NTDS 028

Key:

I: Interviewer
R: Respondent

R: How is that doing, is that okay?

I: Yes.

R: Right. So if you kick off with one of your topics.

I: Okay. So, basically, we can start from the start, sort of if you tell me how you got involved in MEDMI, what kind of role and with what kind of expectations?

R: Okay. Well, I already knew Laura and we'd had several discussions about different kinds of research things. I am an applied statistician with a very general interest in epidemiological modelling. My technical expertise is actually in spatial and temporal epidemiology. So I'm not an epidemiologist, but I do know about how to model spatial and temporal data, and a lot of epidemiological data, of course, is spatial or temporal.

So I think that Laura was interested in me being involved mainly because I have experience in modelling this kind of ecological… we would call it ecological data really in epidemiology because it's group data, it's not so much individuals, yes? Anyway, I have quite a lot of experience in modelling this sort of stuff. I think she just felt that I might be helpful in more than one maybe of the sub-projects that were kicking off because one of them was looking at the suggestion when we were proposing… one of them was going to look at heat waves, another of them was going to look at modelling some kind of… well, it ended up being campylobacter but it could have been other things in space and time. So I think she just wanted me to be involved because of that, as a consultant, almost, on the project. Yes, I was very happy to do that.

So my role, though, it was only a very small proportion of my time on the project, I think it was 5%, yes. However, they are a good bunch of people to work with, so I was very happy to join in with the meetings. A lot of this stuff has been telephone conferences. And, personally, I do not find them very satisfactory, to tell you the truth. I find it really quite difficult to actually get much productive discussion in a telephone conference. However, it would have been difficult to run the project in other ways because… I'm kind of going a little bit off piece now, but let's just run in that direction.

I: Yes, feel free.

R: Because there were several different people involved from different parts of the country, getting them together would have been really quite tricky. Maybe if we'd have got a bit more into group video meetings. I'm talking about simple ones here, I'm talking about, say, using an application like Skype or there's another one out there called Hangouts or something nowadays. I've used those in my own other research projects with people internationally and they can be a bit glitchy sometimes, but they are preferable to a telephone conference. That's my own feeling.

So that was my role in the project really and that's how I got involved in it. Have I been very helpful to the project? I think some of my involvement in the very early days about data structures and what sorts of things to look out for and that kind of stuff might have been useful input at that stage. I don't think I've really… and then I've had quite a bit of input into the modelling of, I think, Demo Project 2, which is the campylobacter. Again, the early part of that pilot project, just sort of checking out the models from a mathematical point of view and making sure that roughly reasonable things were happening. But apart from that, I've been to the project meetings when we have had one-to-one meetings, I've been to a couple of them in London, and I think I've made some positive input, but I haven't changed the world in this research project, if you see what I mean [laughs].

I: Yes.

R: I mean, the drivers for this research project have been the leads of the different sub-projects and, of course, the overall lead, Laura, for the project.

I: Yes. Okay. Would you like to… It's a very helpful start, because you gave me some hints about some kind of (unclear 0:05:19.0) and day-to-day sort of help you've given. Would you like to unpack that a bit more?

R: Okay. So some of the complicated things about the project in terms of… and in some sense would have led Laura… if I'd have been in Laura's position as overall PI on the project, then I think I would have been really quite frustrated about one or two things that were not under her control. First of all, the appointment of the data manager for the project was delayed a little bit, so we were a bit late getting started on that, and then, although this guy, the original guy we got was really quite good, Kerry, his name was, he then was worried about… ultimately, worried about his own employment prospects, so he left the project at one point. He was then brought back part-time for a while. We then couldn't get anyone really to replace him. And finally he's come back on the project again, towards the end of it. So there has been some discontinuity in that very important aspect of the project and, of course, that discontinuity is really quite important because there are people working in the Met Office and you build a relationship, he builds a relationship with them so they now know where he's coming from, then somebody else tries to come into that and the relationship is not there anymore. So I think that's been difficult.

But the second thing that's been tricky in the project was that an important contributor to the project was Public Health England. It just so happened that the project kicked off at about the same time as Public Health England was going through quite a big reorganisation [laughs] and that some of the individuals involved in the project were unable to really get on with the job because they didn't actually know where they were going to end up and that meant delays in them appointing the research fellows that were under their brief, if you like.

I: A-huh.

R: So the reorganisation of Public Health England was unfortunate, I think, yes?

I: Yes.

R: Of course, they are a major data provider, you see, and you get into that sort of situation where well we don't know whether we can supply the data because we don't know who our boss is at the moment [laughs]. That kind of thing. This is not their fault. I do stress this is not their fault, it's just reorganisation is sometimes a little tricky.

I: Yes. So the consequences of this restructuring was delaying appointing the research fellows?

R: Yes.

I: And also delays in making resources available and…

R: In getting permissions, basically, for data.

I: Right.

R: And then getting that data to Kerry so that he could do something with it, yes?

I: Yes.

R: So these things happen. But I think those are important aspects of the project. I think the Met Office was pretty stable as a partner, and the work being done in the Met Office was pretty okay, I think.

But I think the other thing that all of us underestimated was the confidential issues around obtaining data, how difficult some of those were going to be.

I: Right.

R: Well, in the end maybe not difficult but just lengthy. So the original kind of brief of the sort of databases that we are talking about had to be narrowed down, basically, as the project went on, and to some extent I'm not surprised by that. And if you've been talking to the SAIL people you'll know that confidentiality is a big thing for the SAIL people. I think we underestimated some of those issues.

The other dichotomy in the project - this is going on to a slightly different level, but it's again a learning process – originally MEDMI was going to provide a platform which users were going to be able to interrogate and a set of tools which they would be able to use in conjunction with the data.

I: Yes.

R: It turns out that there is, of course - and we realised this from the beginning - that there are another group of users which they basically just want integrated data, they want to use their own tools, because some of their tools are really quite complicated, yes, and they are used to using that. So you are trying to solve two problems at once, if you like. You are trying to provide a kind of usable website for somebody who doesn't want to use their own tools at the same time as provide access to integrated data for much more expert type people.

[Phone rings]

R: Can I just answer that, in case it's the Vice-Chancellor?

Sorry about that.

I: No problem.

R: So yes, you are trying to provide something which does two jobs at once. And I think as the project went on we began to realise that there are difficulties in trying to do that. On the one hand you want a flexible system which is able to say that I want you to give me this data integrated with that data so that I can download it and put it into my own software, you know, and do what I want to do with it in there and other people who are expecting there to be a tool there that just does what they want it to do.

I: That's the analysis.

R: Yes. And these tools that we are talking about, you can't make them complicated enough, if you see what I mean. If you are not careful, all you end up with is a load of visualisations of the data, so you can visualise various different things in the data and you might say well that may generate some hypotheses for the individual using the system, but then when they want to go on and actually follow up those hypotheses, they need to be able to model that data [laughs] and you can't provide a sophisticated enough interface for them to be able to model the data themselves on the website, and if you try to do that, to some extent you are just reproducing what sits around in… is already there in other pieces of software.

I'll give you a more specific example - and this may not be fair really on the project in some ways. I mean, there is a modelling language for statistical modelling, which is internationally accepted, which is called R. It's just called R. You won't find… people in biology, people in epidemiology, people in statistics, in engineering, lots of people use this language. It's in the public domain and it's very sophisticated. Now, you can't possibly reproduce that sophistication on a MEDMI website, it's just not there, right. So if you've got someone who is used to modelling things in R, then actually what they want is the data. They don't want a picture of it because they can draw their own pictures [laughs]. So I think as the project went on, this dichotomy of what you are actually trying to achieve is rather tricky, yes?

I: Right.

R: The other thing about the project, which again we knew from the outset and we knew that it was a learning process, is what level of aggregation you actually provide your data, because you are starting in some cases with individualised data. Now, that's really quite tricky because there are all kinds of confidentiality issues around that, yes? On the other hand, you've got Census data, which is at the level of the enumeration district and even some of those may be blocked out because of small numbers of people for various reasons in various categories. You know how the population Census surveys, they censor certain levels of data, don't they? So you've got your population data coming in like that, you've got some individualised data, you've got your Met Office data which is coming in usually on a grid because they are the results of modelling which is on a so many degrees by so many degrees grid and, of course, you can have that at all different kinds of resolutions in terms of how big is that grid or how small is it and then finally you've got some health data which is not available at individual level because of anonymity issues and so it's only available on the basis of some kind of district, yes, maybe a health district, maybe a postcode, right, and then other data like the campylobacter data which is essentially coming from laboratories, which have defined catchment areas, if you see what I mean, certain tests are going into certain laboratories. So putting all that together and trying to summarise the problems, you have data at lots of different scales of spatial aggregation and at the same time sometimes lots of different scales of time resolution as well – some of the data is daily, like the weather data could be daily, the health data is monthly, it could be annual in some cases, and then the Census data, of course, is… well it's annual, I suppose, by the time you put the predictions in there, but it's really ten-yearly [laughs]. And if you are going to try and provide a platform which kind of integrates all of this data, what is your baseline? What is your baseline? Because you can always aggregate from a baseline but you've got to do quite a lot of fiddling around to actually match up a postcode with a grid square or that kind of stuff. And it is not true, you see… it's the old story… God, I'm boring, aren't I? But it's the old story about the newsprint. Essentially, if you look at it at the right resolutions, there's a story there, but if you go in too close [laughs] it's just a series of dots [laughs] and if you go too far away you can't read it. That's the old story of spatial aggregation really. Some disease environmental things you will see happening at a certain level of resolution, but if you stand back too far you'll never see them and if you go in too close there's just too much variability to see them. So again, having some… well, the difficulty is there… It's a resolution problem - where at what level do you store all this stuff?

I: Yes.

R: When we started out, there was a bit of a discussion about well maybe we should decide upon a grid, a resolution on a grid basis and interpolate everything to that grid. So even though it wasn't coming on that grid, you would interpolate it all to that grid. Now, that would produce confidentiality problems, maybe, because even though you weren't being supplied with confidential data, you may be generating things that other people would think was confidential [laughs] because they…

I: Like?

R: Well, if you get numbers of AIDS cases… well, this wasn't part of the problem. But if you get numbers of AIDS cases in quite a large area and then you interpolate them down to postcodes, then it may be an interpolation, but other people may treat it as the truth [laughs] and then you may have a bit of a problem.

I: Yes.

R: So some of these things are really quite tricky, yes.

I: So is that why it's sort of… so you said initially we thought about it, but I think everything on top of a single –

R: Grid.

I: - grid.

R: Yes, decide on a resolution in other words and then try and put everything on that.

I: And then you abandoned that because of these risks?

R: I think we abandoned that because we just found it wasn't really feasible to do that, it produced other problems, like that confidentiality thing that I was talking about. So then ultimately we were using the data at whatever level it was coming in at and trying in the tools, in the visualisation tools to do some matching between grid squares and postcodes and whatever.

I: What does it mean "to do some matching"? You mean the visualisation tool?

R: Well, you know, if you've got a grid square you can find out which postcodes are in that grid square, for example, and then if you've got data for that postcode you can say something about that data for that postcode.

I: Okay. So basically you would sort of provide browser functionalities with these selections.

R: Yes, to try and do that behind the scenes, if you like.

I: So this would be recorded sort of in a default way of working of the browser application or it wouldn't be options to the users?

R: No, but I suppose what the advantage of doing it that way… it has all kinds of problems when it comes to how you match these things up, but the advantage of doing it that way is the data you've actually got in the system is the data that you received, yes? There isn't a set of gridded data sitting there that you've artificially produced. So when somebody asks for some data or something, you may have had a tool that did the matching, but when you give them the data you are giving them what you actually had, the original data that you are allowed to give them, yes?

I: Mmm-mm.

R: The other aspect of the project which, again, is an interesting one is… and I don't know much about this side of things, but I do know that it was discussed a bit, and that's the question of what levels of user you have to this kind of platform. You have people which can only get so far into things and other people who are kind of signed up as trustees, if you like, who can get a bit further in and then what level of expertise you are trying to cater for. I mean, Kerry at one point was trying to say, "Well, I can provide Pascal interfaces, programming interfaces to particular tools," but then you are expecting people to have expertise in using Pascal. Not Pascal. It's not Pascal, it's…

I: Python?

R: Python. That's right, yes. Now, for certain kinds of users, that may be perfectly okay, but for other people they will just get frustrated by this because they will say, 'Well, why do you want me to programme something I can programme in R in Python for God's sake? Just give me the data and I'll do it in R.'

I: **Yes.**

R: So quite a few problems to address there.

I: **Right. So also these were sort of figured out along the way.**

R: Yes, we've ended up somewhere. We've ended up with a demonstration website, which will be fine, I think, as far as the overall project is concerned, but all I'm saying is that it's…

I: **And then there's sort of the privileged access users.**

R: Yes. But then the other thing is the longer term. [redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

R: No, remember, this project started three, four years ago. I think if it started again now then you would have to have serious thoughts about what you are… I'm talking now about the basic computer technology; I'm not talking about the idea of the project. You would be more interested in Cloud-based kinds of activity, I think. But yes. Ever since I started working 100 years ago [laughs] on modelling what you might call spatial epidemiology, the linkage between different data sources has always been the biggest problem really, trying to get links between health data, which is inevitably collected in areas of some description or other, and climate data and environmental data, which is collected in completely different ways.

I: Yes. So you haven't seen new issues in that respect?

R: And it's always been an ad hoc kind of solution to the linkage problem and then you can do some modelling, you get some results. But that automatic linkage has never really been there.

I: Right, yes.

R: So I don't think anybody has got this problem cracked. If you are an organisation like NERC, you've got databases on sea levels, you've got databases on river levels, you've got databases on flooding, all this kind of stuff, and it's all under your control, right, and you stand a chance of being able to integrate, and that's what they are now doing. If you work in the Met Office, the same thing is true, but if you are working across different agencies then you've got problems [laughs].

I: Yes. So have you, in this respect, other issues related to this project of the linkage that are kind of new in respect of the issues, for example, data linkage that you said 'I've seen this throughout my career'?' So was there anything surprising or anything that you haven't seen happen before?

R: No. I mean, that may sound… yes, that may sound a little arrogant, but it's not the fault of this project. I mean, I think this project is extremely valuable because it highlights some of the difficulties. Even within the restricted domains that this project was trying to address, yes, what it does is to highlight some of those difficulties and how difficult it is to resolve them, and that in itself is a good thing, right?

I: Yes.

R: I mean, there are mechanisms, there are models… you could grid the whole of this country on a 1km grid, yes, and all of the public domain data that is available could somehow or other be interpolated down to that 1km grid, if different agencies decided that that's what they were going to do. Think of the Census data, for example. You look at the Census data for an area in this country like Wales, or parts of Wales - you might think this is a really stupid comment - and it tells you that there are a certain number of people in certain deprivation in this area, but it's just not true, the people live in the valleys, they don't live on the hills [laughs] and that resolution is not there in the Census data because you've got all these enumeration districts and some of them are huge because there's nobody lives there and some of them are very small. And that just completely ignores the geography. And it's the geography sometimes that's actually rather important. I mean, this is taking it to a different bit of work. But a lot of my work in the last five, six years has been on Vector-borne disease in South America, in particular dengue fever. There's very good data in South America on dengue at a municipality level. There are 5,000, nearly 6,000 municipalities over the whole of Brazil, right, but the mosquito behaviour, that's to do with drainage and it's to do with water and it's to do with river flow, right, as well as rainfall and temperature, and that drainage doesn't respect municipalities, you've just an arbitrary line on the map and said, 'Well, we'll count the dengue cases on this side of the line and we'll count them on this side of the line,' and the geography of the drainage doesn't correspond to that at all. Now, that's an extreme example maybe of this kind of problem, but you can see that even, you know, in a country like the UK, some of these issues are masked by the arbitrary aggregation of different sorts of data to different kinds of administrative units. So the water people have their way of doing thing and the health people have their way of doing things and the Census people have their way of doing things. There is no kind of unifying principle that allows this stuff to be brought together very easily.

I should apologise because I'm really saying the blindingly obvious now. I mean, we could have said that at the beginning of the project. In fact, we did say that in the beginning of the project. Has the project really enabled… well, the project has tested the water as to what extent you can work with existing data sources, yes -

I: Yes.

R: - but it's also highlighted the fact that you are still not really solving the problem because you produce something which is a lot of work and ultimately you can't maintain it, it doesn't have a life beyond where you are.

I: Could I ask you to also go back a little bit when you were saying you had input on data structures and modelling and opened these a little bit more on the historical side of things? It would help me also when I talk to other people to connect these experiences and things like that. Because this has been extremely useful anyway to sort of sort out the more overarching issues.

R: Well, if you take the campylobacter stuff, essentially what you are talking about there is individual cases of the disease, yes?

I: Yes.

R: So that's almost like point cases, and then you've got other things involved and it's just what sorts of statistical models are sensible to fit to those kinds of relationships? What sorts of probability distributions might be involved? Now, they happen to get a good… eventually on that project they happened to get a good postdoc who is statistically quite aware and my involvement in that was working with that guy, because he'd come from a certain background. I was working with him a little bit in order to say, "Well, why don't you do this?" and then he would say, "Well I've already done that," and we could get towards a sensible sort of modelling framework, and a paper has now gone off which sort of reflects that. I don't know whether that paper will be accepted or won't be accepted, that's the luck of the draw, but at least I think that the models that he's using are reasonably sensible and not…

I: Yes.

R: Yes. And that's an example of what we were talking about. I didn't do much on the heat waves because the guy who was running that project, he's pretty statistically aware anyway, has done this stuff in the past and so he was applying the similar sorts of models to the data he was getting in MEDMI and he was mostly using time series data and not space time data, so I didn't really need to get involved with that very much.

I: Right. And instead your input on data structures, was that involved in the interpolation grid?

R: Well, that was when we were having initial discussions about whether we could grid this stuff or whether that was the way to go, whether we should try and put everything on a grid so we had a single resolution for everything and then you've got flexibility to aggregate it to other things in other ways, yes, and discussing some of the difficulties around that. Also talking to Public Health England about whether it was better to have the certain kinds of data on a daily basis or whether it was better to have it on a monthly basis, those kinds of things.

I: Yes. Why couldn't you have both things, like an interpolated grid and the original data?

R: Well, because it's a hell of a lot of work to put it on a grid in the first place [laughs], so that would have slowed down other things, I think, ultimately.

I: Right. So in terms of timescale, where do you place these events, how long did it take for you to discuss the interpolation grid option, then to change, to move on?

R: Well, Brian Golding was always very useful in this project because he was clear from the outset that the most important thing to try and establish was the data structures and he wanted that to be properly documented and well understood. So really that's where he wanted the project to start, and there were some delays, as I said, in trying to get Kerry in post originally and then, of course, Kerry had his own ideas that he was then bringing along. But there was quite a bit of discussion about data structures in the first year of the project and I think it would be fair to say that really the pilot projects, yes, the pilot projects didn't really get off the ground until a year and a bit and the decisions have still really not been made about the final data structures and things like that when these people were getting on board. All projects are slow to get going [laughs], particularly if they are complex and they involve interagency institutional work. Yes, they are tricky, some of them.

I: **Right. So if there were MEDMI… two kind of projects, what would you do differently or sort of…?**

R: Well, now, you'll get very different views on this from different individuals involved in the project, but I would have thought that if there was going to be MEDMI 2 then you'd sit down and you'd say, 'Well, look, the problem with MEDMI 1 is that has illustrated this whole range of problems and now what we have to do is to put in place a project which is going to build a system which has long-term sustainability rather than is going to just illustrate the same things again, yes?' That means that we need to be really quite careful about the scope of different sorts of data that we are going to put into the system and we have to think about new technologies like the Cloud and how that works and we have to think about this business of resolution and how you match different kinds of data together, and we have to solve all those problems before we actually start trying to put all this stuff together, because you are going to have to make a bid to Public Health England for a certain set of data, you are going to have to negotiate confidentiality around that bit of data. Unless you are asking for the right thing in the first place, yes, you don't want to keep going back for different things. You've really got to think about sustainability, I think, maybe that's one way of putting it and the sorts of technology.

I: **Yes. This is the last question. Was there kind of also relationships within these kind of issues that made them, in a way, worse, like delayed access to data, did that sort of interfere?**

R: I think Laura… okay, I think relationships… I think Laura did a very good job of trying to manage those relationships. It's a very… she's very good at that, right, so she never lost her temper and she never allowed frustration to come into the game, she was always very supportive of everything that was going on, [redacted]

I: **Yes. But I also was thinking about, for example, if I can't access quickly the data from the health site, for example, I can't make good progress on the grid option or things like that. Was there kind of consequential chains or… you don't recall that?**

R: I don't recall the specifics about that, but remember that I've not been so much involved in the detail on the data side, what data is actually coming in at different times and what people are trying to do with it.

I: **Sure. Great. Amazing. Thanks very much, I think we've covered all the issues I wanted…**

R: Was that of any use?

I: Yes, absolutely. Very useful, yes.

R: Yes? Okay, good.

I: Absolutely. Thanks.

(End of recording)