Plymouth, 11/20/2015

I: Interviewer (Gregor Halfmann)

R: Respondent (Esther Hughes)

I: I want to talk to you about DASSH and as far as I know you are the data manager of DASSH. The first thing I want to know is, what exactly are your tasks as a data manager? And what takes the biggest amount of your time in working for DASSH?

R: Yeah, well, I work on a wide variety of projects and I work across quite a lot of different sort of disciplines within marine data. So I would say the thing that takes up most of my time is actually QA of data, so quality assuring datasets that have been either sent to us by a data provider or that have been transcribed by data officers in DASSH. So for example, the JNCC commission surveys offshore marine conservation zones and we often have short contracts given to us by them to input this data into Marine Recorder which is an Access database. And our data officers are tasked to enter this data into Access. And then I am responsible for QAing a certain percentage of the samples. So I do spend a lot of time on that and reporting on projects, but I also spend a lot of time on verification and validation of marine records. So this can be from ad-hoc sightings from the public recording things on the beach that they find, right up to scientific surveys carried out by Natural England and Universities and students. So I am responsible for marine records but also more specifically macro-algae records, because my expertise is in seaweed. So I am responsible for that. I am also responsible for validation of marine non-native species that get send in through the GBNNSS, so the Great Britain Non-Native Species Secretariat. That are things like invasive algae, animals like crabs, sea squirts and things like that. I am not always the expert in that particular group, but then I will send it on to other experts that can confirm the record. So I work as a [unclear] between the kind of things. Does that answer your question?

I: Yeah. So you are a seaweed specialist. What is actually your background?

R: I did a marine biology and oceanography degree in Southampton. When I left school I was really interested in biology, I also spent a lot of time when I was younger on the beach, I am from Plymouth. So I spent a lot of time on Devon and Cornwall beaches and rocky shores wandering around there and looking at plants and animals. I’d say my interest in seaweeds came from my mother who is a botanist. So she taught me land plants. And when I started university, I then became more interested in seaweeds. And I have done a few courses, taxonomy courses to learn seaweeds of the UK and it just kind of built up from there really. And I have done some rapid assessment surveys for Natural England as well.

I: You mentioned that you have more than one discipline, actually. So working for DASSH is only of your many tasks if I understood correctly. Or one of the many things that you do …

R: Yeah, it is my main job to manage data coming into DASSH and to make sure that it is archived properly so that they are safe and secure, and so that it is formatted and standardised so that it can be reused in the future. So that includes making sure that it has got all the right metadata with it, making sure the contacts are correct, making sure that all the information about the methodology is included, and QAing the dataset. And then making sure that it is archived securely so it can then be disseminated, if it is requested. So I will get involved with that Access request as well, when people ask for a particular dataset, and [unclear] the methods publicly accessible.

I: I will come back to that later. I am also wondering … so you are the data manager, but how is DASSH structured? You are not the overall …

R: No, so Dan Lear is the overall IT and data manager, so he is the head of IT here at the MBA, but he is also my manager, so head of data as well. But within the office I am the data manager in terms of being responsible for managing the datasets. We also have a GIS officer who is responsible for all the mapping for the MBA as a whole, not just for the data team. And below me I have two data officers, a data officer and an assistant data officer who works part-time. So they are responsible mainly for data entry.

I: Yeah, I was going to ask how many people belong to DASSH?

R: So there are me, Dan, the GIS officer, the two data officers, and we have a volunteer who comes in on a Wednesday, she does kind of primary data entry. We also have a web developer, he works in the office, but he is actually not part of the data team, he is part of the IT team.

I: Okay, we are not in your office now, but I was going to ask you whether you can give me an idea what DASSH kind of looks like. What does it look like to you when you are at work?

R: Well, the way I think of it in my head is that it is like a bird’s nest. I get this idea because there are shelves all the way around on every wall of the office and there are shelves that go like right up to the ceiling. So there are shelves all around you and there are a lot of boxes and archive data and lots of papers and books and stuff on it. So you feel like you are in a bit of a bird’s nest. And then we have really good view. We have a really good view of the Sound and I sit … So we have desks all around the outside walls and then we have a main middle table where we put the cakes that we make for each other every now and again. In the MBA we are known as the cake office because one of the older data officers loved baking. She used to work in DASSH a couple of years ago and she just loved trying out her cakes on all of us. So we got known as the cake office. But we do have a lot of documents, a lot of boxes, and sometimes it does look a bit untidy. But we do have a good view of the Sound and the sea.

I: But I guess most of the time you have to look on your computer right?

R: Yeah, so all of us are desk-based and we are all looking at our screens for most of the day. So it is a typical data office, I guess, in that sense.

I: So DASSH has a distinct space here in the building?

R: Yeah, we are known as the data team within the MBA.

I: And you also have your own … Some of these things I read on the DASSH website, so if I understand correctly, there is a specific secure network for DASSH, is that correct?

R: Yeah, so we have a secure server to hold all our datasets and all our files associated with these datasets, like images, video, and original documents, original reports. We also have a physical asset space, which is fireproof. We call that the fireproof cabinet and that is where we keep old record cards and paper-based documents and hard drives as well.

I: Can I take a look at that, too, later?

R: Yeah, yeah. That is just below here, actually.

I: Okay, cool. But are you expecting to run out of space at any point or do you think there is enough room to still grow?

R: For archiving or as a team?

I: No, I mean the physical space that everything takes.

R: Well, the file cabinet has got space in it, so I don’t think we are going to run out of that anytime soon because nowadays the datasets are digital. So really we are only talking about historical things, when we are talking about assets in the fireproof cabinet, unless we find out about a very old historical dataset that has been set in a cupboard somewhere in the UK and they want it to go to DASSH, I think we are okay for that at the moment. But then we do have funding available for that kind of thing, for buying new space. When it comes to servers and things like that we do need to always keep looking out for funding and encouraging Defra and MEDIN to fund us for more space to store digital information.

I: So DASSH is mainly funded through Defra?

R: Defra and Marine Scotland.

I: And for how long is that going to be?

R: Five years, yeah, so I think last year we found out that we are funded for another five years, so we have got four years left in the hope that they will then fund us for another five years after that. But we can’t ever be certain of that, so …

I: I mean, a data archive usually should be funded like forever, you know …

R: Yeah, I am not a 100% sure, but I am pretty sure that it is five years at a time and it has been like that for a long time. So DASSH was set up in 2005, so we have been going for ten years now.

I: Was there anything similar to DASSH before it started?

R: Yeah, there was what we call the MarLIN team, which is the Marine Life Information Network, and that still exists at the MBA, but it was divided. So it was kind of like a data repository, I guess, but it was not part of the MEDIN umbrella. So DASSH is one of the MEDIN DACs in the UK. And we look after species and habitats data, so biodiversity data, but then there are eight other DACs that come under the MEDIN umbrella. And they store things like geological data, oceanography, seabed bathymetry, archaeological, fisheries data. So we are the only biodiversity archive centre for marine species and habitats. So yeah, we were originally part of the MarLIN team, but then that got divided into the education team and the data team. So now we have a separate education team who we do like to talk to and liaise with and we do work on projects together, but we just are separated now.

I: What does it mean for DASSH now to be one of the DACs?

R: It is a good thing.

I: What are the advantages?

R: I would say the advantages are that we come under the whole MEDIN umbrella, so we have got more support nationally and internationally for data standards and archiving. We are also part of a wider team. So we talk to one another, we agree certain things, and we are also funded altogether. So I guess we have more sort of secure funding. And I think we just feel part of a team, I guess. We are not just all on our own at the MBA. But yeah, I think there are probably more advantages I cannot think of, but … I guess security of data.

I: Okay, I wrote in my email that I wanted to know a lot about data users, but first I would like to talk about the data going into DASSH. So, data is submitted by scientists or by the public. Is that done voluntarily or do you ask them for the data? How would you describe the process?

R: Well, it does depend on the data provider. So if it is a member of the public, usually they take it on themselves to submit their own records. So they have an interest in marine data or they find something interesting and they don't know what it is, so they ask us to identify it. And that is how records come in there. And then you have got retired sort of experts as well who are continuing their work of recording. So we get quite a lot of records in through them. But then when it comes to scientists, like scientists from universities, we contact them. So MEDIN contacts them on behalf of us and tells them about MEDIN and tells them about the advantages of storing their data in MEDIN or in the data archive centre. But it has been quite slow on the uptake because, as you can imagine, a lot of scientists in universities think that they have got their data securely stored and they have got it how they want. And they are also worried that their data will be made publicly available. So when they have got papers to publish and things like that, they are less convinced. But part of MEDIN’s work is to sort of work with these people and show that any data is not made publicly available if they don't want it to be and that it is securely stored. The other group are the statutory agencies like Natural England and JNCC, who are required. Because they are funded by Defra, they have to make their data publicly available. So they are obligated to send their data to a MEDIN DAC. Yeah, so we receive a lot of data from them and the Wildlife Trust as well.

I: So I see that it is very diverse.

R: Yeah.

I: That’s very interesting. So are researchers aware of such data archives or do you have a strategy to make yourself visible in the community?

R: Yeah, MEDIN has only just started doing this really. They have started to write letters to universities to encourage the lecturers there to tell their students about DACs and educate them about MEDIN. And we also run MEDIN workshops to teach people about how to create metadata and how to store and archive their datasets properly. Those are free workshops that run all around the country and anyone can attend them. So they are quite good. And they get send all around the statutory agencies and to universities and to consultants. So I think the message is getting out there. It is quite slow. For DASSH, we have got a link to our page through the MBA website. So anyone visiting the MBA website can find out about the data team. We are generally … If you go to a university and you talk about MEDIN to a marine biology lecturer or something like that you are not always guaranteed they know what it is, but slowly people are starting to know what it is about.

I: I am wondering how much of the data that you get is newly created data and how much is like old records. Can you maybe estimate that?

R: So we get a lot of ad-hoc sightings from members of the public. That is appearing all the time and pretty much every day, at least every week.

I: Like how many, roughly?

R: Roughly around five or six records are emailed in and then once they are submitted through our recording scheme, which is Sealife Survey, I would say I get like ten a week, roughly. It can be more because just yesterday someone wanted to make sure their records get into the new snapshot for the JNCC and the NBN, so they have submitted hundreds of their records just yesterday. So yeah, a lot of new data is coming in from there. New surveys, like all the surveys for the marine conservation zones and the marine protected areas, all monitoring sites, come through Natural England and JNCC and they are always coming in constantly, in particular JNCC’s offshore and sea sets. We started receiving them in January 2015 and we have slowly been getting a new survey every so often. They send them to us by hard drive and I think we have got about fifteen so far. They are kind of expecting around sixty or seventy surveys in the next year or so. We are receiving a lot of data from the marine conservation zones.

I: So would you say that the rate has been increasing over the years?

R: Yeah, definitely. And it is set to increase a lot more apparently.

I: So sooner or later you are probably going to need more people?

R: Yeah, possibly. The only things we are not getting more often really are academic surveys, so information from universities and [unclear] marine institutions that are academic-based. It will increase but we just need to make them aware. I think the awareness there is quite low still.

I: Do you also get data from people who work directly at the MBA?

R: Yeah, again it is quite difficult sometimes, because they want to publish papers. But the education team gives all of their survey results to us. And we get all our bioblitz surveys every year from the education team and all their schools’ events, where they collect data, we get them immediately, so that is really good. But an academic at the MBA … You can’t really ask them until they are finished publishing their papers. But we have just received a load of non-native species data from the ship group which is a group that focuses on non-native species, biozones, and ascidians. That is a really good dataset and that was funded by Natural England, so it was mandatory for them to send their data to DASSH. Most of the people know about DASSH and know about the data team. They come to us if they need any data or [unclear] records of certain species.

I: Okay. So I read a bit about the quality assessment, right? Or quality assurance?

R: Quality assurance.

I: Okay, about the quality assurance procedures on the website of DASSH. And if I understand correctly it is basically divided into three strands, the permission side, the metadata side, and the data side. Is that roughly how it works?

R: Yeah, as soon as we receive a dataset, so when someone will send us something saying “Oh this is a dataset from my project”, first we would ask them to send us formal written permission. So we will send them a template with conditions that they can tick boxes on. Some of them might be just to archive the data in DASSH. The next level above that is to archive and disseminate it with low resolution. The next level above that is to archive and disseminate it on the NBN. And the top level is to archive in DASSH, disseminate in the NBN and to make the dataset publicly available through view and download servers, so it can be immediately downloaded without permission. And most people who send data to be archived in DASSH take the top one, so it is fully publicly available. Then they have to write down the dataset they are providing and sign the form. We get that back, we archive that and make sure we scan it in to have it as a digital copy as well. And then the next step is creating metadata for that dataset. So it is just some useful information that will make the dataset discoverable to someone who is looking for that data. So if it is about like microalgae in Plymouth Sound, we will create metadata with the year of the survey, the location, so Plymouth Sound, what the survey was, whether it was like a transect survey or a quadrat survey, and what type of information it has collected. We wouldn't [unclear] any data, we just have metadata and then we publish that to the MEDIN portal with a link to the dataset which often goes back to the DASSH website. And then once we have done that, we then QA that metadata to make sure it is right and then we would QA the dataset and we release it. So it is QAed, it is entered, it is checked … twice effectively. I QA it and then I get one of the other data officers to check it again or ask Dan, my boss, to sign it off as well.

I: Okay, so I saw that there is some kind of QA matrix, which you have online.

R: Yeah.

I: How exactly is that applied?

R: So that is the percentage of samples we QA per dataset. So if a dataset consists of between one and ten samples, we QA between fifty and hundred percent. Most of the time that is hundred percent, so we QA every piece of information within that sample. So that would include the location of the sample, who took the sample, who determined the species and the habitats. We QA the species list and the biotope or habitat that has been assigned to that sample, and any other information, physical information like the temperature or salinity, the depth of the sample, and sediment information or substrate information. For example in those [unclear], we would QA at least fifty percent of any dataset with ten or less samples. Anything more between ten or twenty, we would then QA between thirty and fifty percent, and then between twenty … no, I can't remember now … between forty and eighty, I think, or fifty and eighty. Then we QA thirty percent? Right down to a hundred samples, where we QA at least ten percent, often twenty percent. So the minimum QA for any dataset is ten percent of the samples.

I: How does that actually work when, say, you have a record with temperature and depth, how do you check that the quality of that is good?

R: Ahm … it depends on what we are QAing. For example, if Cefas sent in a dataset from a marine conservation zone, all that information will be recorded, with Biologger or [unclear] sampling equipment. We are then QAing the transcription of that data essentially, not the data itself because the data itself has been internally QAed by Cefas. So we are QAing the data that has been entered into an Access database from their [unclear] and raw data. But we will do sort of general level checks like looking at the spread of numbers for depths, or salinity, or even species distributions and things like that, and coordinates looking at all the samples. So we start off by plotting all the coordinates for all samples just to check whether the survey is in the right location. And then we do general checks on dates for the samples, surveyors, depths, salinity, temperature, and look for anomalies and spikes in that data. So it does involve looking at the general spread for every single sample. But then when it comes to QA of a sample, then we only do at least ten percent for samples over a hundred.

I: And in the end you say that … I mean what is the end product? Do you give like a quality estimate or something like that of the entire dataset or do you put a flag on every single sample that you have checked?

R: Yeah, so when we get a dataset in a spreadsheet we make a copy of that spreadsheet and we rename it as the name of the survey code with underscore QA underscore EH for my initials, because I am QAing it. And I will fill the row in colour green when I have QAed it. A thing that I come across that doesn't seem right I highlight yellow and put a comment in there. And then I send it back to the data enterer to double check. And if they don't know, it goes back to the data provider to ask them, for example, Cefas. And then they will come back with the answer or [unclear]. So yeah we do keep a copy of the QA and any changes that we have made on the original data as it was.

I: But how would someone see that? When I would be looking for datasets online is there something that tells me whether the quality of that is overall good, or … ?

R: So in our MEDIN-formatted data, any data that we disseminate to the public has been put into a spreadsheet which standardises the data. That will include validation checks, so you will be able to see how many samples were checked. When it comes to confidence in the data, that is often set by the internal QAer at the organisation. So for example, Cefas will ask another organisation independently to check their data. And once they have checked the species they will assign a certainty to that. I think it is a “likely”, “probable”, “certain”, “definite” kind of thing. So you see whether the species or habitats are … They have like a measure of confidence but that is not done by us, because we have not been able to check those. So we don’t check the samples.

I: As a data archive, what do you actually consider as raw data? Is it all the data that has not been curated, or …

R: Raw data is the data that has been put into a file, a document, as it was at the time of the survey. So that might be an image, or a video of the seabed, it might be a spreadsheet that the surveyor has input raw data in as he has been looking at the species from the quadrant or whatever it was he was looking at. It is not anything that has been processed or changed in any way.

I: So if you would get a dataset that has already been processed in any way, you would not consider that as raw data inside DASSH?

R: Ehm …

I: Because you said, you also keep raw data, right?

R: Yeah, it kind of depends, because some people would just send us a spreadsheet of their raw data that they have created in the field. But then some people would send us a spreadsheet of the data that has been checked internally at a later stage after the survey. So we still consider that raw data in a sense that it has not been analysed in any way. It has not been statistically checked. Yeah, it is a good question actually. I think my understanding of raw data is simple figures that have come from species counts or habitat characterisation, rather than averages or any percentages that have been created after the event, after the survey. I am not sure that has really been discussed that much in DASSH, or what the DASSH perspective is on that. We tend to call it “original file” rather than “raw data”. So when we get a dataset in we put that dataset into its own folder. And within that folder, there are separate folders. And one of them is called “original files”. So that are all of the files as we were given them. We have not changed anything, we have not QAed anything, and that goes in there. And then copies of that are made when we QA, and that goes into “processed data” or “QAed data”. Then we have a folder for correspondence and a folder for MEDIN spreadsheets, so standardised data.

I: That is also interesting and one thing I wanted to ask. How is DASSH kind of structured on the hard drives?

R: Yeah, I can show you when we visit the office, how it looks. It is quite simple in a sense, because we have an Access database called our data management system, where we can search for particular datasets and find where they are stored in the archive. So that is a really useful database that we use all the time for finding what we need.

I: Okay. I also read on the website that there is like a tracking system that keeps track of every step in the QA procedure. I was going to ask you if I could see such a tracking record.

R: Yeah, I can show you.

I: Can you remember any cases where this has been especially helpful or useful? Or how often do you need to look into that record?

R: I think it is most useful for when we have contracts, when we are getting paid to digitise datasets. For example, we digitised around 110 surveys for Natural England over the last year. They require a very strict QA procedure to be undertaken on all of the digitised data. We are manually transcribing data from scanned reports and put the records into an Access database. So we have to show them that we have checked this data. It has been really useful for that, because we have been able to send the log sheets of the QA to them. It is some kind of evidence that we have done that. So if, for example, one of the samples comes up with a species that is wrong or something is incorrect, then we can go back to the QA document and see if that sample was QAed. And if it was then we know that we have made a mistake in the QA procedure. But I can't think of anything like any [unclear] in particular. I think it is really useful for training, definitely. I have got two data officers at the moment and I have trained a number of data officers now, I think five or six data officers and volunteers as well. And if you can show them that you have QAed their dataset and you found mistakes in the dataset, then they know they need to keep improving. They need to do a good job because if they did not focus, I will find their mistake. So I think it is quite good for that in training. Basically, your data officers are always getting better in entering data because they know where mistakes are made. So for example, if someone consistently makes a transcription error with the date, sort of a [unclear] error, then I point that out and they can see the QA documentation and say “Ah, I have done that a couple of times, I should make sure that I don’t do that again.” So it is quite good in that sense. And that has been really useful for me as well. So yeah, I think it is quite good for that.

I: The training aspect is quite interesting. How long, would you say, does it take to become a data officer?

R: Well, that is a good question. I think it depends on the person, obviously. Some people learn quicker than others. But for our full-time data officer, I think it has taken about a year, basically to train him. He joined in November 2014 and about now I feel like he can take on things for himself. He knows all of the parts of DASSH, how to run DASSH, basically.

I: So when somebody is in training, do you have an additional amount of cross-checking and all that?

R: Yeah, in the beginning, I spend a lot of time … I spend a good two or three months of training on how to QA data, how to enter data into all the different formats that we use in DASSH. That would take up quite a lot of my time. It was a little bit frustrating sometimes because I obviously had to work on other things and at the same time I knew how important it was that he learned it right the first time. Because it is very time-costly, if you don’t train someone to create a dataset properly first time, then that whole dataset needs to re-doing. So it is quite costly if mistakes are made.

I: It sounds like it must be … I mean I have no idea how long it takes to enter a dataset into a database. Of course this varies, it depends on the size. but it sounds like it would take very long …

R: It depends. For example, one of the Natural England datasets took us … I think it was meant to take us 33 days. We estimated it to take us 33 days to enter it into an Access database, because we only had a paper record of the survey. So we actually had to read it and transcribe it. So I contacted Natural England and said “This is going to take us 33 days to do it. What do you think about that?” Just to [unclear] is going to cost 33 days’ worth of time. So they agreed that only a part of the dataset would be entered, so only the species and habitats, rather than all of the PSA as well. So we still have that to enter. Either a volunteer has to enter it or we get another piece of funding to enter the rest of their dataset. As it happens, they were only interested in the biotopes mapping for their marine conservation zone monitoring. So it did not matter that the PSA was not in there. But yeah, it can take a long time. Like offshore and sea [unclear] stuff, entering that into the Marine Recorder tends to take … A survey with like a hundred grabs and twenty video tows can take a couple of days to enter into Marine Recorder. Because you are entering PSA, so particle size analysis, and then you are entering all the analysis for each delve, for each video tow. So a video tow can can be like a hundred metres long and a still is taken every fifteen seconds or thirty seconds. So you are inputting each still in as a sample and then each sample has a species list and a biotope assigned to it. So it does take a long time. But then, if you get a dataset already in an Excel spreadsheet, that doesn’t take that long, because you are just transcribing it for us to standardise it. So it really, really depends on the dataset.

I: So it depends on the format, in which it comes …

R: Yeah, whether it is in paper or already in spreadsheet or in Access database.

I: I wonder how often do you encounter like unusual records and have to check back with the data provider.

R: Quite a lot …

I: Yeah?

R: Yeah, quite often we have species lists that come in and we taxon match it through a database called WoRMS, which is the World Register for Marine Species. And quite a few of these records require the authority to have been recorded, because some species have more than one authority, depending on where it was recorded and when it was found. So we have to go back and check authorities. Most mistakes, I think, are made by volunteer sightings from members of the public who find a crab on the beach, for example. And then they google “crab” and find one that looks like this, and then they will describe the crab [unclear], and then they will record as the species that they found first on google results. So we had quite a few unusual records. For example, only a couple of weeks ago we had a starfish that is actually known as a cushion star here in the UK and it is about that size, about two or three centimetres across, the biggest really. But there is another species that is in the Pacific that is also called cushion star but that is a totally different species. So it was recorded as the Pacific species. So that is why the verification step is really important, because it means that I get the record and I see the photo of the record. And I check that with what they have called it. And then I will have to match up, and if it doesn’t look right, then … If I don’t know it, because I am not a starfish expert … I knew this one because I knew it was not that starfish, because there are only a few cushion stars in the UK that you could assign it to. But with other species like tunicates or anything like that, I wouldn’t really have that much of an idea. So I would send it to our tunicate expert, who has an office downstairs, and I would ask him to check it. He is really knowledgable and he would verify it for me.

I: So if you need expertise from someone else, you just send it via email to them?

R: Yeah.

I: And they usually tell you …

R: Yeah, they are usually really good. So we have quite a few experts here in the MBA, who are really good. But I also have contacts at the Natural History Museum, the Welsh NRW, Natural Resources Wales, the Wildlife Trusts, and Scottish Natural Heritage, for checks of regional records. So if it is a Scottish record, I tend to check with them. There are a few groups that are really difficult and there are not really any experts out there or they are quite difficult to get a hold of. For example pycnogonids, which are little sea spiders … A lot of knowledgable marine biologists are able to put the species down to a family or a group, but they are not able to assign any sort of certainty to that record. And the guy who was the pycnogonid expert in the UK died in 2010. So we now have a lack in that expertise. That is a little bit of a problem for that group. But most groups have someone, who …

I: Do you have like a list with experts?

R: Yeah, I have a list and I build on the list all the time. If I find out that someone is really good at something I will write their name down and the email address and add it to the list. It is really good.

I: I am wondering exactly what the Marine Recorder software is. Maybe you can give me an idea. Because I have the understanding that for some things you use it and for some things you don’t? Is that correct?

R: Yeah, so the statutory agencies use it for all of their surveys. It is for storage of species and biotope information. So it is a software that is used to put species and biotopes information, as well as some physical information like sediment information into an Access database. So it is very much like an Access database, it is an .mde file. It is kind of like a pre-structured form, so you enter the data into the form one at a time. We don’t use it unless we are contracted to use it. So all of our data is standardised into MEDIN spreadsheets but JNCC and Natural England require all of their surveys to be for Marine Recorder. So yeah, we only use it when they ask us to do that.

I: Would you say that it has any advantages or disadvantages? Or is it essentially the same thing with a different interface?

R: It has a lot of advantages. Its advantage is that it has a good audit trail because you log in. You know exactly who has entered the data and you know exactly who has QAed the data. So it is really good for that. It is also good for the fact that it is free. It is freely downloadable from the web. It is an Access database that is query-able and you can merge databases as well. And it is really good for storage. You can keep hundreds and hundreds of surveys together in one database. And it is one of the NBN’s exchange formats as well. The NBN, which is the National Biodiversity Network, will accept Marine Recorder snapshots. So it is really good for that. But the disadvantage of that is even though you can import species data with an import spreadsheet, you have to manually add sediment information and biotope information into each sample. So that takes a long time. It is very time-consuming.

I: So in the end it is more time-consuming when you have to use it …

R: Yeah.

I: Okay. So just so I get this right. You mainly have species data and habitat data, right?

R: Yeah.

I: So, species data is a location and a time and whatever species was found.

R: Yeah.

I: And habitat data is also space and time probably?

R: Yeah.

I: And then is it like a set of parameters describing the habitat, or is it a category, or …

R: It depends. In the marine environment they are biotopes. So they are descriptions of habitats. That is linked to the substrate type and the collection of species that are found around it, and also which marine zone it is in, so whether it is inter-tidal, sub-tidal, deep sub-tidal, littoral, or shore [unclear] zone, for example. Biotopes that are found in the littoral zone start with LR and following that, it has a species code, so what kind of species are in that environment, whether it is red algae dominated or green algae dominated, and what kind of fauna live within there. So it is kind of like a coded system based on substrate, interaction between species, and then species.

I: So they are really interrelated.

R: Yeah.

I: I didn’t know that.

R: Yeah, so certain species only live in certain habitats.

I: Yeah, I mean the data as well. The habitat kind of value that you have depends on what species live there.

R: Yeah. But then some habitat information will just be descriptions. In particular older, historical datasets just tend to describe the habitat, so “rocky shore with algae, kelp-dominated with [unclear]”. So they could technically be assigned a biotope but because it is old data, it just has a description.

I: So do you do that sometimes?

R: Sometimes. I tend not to do that for old data, because you are kind of making assumptions on the description. So that is not a good idea. But where I have an image or a video and I can actually see the habitat, or where I have been given a species list and detailed substrate information, then I can assign it a biotope, even if it is sort of low-level confidence.

I: Okay, we already talked about the data sources. What are the actual ways of transferring the data? You get a lot of data via email, you said.

R: Yeah.

I: But also, well, is that the sort of main way?

R: That we send data?

I: No, that you receive data.

R: It depends again on the data provider. Statutory agencies tend to send data on hard drives, because they have got such large volumes of data. They send us a hard drive via post or we meet and exchange. Then we download that one to our server and send the hard drive back. Volunteer sightings or public sightings tend to come through a recording scheme. So someone will find a shell on the beach and they will want to know where it is, so [unclear] record “shell” or something like that and often they will come across some kind of recording scheme. For example, our recording scheme is Sea Life Surveys, so that comes up at the top of the results of a google search for like “recording marine life”. It is just a form-based format where they enter their name, their contact details, what they think their species is, and they can upload a photo if they have a photo. And then that comes to us directly through a warehouse. It is based at the Centre for Ecology & Hydrology. And we then verify or validate it and we email them back saying “thank you for your record”. A lot of those come through recording schemes. Other recording schemes are iRecord or iSpot which are terrestrial and marine … And then there are non-native species recording schemes as well for recording species that they think are non-native species. A lot of academic stuff tends to be emailed. And then we also get data directly from the NBN gateway. We can download datasets from there, for example marine sightings that we have not verified, but there is another organisation like maybe the Scottish Marine Biological Association that has verified them. We can download those through the NBN gateway. We just receive it as a csv file and we can store that.

I: Okay. So if I understand correctly, you have a particular metadata standard and maybe you can talk a little bit about that. I am not really familiar with different metadata standards, so what makes this standard particularly useful, or … I don’t know …

R: Yeah, I know what you mean.

I: It is a bit difficult for me to …

R: So under international standards, you are required to … There is a standard called Inspire and that kind of sets the standard for discovering data. And that might be any data, anything, not just species or habitats or marine, anything from systems or education or health or whatever. So that all comes under a particular metadata standard called Inspire. And then within that, we have a UK version of that called GEMINI which is just a little bit more UK-specific. And again that is for any type of data. And it is just sort of like a set of fields that are required to make that dataset discoverable by the public. So things like “Does it have a title?”, “Does it have a date?”, “Is it available?”, “Can you download it?”, “What access and use constraints are on the dataset?”, things like that. So MEDIN has created its own standard that sits within GEMINI to make it marine-specific, so for marine data only. So it is based on GEMINI and Inspire, it has the same kind of things like the title, the date, who created the dataset, who funded the dataset, things like that. But then it has got other fields like keywords and data format type and control vocabularies that make it more marine specific, to describe in sort of marine [unclear]. And then all that gets exported to a portal in the web where you can search for metadata and data through that.

I: Okay, so it is not like … Sometimes when I see this or that has a certain metadata standard, it is not like there are many different metadata standards and this has this particular one for some reason. It is not really like that, is it?

R: No, I think the MEDIN standard is just marine-specific, for marine data. I think there are lots of metadata standards around the world, but we have like a European and a UK version, Inspire and GEMINI are kind of set in Europe and the UK. I can’t talk for like Australia, [unclear] they all have their separate things. But it will be based on the international. They call it an ISO standard. I think most metadata standards are based on that, so sort of very general, unspecific fields.

I: Do you sometimes receive datasets where you have even more information than just what is required by the metadata standard?

R: Yeah …

I: And you keep all of that or as much as you can, or … ?

R: Yeah. So the metadata is formed of xml which is extensible markup language, which is computer-readable, and essentially you can put anything that you want into that. Within our metadata standard, we require certain fields to be filled out in order for that metadata to validate, so things like the title and start date and who created the dataset are mandatory fields that they have to fill out. If they don’t fill out those things, then the metadata won’t validate and therefore it can’t be harvested by the portal. But then you have conditional fields and optional fields. Conditional fields like the end date are required if you have an end date, for example, and the data format type, and then optional fields are extra information that you fill out if you have that information and have the time to put it in there, so things like citations [unclear]. What else is optional? … Website names and website links, data specifications and things like that. Not a lot is optional, but there are some. And you can add in extra information into the xml. But we have an online form that is used by anyone wanting to create MEDIN-compliant metadata. And they have all the fields in there for people to complete and submit it and validate it. So it is quite an easy way for people to create metadata without actually having to write the xml themselves. And I can show you that in the office.

I: Okay, so maybe now we can talk about data access and use. So maybe you can show me what you brought for me.

R: Okay. So currently the number of datasets DASSH holds is at 2,622 datasets and they are all on the MEDIN portal, so you can discover them via that. The total datasets, which we have in our view and download server, it is basically where you click a link and download it directly, is 595. But then we have a further 43 which are in our web-accessible folder. We have another 40 on Data.gov and we have over 500 on the NBN gateway, which are also publicly accessible and downloadable. So our total downloads from the view and download this year was 1,506 and the total records downloaded are 581,212. And that is all divided up into types of requests, so you have personal interest, educational, research and science, conservation NGO work, professional land management, data provision for commercial and non-profit, and statutory work. So most of the records downloaded, 405,000 of those, are for personal interest. And then you have got another 20,000 for statutory work.

I: Wow, that’s great …

R: Yeah, so we make our data available through a number of options. We have our web-accessible folder, we make all of our publicly available data on the NBN gateway for species, and then we also have our habitats data on the MESH online tool. We have our view and download, which is through DASSH, and we also send datasets to EMODnet biology. And there are direct links from the MEDIN portal. So yeah, there are quite a lot of downloads. And then from the NBN Gateway statistics, you can access how many downloads we had since 2013. So you can see which organisations have downloaded them. JNCC are the top ones, they download this data for use in their evidence for marine protected areas, these are 836,000 records.

I: Wow.

R: We download our own data, Natural England also, Buglife, which is quite an interesting one, and Scottish Natural Heritage has downloaded just over 3,000. And then that is for this year alone, so it is quite a lot for this year.

I: So only you can see these statistics, right?

R: Only me, Dan, and our deputy director.

I: Okay. So are these statistics taken automatically?

R: Yeah.

I: So it is not something that you have to enter manually?

R: Yeah, if someone is searching for a particular species, they have to log in and create a user account to download those species. So those datasets are logged, when they download it. That is how they get the figures. And all datasets have a count of the number of species or the number of records they have got in each dataset. So that is how it works. And then I have to log in to view these.

I: So, let me take a have another look. So you have 2,600 datasets, but only 600 are downloadable.

R: Yeah.

I: What is with all the others?

R: All the other ones are either in Marine Recorder or they have not been made publicly available yet. The reason is that quite a lot of them are historical datasets, so they are on record cards in paper or in scanned documents and we have not been able to secure enough funding to digitise them. So they are available. If someone wants it, we can send them a copy of a scanned record or whatever it is, but it is just not digital. So these are the datasets which are digital and standardised in MEDIN-formatted spreadsheets.

I: But some of your data are also restricted?

R: Yeah.

I: There was somewhere a list of reasons for why some of them might be restricted. So what is like the main reason for why data …

R: One of the main reasons is that the data have not been published yet by academics. There is a wait on it. It can be like a two-year wait or something like that until they publish the data. Another reason is that it is commercially sensitive, so things like oyster beds, where oyster beds are kept, they are commercially sensitive species that are threatened. So we keep those under low resolution coordinates. Then you have got things like industry datasets from oil companies that are required by law to do an environmental impact assessment or whatever. We will have that dataset of species, but they cannot submit it either at all or only at low resolution, because again, it is commercially sensitive. So these are the main reasons.

I: Okay. I saw all these different organisations or institutions …

R: Those are just places where the data can be harvested. The WAF is the web accessible folder which we hold at DASSH. So we can provide an organisation with a login to access datasets. The NBN is a species repository and you can download datasets through there. The MESH is the habitat matrix, that is another place to put datasets. And then our view and download is like a database based on GeoServer which is just kind of a way of downloading a csv file or an Excel spreadsheet [unclear].

I: So I am wondering what is the main way by which different groups search for data …

R: NBN, I would say, it has the most … Because all the statutory agencies go to the NBN for species and academics also use the NBN and consultants … It is kind of the main port of call for species data. But then MESH, if you were looking for habitats. But that is used more by the statutory agencies like JNCC and Natural England. Yeah, NBN is the main one.

I: Okay, can I keep this?

R: Yeah, yeah.

I: Great, thank you.

R: And you can also have my annual report as well for the data team. That might be useful to see what we have done in 2015 in a very brief overview. It is not our full annual report but just a summary of what we have been doing.

I: Thank you. Let me see if I have more questions on data use. There is also a great variety in users. Does it happen that you have to communicate directly with some of these?

R: Only if the dataset is restricted. For those stats of the NBN, you can just log in to the NBN and download that dataset without having to talk to anyone because the dataset is publicly available and it is free for use. For whatever reason there might be a certain restriction like usage constraints on it. You might have to quote the originator or something like that or maybe you only have to use it for educational purposes, but apart from that they can just download it freely. If it is restricted, like if it is a lower resolution, for example if someone wanted to know where all the oyster beds are in the UK, then they will put in your request and then we will receive their request. And we will either grant or deny it. So we will then contact the originator and ask them if the person can have the dataset for whatever reason and then they might say “yes”, or “yes, but they have these restrictions”, or “no”. Then they have to give them a reason why they cannot have the data. It is because it opens up potential exploitation of species or habitats, so we have to protect those.

I: Do you also aggregate or pool datasets together if somebody wanted that?

R: Yeah, we can do that. People will quite often ask us through contract to provide them with a certain set of species or a certain set of habitats. For example in the moment, I am collating all non-native monitoring species for Defra. They have awarded us a grant to collate all of those together. So I have to pull those species out from every single dataset we hold. And also I need to ask for data from other institutions like the Wildlife Trusts and other local record centres and things like that.

I: And then you would put that together as one new dataset …

R: As a collection of datasets. Yeah, we don’t really create a new dataset from datasets. We keep them as they are but remove other information that a person does not want. For example, one of the non-native species is sargassum, which is a brown algae from the Sargasso Sea in the Pacific [sic]. So it is quite invasive and it kind of smuggles a lot of other algae on the rocky shore and sub-tidally as well. We are collating all the records for that. So first of all, do a search of our own datasets and then we will do a search for all the datasets on the NBN and pull those in. And then we are also going to be asking for additional records from other agencies, other record centres, for places where we haven’t got enough data. Then those records will be tagged with that particular dataset. And we will just put all the records with the tags on with the dataset where it has come from and then remove all the other records they are not interested in. So yeah, we essentially create a new dataset but we keep the links to the original datasets. Every dataset has a code associated with it and that code is also associated with the data provider and the data holder.

I: So besides having the web interface where everybody can go and download data you also do this contract work where you provide datasets you put together …

R: Yeah, so we will ask permission and we will have to gain permission …

I: How often would you say does that happen?

R: We get a contract like that maybe once a year or once every two years.

I: So these are bigger things …

R: Yeah it is kind of similar … It is usually for monitoring work, so it is usually Defra-funded, or we have been asked by Natural England or JNCC. Occasionally … quite often, we get a request for certain species from just an academic or an interested member of the public. And we will point them to all the datasets that we know. If they want extra information then we will have to charge them to collate other species information. We will point them to the correct datasets for free if we have those available and if we knew their location.

I: I have one more question about the tracking system for requests. So this is via the NBN gateway, but do you have like another central tracking system for all requests, like within only the DASSH database?

R: For the view and download, we have … these are from our view and download from the DASSH website and for records that are just held in DASSH. But then we also have requests and enquiries, where people will phone up or email. And we log all of those enquiries, so we do get an idea of how many … But this is not automatic. We have to put it into an enquiries database on how many are requested. For our view and download you have to log in, so it tracks how many datasets that person has. And then you have to give a reason, like whether it is for personal interest, or … They are the same reasons that the NBN has.

I: So people who work at the MBA would also download data, as we can see here …

R: Yeah.

I: So they would also use these interfaces or would they just come by your office and ask if you have this or that …

R: Quite often, I think most of the time, they come to us and then we will say “You can download this from the NBN” or “We will pull these records out of our Marine Recorder databases.” If they want data from restricted datasets then we have to go through the same process with them. But yeah, quite often they do come to ask.

I: Okay.

R: A lot of the younger scientists don’t know about the fact that they can download data from the NBN, so you have to kind of tell them that they can do all these things.

I: Yeah, I think I read somewhere on the website, I think in the data policy, that DASSH is committed to promote the use of the data, and I was wondering how DASSH actually promotes the use. But we kind of already addressed that, I think.

R: Yeah, I think so. We are quite keen on open data, especially data that has been publicly funded, as is Defra. We have commitments under Defra to provide datasets that have been funded by the public. So yeah, we are keen on that. Holding on to datasets and not allowing the public to see data that they have paid for is not really fair. We are quite proactive about that and modern thinking.

I: I was going to ask about your perception of how researchers deal with open access and open data, especially with respect to the types of data that you are dealing with.

R: It depends on what they want to do with the data. For example, we come across a little bit of resistance from local record centres in particular, because their aim is to make money from providing their data. So every time they get a request through a consultancy, say someone has to build a house and they have to do an environmental impact assessment, then they need to know about any surveys they have done in that area, in terms of protected species, before they are allowed to build that house. They pay a local record centre to provide them with their data, quite often it is like a one-off [unclear], but they are still making money out of it. So when we are providing similar datasets free on the NBN gateway, publicly available, then that obviously goes against their best strategy in a way, because they are trying to make money out of their data. It is not the same data, but it might be similar data. Whereas we are public, we are funded by Defra and by the government, and by the Scottish government. So we don’t have to rely on people paying us for data. I guess there is that side of things. I think academics in particular want to publish papers, so they don’t want their data going out there and someone else publishing. Because essentially, if it is publicly available, someone else can look at their dataset and make their own theories about it and publish a paper on it before they do, when they have collected the data. But at the same time, most of these datasets are funded by the government and by the public, so they are obligated to provide their data at some point, usually within two years, I think. I think most academics understand that but some get sort of frustrated about that.

I: Is it perhaps also a factor that, at least that is my impression, especially these ecological datasets are very much a product of manual work?

R: Yeah, yeah, all that hard work …

I: They have a kind of high value, I would say …

R: Yeah, definitely. Some surveys have thousands of pounds, large amounts of money spent on their survey, millions in the case of marine conservation zones. Yeah, just for a ship to be sent out offshore is a lot of money per day. It is very valuable.

I: Yeah, in these last questions I am asking more about your general view on the scientific field, but from your perspective of course. Again, maybe one more question about metadata, I think you have also been involved in the development of metadata standards. Would you say that there is something that scientists in marine ecology generally miss or do wrong regarding metadata? Or is there something that you perhaps would change in the standard if you could? I don’t know if there are such things, I am just speculating …

R: Yeah, quite often they don’t give enough information because they are time-limited. They will just do the complete minimum of what is required, which is fine and sort of okay, but at the same time it is not really that helpful. It is not the most helpful you could be. That is quite frustrating because it would only take a few more minutes. It is not that it would take hours of time, it would be probably just a few more minutes to copy and paste that information across. So that is quite frustrating. Another thing which they don’t often do is fill out the title correctly. We have a recommended format which has the year, the organisation that funded the dataset, where the survey was, and the type of survey, but some people would just write “transect survey” and that’s it. So it makes discovering the dataset really … It is not very helpful and it kind of removes the whole point of metadata. When people do that, I am tracking it all the time, so I will often write to them and say “It would be really helpful if you could do it like this”. And often they are really good and they will do that. But then sometimes they will say “I don’t have time”. So yeah, it is a bit of a time thing.

I: Also, DASSH has in its online FAQs a quite inclusive definition of data, and what data can be. So it says like habitat lists, biotope lists, species lists, but also maps, figures, images, and so on, basically all kinds of information. Would you say that scientists are aware that data are so diverse?

R: Nay …

I: Probably not?

R: Probably not. I mean, I [unclear] … I think a lot of them are, but I think a lot of them are within their own fields and they think that is the only type of data. I think it is also a generational thing as well. Older scientists tend to think more about species and habitats, whereas younger scientists are putting different data types together more often now, like oceanographic and bathymetric data together with species and habitats, so they kind of see a bigger sort of ecosystem picture. What was the question again? Sorry.

I: Whether they realise the diversity of data, or of data formats and types.

R: Yeah, a lot of people, I guess, don’t think an image or a video is a piece of data. They are like “Oh, why is that there?” And I am like “Well, it is data because it tells you something about … , you know, it is a snapshot of the seabed”, for example. Yeah, I think a lot of the older school people think that data is sort of a spreadsheet or in a database and kind of forget the images and video and the mapping side of things. But when it comes to a dataset, that is often a bit of a confusing one for people. They often wonder what exactly a dataset is.

I: Okay.

R: Yeah, that’s another question …

I: How would you define a dataset?

R: It is a bit of a tricky one. It is usually a specifically funded piece of work, usually it is easily extractable via third party, it is usually done in a set piece of time or for a specific reason. Anything more than that could be a series of datasets, it could be one, two, or more datasets. And sometimes people get confused by sections of datasets. So an organisation will go out and do a multidisciplinary survey and they will measure sea floor information and substrate, species information, biotope information, water column information, all the different things like currents, weather and everything. But then certain people will submit just a section of that and call that a dataset. They just submit the species and habitats part and they think it is one whole dataset and the rest is another dataset, when actually, the whole thing is a dataset in itself because it was done in the same time and in the same place. So yeah, that is quite an interesting one there.

I: Yeah. This is the last question, I guess. It is kind of related to what you said earlier. If I may ask, how long have you been in data archiving and do you perceive any change in the overall appreciation or interest in data archiving work? Like from the public, or the government …

R: I have been in the MBA for four years now and with the data team for around three years. So I would say that it has changed. When I first started we were still setting up the data archive centre. It was still in progress, but I would say now we are fully functional and we are ready to receive datasets. We are not setting up any more. I think more and more organisations are starting to realise that archiving and storage of data is really important, especially now with something like marine protected areas and the amount of data that is coming out of those surveys. For example, the marine conservation zone areas are like terabytes of data per survey, tens of terabytes of data. And they are starting to realise that this data has to go somewhere because so much money has been spent on it, like millions of pounds. So it has to be stored securely. Whereas previously, it was just kind of put on a hard drive and stuffed into a drawer. Or it is just put on some paper and, you know, they cared less that they had to protect it from damage, loss, or fire.

I: Okay. I think that is all.

(end of recording)