NTDS_026_001

Key:

I: Interviewer
R: Respondent

I: What I would like to talk then is generally your experience in MEDMI. I understand that you've covered different kinds of roles at the same time and you've developed part of the infrastructure, the web processing application and web analysis application, or you contributed to that project. Then also more in the data use and consumption side, researching the UV project with Marjorie and Nick.

R: Yes.

I: So we would like to talk about both of them. So I would like to start with asking you first if you can quickly introduce what role you cover and what's your research interests and perspectives beyond MEDMI in a more general way. Then if you can start by talking about how you've been involved to MEDMI, how were you brought in the project and what kinds of perspectives and interests.

R: Okay, that's fine. So my role at the Met Office is really to support the health programme at the Met Office with research. I've taken more and more of a coordinating role within the science at the Met Office to bring about this research. One of the things I've discovered over the years is that there are a lot of detail requests for environmental details that come from the health sector. These tend to be quite specific, and typically that requires quite a lot of effort to both extract and process the environmental data into a suitable parameter that can be used for these different health projects. So one of the key reasons for getting involved in MEDMI is really to facilitate that extraction and processing to make it a lot easier to provide the correct data to health researchers. So that was the key benefit from the Met Office perspective. Initially my involvement in MEDMI was to work on two demonstration projects.

One concerning heat, the thermal environment, and air quality and its impact on health, specifically mortality. The second demonstration project was looking at sensitivity of infectious disease with different environmental parameters. That was looking at infectious disease records. It became apparent after probably about a year that we hadn't the resource to deal with the vast amounts of environmental data and process them. So I then started working much more on the data end of things, structuring a database of environmental data, and now infectious disease records as well, and creating the computational tools to process this data as efficiently as possible, and creating the basic data tools to facilitate extraction processing, which is still the key objective as far as the Met Office is concerned and as far as the wider MEDMI project is concerned. So we are dealing with about 9.5 billion data values, so the task isn't easy.

I: So these data values in terms of fields in the tables?

R: In terms of actual values, yes. So they're individual cells, individual environmental observations.

I: So you had not enough resource in what terms to deal with these requirements?

R: I think after a year we realised that the specific skills that were required to deliver on this database management side we simply just couldn't recruit. We went on two recruitment rounds and we didn't find anybody suitable, unfortunately.

I: That seems to be often the case with these data projects. There's so much demand, isn't there?

R: Yes.

I: Then the work that you were trying to do, which I understand then basically you weren't able to hire, so you had to be more involved yourself, is that what you were trying to say?

R: Yes. I wanted to make sure that by the end of the project we had the basic tools that gave us access to the data in a useful way. Until then we just had 9.5 billion data in files and nobody was accessing them because nobody knew how to.

I: Could you open this up more and tell me more about these tools, this work you said on data structure inside and then preparing these tools for data processing?

R: Yes. So there are several aspects to it. One aspect is how the data is stored to facilitate speed of access. Initially we decided not to concern ourselves with that. So we just stored the data in their original files. We then discovered that it didn't help with the speed of processing, so we decided to change our approach and start structuring the data into different directories depending on date of observation. That, in itself, would allow the processor to easily find data for each date. So it kind of goes through the dates in chronological order, and I put the processed data as it was processing it. So that's one big advantage. The second aspect that is worth mentioning is regardless of how the data is stored. One of the objectives was to link data with each other, so linking the environmental data with infectious disease records, for instance. There the challenge again is to do that efficiently.

So initially what I had to do because I was quite new to the computational tools, or computational language, was to find ways of extracting and processing data quickly from the data end without... What I hadn't taken into account was making the code efficient at the point of linkage. So at the same time of changing the approach in terms of data storage, we changed the approach in terms of coding for processing where my current work is to start from the linkage aim, determine what data is required and how it needs to be processed before doing any of the processing. So the code knows what it needs before it does any work. Then it extracts the data in an order that avoids repeating any data extraction and repeating any processing. So a key advantage is it does tasks only once instead of doing it many times. So that's something I'm still working on. Restructuring 9.5 billion data points is currently taking the computer about a week.

It started taking months and I had to rethink how it was going to do it, so managed to reduce that to about a week, which is reasonable, of computer time. So now that I've sorted that aspect of things out, I'm working on changing the approach on the extraction processing and linkage, so starting from linkage to data processing extraction before even doing the extraction processing and linkage. Initially it was building tools for extractions and building tools for processing on top of that, and building tools for linkage on top of that, and the linkage did not take into account... or the linkage was not efficient, so one project was to link 1.5 million infectious disease records, and that was taking about 20 days, which is not at all sustainable. So we changed that to trying to find ways that would be a lot quicker.

I: So I'm struggling at the moment to figure the way the infrastructure works. So are we talking about requests for data that come from an API or the web?

R: Yes. So at present the database system is totally separate from the browser systems. The browser systems are still closed to MEDMI researchers only, so there are no external users, or guest users. By speeding up the database processing tasks, the current aim... initially we wanted to link the browsers to the database directly, but even if the data extraction and processing takes 30 seconds or a minute, that's still too long for user time to use the browsers. So what we've settled on in terms of objective by the end of the project is to create a browser that allows a user to select data and then that will launch a process that will extract and process the data and save it to a file, and that file can be either emailed to the user or picked up by the user through an FTP system.

I: So one can send the request through the browser, but then can go away, do other things and then when it's ready?

R: That's right. So the processing, I suppose, would happen offline, or perhaps it will have a status bar or something on the browser saying how far it's got to in terms of processing it. That's details to be determined.

I: This is different from what I saw. I navigated and found the demo of the browser as it is now it's actually different from what it will become because it feels more... I understood it was more like a demo of real-time analytics, even with a small test case maybe.

R: There are three different browsers. All three rely on pre-processed data. So it's getting the data to a point that satisfies the requirements of the browser. That means it limits what the browser is able to do. We currently have, in terms of just weather and climate data, we have 400-500 different parameters, and obviously all of these are not listed in the browser. So the new browser hopefully will list as many of these as possible rather than currently I think we've got about three environmental parameters in the browsers. Three or four. So it's a bit limited.

I: Then instead the application runs offline. That is not related to the web, the proper linkage that the researchers are using?

R: Once they've extracted the data, yes. I think that's the current idea. I think it should be able to request linked data in the same way. So that's what I'm working towards, is a generic piece of code where you can request a list of parameters, that are linked with each other, including a key parameter that holds all the parameters within it too. So if you want infectious disease records for campylobacter, male patients only, in the region of England, then you should be able to specify... the dataset would be the infectious disease dataset. You would be able to filter for campylobacter and filter for 'male only', give a range of latitudes and longitudes to select the laboratories that are within that range, and then you might want to have maximum temperature, daily maximum temperature, daily minimum temperature, and daily cumulative rainfall, with a one-day lag, and these will be estimates for the previous day for the location of the laboratory.

So the infectious disease record determines the locations, because, of course, the recordings of maximum, minimum temperature and rainfall will be at those linked sites. That will be elsewhere in the laboratory. So you need to do some special processing to determine the estimate at the laboratory of these environmental parameters.

I: So that's the core of the linkage?

R: That's correct, yes. So if you specify these things, each time you have a record that has the same laboratory and the same date, then the new approach will do the processing only once for all of these records with the same laboratory and date, while the current linkage does it over and over again. It goes and extracts maximum temperature for that date, processes it spatially and links it, and then when it's got the same thing again it will do the same thing again, which makes processing very slow. So the new approach is to avoid these repetitions. So the objective in terms of selecting from the browser and saving to file and emailing that file, then potentially the linked data can go into a file and be sent to the user as well. It then becomes a question of how you design the browser to allow users to select linked data or not.

I: You mean the totality of the data that the MEDMI sort of... are you talking about to design the web application as such that the user can browse through the entire scope of the data?

R: That's right. It will be limited by the design of the browser, how much the user can access the data. On the other hand, if the user can use the code itself, then the access to the data is limitless.

I: I was thinking I have seen that there are several datasets from the environmental side and then several datasets from the health side. So obviously there is some pre-difficult work to do to link them as to find out how to link them because they are based on different topological orders, right?

R: Yes.

I: So then have you had to work on every single combination to establish the linkage of one health dataset with another health dataset and another health environmental dataset, and how to establish the ways... say locations that are recorded in different ways are overlapped?

R: The original approach established immediately that we needed to stick to one special coordinate system. So that was latitude and longitude, which is a most general one, which doesn't help creating nice grids because the shape of the latitude, longitude grade on the UK is a bit trapezoidal. So a lot of data comes in national grid references and so these need to be translated into latitudes and longitudes, which I haven't done yet. That's one of the tasks I've postponed until I've finished working on the tools. The other thing we decided quite early on is because of the programming language used, which is Python, I decided, it's a small decision, that the time coordinate would be either for the time at which the measurement was taken, or for the beginning of the time at which the measurement was taken. So a lot of measurements are taken over a period of time. So, for example, maximum temperature will be maximum temperature over a day.

So, regardless of how it's recorded in the actual data coming in, it will be defined as the start of the time period. So there are a number of conventions that had to be set-up at the start. What the new approach is different to the old approach is that in the old approach we wanted to make the tools intuitive for users with a number of defaults. If a user wanted maximum temperature, it could get the best value of maximum temperature, meteorologically-speaking, which is a nine o'clock to nine o'clock value. That has specific processing methods that are purely meteorologically ones. So that approach meant that we had to go through each parameter in turn to determine what was the best way of processing it. So, if somebody wanted rainfall, then the database had both hourly info and daily info, but it did not have weekly info. The question was, 'If we want weekly info, it might be best to use the daily info to calculate the weekly info.' Then the daily info might be from nine o'clock to nine o'clock.

So if you wanted a midnight to midnight value you'd have to go back to hourly info to calculate it. So we tried simplifying things, which meant a lot of coding to say, 'Well, actually, this particular parameters has these characteristics and will need to be processed using these specific tools before it's released to the user.' So that would be a default setting, unless the user wanted something different, and so they could go in and change the settings, change the defaults.

I: So it stays configurable.

R: That unfortunately meant that we'd potentially have had to go through each of the 400-500 parameters that are in the environmental datasets and determine what are the sensible defaults. We found first of all users were not going into the code to use the code, simply because they are not used to that, I think, in the health sector. In particular, coding is not a huge skill. We also found that how the data was being processed, these defaults were not transparent enough. So users were still not really understanding what was happening to the data before it was being released to them. So the new approach will get rid of all that and we would simply say, 'All of these data are available. These tools are available to process the data. You need to say exactly what you want.'

I: Does it mean you got rid of all these conventions and just gave the users all the green light you've got?

R: Yes. So the new system will allow users access to all of the data. Not from the browser but from the coding, to allow users all of the data, and allow users to use any of the processing tools on any of the data. So the current tools are things like arithmetic mean, minimum and maximum. It will do arithmetic mean but temporal as well as spatial. So there are two tools, essentially. So you could have an arithmetic mean over a week. You can get an arithmetic mean over a range of latitudes and longitudes. For wind one of the things related to the pollen dispersion work was we needed a wind estimate. So wind is a vector. So there are various ways of getting an arithmetic mean of wind.

So you can get a complex mean, which is a mean of the vectors, as opposed to a mean of the magnitude. So you can just take wind speed and just get an arithmetic mean of that, or you can take wind vector and get an arithmetic mean of the wind vector. They end up being two different bands.

I: Is that because you need to calculate the directions and make them clash?

R: Yes. If you have two vectors of the same magnitude in opposite directions then the mean will be zero. While obviously if you just take a mean of the magnitude it will just be the magnitude. So depending on what the users want and hopefully that will maybe help the users think, 'Actually, for my application, what is it that I really want? I didn't realise there were two ways of working out the arithmetic mean of wind. Do I really want the mean of the vectors, or do I want the mean of the magnitudes?' For wind it's quite interesting because wind, if it's an atmospheric dispersion question, if you want wind combined with pollen, then you want the mean of the vectors because you want to know where the pollen is going. If you want wind as an exposure value for somebody then the person is exposed to the mean of the magnitudes. If it's windy in every direction, as far as the individual is concerned their exposure is not going to reduce to zero.

So while for pollen, the pollen grain will be moved this way when the wind is in this direction, and it will come back if the wind comes back. So that comes as if it's a wind of zero. So it really means that the user really needs to think through, 'Actually what is it I want?'

I: So basically now then the new system would get all the data at the granularity we've got and then here are the various ways you could think about operating it? So that's the new system?

R: Yes.

I: So you provide a set of methods. You suggest a set of methods to think about these variables?

R: Yes. So there will be a list of tools, and at the command line level it's, 'I want this dataset processed with this spatial tool, this temporal tool and maybe this spatial tool again,' or something.

I: So in terms of this, what the system is about, is it a library of commands, of scripts? What is the application for the offline for the researchers?

R: At present it's a Python module which is essentially an object-oriented programming language. So it's a list of objects that make use of a number of functions or methods in Python language. So the number of objects currently, in the old module there are three objects. An extract object, a dataset object and a record object. The record object was used to link the records to this environmental data, and that's going to be taken out because it's just not efficient, and the dataset object should do the job of linkage really. So the dataset objects use user-defined data for user-defined locations and times. So it does the spatial processing and temporal processing from the extracts. The extract object is simply to extract the data from the database.

The new system has an import object which is an object that allows the structuring of the database, which is more for the database management side of things. It will have an extract object, I think. It will have a dataset object tool, to allow the linkage as well as the processing.

I: Why the record object was not efficient?

R: The record object was inefficient because it was doing the same thing over and over again for the same locations and dates. Really actually the dataset object should... if you're really wanting to go generic with the tools, the dataset object should allow the linkage.

I: Staying with this more technical side of things, a few minutes ago you were saying that because of Python the choice of the programming language, that was related to this number of conventions that you had to prepare and stuff. So I wanted to understand better in what ways Python allowed or also required you to make decisions.

R: So Python is I understand the underlying code for GIS systems. It's also now the scientific software programming language that is used at the Met Office, since probably about two years ago. So that's why we chose Python. Like any programming language there are a number of conventions. The choice is whether to create our own conventions and thereby requiring to code these own conventions within the software, or just use the conventions within the programming language as they are. So it's simpler just to use what's there. So, for example, potentially a date is from midnight to midnight, but in Python if you take the end point of that period, it automatically returns the next day as a date.

I: Is that 23:59?

R: No. It would be from zero to zero. So it's simpler just to take the first midnight, that's the start of the period, because that in Python will automatically return the date for that whole period. So that's the reason for the conventions.

I: So basically because you were working with Python and then in some way you were trying to find solutions that would exploit and accommodate in the ways Python uses conventions.

R: Simply, for example, all of our environmental data, the time field is very often different from parameter to parameter. Sometimes it's the start of the period, sometimes it's the middle of the period, sometimes it's the end of the period. So to remove any confusion we had to decide on a single definition of what the time field should be, and we just went with that definition that already exists within Python.

I: In terms of data discovery resources, I understand there could be a man file. How would that know how to browse through the datasets and the resources, find how (unclear 0.37.18) and time stamps are.

R: There is a spreadsheet that lists all of the datasets, which I often circulate round. I can send that to you if you want.

I: Yes, but how can you discover what's within a dataset, for example?

R: Using the extract tools and that kind of thing.

I: So the extract tools guide you?

R: Well, no.

I: Do you prepare that documentation? How would the end user of the system learn, 'I can...?'

R: So the metadata spreadsheet will give you basic information, such as the temporal frequency, the spatial distribution and the coverage, the number of values within the dataset, the start date and end date, and which file it's saved as and that kind of thing. What it will not give you is specific detail about whether there's any missing data. So it will give you start and end date but if there's a month missing it won't actually tell you that kind of thing. It's just too much data to start listing all of the times when there is no data. In some ways the missing data is not data that's missing, and that's something I've had trouble with some researchers in the past, because it's just data not recorded. It's not that it's been recorded and it's disappeared somehow.

I: Right, because they think that it's disappeared?

R: Yes. So the definition of whether data is missing or not is... if it was never recorded in the first place it's not really missing because it never was there in the first place.

I: Is it difficult to provide evidence that there is traceability of the process?

R: There is traceability, yes. So you could look back and say, 'Well, actually, this piece of equipment was out being repaired for three months, which is why there's no recording.'

I: I think they need the room.

R: Ah yes. Okay. We might need to go...

NTDS_026_002

I: So one question that I wanted to ask you was normally in the other systems that I'm seeing, like for example, SAIL, a lot of very sophisticated solutions need to be found for protecting the security of the datasets confidentiality and all that. So that's also something that I'm going to ask about MEDMI, what are the requirements and what are

the solutions relating to information security. So how demanding was that on this kind of project?

R: I haven't really been involved in how that aspect of things works. What I'm aware of is that on the one hand the data currently is for user access only for a set of users requiring a login. One dataset in particular is sensitive and was restricted to me, one of the infectious disease datasets. It's on the server but it's read-only to me. So nobody else can actually look at it. Somebody did try looking at it once, which was somebody who owned the dataset in the first place, from Public Health England, I don't know whether they were expecting to access it or not, but because it was limited to me only, and that was their own requirements, they couldn't actually read it. So it demonstrated that that actually worked, restricting the dataset technically was actually...

I: That's the research work for your research project?

R: It wasn't for my research project but it was my task within the research project to do the linkage against that dataset, and that dataset was particularly sensitive so we limited it to me so that I could do the linkage.

I: So the rest of the team members can't access that dataset directly?

R: No, that's right.

I: Only through you.

R: That's right, and then I sent the linked data back to Public Health England, who are doing the analysis.

I: How would the rest of the members access the data? Would that be anonymised?

R: At this stage there isn't a requirement for doing so, for that particular dataset. So it's just there. If somebody wants some work done with it, then they would have to ask me to do it.

I: I guess it's computer variables, or defined variables.

R: That's right, yes.

I: That's what they need in this project and that's what they access.

R: Yes. It still needs to go through a human intervention in the sense that they would have to ask me to do the work. So there is nothing automated. So currently there is no system where a user can ask to derive data from a restricted dataset that they have no access to.

I: Only for those that they have access to?

R: That's right. So if they want something from a dataset they do not have access to then they would have to ask a user that does have access to do the work.

I: The manual work?

R: Yes.

I: Okay, so basically the system automatically sorts out... I guess these tools that you were talking about for data analysis can be applied to the datasets, and then user access privileges also come in so that they can point the tools only to specific datasets I have access to?

R: Yes. Any user can use any of the tools. It's just if they attempt to get user tools on restricted datasets the tool would not be able to read the data simply because the system would recognise that it's not the correct user. It's not so much the tool but the system itself.

I: So it's between the tool and the data, there's a filter.

R: Yes, and that's set within the computer system rather than...

I: I'll talk to Cherry a bit more in a few days' time. So he's the person I should open this up more?

R: Potentially, yes. The way Linux systems work is simply I just change the permissions to restrict it to me, and that's how the dataset is restricted.

I: So the administration of the datasets is distributed to various people that are in charge of the specific dataset, like you were in charge of that dataset?

R: Currently I'm the only one managing the datasets, all of them.

I: So you are the administrator.

R: By default, yes. Not that I ever wanted to be.

I: So you are the person that does this. So if there's a new pilot project you will be the person who set-up this person accesses this?

R: I don't give access to the system. That's done by the university. What I do is make the datasets available on the system.

I: To specific users?

R: No. I wouldn't know how to do that. I would just make it available to all users. If there's a specific requirement for restriction then I would have to liaise with the university first to work out how they're going to restrict things.

I: So there's basically a binary set-up. So the dataset is either available to all people, or restricted?

R: Yes. So every user can restrict files to themselves and that's essentially what I've done with this specific dataset that was restricted to myself anyway.

I: Then also the datasets that are available to me participating on the (ph 0.07.34) *POLEM* project would be the same ones available to another person participating?

R: Yes.

I: So the datasets are shared within the MEDMI research community or groups?

R: Yes, that's right.

I: The last question is just looking in retrospect. I learnt from Laura there's been lots of learning during the project. So what would you like to change in the future? What improvements, and what would you have done differently? You can talk at the general level, not only about yourself.

R: The one thing I feel everyone has learnt on the MEDMI group is that processing environmental data for linkage is a lot more complex than people envisaged. Certainly from my point of view I learnt it's a lot more computationally challenging when dealing with large datasets. So there is both the scientific methods, in terms of how do we process data to make it relevant to specific research projects, but it is also if we want to process all of this data then if it's not efficient on the computer then it's going to take weeks before we get anything out of it. I think a lot of the MEDMI researchers are starting to realise that it's not as simple. That's encouraging because before MEDMI we were getting a lot of requests from the external requestors' point of view looked simple but from a meteorology expert, the question was, 'What exactly do they really want and how is it relevant? How do you make sure that what you provide is relevant to what they're doing?'

So that I think is what we've learnt. In the future I think while perhaps we might not have a solution to an interactive browser database system, I think we'll have a good set of tools that can be used to facilitate providing good data to health stakeholders, and data that is relevant to the actual applications that the health stakeholders might be looking at. So I think that's a big benefit for the future.

I: Do you think that's more a priority to work in respect to the web application? Or is it also a problem maybe related to complexity of the web application?

R: I think users of the web application would expect data to be written very quickly. Until that is resolved I'm not sure that it's a really... the risk is providing a web application that is slow but faster than anything that exists. Still, from the user point of view, they give up before they get the data returned. The problem is if 95% of users give up when using a web application then it's not sustainable. I think it's a kind of human interaction with web applications issue. Personally I know what a computer needs to do to process the data, so if I was using a web application I would expect it to take minutes, at least, and so I would wait, but I'm too involved, so I know what it's doing.

I: I find this very interesting and insightful. It's a human and computer interaction problem. It is related to expectations and perception of time. The desktop applications often you see the application is reacting so you perceive that it is fast, whereas the web maybe it's already started crunching but it's just less interactive in terms of operating system integration. It feels slow.

R: Yes. So one of the browsers in particular, which is a web browser, even though it's using processed data, so it's as fast as it can be, there are defaults in the web browser system, I think it's Firefox. It needs to use Firefox rather than Internet Explorer because it doesn't work in Internet Explorer. Even in Firefox, there must be a setting in Firefox itself that if something is running too long it will pop-up saying, 'Something is taking too long. Do you want to cancel?' So even within the software there is an expectation that the application should be returning something a lot quicker than it currently is, and that's without doing any significant amounts of processing.

I: That's interesting. It's like browsers have not been designed and they've been configured for other kinds of applications.

R: Yes.

I: They're not ready for it. Great. Thanks a lot.

(End of recording)