**NTDS 044**

Key:

**I: Interviewer**
R: Respondent

**I: You want maybe first to introduce yourself and the kind of work you do and stuff so that we can, you know. You can then explain your involvement in MEDMI.**

R: Okay. So my name is Neil Kaye and I'm a Visualisation and GIS Specialist. And in the last few years I've been developing web tools to help communicate climate science, and my involvement in the MEDMI project was to develop a web tool to integrate climate data and health data. The health data that we had is… well, it was actually originally daily data which I've converted to weekly data and it's number of cases across approximately 200 odd laboratories over the UK. So essentially the data comes from a specific person and then… well not data, like if you go into a doctor and you get a sample taken, then it goes to a laboratory and gets tested and they tell you if you've got the flu or whatever you've got. So there's various different things. Some of them are much more prevalent or common than others. So, for example, campylobacter, since 1989 there's well over a million cases, and things like salmonella, for example, is half a million of those. And you can select different diseases, and then what it does, it goes and shows you for each week from 25 years from 1989 to about, I think it's 2012, where the diseases occur. So the map you see here saying… if you hover over it it says 16 cases compared to an average of 6. So red means there's more cases and blue means there's fewer than you might expect. So you can kind of go into the map, and then as you scroll round here they change, you do it slowly with the scroll, then you can see that. And then these are… at the moment, it's minimum, maximum temperature and precipitation. The first one on the left shows that 1989 to 2010 average for the week and then that's saying for April 8th 1991 to April 14th 1991, that's the temperature and that's showing that it's warmer than the normal for the week, because that's why it's that red colour. And it's the same with the maximum temperature. So if you went, for example, to December 2010, which was exceptionally cold. I don't know if you were here at that point but…

**I: Yes, I remember.**

R: You may remember that. I think that's the coldest month we've had for the last year. Certainly… where are we going? Sorry. Where are we? What's it doing? Oh sorry, I need to go…

**I: December '11, right?**

R: Yes. December 2010 is when it was exceptionally…

**I: Yes. Everything frozen.**

R: There's one week, yes, that was… I think over Christmas was exceptionally cold, so you kind of see that from here, and you can also see… what you are seeing here is the relationship between the number of salmonella cases and

the temperature. So there's generally a relationship that as it gets warmer, as you'd expect, there's more salmonella. Because of a couple of tools you can increase that relationship, so there's an R squared value of about 0.3, 0.29 or whatever for mean temperature. If you change it from… You'll see that the overall trend in salmonella is it's decreasing over time but you can change it to proportion of the year and that immediately… so that means that that week will say 2% of the total or 5% of the total for that year. This means that the bumps are all…

**I: So this is compared with the same proportion, the same time of the year?**

R: Yes. So it means for 2010, instead of it showing you the absolute numbers, it just shows you that week there were, say, 83 cases and that was the equivalent of, say, 3% of the year's total, whereas obviously earlier on there's a bigger pic… but it shows that the relationship increases. And you can also do things like average the climate data. So if you average it to six weeks, then the relationship is even stronger, so it's suggesting that there's some kind of lag going on. So the idea of the tool was that you could maybe explore some of the relationships with… yes, with different diseases. Some of them are far less prevalent which is why this is only going up about 20 in a week and then you can… yes, I don't know, you can select them. And some of them obviously have a stronger relationship than others, and that one seems to have a slightly negative relationship temperature.

**I: Can I ask you again, the proportion of the year selector, does it compare then to the same week in the other years, is that what you were saying?**

R: So what it does, it's not the same week in different years, it's just taking…

**I: So unchecked?**

R: Unchecked, it's just the raw information. So it means that's the number of cases. So something weird is going on. I don't quite understand why there seem to be about 50 cases and then suddenly it seems to ramp up to well over a 1,000. That could be a reporting thing. It seems incredibly unlikely that it would go like that, because obviously that looks like… until that point, it's a bit kind of random and then there looks to be some kind of… it looks like there's a bit of a cycle and that, a vaguely seasonal cycle which would probably tie up with the relationship to… a slightly negative relationship between temperature and figure or whatever. Well that's influenza actually. You probably expect more in the winter.

**I: Yes. But then if you check the proportion of the year?**

R: So that means that… well, if I do one that's a bit more obvious, like campylobacter, the reason… So if you… Oh maybe not. There was one that was… it was crypto… was it (unclear 0:08:00.7)? Let's have a look. So, for example, this one, this starts off low and then it seems to come to a peak and then go down. So what it would mean, if you just left it like that, the relationship… no matter what time of year, there's going to be more in this period because it's generally higher, so the idea is that if you click the proportion of the year it gets rid of the fact that there's a higher relationship just because that particular year is higher, although…

**I: Okay.**

R: Does that make sense?

**I: So basically it harmonises within here sort of data succession of years.**

R: Yes. It takes all the years, like 2006, 7, 8, 9 and then it just says that for that week that was 3% of all the cases for that year. So that's what that does, and then the map shows you where around the country the cases were, because there's some weeks where – I'm trying to think – where there's quite a focus in a particular area. Where's that one? I don't know. You can kind of… they are kind of well… but then there was one. I'm trying to remember it, but there was one where it was all like in one location and that became obvious because when you sort of went through it it was a real peak, a massive circle showing that that was that particular place. And clearly for this rotavirus there's a strong negative relationship as well for winter. It doesn't make much difference when you… That one seems to be quite static where it almost disappears over the summer, autumn and then comes back again. Yes. It's obviously over the winter. Anyway. So that's pretty much -

**I: That's really beautiful.**

R: - what I was involved in doing.

**I: So is this part of the publicly accessible browser?**

R: I think the idea is it will be. At the moment, it's on this tools data mashup thing. Well it's actually the data mashup thing. I'm not sure if it's… see, this is the MEDMI website, projects or something. It will probably be there somewhere, view the tools, but I don't know that it's linked yet because it's… It might be. I think it's probably this one here. But it just seems to go round in circles. Yes, it doesn't…

**I: Yes, it's not that one.**

R: But I think it just needs to be added in.

**I: Yes, there you go.**

R: Yes.

**I: So how long did it take you to develop this?**

R: It probably took a few months, I would say. It's kind of based on work that I'd done before for other projects but it's probably a few months work to do it. And I looked into other stuff that I'd been looking into but hadn't quite worked out how to do was normalising the data, because at the moment it's like absolute values but it would be useful to know say per 100,000 people. But the problem was that the laboratories massively overlapped and they would change through time the location of the laboratories, but if you went to somewhere in London, for example, it would be… You'd have like the laboratory and then you could see where the people who…

**I: (Unclear 0:12:34.3).**

R: Then it was like that and then you'd have one over here and then it would be like that and then there would be one over here, and like that, and then there would be one here and it would go like that, and one here. And it got to the state where… I calculated the number of overlapping laboratories and it got up to nine. I wasn't quite sure how to assign like a population to the laboratory because it would be that population there, but then there's also this population there that's overlapping and that one there and that one there, so you'd be kind of like…

**I: But is that about the risk of counting the same people more than once?**

R: Yes, exactly. So in the end I just stuck with the absolute value. But it would be more useful obviously because you normally normalise it per 100,000 or something, but that wasn't… because…

**I: But why does the absolute value not work, is that because you could have the reports in different labs that belong to the same person?**

R: No, you wouldn't.

**I: What is the shortcoming?**

R: I don't know that it's necessarily a shortcoming, I just think that, in general, you know when you report morbidity and mortality it's like deaths from cancer per 100,000 people or something like that rather than an absolute number, because I think the issue that if you look at… so just as a demonstration. Oops, not that one, this one. Ooh, why is it doing that? Why does it do that? It didn't do that before. Some bug probably. That's a bit weird. Ah, there. What I was going to say was that if you… let's do one with a few more. Let's go… Where is a lot more stuff? So, for example, these circles, you don't know whether they are bigger because there's three times as many people living there or if it's actually because there's… so if there is, say, 100,000 people in that region and 300,000 in that region, you'd expect there to be three times as many cases anyway, so the size of the circles is kind of… it should really be proportional to…

**I: Yes. Because otherwise London always looks thicker.**

R: Yes, exactly. Although, in some respects, because of the overlapping nature of it, there's so many in London, it might be different.

**I: So every lab has got a circle here?**

R: Yes, every lab that reported anything in that particular week has got a circle.

**I: Okay, yes.**

R: So you can see in this situation that there's quite a lot of overlap going on. I mean, it gives you a good idea of the spatial distribution of where the cases are, I suppose.

**I: Yes.**

R: But there's always things you could improve. So there's obviously that one in London, but then maybe that's because there's a lot of people living there, so. I don't know.

**I: Yes, absolutely. So you were saying that it built on the work that you've done before for other projects. So what are the tools you are using and how much you are able to reuse from project to project? (Unclear 0:16:14.5).**

R: That was something I… I can quickly show you that. The idea of that was slightly… I mean, it might actually be of interest to you because some of it is about perception. So one of the things that's interested me is how people see different colours. So, for example, if you had this red, yellow, blue scheme, the idea I had here was that you could see what somebody who was colour blind would visualise it. So, for example, that might be how somebody who is colour blind would see that particular scale. So red, yellow, blue is actually quite a good one for somebody that's colour blind. But if you, say, changed it to red, yellow, green, that is not so great because this is… to give an idea of what that would look like to somebody who is colour blind then it's clearly the reds all look the same as the green, so it's a horrendous thing to be doing, but people might not be aware of that. I mean, obviously, because it's actual temperature, it's obvious that that's cold and that's hot, but you can do stuff. I changed it so you had a few example maps. So, for example, radar shows basically low to high or whatever.

**I: So are these all potential… So that predefined…**

R: Predefined. They are examples of colour schemes that are used. So, for example, one of the common ones is this colour scheme here, which is this rainbow colour scheme which is being shown to not be the best one to be used.

**I: Right.**

R: But also the idea is you can create your own colour schemes, so maybe you want to create a yellow, green, blue colour scheme and that's what it might look like for a temperature anomaly. And I tried to make it so it could automatically colour it up. That one doesn't work brilliantly. Yes, so you could… I mean, you could do anything you wanted. Essentially, you could create some horrible ones but… Yes, and then the idea is you can save the pallet that you've created and export it so that you could use it in your software. So somebody asked me about a graph, what colours they could use to create a graph. That's what… If there were eight different lines, that's what they would look like if you used that colour scheme or you could use a different one, and that's what it would look like.

**I: Super. That's very interesting. I read something about web design for colour blinds and stuff.**

R: Yes, it's quite interesting. People can't… I mean, I think it's about 10% of the kind of white male population are colour blind of a form of red/green colour blindness and it varies. There's much fewer women because it's attached to a particular chromosome. Anyway, so that was just an example. I mean, I've done stuff that's also… for a Brazilian consortium as well looking at different ways of visualising climate data.

**I: Post Brazil.**

R: Yes. So it's just learning at the moment. Hopefully it'll… there's a few things it has to load up. So you can look at climate data over different time frames and click on regions and then graphs come up and hover over the graph and it tells you information about that.

**I: Right. So it looks, however, also quite different from the other ones. So in what ways are you building on a previous…?**

R: Well it's kind of… It takes some of the things, but this might add extra… I mean, quite a lot of it is new, it's just the general underlying structure might be quite… some of it's similar. But I'm always trying to think of different ways of visualising the data and communicating the data.

**I: Yes. What did you do to know sort of what kind of parameters and visualisations do they need?**

R: Well, we've had various meetings where they kind of… But, I mean, I think originally… I don't know if that's… is that it? I'll just check something. That's something I did. I could just quickly show you… it's just a matter of finding it. Okay. Gordon Nichols. No. I'm going mad. Let's have a look. So I did something, I think, and then Gordon… Maybe it's not that one. Sorry. Yes, back in maybe last August or whatever I was… Oh no. This is probably it. Sorry, I'm just trying to find the… No. I'm just looking for something.

**I: Yes.**

R: [Clears throat] Yes. I mean, where it came from, Gordon had sent through a PowerPoint presentation of the ideas that he had. So he sent through this maybe a few months ago or whenever last summer and the idea was he wanted to show maps and have a big map and have graphs showing various things and a slider bar. So that kind of vaguely got translated into… It was kind of that original requirements and then he kind of made some comments on the tool in December and then I made some changes based on those comments.

**I: Yes. And how was the conversation about… sort of what kind of changes? Was it just updating each other?**

R: Yes, kind of. I mean, we didn't have too many conversations, he was relatively happy with what I'd done and he just wanted a few changes. Yes. I mean, he just provided a rough idea of what he would quite like and then I tried to duplicate it as well as was feasible.

**I: Yes. It would be really nice if it would be possible to have both slides. Maybe I could ask Gordon.**

R: Yes, I did this for Laura.

**I: For the presentation that they are doing in China?**

R: I think so. Is that, yes, Anthony?

**I:** **Yes, Anthony.**

R: Yes, yes. So I just gave him a couple of slides just to show examples of how you could change the…

**I:** **Okay. So I should have that. Actually, I haven't opened that presentation.**

R: Okay. Well that's just two slides. All it is is that just to show that it's a nice… it basically shows the tools, you just click between them.

**I:** **It is a good annotation for me to…**

R: Yes, so that's what that is.

**I:** **It would be nice to see also the requirement.**

R: Yes, I could probably… I could just forward that to you probably.

**I:** **That would be amazing.**

R: Is it this…? Hang on a minute, I'll just. Dashboard for Met data. I'll tell you what… What's your email?

**I:** **n….**

R: Ah you're there already. Okay.

**I:** **Thank you. Amazing.**

R: Is that spelt correctly? It just doesn't like your name.

**I:** **Yes, it's a common spelling. So it's okay. (Unclear 0:26:52.8) spelling.**

R: Okay.

**I:** **So it's pretty quick. Is that all the data all precomputed?**

R: Yes. So it's essentially quick because… because originally it was going to be coming from a database that… Have you spoken to Christophe Sarran?

**I:** **Yes.**

R: Right, it was originally going to be coming from that database but it kind of became apparent that to get the query out to do this would probably take minutes and obviously for a web page that's not acceptable because people expect things quite instantaneously. So the only thing that takes a little bit of time is, if you see here, this is taking maybe two or three seconds, and what that is loading is a… I can show you the data, if that's of any interest to you, or I think I can. Let's see if it's… or the idea of how it's… I'll move it up. Actually, it might be… it might have… Ooh, let's try again. Yes. I thought I could show you. Well essentially what I… I can't seem to find it. Yes, so this one is the biggest file that it has to load, it's campylobacter and it's almost 3 megabytes so that's why it takes a few… maybe five seconds, and what that data is… So in this case there is 200,000 rows but it's been done so the ref ID

(unclear 0:29:08.2) is basically the ID of the laboratory and there's another file that's got the xy coordinate of that laboratory and then this date is just the date at that laboratory, the 28th December '88 and that's for that week, it's saying there's one case of campylobacter. So as you go through it's got the lab ID and then the 17th June '96 there were six cases at that laboratory, 17 or whatever. And these files are what is used to create the points. And then I think there's also… separately from that, there's a weekly disease data which is a much smaller dataset and that just shows you for that week the total number of cases.

**I: So that's the number of the week.**

R: Summarised across the laboratories.

**I: Okay. Summarised across all the laboratories. So it's an absolute number of cases of campylobacter.**

R: I'm not convinced actually that was the latest design; it might be slightly different from that.

**I: And will that information go here?**

R: That information is what you see on this here.

**I: Yes.**

R: And also this graph here.

**I: Yes, it's got all the weeks.**

R: Whereas that bigger file, it's also split up by laboratory, so that's why it's probably a couple of hundred times bigger because it's on a laboratory basis, whereas this is just information for each week, of which there's maybe a few hundred.

**I: Yes. So for each week here and here and here. Yes, of course.**

R: Yes. And it tells you there the number of cases. So what it does is, as you select this, it loads both of those…

**I: Select with the scroll the weeks, then it changes the years.**

R: Yes. I mean, initially, you select this. So you'd expect this to be quite quick because it would be a smaller file. If you look at it, there's only 3,000 cases so that would just load quicker, whereas, as I said, this one to load will be the campylobacter one because that's got much more information. Anyway, to do this using Christophe's thing would probably have taken five, ten minutes. It might have even been… because basically to get it out the database was…

**I: Why?**

R: I'm not sure. It just seemed to be a bit slow. I'm not sure if it was… I don't know if it was the database design or if it was just the way that… because this is just looking at a text file on…

**I:** **This year's (ph: 0:32:00.4) *fee*.**

R: Yes, just opening this year's *fee*, whereas the other way you'd have to go to the database and then do the database query.

**I:** **Would have picked things from one file [overspeaking]**

R: I think so. For this it wouldn't have worked because it would just take too long.

**I:** **Yes. So this data has been prepared by you?**

R: Yes, I prepared it in that format. I kind of tried to work out the way that it would be the smallest and most efficient to…

**I:** **Yes. From basically tapping into Christophe's data, you prepared this.**

R: Well actually no, I think I got this originally from… the original source was a big… this one gigabyte CSV file. So what this CSV file, it's got daily data for all the different variables but then…

**I:** **It's got all the labs, all the measurements, all the [overspeaking] for everything. Okay.**

R: Oh actually if I… you could probably look at it in… I'll just see if that… I'll quickly show you on here. I probably haven't got too much more time because I need to have lunch before I go to another thing.

**I:** **Right. Sorry. Sorry to keep you.**

R: I'll just quickly show you this. Oh, I don't need a bracket. I forget what I'm doing. Actually, if I make this wider. So that's the original data. So it's got the xy location and then there's all sorts of information, there's sex of the patient, the date and stuff. So each of these is a single case, it's a single person, but we only have the laboratory ID. For confidentiality we don't have the person's… So this has basically been changed to something that's on a disease basis and for the big file, this one here, it's kind of been… I've tempted it to shrink it round as much as possible. I did have it even smaller at one point where I just had the weeks as numbers and related to this, but the speed of linking it to something else was -

**I:** **Increasing?**

R: - yes it was increasing because I had to do some links in the java script which took longer than just loading initial files, so in the end I kept it like this. So the job was to take this and summarise it by week and then just get rid of…

**I:** **Everything else.**

R: Anything that was taking up unnecessary space in order for it to be as quick as possible to show. Because once it's loaded the file it's all in memory, so it's the user's machine that's basically doing all the work because it's sitting on the… yes, the work is being done by the machine; it's not being done by any server. That's how it works.

**I:** **Yes. Pretty nice. That was (unclear 0:36:02.3).**

R: Sorry? Yes, that was. Yes, I need to get. Yes.

**I: Okay. Thanks a lot for…**

(End of recording)