NTDS_029

Key:

I: Interviewer
R: Respondent

I: Okay, so, I would like to start maybe from the basics of your involvement in MEDMI. How it all started, how you got involved, and what it meant, and your everyday work.

R: Okay.

I: Yeah. From start to now.

R: I can't remember when it was, but essentially, it was part of a broader project to grow collaboration between the Met Office and Exeter University, which really started when the Met Office moved down here. And there was an initial flurry of activity when we moved in 2003, 2004. Then it went a bit quiet, and I think it was when (ph. 0.01.00) *Julia Swinger* became our chief scientist. I think that there was a move to do something new. There was a general, opening meeting… in fact, a couple, I think, general open meetings, where senior managers in the Met Office and senior staff from the university – I'm not sure how they were chosen – participated, and out of that came the idea that one area we could focus on was health, I think. As broadly as that. And Julia asked me if I would take the lead from the Met Office point of view, and that naturally fitted in my portfolio. So at that time I was Deputy Director of Weather Science, so the senior research director in the Met Office, on – if you like – short-range forecasting, warning, that sort of thing. And my background was in weather impacts, so it tended to be focused on floods, but… health. I covered most things over the last twenty years of my career, so health wasn't outside… and I had had an existing interest in it. So essentially by appointment [laughter].

How did we get to the project? I'm not sure how we got to the project actually, to be honest. I think… no, I really can't remember. I suspect that the origin came from Laura, because she was appointed about that time, and she took the lead, and so it would have been natural for her. From the Met Office point of view, I think it was a natural… well, semi-natural follow-on from work we'd done in the past, which I'm sure Christoph told you about. So for a long period, we had had (unclear 0.03.48 – 0.03.49) one person in the Met Office who did research in collaboration with other places on epidemiology, with the aim of producing some sort of early warning service which would be self-funding. And we'd done a lot of work on COPD, which had been pretty successful, but we'd been having huge difficulties in getting anyone to pay for it, and so essentially the Met Office view of this… apart from… well, there were two Met Office views. One was: A success was if we built up a collaborative relationship with people in Exeter, and more broadly. And then the second one was: If it could come up with predictive representations of the link between weather and health that we could use to provide future services. And I always took that, I suppose, to imply… and to get them to the point where the… whatever level of trialling was needed in order to justify funding

had been done, in the research context, so that then we could go out and find a funder who would…

I: Yeah.

R: Because that had been the sticking point for the COPD – that we never really had the funding to do a full clinical trial. Arguable you don't need a full clinical trial if there are no side effects, but in practice you won't get funding unless you've done all that. So that, essentially, that's where that died.

I: Yeah. So you had these number of projects before… having talked about this with Christoph. And so, is there a legacy that you brought into the MEDMI? Is that relevant [overspeaking] in terms of number of…?

R: A little bit. But that had always been done in a conventional way, but when I came into it, my observation was… I'm particularly looking at the papers that come out of the US in this area. That most of them appeared to be done as PhD projects, and most researchers appeared to take about three quarters of a PhD project getting the data together, and then they would dash off some really rather simplistic analysis of the linkage of the causality or the statistical relationship. So for me, and I'm not sure whether this was… to what extent this was shared by the Met Office at large, but for me, the thing was: Could we produce something that meant that perhaps we took just the first couple of months to get the data together, and then the rest of it was doing some useful science? And if we did it, and if we did a pre-linkage of some of the big epidemiological data sets with the big environmental data sets that we had, then hopefully it would be almost a trivial exercise to bring together the particular data sets that were of interest for a particular project.

I: Right, yeah.

R: Somewhat naïve, perhaps, but there we go.

I: [Laughter]

R: [Laughter]

I: And so, then… so, in terms of the story, the timeline, of the project, how did you start on that? And then we'll go back to the learning, and what did you do…

R: The proposal itself, of course, is a big exercise, and took… I can't remember how long it took, but it was a lot of work, which Laura did the bulk of. But I was quite involved in that. The little diagram we had of how the data came together was, I think, broadly mine. And the acronym was mine, amazingly, because I don't like acronyms.

I: [Laughter]

R: But there we are. [Pauses] Yeah. [Pauses] And there must have been some coming together on the three demonstration projects, but I don't really remember how those… I remember how the third one came about, because that was Laura's particular pet thing, and we said, 'Right, that will be blue skies thing.' The other two, I think, broadly represented the two areas of

interest of the Met Office, so one more warnings-focused, so the natural one was to look at temperature, and / or air quality, because they often go together. And then there had been some existing work in that area. In fact, the COPD is effectively a temperature linkage, so that was an area that was basically familiar but which there were many things included that we didn't know. And then the second one was infectious diseases and particularly those carried by vectors, and in that one we were particularly looking… particularly, the Met Office is interested the climate change issues of those, so although this wasn't primarily a climate change project, it was seen that there was the potential there for bringing in data sets about the future, and so that might be predictive in a climate change sense, whereas the other one might be productive in an issuing warnings sense. And then the harmful algal blooms one was a nice one, in a way, for us, because it linked into the oceanography, which we do a fair bit of. And we'd been doing work on algal blooms, but not the health impacts. So that fitted quite well.

I: So you had… as a result of this work, you had the various relevant data sets.

R: Well, we thought we had, yes. So one of the big things in the proposal was to list the data sets we would use. We listed all the Met Office data sets, which was relatively straightforward. Listing the epidemiological data sets… basically, different researchers came up with the ones that they knew, and we made a list. And what wasn't at all evident was what the constraints would be on accessing them, and that's really been the story of the project. That we have only just got access to our first health data, at this point in the project. And most of what we hope to get, we still haven't. And the third leg, which was the socioeconomic data, which is often what produces the confounders for the relationship… we thought that was going to be simple, but somehow the people who tried to do it, the collaborators, managed to make it so difficult that we still haven't got access to socioeconomic… straight census data. And I think the problem has been that they've always tried to get too much. The same has been true of the epidemiological data. If we'd been happy to have coarse-resolution data, then I think we would have had it ages ago. The researchers wanted to… were used to getting data that was very fine, and having to sign up with their lives, sort of, to the fact that no-one else would look at it. And as a result, we've got stuck between this… okay, one person could do that, but the whole team can't. But we really don't want to back off and just have the coarse-resolution data, because there was the… the epidemiological researchers say, 'Well, that isn't good enough.' If you do it like that, you'll have to do it again in the future.' And maybe that was the line we should have taken. 'Yes, we will accept that we better do it again.' [redacted]

I: So, at what point in either project did you start hitting into this wall of accessing data sets? But had you done… at that point…

R: It was probably after about a year, because at the beginning the epidemiologists said, 'Well, this will be fine. We just need to find what we want, and put in the proposals.' And it took about a year before they had done that. Partly because there was a delay in appointing somebody to do the work up at the London School, I think. So I think there was a six month delay, so nothing got started up that end for six months, and then also there was a delay getting the hardware in Exeter University. That must have been over a year. Well over a year.

I: Okay. Hardware… the servers…

R: That's right.

I: … that were [overspeaking] the data sets, right? Are we talking about servers…?

R: A server to host the data. Yeah. That's right. And the intention was always that it would be a stand-alone… sorry, not stand-alone. It would be a dedicated server, partly so that we had control of the access, which was clearly going to be necessary for the more sensitive data.

I: Right, yeah.

R: And to that point, in the while… in the first year, what had you guys here at the Met done, and how were you getting on in your [overspeaking]

I: The two things that were happening in that period, I suppose, were that we were identifying and extracting some of the data ready to go in the Met Office, but also the demonstration projects were making progress using… not using the MEDMI system, but using, if you like, a copy of what it might look like at some point. And so data were extracted, as a sort of one-off rather than a generic solution, and that was good for pushing ahead with getting papers, getting presentations and so on, but it was bad from the point of view of some of the work has now got completely tied up in the those mini data sets, and can't actually be translated into the generic system. So, the other thing that was happening at that stage was the initial work on the website, and on the – excuse me – on the analysis engine that would sit on top of the data, and that was… the consultant was doing that, and he pushed ahead quite fast. Using a storyboarding approach which he modelled on the Demonstration Project 1 analysis, which was being with one of these. So they extracted a small amount of data, and in fact they reproduced a result that had already been published, but it gave a model for developing the website and the online access and so on, and menuing system, which is very good, but very very limited, because it can't now access the main database.

I: Okay. Yeah.

R: So that… if you like, that took us down a bit of a side turning, which didn't really lead anywhere, except insofar as it gives a good feeling to people who don't know about it. How that research is done, and how it might work.

I: Right, okay, yeah. So, then, I understand that basically… so this research work that was done on the side, basically, was in order to get ahead and do something. And in what ways… you were saying that now it can't be translated on the generic data set, and the web application can't really translate it. So what's been happening there?

R: Well, essentially, what happened was that the consultant produced a mini-database with just those data in. Meanwhile, Christoph was creating the generic database with the environmental data in – because we only had the environmental data – using a different approach, because the consultant was using technology that Christoph isn't aware of, and the consultant didn't have enough effort… in fact, he left, in the end. Because he went off to another job.

Had the consultant stayed, he could probably have translated either what Christoph was doing into his database, or translated his database into what Christoph was doing. It didn't really matter. But because he went, there wasn't the expertise available to do that. The expertise was with Christoph… sorry, the effort was with Christoph, and so effectively we had to switch to using Christoph's approach, which is a perfectly good approach, but it didn't then link to that initial work that had been done.

I: That had been done elsewhere.

R: That's right. So we have the demonstration, and what we're doing at the moment, actually, is very finally to create that link, but in a way it's too late to do anything with it now. Hopefully there will be a follow on, which will be to make use of this capability, but so far as the project itself is concerned, it's too late to be used.

I: Right, yes, yes. And so this kind of work started in the first year…

R: Yeah.

I: … and so, is the relationship between… this kind of alignment between Christoph's and the consultant's work, and also the issues of the constraints of accessing data sets.

R: Well, there is and there isn't. So, one of the things that pushed the separate development of the Demonstration 1 work was that… I think it was Shakoor Hajat at the London School. Might be wrong. He already had some environmental data, but in particular he had an extract of mortality data, which he already had access to, and which they could do work on. The expectation was that that would become redundant within a few months, because we'd have the big mortality data set, and we'd have the big environment data set. Well, we got the big environment data set, but we never got the big mortality data. In fact, I think we're still waiting for it. So, what happened was that that piece of work extended, or it was attempted to extend it, without getting the bigger mortality data. The data that Shakoor already had couldn't go on MEDMI because it had been provided to Shakoor on the basis that he was the only person who would look at it. So these constraints on anonymity and so on, and confidentiality, meant that we couldn't use his data set, and eventually we will get the data for MEDMI, but it's taking that huge time. It's taken the full three years. Which, of course, is… in a way it's not surprising, because that's exactly why I wanted to do MEDMI. Because I knew that it took this three quarters of a PhD's time to get hold of the data, and the aim was to bypass that. And yes, we may yet do, but it will be too late for this particular project. Whether or not it then provides a basis for future work depends partly on whether there is any follow-on, and partly on what the limitations on use are when we eventually get the data. At the moment it looks as though that data will only be usable by the named individuals within MEDMI, in which we may not have gained very much [laughter].

I: Yes. Because the purpose was to make this infrastructure that could be accessed easily.

R: Yes. Not openly, but at least on the same time that it would take for somebody to get funding for a epidemiology project using MEDMI, so… and

it's not clear whether the constraints that are being placed on us now will allow that, or want to allow that. It's…

I: It's still something that you are trying to figure out.

R: That's right. Yeah.

I: I understand. And is this something related to what you've mentioned in the beginning about ways of doing things and approach, or is it different?

R: No, not really. No, the interesting thing about the meteorologist's approach to big data… and I don't know what the definition of "big data" is, but arguably we've been working on big data since the nineteen-fifties, but…

I: Yes. But that's also why we don't use the term "big data". We talk about data-intensive sciences to stress the continuities.

R: Essentially, what the meteorologist does is to say that the dimension of the big data, the first thing we do is to reduce the dimension, because we know something about how the atmosphere works, and in particular it scales above ten to twenty kilometres, the atmosphere is very, very strongly constrained by the rotation of the Earth, and that's understood in a physical, mathematical sense, and so effectively that reduces the dimensionality of the data to very small numbers, so it isn't really big at all, once you've done that. And so that's always our approach. So, you look at the physical causation, and you say, 'What are the laws here that enable us to make sense of this without having to crunch through the data and look for it?' The idea of using statistics to identify the rules, if you like, wouldn't cross a meteorologist's mind. They would always start with the physics. They might then use the statistics to demonstrate that the physics is actually what controls it.

I: Right. Yes.

R: But it would always be via the physics, and of course that's very, very different from… I think, although the epidemiologists will look for causation – at least, they say they will – most of the ones I've come across would believe the statistics over and above the causation [laughter]. They wouldn't say, 'Well, if I can't find causation, I'm not going to pay any attention to the statistics.' Whereas meteorologists, I think, broadly would. Partly because we've so much experience of the statistics being wrong.

I: [Laughter]

R: For years and years people tried to demonstrate relationships between the solar cycle and the weather, and there are well-documented cases where people would find and publish results over two solar cycles – twenty-five years, it worked beautifully. And in the next solar cycle, it would reverse sign, and then it would just go… chaos. So, you… an epidemiologist typically working on much shorter data sets than that, so our experience tells us that if you can't see why the statistics are doing what it is, then probably it's just chance.

I: [Laughter]

R: And then when the statistical people come up and say, 'Well, this is significant at one percent,' we say, 'Well, you may think it's significant at one percent, but it can still be wrong.' [laughter]

I: And now that these conversations come across in MEDMI, what occasions…

R: Well, yeah… it probably didn't. It was just underlying the discussions. So, part of the problem is that we never really got to any new hypotheses. The nearest we got was some exploratory work that a guy at… I think at PHE, actually, is doing on infectious disease. And he's basically screening all the pathogen data for seasonality. So he is actually taking a causal starting point. He's saying, 'Which are the pathogens that appear to have a relationship with the meteorology, because they appear to have an annual cycle that relates to winter and summer?' Well, it's… there's a little bit in there. But there's still no… if you look at even the ones that have the biggest annual cycle, by and large there's no understand of what the process is that causes the change in the temperature, to lead to [overspeaking]

I: It might still not be causal.

R: Yeah. It might not be. That's right. And that's quite frustrating, for a meteorologist. But I think we… the people in the group, by and large, I think, there is an understanding of each other, so the epidemiologists would say, 'Yes, it's certainly our aim to understand the cause, but we recognise that's very difficult, so we'll work on the statistics in the interim.' And from the meteorologists' side, a recognition that many of these things do respond to temperature, and we know that life in general responds very strongly to temperature, to the annual cycle of temperature, and that therefore that's a plausibly group of mechanisms, even if you can't focus in on a specific one. So it hasn't caused any particular problems, but then nor have we overtly attempted to develop big data approaches to this. It's tended to be traditional approaches to epidemiology, and perhaps a little flavour of traditional approaches to meteorology coming in together. And yes, they are very different, but we each come from our own…

I: And in what way are these reflected in the work that you were doing [overspeaking]

R: Very little [overspeaking]

I: So there was… decisions you took about functionality, to put into the work. An application, or into the data structures.

R: The data structures, I think, the challenges have been about efficiency. Not really about how you use the data, although clearly there's a link there. But the problems have been in terms of extracting time series. So most of the analysis has been time series analysis, to date. And extracting time series from the very big environmental data sets has been very time-consuming on the server. And so there's been work to rearrange the data so that it's more efficient, but that's essentially a technical, computational driver, not really a… and on the other side, the actually analysis techniques that have gone into the website are standard epidemiology. We looked to the epidemiologists to tell us what should be in there, and we're happy to take their word for it. And in a

sense, that then drives what the nature of the extraction for the environmental data needs to be.

I: Okay, okay.

R: So it's really a more… if you like, any sense of conflict between the approaches is more at the strategic than the working level, I think.

I: In what way did the epidemiologists requirements drive what you need to extract from the…?

R: Yeah. Well, at the end of the day, they need… well. If it's like the Demonstration Project 1, then you have a series of, I suppose, death rates, or something. Mortality rates, for a location, and you want a series of weather data, or multiple weather datas for the same location. That location, in general, doesn't exist in the weather data, so you need to interpolate to it from a number of surrounding locations that do exist, and maybe which they're slightly different distributions at different times, as stations open and close and so on. And so what you're after then is a long time series for each of those nearby stations, which you can then interpolate, whereas we tend… the basic data tend to be stored as all the data for one particular time, followed by all the data for another particular time, and it's very inefficient, then, to get out. So you've got something like, say, three or four hundred stations, for every hour, one after the other. And what you're really after is a sequence of perhaps two or three variables for maybe half a dozen stations, but for a lengthy period. And it's…

I: I'm aware there's a question there, that is almost maybe a bit naïve, or dumb, but why couldn't it be the other way round? Is that a problem of…

R: I can be the other way round, and I think…

I: … a problem of resolution, or… yeah.

R: Yeah. One of… I suspect there's also an issue that we're used to working on the simple computer, where it doesn't really matter, because everything's fast anyway. And when you translate onto a server, which is only moderately powered, then the… and maybe which doesn't… has a different architecture. Then simply copying the data from one to the other isn't the best way of doing it. You should reformat the data. But the drive… so that's what Christoph's doing now. But of course, the drive on him was to get the data there, initially. And then to get more data there, because the epidemiologists have never been able to get the data at all, so they're really keen on there being something there in any format, and it's only when they then come to use it and find it takes a week to download the data that they say, 'Ah, well, can you do something about this?' So it's not their number one priority, but once they've got their number one priority, then it is tremendously important.

I: Okay, so it wasn't necessary for the state… it wasn't imposed by the state of the data in the epidemiologists' data sets to reduce the meteorologists' data down to…

R: No, not really, I think.

I: Could do that more dynamically.

R: One of the interesting things is that the original data for the Demonstration Project 1 was only, I think, ten, fifteen years of data for one location. Very small data set. And I think no-one quite realised what the implications would be of trying to scale that up to, let's say, sixty or seventy years at five hundred locations.

I: Okay.

R: Although even that isn't really big data. It's… and perhaps, you know, well… some of the epidemiologists say, 'Well, we want to try out every variable that you've got.' That's perhaps twenty or thirty variables in the third dimension, so it quickly becomes quite a big set of data, and then when you've got to interpolate along the way, or do other sorts of processing along the way, then it just becomes a big software job, whereas the starting point was really quite trivial, in terms of data processing. And indeed, in the starting point, all the focus was on this statistical software, the Poisson… fitting the Poisson distribution and so on.

I: I was a few days ago talking to Trevor Bailey.

R: Oh yeah, yeah.

I: Who was also describing this kind of… also learning about first the approach of trying to interpolate everything in the universals… with my words. With universal interpolation, except… and then he said, trying to, as I understand, keep these things more separate and linked dynamically on…

R: Ah. Ah, yeah…

I: No?

R: Well, that's an interesting one. I'm still not sure.

I: No, I'm not saying he was suggesting that. I might be recalling wrongly. But he was just describing how the project overall has moved…

R: Has shifted.

I: … has shifted from an "interpolate everything" approach, to a "not interpolate everything" approach. And he was also recounting that you used to say, at the beginning of the project, that it was most fundamental to get the data structures right at the start, and I think that he meant to say that you were right, but then that there's been some learning in that respect. That there's been some wasted resource, or time. So, will you want to…

R: I'm still not sure whether we've got this right about the linking. The original idea was that we would take the environmental data, and we would pre-compute values on every postcode, or every administrative district. Or on a grid. And we have shifted away from that, towards doing it on the fly. Part of the problem, I think, is that if we did it for every postcode, then the database of the linked… not linked, because it's… of the remapped data, let's call it,

would be very, very big. There are an awful lot of postcodes. Even if we do it on a five-kilometre grid. That's slightly easier, because you can short circuit the access by using the orthogonality of the grid. But it's still a very big data set, and I think part of the reason we've moved away from that is the recognition that simply downloading, say, a daily series over fifty years of those complete data sets in order to get out the individual elements that you want could actually be more expensive in terms of computer time than doing the interpolation on the fly. And this is the problem with computational science, and we're… just to take a completely different area, we're currently rewriting from scratch our weather forecast model because the new generation of massively parallel supercomputers doesn't work efficiently on the old methods that we used, which were physically and mathematically based, and we actually have to re-couch the equations in forms that are suitable for a massively parallel system. So, it's a ten-year job. Huge. And completely driven by the fact that that's the way supercomputer technology is going. So this is something we're familiar with [laughter].

I: So, you mean, this is all so very experimental in the sense that you don't know what the gains are going to be yet, or…?

R: Well, what do know is what the losses if we stick with our current data structures will be. We know that, basically, beyond a certain number of processes, however many processes we add, effectively, take just as long. So we know we've got to do it. What we don't know is how efficient it will be when we've done it. So that idea that the computational constraints actually drive the way you do the analysis is something that we're familiar with. In this case, the constraints are coming through to us via the epidemiologists, so we don't have a particular… it's not particularly our field, to understand why we've got to do what we've got to do. And of course, to the extent that the epidemiologists don't understand it either. We do have a bit of a problem there. We're not in the fortunate position that we are with the weather forecasting model, that we have a whole team of expert computer systems engineers who can tell us what to do. With MEDMI, we have… well, really it's down to Christoph. He is the only technical person on the project, now that Kerry's left. And even Kerry was technical in the sense of the web interface, not the data structures.

I: Yeah. But it's been working awesomely. The structures are…

R: It's been working on the database, but I don't think his expertise are in these things about efficiency with regard to which way you store the data. I don't think that's his field, but I could be wrong [laughter].

I: No, I remember he was telling me about some work he had done to… yeah. To change the way the data are accessed from a time point of view.

R: Oh, okay. Yeah.

I: Yeah, yeah. So, if you look back and you compare… so you tried to make… or take away points and compare it with what expectations you started in, and how you've changed your expectations across the project. And also, if you have changed the way you would work in this collaboration.

R: Yeah. I think if we had had somebody like Harriet at the beginning, to… so, if you like, and administrator who was a leader of the project, rather than an administrator who was a supporter of the project, then I think we would have made a lot more progress. I think that role would then have taken the lead on getting the data sets. We might still have come up against the same issues, but some of the delays at least wouldn't have happened, and I think we would have got to know what the problems were much more quickly. So I think… one learning, then, is how important that side of things is. The person who writes off to the owners of the data and requests access, and when nothing happens for a couple of weeks follows it up. They don't notice six months later, 'Oh, we never got a response from…', which isn't a scientist's job. The scientists are very bad at that sort of thing. We were absolutely right about the technical requirement for the computer for the database and for the server. At the time I thought we should focus on the database, not on the interface, and I still think that, actually. I think…

I: Who is "we" in the "we were absolutely right in the requirement"?

R: We…

I: The MEDMI…

R: "We" was the MEDMI PIs and co-Is. When we were putting together the project, there was one post that we said we must fund, which was the technical post. And it was described… I can't remember how it was described now, but I think the description focussed on the database, but in the end we actually focussed on the web interface, and that's very nice, but I think in terms of the project aims, we would have done better to have focused on the database. That's a personal view. I doubt everybody would agree with me.

I: This makes me think. Other… am I missing out issues relating to the databases that we haven't touched upon? I'm thinking about interpolation on the fly, which… and the access.

R: I don't know. It might be my lack of understanding of how such a server would work. Clearly there is an issue about… it's not just about storing the data. It's about how you access it. And… one thing that… it clearly takes a lot more effort than I had appreciated, is… interfacing a web interface to a database. I have experience of doing it. Of doing it with a graphical front end, not with doing it with a web interface. And I naively assumed that a web interface was a modern package that made it easy to interface to the database. In reality, the web interface makes it much, much more difficult to do. So, back in the nineties… eighties? When I was doing it, it was actually much easier to do this job. What you got was cruder, of course, but the constraints of a web interface, that they place on you, appear to consume vast amounts of effort. I don't know what it's for, but it clearly is there.

I: Right. You mean compared to in the nineties or eighties when you were doing… when you were building applications on the web interface, of the time?

R: No, because the web didn't exist.

I: No, you mean… okay. You mean a graphical user interface.

R: So, you have a graphical user interface, using basic graphical languages, like GKS, at the time. I don't know if that means anything to you [laughter]. The graphical kernel system, which was what I was doing. So you… I mean, you could do menuing, you could do graphical pictures, you could do analysis. You could do all the things. What you didn't have was the ability… well, you didn't have remote access, I guess. And you didn't have the standardisation of how windows work. You could do windows, but not in the sort of standard way they're done now. And I'm still not sure why it is so much more difficult, but it clearly is. And I just didn't appreciate it would be. I misjudged what the effort required would be.

I: I see. But these also are not really related to the data structures. It's more related to the way web applications interface.

R: Well, yeah. And yet, the story we've been told again and again is that translating the web interface that was produced to work with the database structure that Christoph's developed, versus the one that was put together on the very simple data by Kerry, is a huge amount of work. And they're both… they've both got fairly high-level access routines, so to my simple view, it's just a case of slotting in one access routine instead of another one. That's something I would reckon to do in a couple of weeks' work, when I was programming. But clearly again, it's much more complicated than that. And I still don't understand why, but it's… so the database does come into it, and indeed that's sort of the crunch of where we've got to. We have the browser, and we have the database, and they don't talk to each other.

I: Hmm. How are we doing? Okay. So, concluding questions, because I guess I need to close. What would you like to do if you were to do MEDMI again, of if you were to do MEDMI 2? Also, considering, if you had no constraints on funding, what would you do differently instead?

R: [Pauses] I think one of the things you've already got – that I would have somebody whose job it was to do the administrative stuff.

I: Yeah.

R: I think the pilot project idea was a great one. It was something that, as far as I was concerned, came a bit out of the blue. I was surprised that we had so much interest in it. I think it's a great way to expose something of this sort to hypothesis generation and some ideas. Yeah, I would do more of that.

I: More pilot projects?

R: Yeah. But I think, in the core of it, probably extend the number of collaborators. At the moment we haven't really got a university. We've got the London School, of course. Well, we've got Exeter, of course. But Exeter's involvement has been a bit constrained, I think, because of course Laura's leading it, but with the exception of the harmful algal blooms, she's not actually doing any of the research [redacted] so there's no mainstream epidemiological research being done. And I think it would be nice to see more of that.

I: Yeah. In what way would this help? And in what way, also, would having more pilot projects help? Is it more to just make use of the data, or is it more, also, for the very maintenance, and…

R: No, I think maintenance would probably be more work. I don't know. I would like to see… at the moment, we're still… a lot of what we're doing has still not really broken out from what's already been done, and it would be nice to see more exploratory work, I think. Because that's really MEDMI's strength – or it should be. If we actually had the database and the web interface working, that should give you the possibility to try a wide range of associations between health and environment. In quite a simple way, at a superficial level, but it should also enable you to do deeper research down to a certain point. Possibly you'd have to finish it off elsewhere, but by that time you would have sifted out… so the exploratory search needs to be very broad, because you're going to focus in on… some things aren't going to show any association. Some things are going to be messy and complicated and too difficult. But hopefully, at the end of the day, there will be some things that are worth pursuing, and where actually you can see that there's a potential public health benefit of using it in some way, and I think at the moment, with perhaps the exception of climate change policy, these are diseases that we ought to be worrying about, and that we don't at the moment. With that exception, I don't think we've really got to that stage yet. It's good research, but I don't think it's useful yet. And I'm always looking for something that will be useful.

I: Right. I was thinking also, wondering also, if the pilot projects have helped inform your requirements, or in trying to make developments on the technical side, more informed, at least.

R: I think there's been a little bit. I haven't actually read all the reports yet. I'm sure Laura has. And Christoph is more plugged in than I was, because he was actually creating the data sets that people worked on, because they didn't work on MEDMI itself. But in many cases, most cases, he was providing extracted data, so through all that, he is building up a broader knowledge of what the community needs in order to work effectively in this area. And yeah, I think that's very valuable.

I: Just one quick question of clarification: When you say you would have wanted some person like Harriet to start from the start, is it like more a question of getting administrative work, or leadership as well?

R: I think it's the sort of leadership that… so, she will keep focused on problems that we haven't yet solved. She doesn't do what she was asked to do and then forget it. If that didn't really solve the problem that we were trying to solve, then she will come back to it, and make sure that when we have our teleconferences that we come back to it. And I think that's very, very valuable. I suppose the other thing to comment is that management by teleconference isn't ideal [laughter].

I: Yeah. I've heard.

R: We basically would get together once a year. Twice a year? Probably twice a year, and it's not really enough. But of course, with the distances involved, it's difficult to do otherwise.

I: Yeah, yeah.

R: It could be worse. I've got another project where I've got a team of six co-Is… six? Five. Well, another co-PI, and then five co-Is. And the co-PI lives in New

Zealand, and the other co-Is, two of them are in… three of them are in North America, mostly on the west side, one of them is in Australia, and one of them's in Germany. [Clicks tongue] So, when we have teleconferences… well, I'm lucky, because it's the middle of the day for me.

I: Yeah.

R: But we've got people five in the morning, and people eleven at night. And we can't afford to even have one meeting a year [laughter]. So it's fair to say… compared with that, MEDMI is a breeze.

I: Yeah, yeah. It is kind of something that works…in contrast, there are conflicted… but also, the wish to have more collaborators in that sense. There's a balance to be found there, I guess. Right…

(End of recording)