of that to get a feel for it and understand if that’s useful to their needs.

[0:11:00]

 And they also want to be able to keep track of the work and the resources that they’ve used and what sort of processes they’ve applied to datasets. So, the idea there is that the hub will, when you log in, you would have your own workspace on there and it would be able to keep track of your workflows and what you had done previously so when you go back in it’s not completely forgotten about where you were. They’re also keen to avoid what they refer to as duplication of effort through sharing of work that they’ve done on datasets and accessing the work others have done. So, there was lots of discussion around this. So, the idea that if you were working, say, for example, one example is if you’re in the southwest of England and you’re doing some work, you work in a local authority, they were quite interested to know whether or not someone else in another part of the UK had been doing similar work and was there a way of sharing that and some way of having a peer-to-peer learning system built into the hub, which was again something quite interesting that we’d not really thought about.

 They also wanted to know what were the most suitable ways of applying suitable analysis software to data, so what was the best toolkit? Should it be spatial analysis software? Should it be something that will allow them to do coding? Do they want Jupyter notebooks and things? So, what’s the most suitable software that they should be using for the kind of an analysis they want to do? And again, building on previous work done here at the university, something called Method Box that was developed in computer science, where you can put in what your criteria are and it will tell you what the most suitable method that you should apply to that data. Another key requirement as well was allowing users to combine and link their own data with other data as part of the analysis, so an area where they can upload their own local data that they don’t want to share and allow others to see.

 So, what we’ve done with that, and we worked with Open Data Manchester, who ran the workshops with us, we developed this idea of a user journey. And all of our users, there were over 100 people turned up to the workshops from different types of organisations, we could put them into this kind of user journey. Some users would be there to define a problem, others, and some users went through this whole journey but others would only be part of this. So, some users might just be defining a problem, then that gets passed on to another part of the team who are getting data and then analysing that, and others who might take that analysis and then somehow showcase it and then publish it. So, one example of that was a couple of people who worked in the southwest of England in a local authority. They were responsible for telling stories with data to convince the public why a decision had been made, and they were using... one way we can do that is through something called story maps, where you’re integrating analytics within GIS and within mapping tools, but you’re also telling a story around that to help support a particular decision that has been taken. So, one of the things we definitely know we’re going to be doing is integrating what’s called story maps from Esri, the GI software company, on the hub, because this came out a lot around the country. People want to visualise the data in a way that is useful to a much wider audience.

 We’re also able to put our users into what are called personas or archetypes, so all users we’d class as the analyst, the author, the leader, the investigator, the specialist and the data steward, and again we’ve got a report that you can access through our website, about 80 pages long, if you’re really interested in this that digs down into what all of this means.

 We also put all of this data, so this isn’t publicly available, but we’ve got a rich dataset, so all of the users were put into... they’re all anonymised, but we know the different types of organisations they came from, what data they used, which departments they were using data from, and this is an interactive tool that we can use to really understand how users are using data and what they’re using it for. So, yeah, it’s an interactive dataset, with over 100 users from 84 different organisations. These are the people that came to our workshops.

[0:15:41]

 A really important thing that we’re doing is following what are called FAIR principles, so making sure that the data is findable, accessible, interoperable and reusable, and I don’t know if copies of these slides will be made available, but each one of these is hyperlinked here so you can dig much deeper and find out a lot more about the fair principles. But without this, the hub just will not work. You’ve got to make the data findable and accessible in the first place so that users, if you don’t quite know what you’re looking for, you’re still going to be able to find things. So, a good example of this, so NERC have what’s called the data catalogue service. And if you type something into that search, so here we’ve got where will the hottest place in the UK be in 2030? It gives absolutely no results. That’s because it was not designed to answer this kind of question. What you’d have to put in here is UK CP data 2018. That’s the UK climate change programme data for 2018. It would then probably give you a list of data around UKCP18. So, what we have done is taken a different approach and using a large language model, we’ve trained that large language model, it’s Google’s BERT model, on all of NERC’s metadata records, that’s for those 40 petabytes of data. So, in our approach to this, when you say where will be the hottest place in the UK in 2030, it doesn’t give you the answer but what it does at the moment do is give you a list of datasets that will help you answer that question. The next steps around this would be to actually then allow you to click on one of these links and it give you a preview of that data and then make the decision as to whether or not you want to bring that data into the hub for you to do some analysis.

 So, I guess an important thing here is to say that this is not just a data portal. As I said earlier, it will run on JASMIN supercomputer. The fact that the vast majority of this data is spatial means it lends itself to some kind of GIS approach to this. And also we’re obviously integrating this not just with the NERC data but the other environmental, social, economic and health data, and allow users under those FAIR principles to find and explore data. So, we’re building on things we’ve done before. This is an example of work myself and a colleague, Sarah Lindley, developed. We’ve been doing this for about ten years now, and we’re rebuilding this using UKCP18 data and that will be relaunched. It will look very different to this, but will be relaunched early next year with updated climate change projections.

 So, we’re building on things and testing things and trying things and basically developing a series of what we call a set of proof of concepts, so this is to allow users to test different data types and format, test out different types of functionality, and meet some of the user requirements. And we’re doing this through four broad use cases, one around air pollution and health, one around housing and environmental constraints, another one around flooding and particularly flooding in coastal communities, bringing in tidal surge and sea level rise, and then another use case around heat stress. And each one of these, it brings in a range of different data types. It integrates the different data centres, but also having spoken to a lot of these big national organisations and government departments, we know that these are things that they are interested in, so it allows us, the software engineers, to build something and try things out but also produce data products and tools that we know some users are going to find really useful. But that’s not everything. This also has to be quite a general system as well, so using maybe the same methods and tooling and analytical approaches, but you could slot in another use case there depending on what your job role is and what you’re trying to do with data going to that today.

 So, if there are further details, you can get the report from the user engagement workshops that’s now available on our website. There’s an online webinar with Open Data Manchester on 20 November. You can find out more about that if you sign up to our newsletter. That will be a two-hour online event and we’ll talk in a lot more depth about our user engagement process. So, yeah, more on the website, so thank you for listening.

National Centre for Research Methods (NCRM)
Social Sciences
Murray Building (Bldg 58)
University of Southampton
Southampton SO17 1BJ
United Kingdom

**Web** www.ncrm.ac.uk
**Email** info@ncrm.ac.uk
**Tel** +44 23 8059 4539
**Twitter** @NCRMUK