Plymouth 12/06/2016

I: Interviewer (Gregor Halfmann)

R: Respondent (David Johns)

R: … the level of noise with certain people in the team, a number of them are very, very good at species identification, that's what we brought them in for and that's the training they have had. So they can identify the species and they are very interested in that side of it, but they don't have virtually any knowledge about the ecology or what they data goes into, pretty much no concept. And then you have got other guys who are more into the research side of it, they would think “actually, I have identified the species and it is an unusual species and therefore I am going to look into it and try to find out why it occurred at this point”. I have this constant battle with them all the time that they will say “oh, I have found such and such species” and I'll say “did you not think to look that up, because they have never ever been found there before?” And I am like “Well, I know it's such species, but just take that next step.”

I: So you urge people to ask further questions …

R: Yeah, yeah. We have a team meeting every month and we have an interesting plankton section. I have been struggling with people to … They will put down something that they have seen and I will say “okay, first of all, why is that interesting? What is unusual?” But actually, it is not unusual because we find them there all the time. They are just uncommon. Or they will find something really unusual, that we really wouldn't expect to see somewhere. But they wouldn't have taken that next step to actually go and talk to the researchers saying “look, this is worth delving into because we have never found them there before.” So yeah, I want them to do that, I want them to add something to it, because that kind of builds stories up for papers.

I: Yeah, that's interesting. So, I kind of have to check which questions I still have. First, I always thought that for the QA procedures there is some kind of software running quick checks. Is that actually happening?

R: It used to, but we got rid of it. So this is where everybody has entered their data. These are all the species going down here. And these are the actual blocks, the samples they have looked at or that route. And these are the analyst IDs, these relate to the people who have looked at them. It's a bad example, because that's me and I haven't done mine. I'm the only one who hasn't done theirs. So this is how it is now but it used to be all done on paper and then it was entered into a separate programme called Check-[unclear]. And that would look for those anomalies. So here it would say that's a high count because it got two lower counts on the side of it. Or let's see if there is another example … maybe not. But it only worked on anomalous counts. We did think at one stage of putting in an area code for every taxa, so we could say “we find it in the North Sea” or “We find it in such and such” and then it would run this automatically and say “actually, that species X has never been found there before, flag that up.” But now it just goes on expert judgement, really. People just look at it.

I: So why did you stop using that?

R: Because in that system, there was potential of error. We would write the results in our book, then we would transfer them to a paper sheet. Then somebody would enter that into this other programme and then it would go from there to another programme. There are just lots of steps. So we said “let's completely re-do it.” So now people write it in the book and then two of them will sit together, enter the data, and that's it. It's done then. It doesn't need to be looked at again by those people.

I: And since when have you been doing that this way?

R: Quite a while now, ten years at least, I should think … So that automated checking has gone.

I: If we already talk about the QA, I'll just ask more questions I had on that. Last time we talked you mentioned that you were working on the ISO application.

R: That's right.

I: So how is that?

R: [laughs] We have got pretty much everything in place to do it, but it's just … It's expensive and the real focus at the moment is getting funding in, because funding is just drying up really bad for marine science. So we moved it to the back burner. And the other sort of driver is that we thought ISO accreditation would help if we were tendering for any sort of analysis work. So there are some people in Ireland who wanted some work done and we put a tender for that but they are not interested in ISO at all. They are interested in NMBAQC, which I probably mentioned to you last time. So they are interested in that and if we take part in that. And we are leading the zooplankton work for that. We have got staff who take part in the phytoplankton work, so we are kind of covered for it, really. But we have got everything in place if we do decide to go for ISO. I kind of hope we don't because it just creates so much work.

I: Okay.

R: So as I was saying, it is kind of expert judgement and how people look at this. That's documented, what they would look for. There is a clear criteria. If it's above a certain number, if it's below a certain number, don't check it. So that's all written down.

I: And what would you say, which stage of the analysis, the phytoplankton, the zooplankton traverse, or the eyecount, is mostly subject to the check-block procedure?

R: Which has the most potential errors?

I: Yeah.

R: Which would get picked up? I would say the large zooplankton, the eyecount. And that is probably the most representative of the sample as well, because for that, everything is picked off. For the phytoplankton and the small zooplankton, they have sub-samples. So there is always going to be some element of error. For the zooplankton, you look at, say, a fiftieth of the sample. So there is always a potential there to be a bit of a disagreement from what people see. But for the eyecount, yeah … And there are a couple of species that are very, very similar, for example the calanus species. And they co-occur, they are like three millimetres, and the only way to tell them apart is to look at their tiny little last leg and the teeth on the tiny little last leg. They are the focus of so much science that we really want to make sure we get it right. But there is kind of a tendency that if somebody finds one in one area, they'll say “oh just in case [unclear], because that's the only one you find there.” But actually they [unclear]. So I think that's probably where most of the QA issues come from.

I: Let's see, I think I had a few more questions on the analysis and the database. So everybody is using the same system of notation in their notebooks, right?

R: Yeah, kind of. Everybody is supposed to. We have a standardised procedure written for how you are meant to write it in. But it is meant to be one sample per page, which I never do. That's kind of how I write mine down. Most of them are fairly readable, but …

I: And in the database, you store the actual counts or presence or absence, right? And then there is a conversion to estimates for the entire sample …

R: Yeah.

I: And this all happens automatically, right?

R: Yeah, so what you have seen there. That is actually category values. So when people enter it … Let's get an example here. Eight, this is copepod eggs, that's not eight, that would be category eight. So there are hundreds of them. And that category value gets converted back to an accepted mean in the database. So the data that comes out are all broken down into those sort of groups. It's not actual counts.

I: But this is how they are stored?

R: Yeah. And the database doesn't store … I'm just trying to find an actual … When the data are pulled out, this is kind of what they look like. Each row would be a discrete sample, so this is the 35th Q route and then you get given a latitude and longitude. This is just the standard, but we can pull it out in lots of different ways. And in there is the date and local time. Then we kind of extract the year and month out just for ease. And this along here would be actual species and going down here would be the count for each species. But the database doesn't store zero values, because we have got 250,000-plus samples, 800 taxa, and it would just make it ridiculous if we put zeros in for everything. So actually, it records it as a blank, so that we know this has been a sample there and we know that this has always been looked for, so you can just replace it with a blank. But you see, actually there are zero values in there and that is just an indication that it's present, it is just for presence/absence, and then after that, in this case, these are … So it takes the actual count, it goes into the database, it converts it into a category, and then those are the accepted values, something like 1,500 … So that would be category one, two, three, four, five, six, seven, and up to …

I: I read in one paper that there is a compression of the categories, that you have twenty categories, but in the end you convert it to only ten …

R: That's right for the phytoplankton, yeah, and I am never really clear why that was done. I think it was computer limitations. So yeah, people would look in twenty fields and then actually, if it is seen in only one or two, it becomes a one. If it is three or four it becomes a two.

I: So you just don't change it because it has ever been done that way?

R: Yeah, it would be too difficult. We want to keep that consistent time series. And there are a lot of potential sort of foibles in the dataset. But the fact that it has always been done in the same way … You get lots of people who, it's not an accuse, but who would say “well you under-count certain things”. Well yeah, we do, but they have been consistently under-counted for sixty years. So you can just ignore the abundance values and just look at the trend to see what is happening. So yeah, if you were starting it from scratch, you would do it completely differently.

I: If people request data, what unit are they usually looking for? Because you have estimates per sample and that is probably not what most people would work with, right?

R: No, that is sort of raw data. Most people don't want that. They normally want monthly means by a given area. We have what we call our standard areas which are kind of based on the ICES areas for the North Atlantic. So people could say “I want the monthly mean of calanus in C1 and D2.” Or they could give me a region, a polygon that they want, making sure that it is not too big, because otherwise if average over a big area it just becomes unrepresentative.

I: But what would be the unit of that?

R: Well, if you think of this one for example, you would get data that would be like that. So that then has become the average number of those … You basically have to ignore those old ones because it wasn't counted. That is the average number of organisms per sample per month.

I: In a larger area …

R: In a larger area, whatever area. So that's kind of what they get. This is meaningless, because it's over the whole North Atlantic. It's utterly meaningless, but that's what they would be after.

I: Okay. You mentioned the standard areas and you probably have them in the system, too, right? So that you can automatically …

R: Yeah. This is our data extraction system. So in here you put the years you want it from and then you have an option to select all the phytoplankton, all the zooplankton, all the eyecount or whatever. And then we also created some groups, because people are quite often interested in that sort of bulk, like total diatoms which would be every diatom that has been seen on the sample just summed up to a value. And then you can pull that out as monthly means for these standard areas. Plus, there is like a North Sea general one. And then you can have the actual, proper, finalised, QAed data that we have released. So you can have the stuff that is sitting in there waiting to go. And this is where you have to put in whoever is asking for it. The data has a DOI attached to it and then sent off to people.

I: What if somebody asks for different regions? How much work is it to re-arrange or re-calculate for very specific areas?

R: It depends. Sometimes it can be a fair bit of work but if they want monthly means for a set number of taxa, for a set number of areas, it is just literally pressing the button and going. If they put a polygon in and they ask for a polygon of a region and then they ask for fifteen taxa, monthly means, and then we have to put the polygon area in, put the taxa in, especially if it is a mixture of the different types, and then they come back and say “actually, I wanted a slightly different area” or “I wanted these extra two that I forgot to ask you”, then it adds and it's the same amount of work again. It is not a huge amount of work but it still adds to the time taken. A lot of the time, because the data requests come to me or my colleague, we are the first contact with those guys, so we would try to clear that up right from the start. “What exactly do you want?”, “Give me a ring or email and tell me what you want to do with it. I'll try and see if it makes sense.” And then we try to sort of minimise that.

I: You just mentioned the DOI for each data extracted. What is actually the benefit of having that?

R: Well, apparently it makes it traceable. My IT guy deals with that, but he says it means that people reference the actual dataset in the publications with their DOI and that means you can track the DOI to see where the dataset has gone. But yeah, I am taking his word on it. We have only been doing that for a couple of years. I think it means that if somebody wants to reproduce the work in a paper, they can come back and request that dataset with that DOI and we cam just send it to them.

I: Yeah, that makes sense. I have one more question about that category system. The more plankton there are, the categories cover larger ranges.

R: They get bigger, yeah.

I: I am just wondering, why is it so important to distinguish between one and two organisms? Why is the precision so high for the low counts and …

R: I suppose that if there are a lot of them, once you get over a certain number, people wouldn't count them all anyway. You would sub-sample and were we really that interested whether it is 500 or 700? Probably not. But I suppose for the rare organisms, ones, and twos, and threes, and fours are a bit more interesting … I suppose, but you would have to ask the stats guys.

I: Okay, I was just wondering …

R: Yeah, some of these things I just take on faith and are never asked. [laughs]

I: Okay, we have already talked about the times of day, I had that on my list. I was shown a sample that comes from a recent bloom. Do you have any special procedures for those because there are so many …

R: For the zooplankton eyecount, where you are meant to take everything off, if it was covered in large zooplankton, ideally you would still scrape the whole off and then try to separate it into either a quarter or half, depending on how many were there. And they would be put into a Bogorov tray, a tray like this with a line. And you basically start scanning and go up. So I think what we would do is, you would put it into a tray like that, spread it out as even as you could, and then say “okay, I am going to look at a quarter of this”. But normally what would happen is, if it was really busy, it would be one or two species that would be really abundant. So we would sub-sample those and then we would pick through the remainder for the larger things that you could pick out easily, but you don't need to sub-sample, because there are not so many there.

I: Why exactly are they scraped off and put into …

R: Why? It is just so much easier to identify them. You can't do it on the silk very easily. It is so much easier, you take them off, put them into that tray, add some fluid and then you can manipulate them easily, flip them around. Because a lot of them, depending on how they are lying, they can hide their identification features, so you need to kind of manipulate them 360.

I: Okay. Maybe you can mention a few things … We talked about that earlier when I mentioned that the data are not pre-determined by the sample, because you have changing taxonomic entities and all that. Are there any things that you could analyse and count, but you don't because there is just no interest in it, or …

R: Yeah, there are certain things that we take to quite a coarse level. Quite often just because of time and yeah, because maybe there isn't any interest. So time-wise, decapod larvae, we just have it as decapod larvae. That could be anything from a shrimp, to a lobster, to a crab. A lot of those are commercially very important, they are a real pain to do and in fact, we have a person who can do that, but she is probably the only person here who has had training from an old guy who used to work here. But we can't justify it cost-wise for every piece to spend the amount of time doing it. And there are some other things that are probably simple to do, that we just don't do because they are not deemed as interesting, I suppose.

I: Okay. So I have a few questions about the new SAHFOS website. There is now this function to preview the data and maybe you can just explain the idea behind introducing that feature …

R: Yeah, we'll open it up because I actually very rarely look at it … Yeah, we have had some criticism that it is difficult to see our data, it is difficult to get hold of, you can't visualise it. This has got to do with the move to more open source datasets. So I think having something that can be visualised … This is the very start of what we are going to do. We basically want to have at least a certain sub-set of data that people can just download themselves, basically taking away the work from my colleague and myself, so we don't need to do it. This is kind of the very first phase. It is quite labour-intensive to write those programmes and front-ends. It is kind of there to do, but it just needs the funding at the moment. I kind of like some of this stuff, but some of it … I don't like that very well. The problem we have had here … This is quite good, you can put the thing in, and you can put a polygon in, so people are restricted here to … The idea for this is that people would fill this in and it then comes as an automated data request into SAHFOS, it comes into a SAHFOS email address. And then it is sent to my colleague and myself to deal with. So it comes in with their name, it comes in with the area they want, the taxa they want and they can put in a little line about what they are doing. But what we found from the very beginning is that people would just fill in all of those areas and just select all the taxa. So they basically say “I want all of your database”, for the whole time period with no real hypothesis for what they were doing. The first maybe half a dozen data requests that came in were all like that. And I'm like “We are not going to send you every single thing we have ever done.” So it was designed to save time, but what tends to happen is that the request comes in and then my colleague and myself end up having to get in contact with those people anyway to ask them to explain what they really want. So it has got some use, but it's …

I: I was going to ask how smoothly the thing is working now …

R: Yeah, it works okay. If it is me and my colleague doing it, it works quite well, because we know what we are doing. But the other idea of it is that if a data request came in and if it was on a certain topic and we have got an in-house researcher who works it we were expecting them to get in contact with the data requester, spark up a relationship with them, then pass the data on and hopefully be involved with projects with that person. That hasn't happened at all.

I: But do you get more requests than before through this?

R: Yeah, I think so. We have had about eighty odd this year so far from January, which is a little bit more than normal.

I: And besides this, you already mentioned that you are trying to get into collaborations with external requesters and you explained that the last time I was here as well. So these things haven't really changed?

R: Not really. I do okay with this because I collaborate with loads of people all around the world, because I am the point of contact. But I am trying to get other guys of the research team to do it and it hasn't worked. No, it hasn't worked at all. [laughs] And that is ultimately at their loss, because it means that they all have got really rubbish publication records because they don't work with anybody.

I: I see, and you mentioned that it was intended to reduce workload, right? But it didn't really …

R: No. [laughs] I think it has probably helped our database guy. It probably helps him, because it is quite prescriptive when it comes in. So if it is just a straight forward one he can just answer it, do the work, and it takes him two minutes to do. So it has probably saved me a little bit of time, because he can handle some of that now, whereas in the past, he wouldn't do …

I: And here you have only the standard areas covered, right? So if somebody asks from data from the Pacific or something …

R: Yeah, we can't do it through here, no. They would have to email, I believe. So they would probably not fill any of that in, but they would put something in there. We have had a few of those and people said “actually, I want data in the Pacific” or South Atlantic, or …

I: Then I have one more specific question about this. You have this preview function, where you have a map …

R: Is that map data that one? … No, that one?

I: I think so. You can get a …

R: Is that the one?

I: Yeah, I was just wondering about the numbers there. So is that the PCI number there, averaged probably?

R: You know what? I think that was my first question on the website and I don't think I have got an answer. Because I don't know if they … the other ones are log-transformed. I don't think this is log-transformed, it certainly shouldn't be log-transformed. It could be … It probably is the PCI, the average PCI. So monthly averages are converted to an annual average. Yeah, I'd probably figure that it would be just under one.

I: And for the others, these are log-transformed. I used to study physics, but I am not really sure what is the advantage of having these transformed? Or what does it show?

R: Doesn't it slightly normalise the data? Because you are going from null to thousands. I am sure if you log-transform them it kind of normalises … I am not a stats person. A lot of people here don't bother, I am never bother loging them. I think otherwise you get some massive outliers from the higher end.

I: Maybe it shows you the magnitude only and not the …

R: Yeah, just speak to our stats person … [laughs]

I: Alright, I think that was it about the website. I have two more questions that are kind of meta questions, I would say. I don't know if you … You don't have to give definite answers, but I was just interested in your thoughts. So one question I came across is whether or not the silk, the material of the silk, is actually part of the sample or not.

R: [laughs]

I: That is a question that came up when I talked with other philosophers about this. What would you say, do people consider it here as part of the sample? Or is it just the carrier of the sample and the sample would just be the organisms?

R: I think people would probably think of it as the sort of filtering mechanism and the actual organism on there are probably the sample. But of course you couldn't really have the sample without it. One thing I am thinking about is our current Director is a virologist, so he is interested in viruses and I am sure the silk itself is probably just as important for the viruses as the plankton would be.

I: And I always think of it as a part of it because it has so many interesting effects on the organisms you will get. You explained last time that some organisms are just under-counted just because of the mesh size and things like that, or clogging effects and so on …

R: Yeah, and the nature of the silk because it is actual silk … It is proper silk and it's fibrous and it has a leno weave. That fibrous nature of it allows it to capture some of the small stuff. So it is very important and it is still …

I: Well, as I said, you don't have to give a definite answer. Do you know why this kind of silk was actually used in the first place? Was it just the one available when it started?

R: I should imagine so. And I think because the way it is woven, it doesn't stretch out very easily. And obviously you've seen there is a kink in it [laughs] … And it is fibrous, so it means it will capture some of the small stuff. Presumably that's why they never … I don't think they have ever bothered looking at plastic or nylon as an alternative.

I: I presume these are the materials that are primarily used now for having nets or …

R: Yeah, it might be some sort of monofilament, I imagine. But yeah, this is proper silk.

I: Okay, and where exactly do you get the silk?

R: It comes from China. It comes across in … I am not certain how … It comes across in sort of rolls as a mesh about this wide and then they prepare it at the back. So, I don't know if you have noticed that the sides are folded in. They sort of put these darts in and fold it over and it gets glued and ironed. Some guys in the back, they … So they ship over a whole load every now and then and just prep masses of it up. So there is a big store of it in the back. So we always got enough …

I: That's interesting.

R: Yeah, I don't think it's too … It is reasonably labour-intensive, but yeah, we have them prepped …

I: So it is not specifically fabricated for SAHFOS. It is standard …

R: I don't know, I suppose it can't be, no … Presumably it comes in rolls that are about that, yeah they must be like that wide. It's just [unclear] …

I: The other question I had was about why the CPR Survey … I mean, I saw different reasons for why the CPR Survey could be named “continuous”.

R: Yeah.

I: Because for example, there is the continuous movement of the silk and that has some consequences for the sampling, I guess, and you also have the continuity of the ship's tow as opposed to discrete stations, and you also have the continuity of the method for seventy or eighty years. Do you know why exactly it's called “continuous”?

R: It is just because of the actual recorder itself. It is just literally because it is a continuously moving mechanism. But those other ones fit. [laughs] I really like that.

I: I was just wondering about the name and thought there might be different reasons for it …

R: It fits all of those, but yeah, I think it is just because the mechanism is a continuously advancing silk rather than a discrete snapshot sample.

I: Yeah, only a few weeks ago I actually read up about this overlap of samples. That's also an interesting part …

R: Yeah, they come back as a long … obviously, a continuous band of silk and then they are cut into these discrete samples that represent ten nautical miles. But of course each one is kind of a smear, isn't it really? When you think of it moving through the machine there is definitely a smear across how far this sample is affected by this sample and this sample was, I don't know … But plankton is really patchy and there is an awful lot of work on the patchiness of plankton. So plankton by their nature are patchy and people have done loads of … There was a paper here lying around … on plankton patchiness, basically saying that actually if you took a sample here on one side of the boat and took a sample here on the other side of the boat, they would be totally different, like nine times out of ten, because plankton is just in these discrete patches around. So we probably don't get too much bogged down in how representative that sample is compared to the other one.

I: I always thought there wasn't going to be these quick changes in the oceans between the abundances, but it is very patchy …

R: It is very patchy, yeah, it can be very patchy. Have you ever heard of warm core rings? So around where the Gulf Stream goes, it flows up around here, you can get warm core rings. There will be an eddy and it will spin off onto the shelf. So this is cold water. And that can be tens or hundreds of metres and they have really localised plankton communities in there, it is basically trapped in there. So you can go from pretty much Arctic plankton to suddenly hitting a patch of sub-tropical stuff. So it can be that sort of patchy.

I: Wow, that's interesting.

R: Yeah, it makes things very difficult, trying to work out distributions.

I: Yeah, I guess. But when you average over large areas these effects also get …

I: Yeah, exactly. We kind of say “our stuff is sort of macro-scale”. By the time you have looked at it over that area, like half the North Atlantic, it doesn't really matter. This is why you kind of … I never say to anybody “actually, you could use our data to look at frontal zones”. They quite often think that. A frontal zone is an area of convergence and it might be high productivity, but our samples are almost useless there, because it is a ten-nautical mile smear. You might pick it up, but it's not a fine enough resolution for that.

(end of recording)