NTDS 033

Key:

I: Interviewer
R: Respondent

I: Okay. Great. Okay, so I would like to start from the basics, from these things, and, well, anything should be easy. If you want to sort of introduce your background and how then you sort of got into MEDMI, at what point you got into MEDMI.

R: Okay. So I joined the project about halfway through. So I started in November 2014. So if it started in May that was not quite 18 months into the project, and I joined because the previous project coordinator had left, they were working on two projects, the other one ended and they needed a full-time job, so they left, the post was advertised. I haven't worked in higher education before, my background is much more in project management, in sort of international development projects, but I had a career break with kids and moved down here and there aren't that many opportunities down here, so when I saw it, although it was a different sector, managing a multi-partner project, I seemed to have some of the skills to do it. It seemed interesting.

I: Yes. So it's been an interesting jump, very different.

R: Yes. I think, I mean, on one level I was right… some of it is very similar, some of the parts of it. Some of it is more… there's an aspect of the job which is more administrative than what I've done for quite a while, which is less exciting, but I think because we had quite early on… when I joined there was nobody in post for the database research scientist position but they were still trying to recruit and we had this situation where [redacted] it had been six months nearly since he'd first left and I think Laura felt there was no guarantee if we kept recruiting we'd ever find someone. So there was a sort of change in plan of trying to take that pot of money for one person, realising you weren't going to find somebody who filled all the different boxes, but to try to break down the role and sort of outsource it to different people, and I think from that point I was quite heavily involved in all of that and that sort of made the role, on one level, more interesting.

I: Okay. Yes.

R: And also I think because…

I: Because otherwise it would have been more involved in the administrative aspects, as you were saying?

R: Yes. I still think… I think because, like any project that's got multiple partners in lots of different places, and particularly with MEDMI where you've got very senior academics on for a very little amount of time and then only a few people who are on for a significant amount of time and they are, on the whole, very technical people in very particular disciplines, and I think to have somebody who is spending more time thinking about it and maybe is looking

all the time at what the different partners are doing and how the bits join together and trying to see where things are being held up and move things forward and just do a bit of coordination and give it some momentum, I think that's the other aspect of the role and I think hopefully having the sort of people skills to be able to do that because I think there was a bit of that that was sort of lacking maybe.

I: [redacted]

R: [redacted]

I: [redacted]

R [redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

I: Okay. So basically, this was really hinting at what I wanted to ask next, which is, so as you got in, what were the main lines of work for you, the things you needed to follow or how you would describe it?

R: Yes, I think a mixture of just organising the meetings and putting in schedules. We had some meetings very early on which hadn't happened for a while and then having more regular calls but with proper agendas and then following up I suppose those meetings and then trying to resolve… because it was in about January, early February when we realised Ceri wasn't going to, so then I got very involved in trying to break that role down into different parts and how we might outsource that - reviewing the budget, which I've done ever since, you know, monitoring it and how we can rejig different expenditure lines and things, although not as much financial management as I've done on other grants because actually a lot of it is done by Exeter, but you still need to keep track of it, that's part of my role. What else? One of the things we decided is we were going to outsource the website, that was an easy thing to do.

I: It was.

R: Yes, that was an easy bit to sort of take off and ask another company to do, so then I was responsible for writing up the brief and researching companies and commissioning somebody to develop the website. So we did that. And I think the other part was trying to understand who was actually doing what and where people were holding each other up because they were waiting for data from somebody else or… they were waiting for replies from people. I think there was a degree that… The project had run out of a bit of steam with both Ceri leaving and the previous coordinator leaving and the fact that data was very slow in coming. There was sort of a bit of lacking momentum.

I: Yes, these are very important things. So why did you need to reconstruct this, of who was actually doing what? Because there was a project coordinator before yourself?

R: Yes. But it was more because there had been Ceri. They left about the same time. So when we broke up Ceri's role, we basically took the pot of money that would have paid him and looked at the job and sort of broke it up into sections. And we've done that successfully on some lines. We went and then tried to find who might be able to help us do different aspects of it. So Christophe ended up taking on much more of the database than I think he'd previously intended to do. We got some help from Exeter University to provide some technical support for managing the servers and things and then we took Ceri back on as a consultant to finish his work on the browser and then we got somebody from the Met Office, a guy called Neil Kaye, to do some technical work. He's developed a visualisation tool, which was to do work on one of the other demonstration projects. So they were all different parts of what had been that technical pot of funding. I mean, it's not a very big… we are not talking about a very large amount of money for the scale of the project really.

And then the other thing we did is there was funding in the project for some pilot projects but nothing had happened about that, so early last year, about this time last year, we put together a sort of scheme and I think we made more money available because we realised there was some more funding

and we did a scheme and we advertised it and we ended up funding six pilot projects.

I: Okay. So the pilot projects, what did you mean that nothing had happened about that?

R: Well just that the… the language was a bit confusing in MEDMI.

I: They had not started.

R: They hadn't started. So there was always talk about funding some pilots, but nobody had got that far. So MEDMI, the idea that it was a sort of proof of concept was that there were in the proposal three demonstration projects.

I: Yes, I see.

R: But in addition, there was opportunities to fund some pilots with other sort of interested parties.

I: Okay.

R: And they were much smaller amounts of money. So we funded six things for about £5,000 each.

I: Okay. Six additional…?

R: Six additional small bits of work and that was things like that… that was Jess Tyrrell doing some additional work on Ménière's Disease. There was a PhD student who was here called Mark Cherrie and he worked on two. I mean, I can give you all the details of this because I've got a summary sheet, if that's helpful. We funded Christophe Sarran and Debbie Hemming at the Met Office to do some work on pollen and then externally we funded some work with SAIL at Swansea, which was with Ben Wheeler, and we then funded lastly somebody at Southampton University, Sujit Sahu, doing some work on air pollution. And again, that was with Christophe. A lot of them linked in to Christophe because that's the data we were holding, the environmental data, so they were all sort of needing access to Christophe and things, to send data and things.

I: Okay. Yes, yes, yes. He was mentioning that. I had not realised until I was speaking to him that he was also gatekeeping.

R: Yes, he has been a gatekeeper.

I: Yes.

R: And I think that's the sort of other thing I've done, is coming up with things like a data access policy because there wasn't sort of anything in place.

I: Okay.

R: So researching what was a sensible policy and what would be our policy to accessing it and what controls we'd put in place.

I: Okay. To access what exactly, in the sense that…?

R: To access any of the data on the MEDMI server, what would be our policy and how would people do it? Because we haven't been inundated with applications but we've had a few and we also realised that we needed to have a procedure of approvals as to who would be allowed to access it and that we will keep records. So again, I need to update that. But again, there wasn't anything in place for that.

I: Before, yes.

R: Before that. And then later last year we organised this workshop with Biobank, so we had a big meeting with them and we invited lots of other owners of environmental data. So we wanted to inform people, get people to find out more about Biobank, more environmental… people on the environmental side, but we also wanted to tell Biobank and some of the other health researchers what other environmental data was available, so we got people from NERC and from CEH and from the Environment Agency and from Defra, and that was a whole day workshop, so I sort of organised that.

I: Yes. Okay. It's a lot of things. So the data access policy, going back to that, all these pilot projects need access to the data that Christophe is managing.

R: Yes. So the way it works is the data is sitting on the MEDMI server at the University of Exeter and in order to access it you need a login and Christophe… in fact, he doesn't. The University of Exeter's IT staff set up that login, but permission needs to come. Basically, Laura needs to give permission as the Senior PI. But those sort of requests come through me, I keep a record, and what Christophe does is Christophe has got a mailing list which he set up through the Met Office for everybody who has been given access so that he updates people as he's developed his extraction tools or if he needs to tell people that the server is having work done on it. So he looks after that. And really to date it has been primarily either direct MEDMI staff or researchers or expanded slightly to include the pilot projects. So we gave other people access and then I think we've had applications from one or two other people, but they still work within some of the MEDMI partners, so like at PHE.

I: Okay, yes.

R: So I don't think we've had… We might have had one application from somebody from a completely different university but they were still an established researcher. So those were our sort of criteria. And at the moment, the data that we've got on the MEDMI server is it's allowed to be shared. It hasn't got constraints on it, but what we had to discuss was that we've applied for some ONS data, which they said we could have, but we still haven't got, but some of that data was very strictly allowed only to a handful of researchers and to the point where it might even not go on the server. But it could do. What you could do is make… You could use the server to be the control… you know, you can have secure access that that pocket of data is only available to certain people. But we've yet to have to deal with that because we haven't yet got the data that has those restrictions on it.

I: Yes. Okay. Yes, yes, yes.

R: So it was sort of thinking ahead of if we were to get that sort of data, how would it work?

I: So there's all the sort of controls prepared and stuff.

R: Yes. For demo one, for the browser that Ceri and Shakoor have developed, there are two versions of that - there's a public version and then there's a login version. And again, it was set… so those were all part of the sort of data access policy as to who would control… who would be given access to that.

I: Okay. That's also specified basically there.

R: Yes.

I: Yes, that's also why I was asking what data it regulated because obviously as I was doing these interviews I was coming to understand that there's these different locations, so there's the system that Christophe talks about, this different system basically from what Ceri has been working on mainly.

R: It is. And I mean Ceri's browser is on actually a very limited dataset that is totally ring-fenced as a unique dataset. And in fact, there's two versions of that because they added some additional data, which included pollen and HES data, and so that's why we've got two versions of it.

I: Okay. Okay. Because my understanding was that it had been developed with these ring-fenced data because it was already available, so it started on this data and they were waiting to point these applications basically to the incoming data that Shakoor was eventually getting.

R: Yes, so…

I: Something like that.

R: Yes, so if on the memory server we've got the main data, which is basically the Met Office's environmental data, and then we've got some health data from Public Health England, which is the SGSS, which is the infectious diseases data, the data that Ceri and Shakoor used in the browser, it is some of the same data that's on the server, but that's not how they get it and they don't… the two are not connected, so Ceri may… he had to sort of set up a data… I suppose a small database and a dataset for his browser to run off.

I: Yes. Yes, okay.

R: And that also was using ONS data, mortality data, which Shakoor had. So they had the data already.

I: Yes.

R: So that sort of sits in a unique bit.

I: Yes, yes, but it's also present on the server.

R: Yes.

I: Okay. Actually you were saying also that the University of Exeter stepped in with the IT services, well the IT services of the University of Exeter sort of stepped in in a sort of bigger role as you were partitioning Ceri's or…

R: I mean, it wasn't clear when I joined what the expectation with Exeter was, but the servers have always been sitting… well the project before I started, the servers were bought. That was one of the first things that happened. It took quite a long time by all accounts. But I don't know what the thinking had been as to who managed the servers, whether that was something that was going to be expected of Ceri. But they are sitting up in Exeter, so one of the things we did was speak to the Exeter IT people in, I think it's their infrastructure team, so Bill Edmunds and Dave Barker. I don't know if you know them.

I: No.

R: And we spoke to them quite a lot to see whether actually they would be able to help us with some aspects. Again, I don't know whether this is stuff that Ceri might have done but, for example, so every time we do have somebody give permission, they are the one who set it up on the server. But also recently we've had to do quite a lot of upgrades to the software on the server.

I: Okay. And you needed…

R: And they've done that. And we had a bit of a nightmare when it came to loading the website onto the servers because the configuration wasn't right, the versions of the software on there, so they've had to deal with all of that.

I: Okay.

R: And that's been absolutely critical. I mean, if nobody had done that the whole thing would have sort of ground to a halt. I mean, at the moment, we don't have anybody on the project who has got the skills to be able to do that, if we didn't have that support from Exeter.

I: Yes, yes.

R: Does that make sense?

I: Yes, yes. Yes, absolutely. [redacted]

R: [redacted]

I: [redacted]

R: [redacted]

I: [redacted]

R: [redacted]

I: Yes, I've read the papers, a couple of papers when they published, I think.

R: Yes, and they are just waiting. They put in a publication for one, but they were waiting for permissions from Biobank for that.

I: Okay. And also you mentioned, as you came in, you had to sort of make yourself a map of who was waiting for others to do bits and also maybe sort of the state of accessing the datasets and one day become accessible or who was in…

R: And who was doing what. So in terms of, if you have your senior staff in each of the institutions, but then the people who are actually… and they are like 5%, so they come to meetings, they give direction. But Christophe was on 50%, so he had a big slug of time and then Majeed, Abdul-Majeed who is at PHE. I think they would probably… is it fair to say both of them, it wasn't particularly necessarily joined up, what they were doing. So, I mean, Majeed was working on time-series analysis of campylobacter, so they were doing work, but I think maybe feeling a bit isolated. And I'm trying to sort of… Christophe had also been off for a bit of time [redacted] And also, when Ceri was there, I mean, I think there was a long time when people were waiting for data, there had been… So Majeed… and in fact, Majeed had been working on a whole lot of time-series analysis. But the person who has been very good, who wasn't originally on the project is Gordon Nichols. At some point last year Majeed became more under Gordon and he's working with them and they've made better progress with their paper and it's linked up, because there is quite an overlap in certain areas. There's another project which you've probably heard about, which is the HPRU, which is… The Health Protection Research Unit in Environment and Health and it's basically got exactly the same partners as MEDMI and it's got slightly different areas, but there's quite a lot of overlap and a lot of the people working on it who work on MEDMI work on that as well. So Gordon Nichols was on the HPRU, working for PHE, but he hadn't originally been on MEDMI.

I: Okay.

R: But there was an overlap of work and then somebody, the person who was at MEDMI, Anna Cichowska, she left last summer because she went on maternity leave, so funding was given… We'd already given Gordon some funding and then we gave him more funding from MEDMI, and so that sort of joined things up more, and Gianni… I can't pronounce his surname. Have you come across Gianni? [redacted] he works at PHE with Gordon and they've done a lot of work together.

I: Oh, Lo Iacono.

R: Yes, there we go.

I: Yes, Giani Lo Iacono.

R: Yes, yes. So they all worked together on a paper.

I: With Majeed.

R: With Majeed. And that literally was submitted at the end of last week for publication.

I: Okay.

R: And I think for quite a long time they were waiting for data from the Met Office.

I: Okay. Okay. Right, yes. Okay. Why were they waiting for data? So were they sort of (ph: 0:30:31.4) pinging you?

R: I mean, we had meetings from May last year where they… The data was there from the Met Office, but it wasn't very accessible.

I: Okay. Was that from a technical point of view?

R: Yes. And I think, I mean, this is where my technical knowledge starts being limited, but the extraction tools were not in place to make the data. I think one of the things about the data that they wanted - I'll get a piece of paper because it's easier to explain it – is the Met Office data, the MIDAS Met Office data is a lot of daily data for a whole lot of variables, but in order to link it with the SGS they wanted weekly data and they wanted it by lab, so what Christophe had to do was he had to interpolate the environmental data so you could join a location, but he's had to do a lot of work to do that. So the data was there, but it wasn't in a sort of useable format. Maybe that's the best way to put it. So they had to change it and then there was quite a while where we had the SGS health data, but there were issues because the GEO coding hadn't been done on that, so the linkage points weren't right, so that went back and forth a few times. And I think that in fairness, this is where Christophe has been learning, he's had to teach himself without Ceri there, he's had to teach himself Python coding to do a lot of these extraction tools. I'm not entirely sure what Christophe would have been doing if Ceri had been here.

I: Right.

R: It's not clear what exactly his function would have been, but what he's ended up doing is he has written a lot of these processing and extraction tools to make the two lots of data linkable.

I: Yes, yes.

R: And it's still being developed. He's done enough to be able to send them some data. But literally he's finished that this year and they've been wanting some of it since last summer. So there was some delay.

I: So were there sort of concerns that there was further work, or were they sort of saying, you know, 'Just let us know'?

R: No, I think there was frustrations. I think they had other things they could work on. But yes, there's been frustrations, I'm sure.

I: Yes.

R: Yes, if you are waiting.

I: And I wonder if they sort of had to… in what ways… I mean, it's always like that in research, you always adjust what you do according to the way things come together.

R: Yes.

I: I'm wondering sort of with them what this meant.

R: Yes, I think for Majeed probably because he was on… he was meant to be on 50%, but he had quite a long time on 100%, and I think… it may be interesting to ask him, but I expect he felt quite isolated, he was working on this thing on his own without much support or time and without access to some of the data that he probably needed and wanted. But he was obviously… So exactly what he was doing, I'm not 100% clear, but it seems to have moved forward. But I think since he's worked more with Gordon and Gianni, they've found other things to do but they also have had a frustrating time waiting. They wanted to do a sensitivity analysis. They wanted to see whether they could… Basically, we could get the laboratory data for SGS, but we couldn't get the residential data to link because of reasons of confidentiality.

I: Ah okay.

R: But for campylobacter they had been given permission for residential data and what they wanted to do was a sensitivity on the locations for the environmental data to see if actually it was a reasonable substitute to use the laboratory rather than the residency.

I: Okay. But they couldn't access that.

R: But they couldn't do that for a long time. They've only been able to do that in the last couple of months.

I: Okay. Then they have eventually been able to do it.

R: They've just been able to do it, very recently, and that was because Christophe has been working on these tools.

I: Okay. Okay. Well, and also because then eventually the accesses were sort of figured out.

R: Yes, I think it was more that until Christophe had done the work he needed to do, it wasn't available. And I think Christophe… I think what Christophe was doing, there's been some confusion as to was he preparing the data so that it could be linked to the browser? I think there was a desire for a long time that Ceri's designed this browser tool that takes this limited dataset they have and makes it interrogatable through a browser and I think there's been, for quite a long time, potentially a misunderstanding or difference of opinion as to whether eventually all the data that we have will be able to go through that browser. And Ceri has been saying for a long time, and we've been trying to pass on the message, that is beyond the scope of the work he could do, and maybe originally if he'd been here full-time that would be possible, but certainly on the amount of time we've had him for that wasn't possible and that's in a way why we've got this person Neil Kaye at the Met Office to do the

visualisation tool because that was going to be able to look at a lot more data and that's why we spent quite a long time trying to speak to Black Swan as a potential option to be able to link a lot more data.

I: Right. Okay, I see. I'm trying to sort of recap in that project for example, then there was both sort of… Majeed was a bit isolated because he was waiting for Christophe to get the tools to work in a way he needed but then also at the same time there was also the sort of waiting for the residential data on health records accessible, so it was like a double…

R: Yes, it was a bit of both; it was a bit of double. And actually, that's why… I mean, the most detailed analysis in the paper they've just published is on campylobacter, and they had some of that data, not all of it. So that's what he was working on. He was doing some very, very detailed times series but the idea was to broaden it to lots of other pathogens and for a long time that wasn't available. But, for example, to do the sensitivity analysis they didn't have the data.

I: The data from an access point of view.

R: They were able… Because they work in PHE they were able to access the data through PHE and that's what I presume Majeed was working on, but the data wasn't on the MEDMI server, so Christophe wasn't able to do any linkage because it took a long time to get that data. Does that make sense? It's like they had a separate…

I: Basically, they did the linkage.

R: They did the linkage themselves, yes.

I: Okay, okay. Yes, because with the other people I've talked to so far this comes always up as something that was a bit of a question mark across the project – when do we get access to the data, and on the other hand also when do we get the tools we need, basically, for the job? Obviously, everybody has got a different idea of what they need.

R: And I think in terms of access to the data, for quite a long time there wasn't the data readily available. So if I think about it, when I joined in November, there was lots of environmental data, but it wasn't very accessible because the tools weren't in place and people didn't know how to get to it. But the health data wasn't even on the servers, so we only got the health data… I mean, if you read the original proposal, there was going to be all sorts of health data that MEDMI was going to obtain, but we haven't got most of it, so the only really substantial health dataset we've got is this SGS infectious diseases data and there were also some delays with that because, well, PHE went through this big restructuring and they've changed that because that data used to be called lab base and now they've called it something else and obviously there was some issues with when they swapped from one system to the other, there was some concerns about the quality of the data, there was still some issues with it. But it was only in the autumn that we managed to get that data and there were complications, there was a lot of back and forth trying to get it, so that data only really got put onto the MEDMI server in the autumn, so two and a bit years into the project. And the ONS data, that application has been pending since I joined, so that's 18 months, and that's

been a very frustrating process, and Majeed has been trying… one of his jobs was trying to chase that.

I: Okay. So what kind of complications were these…?

R: Their procedures for releasing data are very… they are very slow and we don't fit their neat boxes -

I: Okay.

R: - of how, you know, the reasons you want their data. But they have these meetings and you put in an application and they go to a panel, but it takes months anyway and then they come back and they ask questions and you reply and then months go by. But I think that… We eventually had a phone call with them earlier this year and they said that was fine but we needed to send them… they wanted to have a look at the browser, but it wasn't ready to share. But I think why the data hasn't been released, I don't know. But it seems that… I mean, I know other people in the centre are waiting for other data like HES data. I think some of the health data, the release of it… the procedures and the release of it are very, very slow.

I: Right, okay.

R: I mean, this is for me all new.

I: And they go back and forth several times.

R: They go back and forth about what it is.

I: There's not a clear sense on this side of the process of what's the progress.

R: No.

I: Okay, I see. So was Majeed's pilot project the one that sort of encountered the most difficulties?

R: His wasn't a pilot project. So he was on the demonstration project, sorry, language. Probably, yes, because they've been the ones… because actually Ceri's data, he had the data, he was developing the browser, whereas demonstration project two wanted access to the data and they wanted the data to be linked. So they've been most affected by this not being joined up and they've needed Christophe to do the work. So I think… Have you spoken to Gordon Nichols?

I: I'm meeting him next month, yes.

R: Okay. So he will tell you from their side sort of the frustrations of asking for it. And also they want… Christophe has got a list of all the Met Office data, all the different variables, but it isn't all in the sort of same format and Gordon would like to be able to get all of these different variables and be able to link them, and what you realise is that in order for that to happen, Christopher has actually got to do a lot of work to write the extraction tools. But it seems like we are now getting there, but I think that it's been slow because basically

Christophe is the only person who can do it and he has had to learn, he's had to teach himself sort of the necessary coding to be able to write the tools.

I: Right, yes. And (ph: 0:44:45.6) instead, and the converse, what was the demo project that had a smoother life?

R: [Laughs] I'm not sure if any of them have been very smooth. Well there's only the three of the demonstration projects. So that was two, that was the Gordon, and that was the infectious diseases. The first one, the browser, well, in some ways that was more straightforward apart from Ceri was working on the browser and he then left, so we then had to sort of persuade him to come back to finish that and that is now finished. If he'd been working on it full-time I'm sure it would have been finished much earlier. So that's not been the smoothest in the sense of who has worked on it, and I think on that project they had hoped to get some other data from people like TPP. In fact, Shakoor has just been given some permission. That data will not come onto the server but at least he can work on it to do… because there's the sort of idea of MEDMI being this big platform and server full of data and then there's actually the fact that the project, the actual research is using data that's not on the server but they are just doing some interesting data linkage between health and environmental data. So I think Shakoor has probably been held up by a lack of access to data, because I'm not… I'll say this carefully. I'm not entirely sure what Shakoor… because he's the other person… sorry, he's the other person who is on 40%.

I: Yes.

R: Yes. So he was…

I: On that…

R: On MEDMI, yes. It's not entirely clear what he's been doing, but partly because without Ceri there and without access to some of the other data it's been quite hard for him to move forward.

I: Yes. And I have the understanding that that project basically developed a little bit, yes, separate from the rest of the platform because he had access to some dataset that was sufficient [overspeaking].

R: Yes. There's lots of reference in the original documentation to this platform and that being the source of all access, but I think the reality of that, it's not happened as it was envisaged. And then the third demonstration project, which is the sort of one related to oceans and harmful algal blooms, that has been much more delayed because basically they wanted to get a PhD student to work on it and it took them two or three rounds of recruitment. So the student literally started in about February. So that work will be ongoing. And there have been some issues there about accessing data, but I think they are making progress. So in some ways, interestingly, the fact she started late may not be a (ph: 0:48:28.4) *train splash* because I think now that she's obviously… I think things may move forward with the data.

I: It's better to start later.

R: If the data is there. So that will run on beyond the end of MEDMI.

I: Right. So in respect to what you were saying about how it's not really happened as envisaged, of having this infrastructure that then would be sort of the base tool for conducting research and stuff, I think we've already touched all the main areas of why, so sort of summarising why was that, because of the delays that were in several parts of the infrastructure development and sort of ran the demos to basically… the development, a bit more separate on their own because the infrastructure was not ready, or was that also management sort of…?

R: I think there are probably multiple reasons why. One is that we haven't got all the data we originally wanted, and I think one of the realisations is getting hold of health data because of the issues of anonymity is that that's been much harder to get hold of and even adding other environmental databases, we haven't even really tried to do that, but then I think then it comes down to the capacity that… what you need are technical skills to build a platform and the project is very light on technical skills. We had one post and we lost them really and we've divided it up, but once you start dividing up one person's salary and you start trying to use people in a sort of consulting manner, it doesn't go very far in the world of IT. So in order to have achieved this amazing platform that was originally envisaged, you'd have needed a lot more technical resource of database management and software development but then you'd have also been hampered by not having as much data as you'd have wanted. Maybe if we had more technical resources and we'd moved quicker with the data we'd got we would have been then in a better position to go out seeking further data. Certainly environmental data, there is lots of environmental data out there, but we've been busy enough trying to make the Met Office data accessible, that we haven't proactively gone looking for other environmental data.

I: Yes.

R: If that makes sense.

I: Yes.

R: And I think the other factor is that the technical understanding in terms of technical from a very IT technical way, software type technical way of the senior researchers, that's not their area of expertise and therefore understanding what the requirements are and the needs are. I think Laura sort of… she gets the concepts of it, but when we've had meetings when you… there's a point where Ceri or Christophe, or when we speak to Black Swan, they start talking about types of database software and nobody understands that. There isn't somebody who has that area of expertise.

I: And in what way does this sort of affect, is that because then they sort of prioritise…?

R: I think if you don't understand the technology from an IT point of view in what is a data or how you manage, then you are maybe not very realistic about what is required and how long that might take and how complicated that might be. So we've had conversations with people where they've said well that would be easy, it's easy to add all the data to a browser and Ceri saying, "It's not easy. You don't understand what is involved in the amount of programming required to do that," and that is a, you know. I don't know how you come across that with other data projects you work on, but they are very

particular skills, aren't they, to understand how you manage data and how you use this very clever software that exists to deal with this huge quantity of data and to process it and link it, whether you want to use post QSL or whatever or whether you want to be using… Black Swan were using… oh God, it's going to slip my mind.

I: Postgres.

R: No. They had…

I: No SQL?

R: It will come to me. But they use one particular type of software. But then I believe that Hadoop is another one, but it's different.

I: Map produced.

R: I can't think. Anyway, it will come to me. It's just there. But those different types of software of understanding that and what they are capable of and what are the pros and cons of them and how, if you go down one and how you can join it with another, they are very technical questions.

I: Yes, and difficult to… Yes. So all these…

R: Elasticsearch.

I: Sorry?

R: Elasticsearch. That was what Black Swan used. Sorry, that's just a… we didn't go with. But there are obviously software out there that's designed for absorbing large, big data.

I: Yes. How did the thing go with Black Swan? Was this idea their… Obviously, I guess, it was part of this idea of dividing…

R: Dividing it up. We went to talk to them. It was originally through Mike Depledge, who had met Steve King who runs Black Swan and he's been talking about various things, but he's interested and excited by their capacity to absorb and deal with big data. So we went to have a meeting with them and I went along and I immediately could see that that could be useful for MEDMI as we were… that was back last summer, I think, originally met with them, and we ended up having quite a lot of discussions and they could have absorbed our data very easily, they would have used the elasticsearch to extract our data. But when it came to it, they were very used to working with a lot of… almost more verbal data, they used lots of [overspeaking] and unstructured, and the linkage… they could cope with huge quantities of data, but their linkage was in some ways not very sophisticated, there wasn't much modelling of the data, and when they came to present… you know, we had lots of talk with them and I had a lot of discussion with them and with Ceri, but when they actually came to present to researchers to see if we wanted to do it with them, the academics weren't very impressed with what actually… you know, it wasn't good enough for them to just link data without being able to put in any lags or do any analysis or do any modelling. So from an academic research point of view… now, their idea is that if they pull in the data you can then write software to run models on top of it, but we didn't have additional

money, we'd have had to give all our money, our little pot of money to do the first bit, we wouldn't have had any money left to add on the modelling, so we decided in the end not to go with them, but I think… and that was sort of probably right, but they certainly had the scale and capacity to absorb the data that we've got and maybe… I don't know what would have come of it. But I think the fact that their data was very different, it didn't grab the academics.

I: Yes. So it was also maybe a timing and resource kind of…

R: I think timing and resource, yes. And we didn't have much money to give to them, so. But just talking to them, you realised again they were asking… they divide up… they have database experts and they have software developers and then they have the front face interface and that was all of Ceri's role and they have very particular people who focus on those different areas. I mean, I think the expertise they have, you realise compared to the scale of what we are trying to do, is more, but again talking to them in the detail of what they could do and how you might add programming. Ceri saw elements of it being very exciting, but you'd have needed quite a lot of work to add on to it, if that makes sense.

I: Yes, absolutely. I would sort of ask a conclusive question because it's been extremely informative, very rich. Basically, the question is very simple, is would you have designed the process of the whole project differently if you were to do it again or would you sort of…? Because there's also some, I think, maybe timing, obviously, the project was within three years and some of these kind of, yes, pacing between the various units of the projects, things, well, if you had all the time you need, if you had 20 years instead of three then you could say all the demo projects, just wait a second, and finish this and stuff. So what would you do different also considering if you didn't have constraints of resources or time?

R: I think the main thing I'd do differently is you'd need more technical resource on it from the beginning who understand it, who understand what you can realistically achieve with the amount of money you've got. I could say you just want lots more money, but I don't think it's that, I think it's having people who understand the technology enough to know what different aspects cost and if you've got this much, realistically this is what you can achieve. I think there are issues with accessing health data. I think the idea of MEDMI will be constrained by the type of health data you can get. I don't think it's possible in the foreseeable future to set up the sort of platform that was originally envisaged of both environmental and health data because I think certainly at the moment the way the health data is sort of made accessible, the fact you want to link it, it clashes with the anonymity, and I think that some of the big health resources like Biobank, the way they do it is they control it and they do the linkage and you send them your environmental data, and maybe you need to be more realistic about that, but there's still therefore a role maybe to do to make that environmental data more accessible and more in a position to send to the health… the owners of the health data to link it. I think down the line, with the technology that is probably existing, you could… I mean, one of Ceri's issues was that if you did manage to get the health data, you know, your firewalls and your protection need to be very sophisticated and we haven't got that resource but that the ability to… the technology is there to do very… to protect access and to protect access to certain datasets because

let's say you had a huge amount of data but only some pockets were much more sensitive, if you've got the right technology I think you can protect that data, but that's going down the line and using a lot more resource than we've currently got.

I: Yes, resources and things.

R: But that's where MEDMI is in a way a bit funny because it had this idea of a platform but then it always talked about just being a proof of concept and the proof of concept was it would try a few things. And, when you think of that, then it has tried a few things. It depends what your vision was again.

I: Yes, of course. There's this library of tools, yes.

R: But I think if you were trying to create a platform of environmental and health data which was part of MEDMI, that then people could use as their source of data to do some demonstrations, yes, I think just more technical resource. And not just the people… not just more suitable IT skills but also – and this may be very hard to find – somebody who understands the IT but also can talk at a senior enough level with the academics.

I: Sort of intermediary figure.

R: Yes. And, I mean, in a way, doing the job I'm doing, if I understood the technology more I could add more value but then I would being paid a lot… it's a different level of person.

I: Yes.

R: But IT projects have project managers on them for exactly that purpose. The people who can… Also because you realise that in a way it's a bit like… when I first started this project I was slightly overawed by the technical expertise, just academic technical expertise, but then realised that normally when you had a meeting in a room, everyone was very technical but they didn't understand the other person's. So you had an epidemiologist sitting next to the meteorologist. Now, they were very strong skills but very different. So you are always having to talk across disciplines.

I: Yes.

R: And IT, it's just another discipline in there. And within IT you've got your 'are you a database person, are you a software…?' but in order to be able to converse across all these disciplines you need to understand them enough. Does that make sense?

I: Yes, of course. And the IT was definitely, of all desires, probably the one that was less represented because it was represented by one person or two people, whereas the others…

R: Yes. And I suppose that makes the project vulnerable.

I: Yes. Well I guess all the sort of the voice of that side, of (unclear 1:06:10.4) tends to be overridden by other kinds of priorities or issues.

R: Yes. And also that the senior people on the project, they cover the other disciplines but they don't cover the IT discipline or the data discipline, which is a growing discipline I think if you work in a big data project.

I: Yes. Absolutely. My PhD is in information systems, which is an area of that level between management and IT. Definitely, I am very convinced of the importance.

R: Yes, I think so.

I: The importance of that sort of intermediate field.

R: And I think that otherwise… there's a risk if you don't understand it, you are asking people to do things without any concept of… it's like saying can you build me a house, but you don't know… you design a house without having a structural engineer to say what's possible. I suppose that's my analogy.

I: Great. Are there other sort of issues or topics or points that you were thinking about along the interview that you would like to mention? Because I've touched on all the topics that I wanted to.

R: No. I mean, I think…

I: It's been extremely helpful.

R: Yes, I think the project has… it has gathered momentum in the last six months, so we've got this extension now, and I think we'll sort of end up achieving some things quite late. So it's not that… you know, there's moments you've wondered what is it going to achieve, but I think it has done stuff and I think if we can get to where we hope to go with making the data more available then that's the sort of… that's got it somewhere, of satisfaction.

I: Yes, I think it gives it a sense that, yes, this didn't go exactly as in the best way some would have imagined, like having the research happen on the infrastructure and stuff, but when you look at the points…

R: Yes. And the pilot projects, I think they've been good, and I think both demo one and two are now… there is stuff coming out of them, and demo three has got a PhD student, it's getting there, and the data, hopefully if Christophe can finish what he's doing with the environmental data and then the idea is we put a sort of wrapper on that to make it more accessible. Because I think one of the other issues is that the data needs to be accessible to researchers, so it's the format of the data and that at the moment there is this data sitting on a MEDMI server, but we've had various academics ask to use it and when we tell them that you are going to have to dial in with a server and you are going to have to write some command-line code, I mean, apart from one researcher, they've all just… they are completely flummoxed by that, they don't know how to do it. So if that's as far as we get it, we haven't exactly made our data available. So I think if we can… What the idea is is that we create… it would be a web interface that will link with Christophe's… all his, the processing and extraction tools and you'll have sort of dropdown menus, but it won't be a browser where you can just literally click on the buttons and there will be the data, you'll say, 'I would like these datasets,' and it will then behind the scenes process the data for you and then drop it somewhere,

probably some of it will be too big to email but at least that means that a researcher could go into MEDMI and say, 'I'd like rainfall and temperature and I'd like these bacteria for these years and linked,' and it might take a week, it might take a day, but they would get that data, whereas at the moment they ask for it, we say, 'Dial into the server,' and they are none the wiser. Yes.

I: Yes, yes.

R: So to me that would be a good step to get to.

I: Yes. But this possibly could be for the next iteration or…?

R: Well we were hoping we might be able to do that in the next six months, yes.

I: Okay. I know Christophe is trying to…

R: Yes. So Christophe is hoping and then I'm not sure whether Ceri will do it or we'll have to find somebody else to do the interface that links with it. But I think that… yes, well it comes back to the sort of resource. We haven't got anyone who can look at what Christophe is doing.

I: Okay.

R: He's taught himself Python and that is somebody… I think somebody at the Met Office will have to look at that to make sure it's… because there's the… I suppose it's the integrity, isn't it, of your data. So somebody needs to review all this processing.

I: Yes.

R: Sorry, I've talked rather a lot. [Laughs].

I: No, it's brilliant because we touched on very interesting topics, I think, and I like this point about accessibility from a skill's perspective as well of the data, it's not only obviously confidentiality.

R: Yes. And that's why Jess is unusual because she could access the data in its current format.

I: Who?

R: Jess Tyrrell.

I: Oh yes, of course. Yes, yes, yes.

R: Yes. So she was able to just go onto the server and get the data she needed to add to the Ménière's.

I: Yes, absolutely.

R: And I suppose that's the other way, is that you have somebody who can access data for other people but as an ongoing resource that's not… well we haven't got somebody who could do that for someone.

I: Do that for… Yes, so it's a service resource.

R: Yes.

I: Yes. Okay, thank you very much for all these insights and conversation. Thanks a lot.

(End of recording)