Plymouth, 3/05/2015

Key:

I: Interviewer (Gregor Halfmann)

R: Respondent (Richard Kirby)

I: The first part of this interview will be about the Secchi Disk project that you lead. What is your role in the project right now? How are you involved on a regular basis?

R: I conceived the project and created the project and it has been running for two years. It is in its third year now, as it started in February 2013. On a day to day basis, in the moment, the project is primarily in awareness phase, in telling the world about it and getting participants to join the project. Essentially, getting as large an audience as possible for the project. It is pressing marketing, I would say.

I: Can you recall the impetus that led you to the idea for the project? What was the main motivation behind it?

R: There was a paper published in 2010 by three Canadian scientists. They did a hundred-year analysis of levels of phytoplankton, the plant-like cells that live at the surface of the sea. That analysis suggested that in the last fifty years the amount of phytoplankton in the world oceans had declined by 40%. Their paper was published in a top journal, but it proved very controversial amongst other scientists. Some scientists saw no changes, some scientists said they thought phytoplankton increased, and other scientists criticised the way the Canadian scientists analysed the data, because there was a problem with the data set. The hundred years broke down into two fifty year periods over which phytoplankton was measured in different ways. For the first fifty years, there was older technology used and around fifty years ago, with the advent of new technology, scientists switched to measuring phytoplankton in a different way. With no thought, it was just a change in technology, as all methods advance. So the old method declined in popularity and the new method took over. So when you take a hundred years you have to cross-calibrate and some scientists said that the science could be flawed, because you were measuring abundance, the amount of phytoplankton, in two different ways; and lo and behold, roughly at the point at which the methods changed the authors claimed to see a decline in abundance and perhaps your analysis is compromised.

I: What were the methods?

R: The first method, for the first fifty years, was using a Secchi Disk because that was invented and established as a method for measuring water clarity in 1865. It is the phytoplankton that affect water clarity away from estuaries and coasts. So it became used by scientists as an estimate of the amount of phytoplankton in the water. The clarity in the water could tell you how much phytoplankton there was. That is a simplification but that is how it works. Around fifty years ago, scientists started to measure the amount of phytoplankton by using a colour assay, a spectro-photometric assay which measured the amount of chlorophyll in the water. So Secchi Disks became used less and now the chlorophyll abundance became used more. The Secchi Disk is still used, but their representation amongst all the measurements of phytoplankton that are being taken is much less now than it was fifty to hundred years ago when that was the only method really available. So what I thought would do is bring back the Secchi Disk in abundance; and one of the best ways to do that would be to get sailors, citizens, to participate because there are sailors all over the world. They are all over the seas, they are travelling everywhere. The Secchi Disk is a very simple tool. It is a white disk that you lower into the water and with modern technology, with smart phones, you can collect the data and have it submitted to a database with very little effort on the side of the scientist once the project is set up and running. You are not having to input data the whole time, you are not having to manage the database. It is self-managing, essentially. And with the smart phone technology you can also improve the quality of the data, and the provenance of the data. You can ensure that it was collected at sea, for example, because the smart phone can obtain the GPS location. The user does not input the latitude and longitude, that is obtained from satellite, so you know at least that the person was at sea and the data is not being submitted by someone sitting in their living room in Swindon who thinks I will just have a bit of fun with the scientists’ database and make up some data and spoil it. Although you cannot ensure that sailors do not make up the data, you do at least know that it was collected at sea. So if someone is going to the bother of collecting data at sea, then probably they are going to be doing it as well as they possibly can. So that is how the project came about; and then, if we can increase the amount of Secchi Disk readings, then in five, ten, fifteen, twenty years’ time, when there are enough of them, you will be able to add those to the data from fifty to hundred years ago and it won’t matter. As for the gap in the middle, you will be able to say “Well, is there a change?”

I: You said it is a self-managing database. I would like to know more about what is happening with the data. The sailors submit their data and it is being transmitted via the network …

R: The data is stored on their phone whilst they are at sea, they can store lots and lots of data. If they are on a passage for two months and they would take a reading every day, it could store all those readings on their journey. When they get back to shore and establish a network connection, the app prompts them and says “You have data waiting to be submitted to the database. Do you want to submit it?” We give people the option whether they want to submit the data or not, because if they are an English citizen and they are now in America it would cost them a lot of money to do. On behalf of the primary goal of the project, there is no urgency for the data. We don’t need real-time data. We can do things with real-time data but the question the project was set up to address does not require real-time data. So if the database has to wait two weeks, a month, two months for the data to arrive, that is not critical. And it can be uploaded by Wi-Fi and so on …

I: Are there any steps being taken with the data once they are in the database or are they basically just there, as the Secchi depth in meters? I wonder what it looks like, basically.

R: It looks like an Excel spreadsheet of data. Whilst people are at sea, we thought, if they take the Secchi depth, they might want to do some other measurements as well. They are optional, so if they have a temperature sensor on the hull of the boat they can record the sea temperature. Smart phones have cameras, so if they see something unusual they can send in a picture and they can write a little note. There are some additional things but they don’t have to do those. To submit anything, they do have to take a Secchi Disk reading. We are not interested in the temperature without the Secchi depth. The Secchi depth is the one we are interested in. Everything else, if they want to do it, is an add-on. So the data comes into the database and it literally does look like an Excel spreadsheet with latitude and longitude. Every user has a unique user ID, so they can track their own data on the website. They can put the name of their boat in, if they wish to. So other people can see on the website, who has been where and who is collecting where. All the data is represented on a global map with Google Maps on the website and you can click on any particular data balloon and it will give you the boat name, the date, the time, the position, and the depth. That is essentially the database at the moment. And we have started looking at the data and started analysing it but not in terms of the goals of the project. The project has only been running for two years, the data is increasing daily as people are participating daily. Although it is the largest marine citizen science study already, it is still in its infancy. We have very good coverage in the northern hemisphere, less in the southern hemisphere, but one of the questions that are often levelled at citizen science by scientists is “How good is the data?” Can citizens be trusted to collect scientific data? Can they do science? Or is it going to be rubbish? While citizens might not understand the scientific process or they might not know how to set up an experiment with regard to controls, for example, I don’t see why they should be any poorer at taking simple data measurements than a scientist. If it is explained to them how to do it properly and especially if the community taking the measurements is interested in doing so, they would do it well. One of the aspects of the project is that the piece of equipment that the sailors need to use, the Secchi Disk, is something that we ask them to make themselves. It can be made very simply out of a lid of a bucket or a piece of round wood painted white; and they have to buy a tape measure or they can attach a rope to it marked in centimetre intervals. If they go to that bother, we feel that it probably means they signed up and committed to the project and are likely to do it well. That brings a bit of loyalty, I guess, and intent to do it well; but also, scientist now, especially climate change scientists … I heard on the radio the other day, that there was an estate, where they had been looking at the time when leaves appear on trees in the spring. I think it was a two-hundred year dataset and they are saying now that springs are occurring earlier. I don’t know much about that dataset, but for many of those like it, the data two hundred years ago was collected by amateur naturalists or people recording when the bluebells flowered in their garden. Many of those were not scientists; some were, some were like Darwin probably, but many were what we call amateur naturalists. They were citizen scientists and their data now is very useful and scientist don’t ignore it because it is viewed as unreliable. So why would we view current citizen scientists’ data as any more unreliable as the data collected by scientists two hundred years ago? It is just very useful and it is just our modern-day perception or perhaps wanting to put ourselves as a different group of people to others. I think we are different as scientists because we have a scientific method and that, but I think collecting data … Well, you can’t get citizens collecting data at Cern, but there are some very useful things that citizens can do and also they can get a benefit from doing it through engagement, enjoyment, or learning. So one of the first things we have done with the data in the Secchi Disk database is to actually look at how well citizen scientists, sailors, divers, or fishermen are collecting the data; and they are doing it very, very well. This is what the Canadian scientists did: They had two methods of data collection, one was the Secchi Disk and then the chlorophyll measurements, and you can convert between the two. You can use an equation to convert between the Secchi depth and chlorophyll. But nowadays we have satellites that are measuring chlorophyll, SeaWiFS for example, or Modis, that are measuring ocean colour. So you can calibrate the satellites to chlorophyll and you can calibrate the satellites to Secchi Disks; and there is a very good relationship between Secchi depth and chlorophyll, and between satellites and chlorophyll, and satellites and Secchi Disks. So we had a look at that correlation for data collected by scientists using Secchi Disks and satellites and data collected by sailors using Secchi Disks and satellites and the relationships are very good. They hold up and sailors are doing it as well as scientists would do it, which is not unsurprising because Secchi Disks are a very simple thing to use. It helps to give some confidence in the data.

I: You already spoke a bit about what I was going to ask: the usage of the data. You already mentioned that the project is in awareness phase and that you used it in combination with satellite data. What other uses of the data and what other data do you see eventually being combined with the Secchi depth, for example to derive estimates of plankton mass, which would be interesting for climate scientists?

R: The Secchi Disks only collect the Secchi depth. If you are looking at making that useful to other scientists … It will be useful for oceanographers, it will be useful for scientists developing algorithms to compare satellite measures with ocean colour for example. I guess, one of the hurdles to overcome is people wanting immediate justification for doing something. The main question that provoked the establishment of the project requires many years of data to collect. So you have to maintain interest in that time. You are not going to say, “well, l would like to participate in that project for the next ten years.” It almost has got to be a hidden agenda of the project. Although all press articles about the project say what the goal of the project is to achieve, none of them explicitly say how long that will take to do.

I: How long will it take?

R: Five to ten years, perhaps … maybe longer, to have sufficient data to then start adding it to compare with the historical records. Scientists are still collecting Secchi depths, but the amount of the Secchi Disk being used has declined around the world, so we are trying to build it back up again using citizens. That is because it is something citizens can use. Scientists will continue to take chlorophyll measurements now in preference to using Secchi Disks. Well, many scientists still use Secchi Disks, but you will not bring back the Secchi Disk to scientists. You can certainly bring it back among sailors, because it is something they can do. Sailors can’t take chlorophyll measurements easily, some might but very, very few. So that’s the long term goal, but there are also short term opportunities for the project, for example the phytoplankton varies seasonally at any one location. Just like you have spring and autumn in the temperate regions, you have spring and autumn in the plankton in the temperate regions. So we encourage sailors, if they are not ocean going sailors but more day sailors, to set up a sampling location that they might go to every time they go out of their harbour and sail past a particular spot. They might set up a sampling location and take the Secchi depth there, because then if they can get weekly, or two-weekly, or monthly readings then those will build up over the year to give an indication of local changes as well. So that will be useful for coastal oceanography, for example. So there is lots which the project can do. That can feed into fisheries, for example. The phytoplankton underpin the whole marine food chain. So if we start to build up knowledge of changes of the phytoplankton at a finer scale than we have at the moment, as no one is measuring it at very many places, and if it starts to change because we know that phytoplankton are changing, then that could feed into fisheries biology or into other sciences, physical oceanography for example. As well as doing that, knowing about your local environment also will hopefully bring in a sense of local knowledge. So people who are taking part in the project will start to understand their environment, they start talking about it amongst friends, it might get into schools … Who knows where the science might end up through outreach facility. It is difficult to know where the project will go in its entirety, although there are specific goals for it.

I: You probably get feedback from sailors who use the Secchi Disk. What do they typically say? Are there any particular positive or negative things that they point out? Do they report any problems?

R: Overwhelming enthusiasm for the project! I gave a radio interview in Australia about a year ago and the interviewer said to me at the end of the interview after we finished recording “How is it really going?” I said, “To be truthful, it is almost as if sailors had been floating around the oceans waiting for something to do.” They are incredibly appreciative and almost desperate to get engaged with the sea in an environmental or scientific way. Because for many of them the sea is their love, really. So to be able to get involved with it or to help they are too keen. There are some sailors, very, very few … in two years I probably had ten emails from sailors who want to submit data on a piece of paper, who don’t want to use mobile technology. I would say 80% of those are older generation who are not familiar with or don’t have mobile technology. 20% probably, so I am talking about probably two people, are people who are averse to modern technology, seeing it has a sinister side, saying “I don’t want a phone know where I am.” So they don’t want phones. There is almost a conspiracy theory against modern technology amongst some people. They would be very happy taking part if they could submit it on paper, but they don’t want to own a mobile phone. But then, that could just be sailors for you.

I: What is the comparative number to the 200 emails? I was going to ask how many measurements do you get per month, but you said it is still growing, so what is roughly the growth?

R: It is difficult to say. It is very sporadic. Some months we may get three or four new participants and in another months we may get thirty or forty new participants. Obviously the world is a large place and although we have got an awful lot of coverage in the press and online, people still need to see it. So sometimes they might not have seen it and then suddenly it becomes known about in one particular place or through one particular way and you get a burst of new participants or a burst of downloads of the app. It is steady growth. We got, I think, about 2,000 active participants so far. We had about 3,000 downloads of the app and about 2,000 of these people are regular but intermittent users. In the northern hemisphere, for example, a lot of people take their boats out of the water in the winter, so we don’t get winter measurements from everybody. Some people do, but some people take their boat out. So it is still too early to give a definitive figure. I think personally, if the project is still running in five years’ time, then it will be established. I still don’t consider it established, it is still very much in the promotional phase. One of the problems with citizen science is its growth. There are an awful lot of citizen science projects out there and I think there is a chance that the citizen scientists out there could get blasé about them, because many of them will wither and fail. People start taking part and either the scientists don’t realise the commitment they have taken when they started and also the amount of time it takes. They might not find the amount of time, the project is not worth it, so it dies. So I think there is a possibility that people could switch off citizen science because scientists will turn them off by having lots of projects of which people say “Yeah, I’m gonna get involved in that” and then “oh well, I’d be worried in getting involved with citizen science, because the last project that I did, I put in a lot of effort and it just stopped.” So I think there is a problem with that. There are so many websites now that accumulate citizen science projects and there are so many projects to take part in. I think you have to put an awful lot of effort in to prove that you are established and persistent so that people realise “Yeah, this is a project that is going to continue and is worth joining.”

I: Do you have a strategy for that?

R: There are several things we need to do for that. One is to keep up with the press. There are a lot of different user communities for this project; sailors, divers, fishermen, recreational fishermen, small boat commercial fishermen all have interest in the sea. So it is very much making sure that there are press releases almost coming out every month in some magazine or some website. So it is not only about connecting with new users but so that it gets talked about. And it gives you the opportunity to push those through Twitter, to post new magazine articles to the Facebook page, so it demonstrates that the project is actively worked upon continuously by the people who set it up. It is not just to set up and languishing. I make sure that the Facebook page is updated at least two or three times a month with some news. There are four of us in the project, I lead it and then there are the two computer scientists who wrote the apps and Sam Lavender who runs and manages the database. Her company runs the server of the database. The other part of the strategy is: We need to produce a scientific paper as soon as possible. And that is just to show that this data the people are collecting is being used by scientists. Because that is why they are doing it, to be of use to scientists and to do something useful about the sea. So we have to show them that they are doing something useful. They probably don’t understand how long the scientific process takes, so we need to give them something back, fast and as quickly as we can, and something that is worthwhile. What we want to give back is the analysis which I described just now, which is showing that their collecting data is good; and because that is interesting. It is an interesting study to science because it shows that citizens science collects data as well as scientists and it has not been done before. So we can do that and get it into a good journal. But we are also going to open up the writing of the paper. I have spoken to the journal editors and what we want to do is actually write the paper online in an open way. I will do it by the project’s Facebook page, so the participants can see the scientific publication process. We are going to produce a draft paper, finished in its final format, post it online, so that all the participants of the project have an opportunity to read it and comment on it if they want to. We may or may not take any notice of their comments but it gives them the opportunity to see a draft paper. We will then submit the paper and then we will post the referees’ comments on Facebook. So all the participants will get to see the referees’ comments and the paper may get rejected or may get accepted. Who knows? I’m quite happy for it to be … well, I’m not happy for it to be rejected, but being rejected is a result because it shows the public, or the citizen scientists that are engaging in the project, a scientific process. Hopefully, the referees will be positive about it and like it and they will see how we address referees’ comments because that will be posted on Facebook. They see the publication process, and all the participants are going to be authors on the paper, although they will be given a generic name. They are going to be called “SD participants”. So one of the authors will be SD participants for “Secchi Disk participants”. So they will not be named individually because the paper would be too long, but they will be able to say “Yes, I contributed to that paper, I am amongst that author.” And that is all being agreed by the journal. So that is our first goal and we hope to do that this summer. That would be two and a half years after the beginning of the project.

I: What if one wanted to access the data? As far as I know one can ask for access through the website? What are the criteria?

R: There is a map on the website and all the data appears on that map. At first glance, it does not look as if there is a lot of data there, but as you zoom in there is an awful lot of data, as you come into greater magnification. If you click on each data point, then it brings up a little bubble which tells you the depth at that point. So if you wanted to access all the data, you could click on every single one of those and write down the depth.

I: But the entire database is obviously kept and …

R: Yes, and that is available to anybody who participates in the project for free. If you contribute data, then the whole database is available to you freely. So, if you are a citizen scientist, a sailor, and you sent in one depth or ten depths, you can ask for the whole dataset if you wish. If you have not contributed anything and you want the data, then we reserve the right to charge for it, which we think is fair because at the moment, the project is run in our spare time. We fund the project out of our own pockets. We pay for the server time, we do not have any funding for it, no university funding. We do it all in our spare time.

I: So that also counts for other scientists that might want to work with the data …

R: Yes, if other scientists want the data and they have not contributed anything, then they can write into their project proposal that one month’s data for that month of interest would be X number of pounds or euros or dollars. If they want a year, it might be more, because that would be a means that we can bring in some money and we can improve the project or improve the user experience of the project, better website, more interactive, than we have got time or ability in terms of web programming than we do ourselves. Perhaps we can pay somebody to make it a much nicer interface or make it more interactive.

I: Have you had any enquiries yet from other scientists?

R: No, but we have a number of scientists joining the project. We have quite a lot of institutes that have joined the project and are contributing their data. But that is not to get access to the whole dataset. We have got lots of institutes which see the value in the project and see that a database for Secchi depth data is a useful thing to have. It is their second store of their data they are collecting and if they are on board of their boat and have a smart phone they can send their data off and they can see the value of making the data bigger than their individual data. So the Swedish Institute of Marine Science, Stazione Zoologica in Naples, Ifremer in France, a lab in South Africa, another lab in Chile, a couple of labs in America, they have all joined the project, because they have got ongoing sampling with Secchi Disks measurements, so they said “Well, would you like us to send the data into your project?” and the answer is “Yes, please.”

I: I will try to move to the second topic. I want to learn more about your own data practices, so how you, in your work as a scientist, work with data. I am wondering, what are the sources of your data? What types of data do you regularly use? How do you decide what sources you are accessing? What is the first thing that you do, when you think about a certain question or a research topic, and you wonder what data there is?

R: I have quite a few different disciplines. I am a molecular biologist and I am also an ecologist. Sometimes I create my own data. Sometimes I have to go and use other people’s data. That tends to split between ecology and molecular biology. So in molecular biology I tend to create my own data, but I also use other people’s. In ecology, I only use other people’s data. So, in molecular biology, I work with gene sequences, with DNA sequence information. The resource for that is GenBank. I use GenBank a lot and I submit new data to GenBank. So I use data that is already on GenBank to assist my research and then I add new DNA sequence data back to GenBank. So, for example, if I want to work on a gene that has not been obtained from an organism before, I will use GenBank to see whether that gene has been obtained from other taxa; and if it has, I can use the information in those sequences to help me obtain that sequence from a new species. When I have obtained that sequence from a new species I will put that sequence back onto GenBank, in fact doing so is a prerequisite of publication in a journal. You won’t have your publication accepted if the DNA sequence you are using in the paper does not have a GenBank accession number. In order to get a GenBank accession number the sequence has to be on Genbank. You submit the sequence to GenBank, you can embargo its release until after publication. GenBank will give you an accession number, you then put that accession number in your paper and your paper will then be sent out for review. So I use GenBank to guide me, then I put the DNA sequence on GenBank, which then helps other people guide their science hopefully. That is my use and creation of molecular biology data. That data can be used then to look at several things: the evolution of the DNA sequence, to look at the evolution of the organism, or the phylogeny of the organism that the sequence is obtained from. You can use it to look at the evolution of the genome and the evolution of the taxon.

I: I am not very familiar with this GenBank and the sequencing. So where do actual samples from the ocean come in?

R: With DNA I can work on anything. I can work on sea creatures, I can work on terrestrial organisms, it doesn’t really matter. One of the questions in the marine environment which I have worked on occasionally in the last few years in different projects is: The creatures I work on, the plankton … most of them are very tiny. Or the people I collaborate with, the creatures they work on, the plankton, most of them are very tiny, and we can only look at them with a microscope to see them. Our abilities to tell them apart are sometimes not very good. I often liken this to an alien coming from another planet, not being aware of the variety of life on earth and seeing that there are gorillas and chimpanzees or chimpanzees and humans, even. Looking at them and saying “Oh, they got two arms and two legs and they all got a nose in the centre of their face, so they are all the same species. They look the same to me.” It is very similar when we look down the microscope at the plankton. I mean, obviously a small jellyfish looks very different to a copepod, you would not mistake those. But sometimes with two copepods, we might think they are the same species, but in fact they are different, because biochemically, their physiology might be different. Or some copepods, for example, you can tell them apart on the species level by the number of hairs on their legs. So for others, we may not have the ability to discriminate them. It is very often said, in a terrestrial context, that we don't know the biodiversity of the rainforest. You can say exactly the same about the plankton. We don't know the extent of the biodiversity of the plankton because we don't have the ability to discriminate between them. These things we call cryptic species: organisms that appear the same but in fact are separate species or a species complex. Now if you can get the DNA from these organisms and sequence it and obtain the DNA sequence, then at the DNA level, they will be different. So you can use that as a means to give you some idea about the biodiversity in the plankton. You may take two creatures, two copepods, that look identical to you and you find that at the DNA level, they are very different. And that's why they are in fact different species, or different races. Likewise another issue in the plankton is that many creatures of the plankton are larval forms. They are the larvae of creatures that live on the seabed: larvae crabs, barnacles, mussels, shrimps. Their larval forms look very different to the adult animal. A crab larvae looks nothing like an adult crab. Many of these larval forms, although we know it's a crab larva, we might necessarily know what species of crab it turns into. It is possible to culture them, to take the larvae and grow them on, but for some organisms, no one has yet worked out or determined the crab culture conditions. So you have these larvae and they die, they never mature and metamorphose into the other forms. There are some larvae, whose identity is problematic: You may know what genus they belong to, but maybe not what species. There again, knowing the DNA sequence, the data that is within the DNA, can help you determine what adult animal it is, because if you know the DNA sequences of the adult animals, you can extract the DNA sequences from the larvae and you can match them up. You can say “This larvae has the same DNA sequence for this gene as this adult, therefore it's the larvae of this adult species.”

I: But you take these samples from the sea …

R: From the sea, yeah. Or you can take them from a museum. DNA is incredibly persistent. You can extract DNA from early humans. To be using the word data … that data is incredibly long lived. It is fossil data. There are people, who have been collecting marine organisms for the last fifty to hundred years, so there are lots of museum specimens you can use if you want to or if there are certain questions that require the use of museum specimens. There may be a species that is now extinct and you want to put it into its evolutionary context. There may be a museum specimen somewhere that has been preserved in a good enough state so that you can obtain some DNA from the skin or a bone. You can then use that historic data, fossil data, with present data and DNA sequences. So there are all sorts of questions you can answer using DNA sequence data.

I: So going back to your actual sources. So you take data collected from research vessels and …

R: Yes, from research vessels and also from machines that sample the plankton on a regular basis being towed behind other ships for example.

I: And what about the other discipline, ecology?

R: In the ecology, I tend to normally use data collected by other people. There is sometimes a crossover between the two. The ecology is always looking at the abundances of species and their changing abundance over time, over the years and the decades. The question that occupies many people at the moment is how climate or climate change or global warming, however you want to call it, is influencing the distributions of organisms in the sea surface, because the sea is taking up an awful lot of the heat that is accompanying climate change and the sea surface is the habitat of the plankton. So as the sea surface warms, they are moving about. So on the ecology level, we are looking at how communities are moving or changing over the years as the sea surface temperature changes. We are using spatial distributions of plankton that have been recorded, so taking data of what plankton have occurred when and where over successive months or years and looking at how that data is changing with time and correlating that, comparing it to data for sea surface temperature over the same period. And that data for sea surface temperature has been obtained using thermometer satellites and various means over that interval. So we are correlating changing distributions of species with the changing sea surface temperature and seeing whether there is a link between the two. It is not a causal link but it's a correlation. And sometimes molecular analysis feeds into that: There was one recent paper, a series of papers, where we knew there was a large change in the abundance of decapod larvae. Decapod larvae are the larvae of crabs, shrimps, lobsters, and prawns. But because of the problems of larvae identification we did not know what species they were. We got these datasets and compared them with temperature and saw that the abundances of these organisms are changing, but we didn't know what species they were. So, we then obtained the DNA sequences from these taxa, from represented taxa in the dataset, and determined what species they were. So that combined molecular with ecological.

I: More of a practical question about that: Do you store a lot of that data on these computers or are they stored somewhere else?

R: All the ecological data is someone else's. They store it. The ecological data I use about the species, about the plankton, is stored by the Continuous Plankton Recorder Survey. That's an organisation that is just down the corridor from here and you might want to consider talking to.

I: I am.

R: You have arranged it, or?

I: No, I am considering it.

R: Yeah, you should. In fact I can make an introduction afterwards if you want or you can leave it to yourself, it's up to you. The temperature data is all [unclear] data, or HadISST, the MetOffice, they hold all that, NOAA in the States and the MetOffice in the UK. The DNA sequence data is all on my machines and there is a lot more data on my machines than on GenBank. I mean, I don't submit everything. I have some data backed up, it is stored on my machines and it is backed up. I don't think it will be a loss to the world if all that data gets corrupted. Yes, it is an awful lot of work, but we have enough difficulty keeping up with the data anyway. It is all useful data, but GenBank is full of enough rubbish already without having more of my rubbish on it. Although it is not rubbish, it is all data that was usefully obtained and that was part of the learning curve in obtaining the ultimate goal. You need that data to get to the endpoint. And if anyone wanted to get to a similar endpoint, the data I collected on that journey would be useful for him on that process. But that's only if somebody wanted to do something similar, not do the same. If somebody wanted to do something similar, then that data would be instructive, but that is also only useful when I'm dead because if somebody wanted to get to a similar endpoint now, the best thing would be to email me and say “You did this piece of work, is there anything I need to know or do you have any background data?”

I: So you would say it is rubbish, because …

R: It would bloat the databases I think. You could submit everything to GenBank, but then GenBank becomes very difficult to use, because it becomes bloated and unwieldy. Already I find GenBank harder to use than it was when I began work in 1989, 1990, when I was one of the first people submitting DNA sequences to GenBank. Then it was very useful. If you wanted to search for something you would get back something that was useful. Now you want to search for something and you get back so much data it is difficult sometimes and it takes you so long to sift out actually what is useful to you. Whereas when there was less, everything was useful. Now there is so much, it's almost for some things, I would say, difficult to use.

I: Are there any alternatives?

R: Not that I am aware of, no. I mean, GenBank is very good for a lot of people. Size is changing fast. The type of work I do on GenBank takes you longer now using GenBank to get to the endpoint that you wanted to get to, because you have to exclude so much to get something useful. You have to go through it and check, because there is an awful lot of errors on GenBank now. The amount of data going into GenBank is vast. GenBank checks, but when people were committing less, the quality of the data going in was much higher. That is the feeling I get, I have done no analysis but that's how it feels.

I: Going back to your own data and other people's data, would you say that you make a difference between those in your work? Do you treat them differently in terms of storage or documentation? For example, one thing I am interested in that I would have asked you next is about the reuse of data. Do you think of reusing the data that you have for other purposes, with perhaps other projects …

R: The molecular data will get reused often. The plankton data can get reused for other questions. You can ask lots of questions from the same data set. The ecology data set, you could ask lots of questions with it. So yes, you might revisit that but that is not my data. You can always use the DNA data to inform you for projects that if you were going to work on the same gene but another taxon, then yes, you want to use that data again because you will be able to do it more quickly, more simply, or it will enable you to do it in the first place. So yes, you will go back to that data and it is all very well … Well, I know where everything is. No one else would, but I do. No one else would want to. But all my data that is submitted to GenBank can be accessed by everybody using the appropriate search terms. So if you want to find data on the cuttlefish GABA receptor, then you put “cuttlefish GABA receptor” and you'll find my DNA sequence. Anybody who wanted to find the sequence on the cuttlefish GABA receptor would search with those terms, so it is all there for everybody. The final product is there for everybody, not the steps on the way, the bits of data you obtained in order to enable you to get that final bit are not there, they are only on my computers and for most people they would be unuseful.

I: Does the MBA have a policy for data?

R: I'm Plymouth University …

I: Okay, perhaps that and the other institutions you worked at?

R: They may do, but if they do I'm unaware of it. With DNA sequences, some big organisations will maintain their own, the big organisations that just do DNA sequencing. But individual researchers in individual universities, the universities certainly won't collate all the DNA sequence data from their scientists, because GenBank does that. That is what GenBank is for. There are other databases for DNA sequences, but I don't use those. There is the Ribosomal Database Project, for example, that collates all the ribosomal sequences. They will also be on GenBank, but the Ribosomal Database also collates them all as well. So there are many repositories for DNA sequence information but individual universities wouldn't have a policy of collecting their own scientists data. It's not practical.

I: I'd go on to the last topic. I termed this data in biological oceanography, but of course I don't ask you to speak for the entire field, it is about your comments from your perspective. I think a few things that I have in mind you already touched upon. So my first question would be how data production methods in your field have significantly changed since you have entered the field? What were the main innovations or technologies that have really impacted your work as a marine biologist?

R: Well, computers. I would tell [unclear], I guess. When I started out, the laboratory had two computers and you handed your paper in to be typed up by somebody else on the word processing computer. You didn't even type your own paper. Now I can write something in five minutes and submit it somewhere. So the speed of writing, which is words or data, I guess, has been a major advance. The speed of reporting is huge. The advent of the electronic PDF has had a major impact on science. This is probably not the answer you wanted. In the past, a grant proposal, for example, might be eight pages long and you would be asked to submit eight copies, because the grant giving body you are submitting to would not want to photocopy eight copies, so you would have to photocopy eight copies. So you print one copy and printers were very slow then. So it was quicker to photocopy than it was to print, now it is quicker to print than it is to photocopy. Then you would photocopy eight copies and you would staple them or put a paper-clip on, put them in an envelope, and send them off and they would get distributed. Now because of the electronic PDF, you are asked to write twenty pages, much more justification, because somebody can ask you to do it, because it is only electronic. No one has to photocopy it because you can just print it out as many times as you want, because it is in the PDF. So in a way, the whole publishing has changed dramatically. The publishing of data has changed dramatically because of the amount that can be expected or is wanted. A lot of what is expected is not necessary. It's only obtained, I sometimes think, because it's cheap. It does not cost anybody to say “Oh, I put another section on this form that they have to fill in,” because the person requesting that section knows it doesn't mean they have to photocopy another five pages. They just circulate it as an attachment to an email. Almost the electronic PDF is a boon for bureaucracy. But anyway, that's probably not what …

I: No, that's good …

R: So going back, everything has got faster. I joined molecular biology just when a process called the polymerase chain reaction had been … not invented, it was invented in 1976 by a scientist called Kary Mullis. I started molecular biology in 1990, so fourteen years later, but just when the mechanisation of the process had been simplified. Before it could take five or six hours to run the polymerase chain reaction, which was the process …

I: What does that do?

R: It multiplies DNA. You can take a molecule of DNA and you can multiply it up to sufficient quantity to work with it. Initially it was a manual process. It involved someone sitting there with a tube and moving it between waterbaths of three different temperature and you have to have it in three temperatures 35 times and you have to let it get to that temperature. It has to be in each waterbath for about a minute. So 35 times three, that is 90, you know … So it was a long, slow, laborious process. Then around 1990 the first machines came out that could cycle the temperatures fast enough, because the temperature change has to be almost instant, a rapid temperature change. That is why you had three baths set up at 94, 55 or whatever and 72 and you would just sit there and literally move the tiny tube from bath to bath to bath and do each bath 35 times. So that was a hundred baths … So then around 1990 the first machines came out and you just put the tube in, walk away and have a cup of coffee and the machine will do it for you. It would just cycle up and down one block, one tube in the block and it would heat and cool, heat and cool, heat and cool. That made it a lot simpler process. You could do 48 tubes at once and you could ramp up everything. So just at that point the process became useful for large sample sizes. Up to that point the process was really only useful for a sample size of one, or three or four, a few … As soon as it became mechanised the ability to use that technique for ecology or phylogeny to look at the genes across many organisms became feasible within the amount of time that was practical. And then you could start obtaining DNA sequences. But then the slow point in the process was actually obtaining the DNA sequence from that amplified amount of DNA. So the most any one individual could do would probably be sixteen sequences a day. And it might take you a week to get to that point. So you would not be turning out sixteen sequences every day, it would take you a week to get to one day where you might get sixteen sequences. And then those sixteen sequences would be about 300 base pairs or 400 base pairs long each. So 400 letters long each. And you run these samples on a big jaw [unclear], you get the typical step-ladder picture of bands of DNA on an x-ray film you have probably seen. Then you sit with your computer and you would read them in one at a time and you type them in, GATC … and you read them down a light box. You would have to do that twice, because you might make a typing error as you're doing it. So you do it twice, you compare the two, and it was a very slow laborious process, but much faster than twenty years ago when you could not do it at all. It was fast compared to twenty years ago when you couldn't do it or thirty years ago and previous to that, but it was still a slow process. But it was also a very accurate process. Then obviously as the process got more developed and more advanced, more machinery became involved, DNA sequencing became much more automated, the person became removed from it largely and now we have sequences that can generate … I am not sure about the figures, but it took a very long time to obtain the first genome of an organism. Now you can obtain a genome without any problem at all because the sequences can do it very quickly, provided you have the money. So the amount of DNA sequence data we can generate now is phenomenal compared to what could be generated just twenty years ago. And keeping pace with that is a problem for scientists, I would say.

I: How do they do it?

R: Well, you do it with GenBank. That is why GenBank has got so bloated. It has got so much data on it, but also I would say that, and this is where the quality issue comes in, while there is an awful lot more data I would say that the quality perhaps is not as good as it was when there was much less data. But that is just my personal perception, I've got no analysis for that. Someone must have done the analysis, but …

I: And the ecological data?

R: The Plankton data?

I: Yeah, it is still taken from the sea, basically?

R: Yeah, it is still taken from the sea and it hasn't really changed much at all. It is as time-consuming as it ever was and pretty much as slow or rapid as it ever was. The temperature data is now collected by satellite, so that is done very, very fast, provided it's not cloudy.

I: I have asked about methods of data production, what about the methods of data dissemination? Obviously the internet has …

R: Yeah, the internet … GenBank for DNA sequences makes my data accessible by anybody. You don't need an access password or [unclear]. You just go onto GenBank and it's there for the whole world to access. For the plankton data I use, I access other people's data so that's for them to determine how it is accessed. The temperature data is widely available. I try to make all my papers open access, so that anybody can read them for free. I believe in the open access model. The problem with the open access model is that it is expensive, but there are ways around that. It is expensive for scientists. It costs 2,000 to 2,500 Euros for a scientist to get open access in a journal. For me that is prohibitively expensive. I am not going to pay … I can't afford to pay that. I can't afford to give that gift of money to everybody. I need that money to do work. However, there are ways of encouraging journals to give you open access. If you can be persuasive, then I would say you can argue for open access and I think open access is very important. I also try to get all my publications, all my science papers, into the press, into national or international media, whether that might be the Daily Telegraph or the Wallstreet Journal, The Financial Times, or the Daily Mail Online. I'm always trying to get my science commented on by the press because if it just stays within the scientific domain, even if it is open access, very few people will find it and read it. If you can get press coverage for your work, then hundreds of thousands will … not read your paper, but notice your work and maybe gain something from it. I think it is very important to get your work or the results of your data analysis as widely read as possible because otherwise I am just doing it for myself.

I: I was going to ask you a question about open access. How would you say is it being discussed in your field? Is it a topic? Do you talk about it with your colleagues?

R: No, I don't talk about it with anybody, not open access. I only talk about open access with my co-authors. With the people I am writing with I always try to get the paper open access, so I have the conversation. The Natural Environmental Research Council, for example, has an open access policy, that all papers from now on or from a year ago or so will be open access. And you have to cost that in your grant proposal. Well, that is fine if you get a NERC grant. Then the fees for open access are covered. I do worry somewhat, a little bit, about public funding for open access through the public purse for the NERC, because then it enables private companies to charge what they want, essentially. Because if the NERC, a public body, says “because our science is publicly funded, we want it to be open access, so you have to charge this,” then the journal can say “oh, well if they are going to fund it, then we'll charge 3,000 Euros, because we'll get the funding anyway.” So I always think public funding of things that go to private enterprises is not necessarily a good thing because it can lead to inflation. There is no incentive for anybody to keep the cost down because the decree is out that is has to be open access. There is a drive for open access. At the top level, yes, you are aware of open access being pushed but at the scientific level, unless you get funding for it from your grant, I don't think many people will push for open access because a lot of people don't see the relevance. I think it is very important but then I would say that I am in a growing minority. I am in a minority, but there are more and more people who think like me, I would like to hope, anyway.

I: I am coming back to a bit of a broader question about the changes in your field. Do you see changes in, I call them agencies, who are particularly interested in biological ocean data. I am thinking about not only scientists, but also the public, politics, industry, economics. Do you think that has changed since you have entered the field?

R: Yes. I don't know whether that is because I am becoming more engaged myself, whether the interest was always there but I never tapped into it, or whether because of the work I do, people are getting more engaged. I know for a fact that through the press and outreach work I do around my science, my science is noticed more. Therefore people talk about it more and therefore people come to me more. I know that there is a growing realisation amongst the industry that I am most closely related to, which is the fishing industry, about the usefulness of knowing about the ecosystem that I contribute an understanding towards. Those bodies are now engaging with me. I don't know whether that is an increasing awareness on their part or simply because I have made it available to them. I don't know but my perception is, there is more interest in science, in the data that science collects, in how it can inform, and how it can be used in a meaningful way by people that are engaging with or people whose livelihoods are influenced by the ecosystem, whether they would be farmers or fishermen or whatever. They can see a need and a usefulness for that information. I don't know whether that is because the information is now becoming more public so they can see it, whereas before it was hidden away and they didn't know it existed, so they didn't know its benefit or use. Or whether it is because I am now at a point where I am engaging more anyway and it was always there but I never contributed to it. I have a feeling that it is because scientists are becoming much more open and much more giving and explaining. So the people can make use of that or can see that “Oh, there is this information that is useful to us,” and so therefore they are engaging. That's my feeling, whether that is the correct interpretation of what is going on, I don't know.

I: Is there perhaps a difference or a shift between information and data? Are they particularly interested in information and knowledge or are they interested in data?

R: Only in information. They are not interested in the data. They are interested in what the data says. They don't want to know the data, the data is irrelevant. It is the patterns that the data show. Data is not irrelevant, of course, but it is irrelevant to them. They don't want to know the numbers. They want to know whether the numbers say that they can catch more fish this year, for example. That is not the question that I am considering, but that would be an example. They don't want to know whether the number is ten or five, or fifty or hundred. They want to know “Is my ecosystem healthy?” and you say “Well actually, there is much less than there was by this amount.” So you can inform them. I don't think they are that interested in the numbers, not immediately. Once they get into it they might delve deeper and say “What are the numbers?” But initially, they want to be informed, “What does the data tell us?”

I: Within the scientific community, would you say that the status of data and data management as important issues to be managed have changed over the past decades?

R: Data has become a currency, in that … Data has a value. So now people extract value out of the data. There are now jobs collating data, building databases. So data itself now consumes resources, consumes scientific research budget. The management of data, where people would now say “We need to bring all this data together,” now consumes scientific resources. Data is actually, I would say, now impinging upon the collection of new data by consuming money. There are now jobs managing data and those jobs are salaries that are all science money. We've got a new tier of scientists, the data scientist, the data manager, the database scientists. So there is a new tier of funding and if the total pot has not increased then there is less money for collecting data. So the management of data, you could say, is reducing the collection of new data. So the management of old data could be affecting the collection of new data. Also, when money is tight or short supply for collecting new data, then a simple thing to do is analyse old data. So there are now an awful lot of scientists using the collected data and analysing, looking for patterns in there, which can be useful. But it's an opportunity that having large datasets has given, which didn't exist before. So now we have scientists that are analysing large datasets. So there are salaries, which would not have been there before. So if the science budget has not increased, then there is again less money for collecting new data. Because we would have money taken up managing old data and we have got money taken up analysing old data. Both of these are useful things, but if the total pot hasn't changed for science, then you now have got less money to collect new data. And personally, I think one of the worries is that in thirty, forty, fifty years' time we will look back and think “We didn't collect any data from then because we were too busy and all our resources were taken managing or analysing the old data, because that was cheaper.” So that could be a problem going forward. Data could actually become … not a bad thing for science, but it could be … well, I think you know what I mean. The actual data itself could reduce the amount of new data being collected by consuming resources.

I: Does it also lead to increased collaboration?

R: Oh yes.

I: On all levels?

R: Yeah, data definitely leads to collaboration. Because of the ability to handle large datasets, the questions you can address are greater and you can now create datasets from components owned by different people and analyse them whereas before it wouldn't have been possible. So yes, it definitely increases collaboration.

(End of recording)